UNIVERSITY OF CALIFORNIA SAN DIEGO

Attention-based and Causal Explanations in Computer Vision

A Dissertation submitted in partial satisfaction of the requirements

for the degree Doctor of Philosophy

in

Computer Science

by

Kamran Alipour

Committee in charge:

Professor Jurgen P. Schulze, Chair
Professor Manmohan Chandraker, Co-Chair
Professor James Hollan
Professor Hao Su
Professor Nuno Vasconcelos

2022

The Dissertation of Kamran Alipour is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my parents, Samila and Eghbal, for their dedication and love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my advisor and mentor, Dr. Jurgen Schulze, for giving me the opportunity to do research and providing continued support and guidance throughout this research. Dr. Schulze has become a role model for me moving forward, not because of his level of professionalism and integrity but also because of his kindness and compassion towards others. I'm immensely grateful for what he has offered me and his unconditional empathy and friendship.

I want to thank my family for their support and sacrifice throughout these years, as I could not imagine this journey without them by my side.

I am thankful to my committee members for their valuable comments and guidance during my research. I also would like to thank Dr. Giedrius Burachas from SRI International, Dr. Michael Pazzani, and Dr. Babak Salimi from Halıcıoğlu Data Science Institute at UC San Diego and Dr. Ehsan Adeli from Stanford University for their mentorship and guidance throughout this research.

Chapter 1, in part, is a reprint of the material as it appears in A study on multimodal and interactive explanations for visual question answering. Alipour, Kamran; Schulze, Jurgen P.; Yao, Yi; Ziskind, Avi; Burachas, Giedrius. In SafeAI workshop at AAAI conference (2020). The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in The impact of explanations on AI competency prediction in VQA. Alipour, Kamran; Ray, Arijit; Lin, Xiao; Schulze, Jurgen P.; Yao, Yi; Burachas, Giedrius. In 2020 IEEE International Conference on Humanized

Computing and Communication with Artificial Intelligence (HCCAI) (pp. 25-32) (2020). IEEE. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Improving users' mental model with attention-directed counterfactual edits. Alipour, Kamran; Ray, Arijit; Lin, Xiao; Cogswell, Michael; Schulze, Jurgen P.; Yao, Yi; Burachas, Giedrius. Applied AI Letters, e47 (2021). The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, has been submitted for publication of the material as it may appear in the European Conference on Computer Vision (ECCV), 2022, Alipour, Kamran; Lahiri, Aditya; Adeli, Ehsan; Salimi, Babak; Pazzani, Michael. The dissertation author was the primary researcher and author of this paper.

VITA

2011	Bachelor of Science in Aerospace Engineering, K. N. Toosi University of Technology
2013	Master of Science in Aerospace Engineering, Sharif University of Technology
2022	Doctor of Philosophy in Computer Science, University of California San Diego

PUBLICATIONS

Alipour, K., Schulze, J. P., Yao, Y., Ziskind, A., & Burachas, G. (2020). A study on multimodal and interactive explanations for visual question answering. SafeAI@AAAI 2020.

Alipour, K., Ray, A., Lin, X., Schulze, J. P., Yao, Y., & Burachas, G. T. (2020, September). The impact of explanations on AI competency prediction in VQA. In 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI) (pp. 25-32). IEEE.

Alipour, K., Ray, A., Lin, X., Cogswell, M., Schulze, J. P., Yao, Y., & Burachas, G. T. (2021). Improving users' mental model with attention-directed counterfactual edits. Applied AI Letters, e47.

Ray, A., Cogswell, M., Lin, X., Alipour, K., Divakaran, A., Yao, Y., & Burachas, G. (2021). Generating and evaluating explanations of attended and error-inducing input regions for VQA models. Applied AI Letters.

Hariri, A., Alipour, K., Mantri, Y., Schulze, J. P., & Jokerst, J. V. (2020). Deep learning improves contrast in low-fluence photoacoustic imaging. Biomedical optics express, 11(6), 3360-3373.

Alipour, K., & Schulze, J. P. (2019). Real-Time Photo-Realistic Augmented Reality Under Dynamic Ambient Lighting Conditions. Electronic Imaging, 2019(2), 186-1.

Zhang, M., Lucknavalai, K., Liu, W., Alipour, K., & Schulze, J. P. (2019). Arcalvr: augmented reality playground on mobile devices. In ACM SIGGRAPH 2019 Appy Hour (pp. 1-2).

Ghaffari, R., Alipour, K., Solgi, S., Irani, S., & Haddadpour, H. (2015). Investigation of surface stress effect in 3D complex nano parts using FEM and modified boundary Cauchy–Born method. Journal of Computational Science, 10, 1-12.

ABSTRACT OF THE DISSERTATION

Attention-based and Causal Explanations in Computer Vision

by

Kamran Alipour

Doctor of Philosophy in Computer Science

University of California San Diego, 2022

Professor Jurgen P. Schulze, Chair

Professor Manmohan Chandraker, Co-Chair

Despite their potential unknown deficiencies and biases, the takeover of critical tasks by AI machines in different fields has created a demand for transparency alongside accuracy for these machines. Explainable AI (XAI) approaches have provided solutions by mitigating the lack of transparency and trust in AI and making these machines more interpretable to the lay users. This dissertation investigates the role of explanations for deep learning models in computer vision. This research explores new methods to produce more effective explanations for such models and

techniques to evaluate the efficacy of such explanations. The evaluation methods rely on extensive user studies as well as automated approaches. Throughout the study, we implement such XAI systems for complex tasks with potential for bias, such as Visual Question Answering (VQA) and face image classification.

We present explainable VQA systems that generate interpretable explanations using spatial and object features driven by attentional processes such as transformers. Our user studies show that exposure to multimodal explanations improves lay user's mental, mainly when AI is erroneous. In these studies, we demonstrate the role of object features in enhancing the explainability and interpretability of such models.

Furthermore, we examine automated techniques to provide controlled counterfactual explanations more successfully than merely displaying random examples. To provide counterfactual examples, we compare an automated generating method versus a retrieval-based approach. Results indicate an overall improvement in users' accuracy to predict answer change when shown counterfactual explanations. While realistically retrieved counterfactuals are the most effective at improving the mental model, this study shows that a generative approach can also be equally effective.

For the task of face image classification, modern models tend to be prone to potential biases that can cause ethical issues in different applications despite their high accuracy. We introduce a novel method to search for causal yet interpretable counterfactual explanations using pretrained generative models. The proposed explanations show how different attributes influence the classifier output, with contrastive counterfactual images as local explanations and causal sufficiency/necessity scores as global explanations.

INTRODUCTION

The inevitable growth of AI in various fields, from medical to autonomous transportation, demands a clearer perception of how these machines operate. AI has successfully demonstrated strong competency on multiple tasks over the past years, beating its human counterparts by many standards. In this direction, AI users gradually demand a high level of trust and clarity to delegate their tasks to automated systems. On the other hand, AI technology embodies more complicated mechanisms and methods to meet the needs of their intricate jobs.

AI algorithms can be so complicated that even their creators cannot explain their decisions. On the other hand, producing AI machines as black-box solutions raises questions about their fairness and bias. Under the European Union's General Data Protection Regulation (GDPR), a business using personal data for automated processing must be able to explain how the system makes decisions (See Article 15(1)(h) and Recital 71 of GDPR). Individual data subjects (e.g., the person who was rejected for a loan by an AI banking system) have the right to ask the business why it made the decision it did. The businesses must then explain how the system came to its conclusion. If companies didn't explain their decisions in response to an individual's request, they would not be compliant with the GDPR. In the US, as part of the American AI Initiative, Federal agencies will foster public trust in AI systems by establishing guidance for AI development and use across different types of technology and industrial sectors. Insurance companies are good examples of these regulations as they are required to be able to explain their AI-based decisions like rate and coverage.

The issue of black-box AI has previously exposed itself in many failure scenarios, and among them, the use of biased AI as risk assessment tools in the US criminal legal system is a

well-known case. An AI algorithm called COMPAS is supposed to help judges determine whether a defendant should be kept in jail or allowed out while awaiting trial. It trains on historical defendant data to find correlations between factors like someone's age and history with the criminal legal system and whether the person was rearrested. As the law requires, COMPAS doesn't include race in calculating its risk scores. In 2016, however, a ProPublica investigation(Angwin & Larson, 2016) argued that the tool was still biased against blacks. ProPublica found that among defendants who were never rearrested, black defendants were twice as likely as white ones to have been labeled high-risk by COMPAS.2.

This research study is a step toward a deeper understanding of efficient decision-making explanations, particularly in deep learning-based AI machines that act as black boxes. Research in this area is critical to understanding how to address the challenges of using AI as a black box in sensitive applications such as medicine. These issues are fundamental to address because, in high-risk tasks, the AI machines should be able to explain their decisions and be transparent about their processing as required by ethics and the law. We cannot create trust in using AI machines for critical tasks without transparent AI. Moreover, we cannot guarantee unbiased and fair decisions by AI models in the absence of explainability, and as a result, we cannot deploy them in sensitive applications.

Among advanced AI technologies, deep learning models are notoriously opaque and difficult to interpret and often have unexpected failure modes, making it hard to build trust. While various explainable AI (XAI) approaches aim at mitigating the lack of transparency in deep networks, the evidence of the effectiveness of these approaches in improving usability, trust, and understanding of AI systems is still missing. This research investigates building, testing, and analyzing explanation systems for deep learning models with complex visual question answering

(VQA) and image classification tasks. In VQA, AI agents attempt to answer a question in a textual format based on an input image. This type of AI model is specifically an excellent target for XAI research due to its complicated architecture and feature space as they fuse word and image features and process them to produce accurate answers. For image classification, AI models are usually prone to hidden biases that lead to ethical issues and even socio-political concerns. We particularly look at the problem of face image classification, which has been a popular research topic in ethical AI.

Our methodology includes devising novel explanation techniques and then implementing experiments to evaluate their effectiveness. We study the impact of explanations in human-AI collaboration and investigate the effectiveness of XAI systems in building trust and understanding for human users. This research effort addresses the challenge of modeling human understanding of AI and improving human interaction with AI. Uncalled trust in AI in cases where the machine does not have the competency to complete a specific task can lead to catastrophic results in many applications such as clinical diagnosis and military. This is a significant issue since, in most XAI systems, the output is only meaningful to AI experts and not the actual lay users of the system, which can be a radiologist or a soldier on the battlefield. Building the XAI systems requires an extra step of modeling the human response to it and optimizing the result for maximum human-AI collaboration efficacy.

Unlike previous post-hoc saliency approaches, we generate a combination of explanation modalities based on AI attention and/or causal attributes that impact AI inference. We also investigate new modalities of explanation by combining AI attention with human-annotated data to produce more intuitive means of communication between black-box AI and lay users. We introduce the concept of explanation helpfulness to build users' mental models and measure them

in a human-AI collaboration interface. Our work proposes quantitative groundings to measure explanation helpfulness. Our results show the strong correlation between these measures and the actual performance of subjects in predicting the AI agent behavior. We conduct extensive between-subjects and within-subjects experiments to probe explanation effectiveness in improving user prediction accuracy, confidence, and reliance, among other factors. In the context of counterfactual explanations, we quantitative global explanations by generating sufficiency and necessity scores for causal attributes. Our explanation evaluation metric showed a strong correlation with the actual performance of users in our studies.

Our results indicate that the explanations help improve human prediction accuracy, especially in trials when the AI system's answer is inaccurate. Furthermore, we introduce active attention, a novel method for evaluating causal attentional effects through intervention by editing attention maps. Our research also assesses the impact of explanations on the user's mental model of AI agent competency for the task. We introduce object-level attention mechanisms in the AI architecture that can invoke higher AI-competency awareness for the users.

For the task of image classification, we aim to produce explanations for the causal effect of interpretable attributes on the final classification results. We present a method for creating contrastive counterfactuals for a classifier using just pretrained generative models. To explain the black-box image classifier, we offer contextual, contrastive, and causal explanations in the form of sufficiency and necessity scores.

The rest of this dissertation is structured as follows. In Chapter 1, the experiment design evaluates multimodal explanations for the task of VQA. The experiment is expanded in Chapter 2 to include newly postulated attention processes and the evaluation of competency explanations. Chapter 3 introduces and evaluates a unique way to make counterfactual changes as part of the

VQA job. We focus on the causality component of counterfactual explanations for face image classification tasks in Chapter 4 and present contrastive counterfactuals using pre-trained generative models. In the final chapter, we conclude the contributions in this dissertation and propose potential future directions under each discussed topic.

CHAPTER 1 ATTENTION-BASED EXPLANATIONS

1.1 Abstract

The explainability and interpretability of AI models are essential factors affecting the safety of AI. While various explainable AI (XAI) approaches aim at mitigating the lack of transparency in deep networks, the evidence of the effectiveness of these approaches in improving usability, trust, and understanding of AI systems is still missing. We evaluate multimodal explanations in the setting of a Visual Question Answering (VQA) task by asking users to predict the response accuracy of a VQA agent with and without explanations. We use between-subjects and within-subjects experiments to probe explanation effectiveness in improving user prediction accuracy, confidence, and reliance, among other factors. The results indicate that the explanations help improve human prediction accuracy, especially in trials when the VQA system's answer is inaccurate. Furthermore, we introduce active attention, a novel method for evaluating causal attentional effects through intervention by editing attention maps. User explanation ratings are strongly correlated with human prediction accuracy and suggest the efficacy of these explanations in human-machine AI collaboration tasks.

1.2 Introduction

With recent developments in deep learning models and access to ever-increasing data in all fields, we have witnessed a growing interest in using neural networks in various applications over the past several years. Many complex tasks which require manual human effort are now assigned to these AI systems. To utilize an AI system effectively, users need a basic understanding of the system, i.e., they need to build a mental model of the system's operation for anticipating

success and failure modes and develop a certain level of trust in that system. However, deep learning models are notoriously opaque and difficult to interpret and often have unexpected failure modes, making it hard to build trust. AI systems users do not understand, and trust is impractical for most applications, primarily where vital decisions are based on AI results. Previous efforts to address this issue and explain the inner workings of deep learning models include visualizing intermediate features of importance (Selvaraju et al., 2017; Zeiler & Fergus, 2014; Zhou et al., 2014) and providing textual justifications (Huk Park et al., 2018), but these studies did not evaluate whether these explanations aided human users in better understanding the system inferences or if they helped build trust. Prior work has quantified the effectiveness of their explanations by collecting user ratings (Chandrasekaran et al., 2017; Lu et al., 2016) or checking their alignment with human attention (Das et al., 2017) but found no substantial benefit for the explanation types used in that study.

To promote understanding of and trust in the system, we propose an approach that provides transparency about the intermediate stages of the model operation, such as attentional masks and detected/attended objects in the scene. Also, we generate textual explanations that are aimed at explaining *why* a particular answer was generated. Our explanations fall under the category of *local explanations* as they are intended to address inference on a specific run of the VQA system and are valid for that run. We offer extensive evaluations of these explanations in the setting of a VQA system. These evaluations are made by human subjects while performing a correctness prediction task. After seeing an image, a question, and some explanations, subjects are asked to predict whether the explainable VQA (XVQA) system will be accurate or not. We collect the data on subject prediction performance and their explanation ratings during and after each prediction run.

We also introduce active attention - an interactive approach to explaining answers from a VQA system. We provide an interactive framework to deploy this new explanation mode. The interface is used to conduct a user study on the *effectiveness* and *helpfulness* of explanations to improve users' performance in user-machine tasks and their mental model of the system. The efficacy of explanations is measured using several metrics described below. We show that explanations improve VQA correctness prediction performance on runs with incorrect answers, thus indicating that explanations effectively anticipate VQA failure. Explanations rated as more helpful are more likely to help predict VQA outcomes correctly. Interestingly, the user confidence in their prediction substantially correlates with the VQA system confidence (top answer probability). This finding further supports the notion that the subjects develop a mental model of the XVQA system that helps them judge when to trust the system and when not.

1.3 Related Work

Visual Question Answering. In the VQA task, the system provides a question and an image, and the task is to answer the question using the image correctly. Combining both natural language and visual features, the multimodal aspect of the problem makes this a challenging task. The VQA problem was initially introduced in 2015 (Antol et al., 2015), and since then, multiple variations have been proposed and tested. A common approach is to use attentional masks that highlight specific regions of the image, conditioned on the question (Fukui et al., 2016; Jiang et al., 2018; Kazemi & Elqursh, 2017; Lu et al., 2016; Teney et al., 2018; Xu & Saenko, 2016).

Explainable AI. The effort to produce automated reasoning and explanations dates to very early work in the AI field with direct applications in medicine (Shortliffe & Buchanan, 1984), education (Lane et al., 2005; Van Lent et al., 2004), and robotics (Lomas et al., 2012). For vision-

based AI applications, several explanation systems focus on discovering visual features essential in the decision-making process (Hendricks et al., 2016; Jiang et al., 2018; Jiang et al., 2017; Selvaraju et al., 2017; Zeiler & Fergus, 2014). For visual question answering tasks, explanations usually involve image or language attention. Besides saliency/attention maps, other work has studied different explanation modes including layered attentions (Yang et al., 2016), bounding boxes around important regions (Anne Hendricks et al., 2018), or textual justifications (Huk Park et al., 2018; Shortliffe & Buchanan, 1984).

This chapter proposes a multi-modal explanation system that includes justifications for system behavior in visual, textual, and semantic formats. Unlike previous work that suggests explanations primarily relied on information produced by the AI machine, our approach benefits from combining AI-generated explanations and human annotations for better interpretability.

Human studies. As an attempt to assess the role of an explanation system in building a better mental model of AI systems for their human users, several previous efforts focused on quantifying the efficacy of explanations through user studies. Some of these studies were developed around measuring trust with users (Cosley et al., 2003; Ribeiro et al., 2016) or the role of explanations in achieving a goal (Kulesza et al., 2012; Narayanan et al., 2018; Ray, Burachas, et al., 2019). Other works measured the quality of explanations based on improving the predictability of a VQA model (Chandrasekaran et al., 2018).

Despite their tremendous insights into the efficacy of various explanation modes, previous studies do not interactively involve human subjects in producing these explanations. In our research, we design an interactive environment for users to evaluate our multi-modal explanation system to help users predict the correctness of a VQA model. Moreover, the users generate explanations and receive online feedback from the AI machine.

1.4 The VQA Model

VQA deep learning models are trained to take an image and a question about its content and produce the answer to the question. The core model extracts features from natural language questions and images, combines them, and generates a natural language answer. Among various methods to train VQA systems to accomplish this task, the attention-based approach is specifically of our interest.

We use a 2017 SOTA VQA model with a ResNet (Szegedy et al., 2017) image encoder (Figure 1.1) as our VQA agent. The model is trained on the VQA2.0 dataset and uses an attention mechanism to select visual features generated by an image encoder and an answer classifier that predicts an answer from 3000 candidates. Moreover, we replaced Resnet with a Mask-RCNN (He et al., 2017) encoder to produce object attention explanations, similar to the approach used by (Ray, Burachas, et al., 2019).

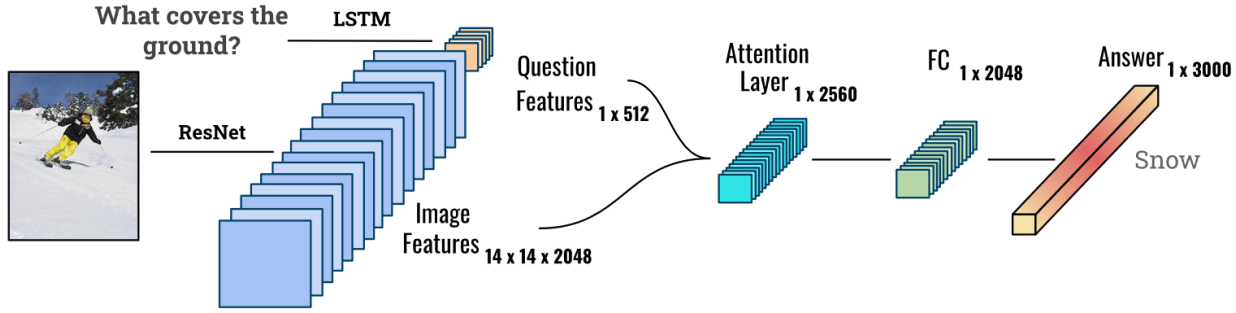


Figure 1.1 Attention-based VQA architecture

As illustrated in Figure 1.1, our VQA model takes as input a 224×224 RGB image (I) and question with at most 15 words. Using a ResNet, the model encodes the image to reach a feature representation (φ):

$$\varphi = CNN(I) \quad 1.1$$

where φ is a three-dimensional tensor with $14 \times 14 \times 2048$ dimensions. The model also tokenizes the input question(q) of length p into word embeddings $E_q = \{e_1, e_1, \ldots, e_p\}$ and then encodes these embeddings to a feature vector of 512 dimensions (s) using an LSTM model based on the GloVe (Pennington et al., 2014) embedding of the words:

$$s = LSTM(E_q). \tag{1.2}$$

The attention layer takes in the question and image feature representations and outputs a set of weights to attend to the image features. Inspired by the Stacked attention structure (Yang et al., 2016), we compute the attention distribution over spatial dimensions of the image features:

$$\alpha_{c,w,h} \propto exp\, F_c(s, \varphi_{w,h}) \ni \sum_{w=0}^{W} \sum_{h=0}^{H} \alpha_{c,w,h} = 1 \tag{1.3}$$

where the attention mechanism F_c is modeled with convolutional layers, and each image feature glimpse x_c is the weighted average of visual features over the spatial locations $< w, h >$:

$$x_c = \sum_{w=0}^{W} \sum_{h=0}^{H} \alpha_{c,w,h}\, \varphi_{w,h} \tag{1.4}$$

A model G with fully connected layers uses the weighted image features and the question representation to predict the final answer from a set of 3000 answer choices (a_i):

$$P(a_i \,|I, q) \propto G(x, s) \tag{1.5}$$

1.5 Explanation Modalities

Our XVQA system aims at explaining the VQA agent's behavior by combining the attention features generated in the VQA model with meaningful annotations from the input data. These annotations include labels, descriptions, and bounding boxes of entities in the scene and their connections with each other.

Our XVQA model either visualizes information from the inner layers of the VQA model or incorporates that information with annotations to explain the model's inner work. The explanations are provided in different combinations to the subgroups of study participants to assess their effectiveness for accurate prediction.

1.5.1 Spatial Attention

The primary purpose of spatial attention is to show the parts of the image the model focuses on while preparing the answer. Attentions maps here are question-guided and more weighted in the areas of the image that make a higher contribution to the response generated by the model. The model computes the attention based on image features in ResNet (Szegedy et al., 2017) layers and the question input. The final values in the attention map are a nonlinear function of image and question feature channels (Figure 1.2).

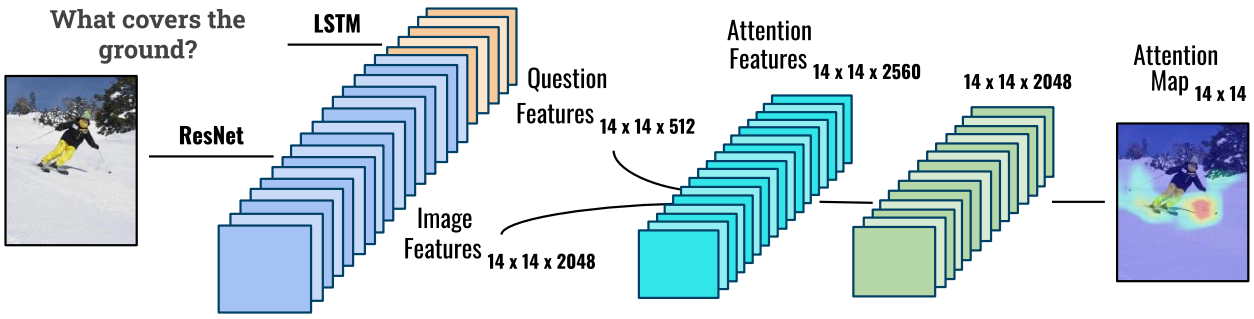
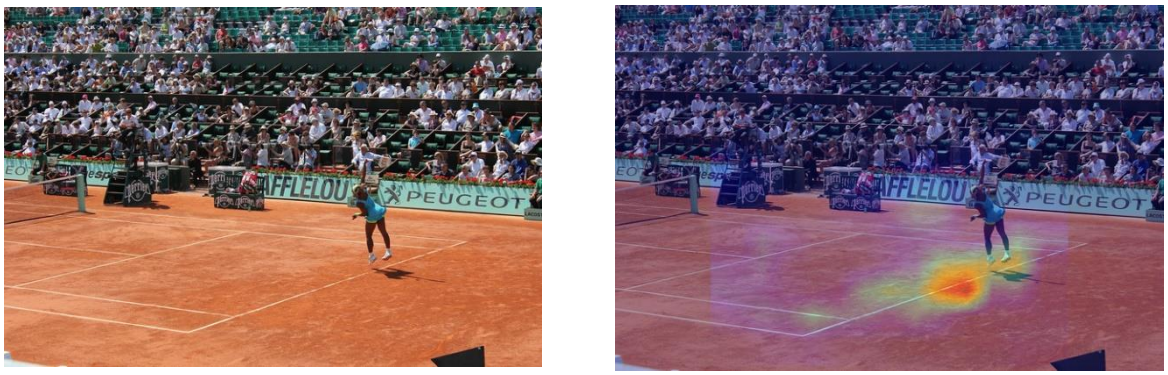


Figure 1.2 Attention maps generated based on the input features in the XVQA model.

Users try to understand how the model analyzes an image based on the question by looking at these attention maps (Figure 1.3).

What sport is the woman playing?



What are the kids riding?

What are the girls skating inside of?

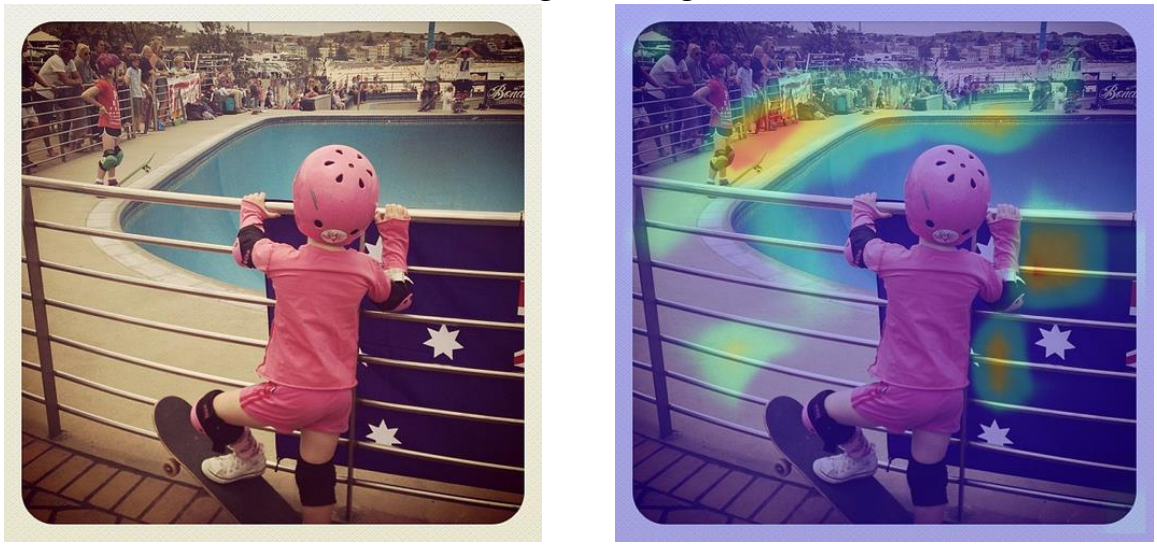


Figure 1.3 Spatial attention explanations (Left: Original image, Right: Spatial attention).

1.5.2 Active Attention

Our model provides this explanation mode for the users within a feedback loop. Users can utilize this feature to *alter* a model's attention map to *steer* the model's attention and the way the answer is generated. In this feedback loop, users first see the model's answer based on the original attention map, and then they modify the attention to create a different response. Users provide the alternative attention by drawing it over the image in the interface. This active attention is then normalized and applied as an intervention in the feedback inference process:

$$x_c^{feedback} = \sum_{w=0}^{W} \sum_{h=0}^{H} \alpha_{w,h}^{act} \varphi_{w,h} \tag{1.6}$$

$$P(a_i^{feedback} | I, q, \alpha_{w,h}^{act}) \propto G(x^{feedback}, s) \tag{1.7}$$

The active attention trial has a two-step task to complete. The first step is similar to spatial attention trials, where users make predictions based on the attention map generated by the VQA model. The subject then observes the prediction results and realizes whether the system is accurate or not. The subjects are asked to draw a new attention map in the second step. Using the manually drawn attention map, the model processes the image and question one more time and produces a second answer.

In the feedback loop, the model directly multiplies the user-generated attention map into the image feature map (Eq. 1.6 and Figure 1.4). This operation accentuates the image features in the highlighted areas and mitigates the features in irrelevant image regions.

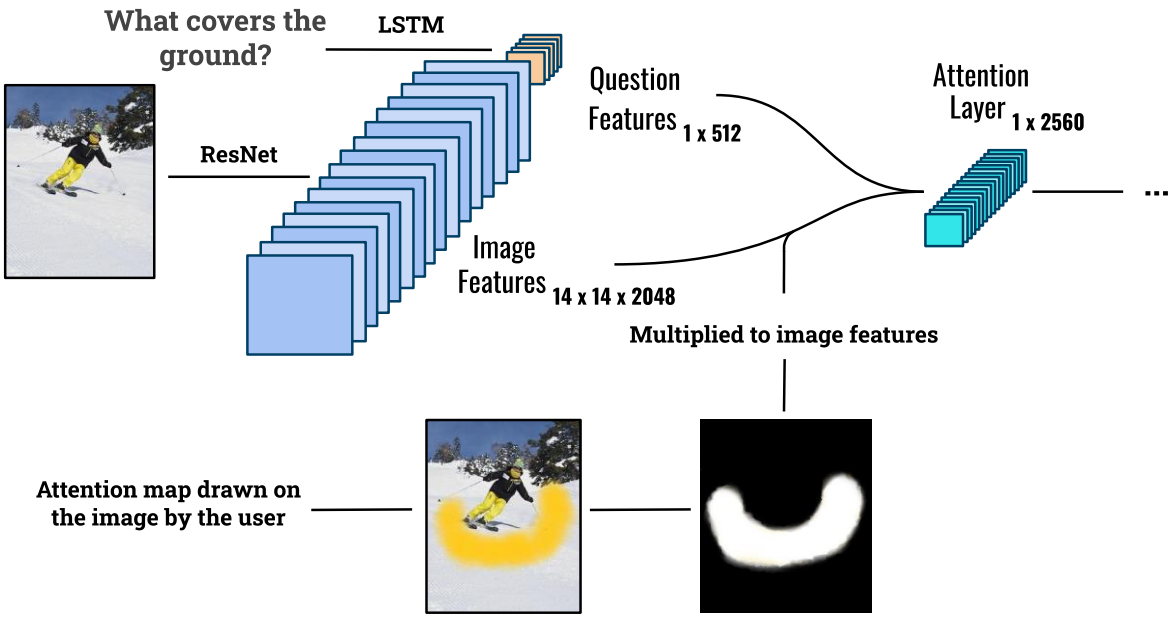


Figure 1.4 The architecture of the active attention loop within the XVQA model.

This operation aims to allow the subject to get involved in the inference process and provide feedback to the model interactively. In cases where the model answers the questions incorrectly, subjects attempt to correct the model's response by drawing the attention map. Otherwise, for those cases where the model is already accurate, subjects try to create a different answer by altering the attention map.

1.5.3 Bounding Boxes

The bounding boxes in this model are generated based on the annotations in the Visual Genome dataset and can carry important information about the scene. A combination of the attention maps created by the model and these annotations can produce explanations of the system behavior on a conceptual level.

Figure 1.5 Bounding box explanations generated based on spatial attention weights for different questions.

We calculate the average attention weight of the bounding boxes in the image based on the spatial attention maps and keep the top K ($K = 5$ in our studies) bounding boxes. Hence, these bounding boxes indicate the most related objects in the scene contributing to the system's answer (Figure 1.5).

1.5.4 Scene Graph

The bounding box annotations are completed by the scene graph information, which illustrates the relationships between different objects in the scene. The connections are in *subject-predicate-object* phrases and can indicate object attributes or interactions. In the Visual Genome (VG) dataset, the object labels, their bounding boxes, and the scene graph connecting them provide a structured, formalized representation of components in each image (Krishna et al., 2017). For each question, we filter objects in the scene graph based on the attention weights of their bounding

boxes (Figure 1.6). The users can interactively locate active objects of the scene graph and see their bounding boxes in the input image.

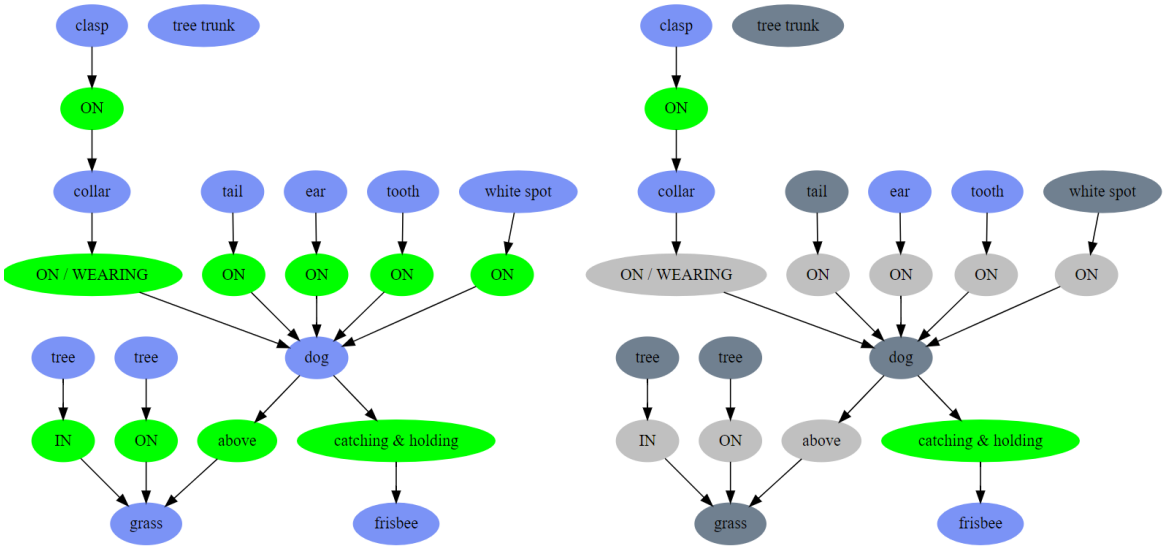


Figure 1.6 **Left:** input scene graph. **Right:** scene graph filtered based on the attention map weights generated by the model in response to a question.

1.5.5 Object Attention

Inspired by previous work (Ray, Burachas, et al., 2019), we added a MASK-RCNN image encoder to our model to produce explanations at the object level. This encoder is used explicitly by the XVQA model, as the VQA model still uses the ResNet encoder to deliver answers.

The model creates object attention masks to highlight objects with more significant contributions to the inference process based on attention modules. As opposed to spatial attention explanations, object attention can segment certain entities in the scene to illustrate a more meaningful explanation for system answers (Figure 1.7). For more details on the implementation of this technique, please refer to (Ray, Burachas, et al., 2019).

What color is the train?

What is this person holding?

What type of room is this?

Figure 1.7 Comparison between spatial attention (**Left**) and object-level attention (**Right**).

1.5.6 Textual Explanation

We also integrate natural language explanations with visual explanations in our XVQA system. Our technique is derived from the work done by (Ghosh et al., 2019), which uses the annotations of entities in an image (extracted from the scene graph), and the attention map generated by a VQA model while answering the question.

For a given question-image pair, our textual explanation module uses the visual attention map to identify the most relevant parts of the image. The model then retrieves the bounding boxes of entities that highly overlap with these regions.

The model eventually identifies the most relevant entities to the answer based on their spatial relevance to the image and NL representation. The region descriptions for the most pertinent entities form the textual explanations. Figure 1.8 depicts a sample output obtained by this approach.

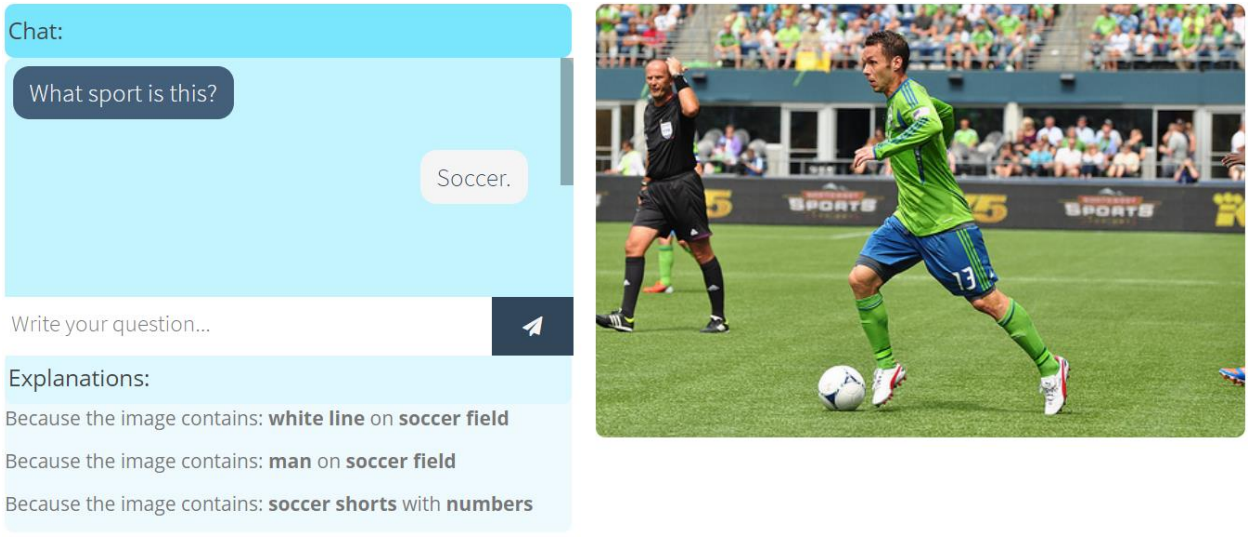


Figure 1.8 Sample results from the NL module producing a textual explanation for the model's answer.

1.6 Experimental Design

For a careful evaluation of all mentioned explanation modes, we implement an interactive interface where users can participate in a user-machine prediction task. The test starts with an introduction section and continues in the form of a series of trials where the task in each trial is to estimate the VQA system's answer correctness.

Within the introduction section, the subjects are also informed of their interaction with an AI system without any implications of its accuracy to avoid any prior bias in their mental model

of the system. The subjects are also provided with instructions to perform the tasks and work with the interface effectively.

1.6.1 User task

On each trial, users enter their prediction of whether they think the system would answer correctly or not and then declare their confidence level in their answer on a Likert scale. Afterward, the subjects view the ground truth, AI's top-five answers, and their probabilities in order. The system also provides its overall confidence/certainty based on normalized Shannon entropy of the answer probability distribution.

To prevent the effect of fatigue on performance in groups with longer trials, the test for each subject is limited to a one-hour session. Participants are asked to go through as many trials as possible within that period.

1.6.2 Trials

There are two types of trials in the experiment: no-explanation trials and explanation trials. In no-explanation trials, subjects estimate AI accuracy only based on the input image and question.

In explanation trials, the subjects first see the inputs and system's explanations. Before estimating the correctness of AI's answer, subjects are asked to rate each explanation's helpfulness in better predicting AI accuracy. At the end of each explanation trial, subjects rate their reliance on the explanations to predict AI accuracy. Figure 1.9 depicts our evaluation system's order of actions in a trial.

Each test session starts with a practice block consisting of two trials. The practice trials are only purposed to familiarize the subjects with the test flow and are not considered in any of the results. The rest of the test is carried out in blocks with five trials.

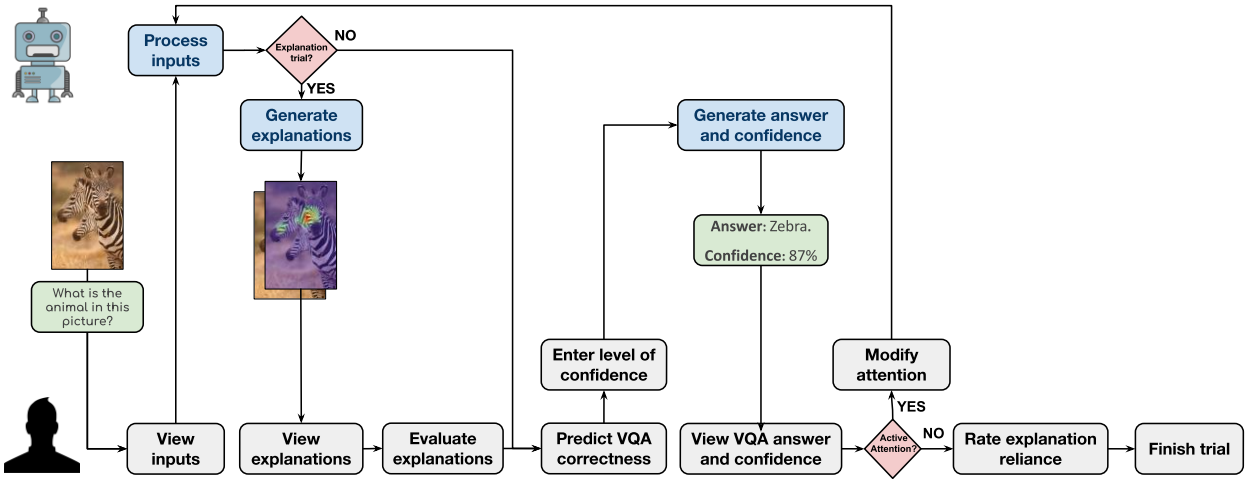


Figure 1.9 Flowchart for a prediction evaluation task. The "Explanation trial?" conditional defines the type of trial as either explanation or control. The "Active attention?" conditional activates the feedback loop in the case of active attention explanations.

1.6.3 Study Groups

The study involves six groups of participants. The control group (NE) does not see any explanation modes, so its task is reduced to predicting the system's correctness in trials. The explanation groups are exposed to either one or a combination of explanation modes before they predict the system's answer.

The control group (NE) only sees a block of no-explanation trials throughout the whole test. The blocks toggle between explanation and no-explanation modes for the groups with explanation modes. The no-explanation blocks in explanation groups act as control tests to assess prediction quality and mental model progress as the users see more trials (Figure 1.10).

The explanation blocks view the explanations generated by the model before the users make their predictions and show the system's answer and the system's confidence afterward. The no-explanation blocks only ask for the user's prediction without exposing any explanations beforehand.

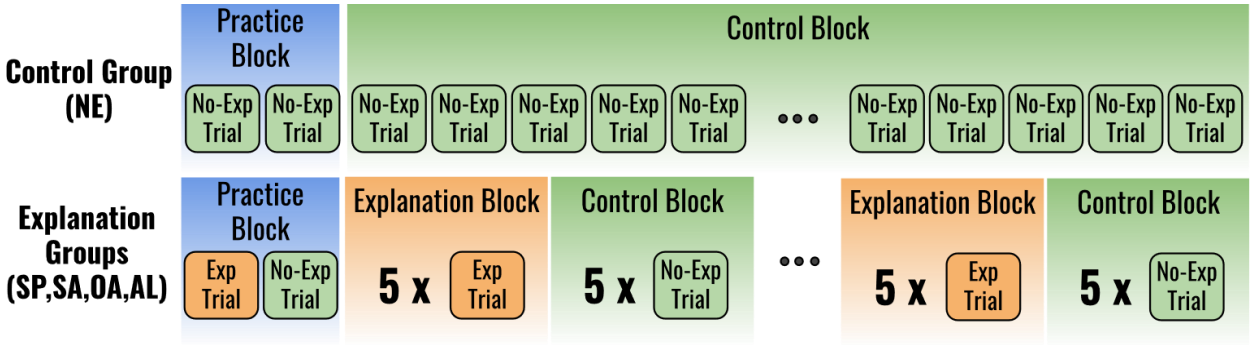


Figure 1.10 Structure of the test sessions in the control group (NE) and explanation groups.

Group SA has an interactive workflow within which subjects first go through the spatial attention explanation and then modify the attention in a feedback loop. Each explanation group is dedicated to a specific explanation mode, except group SE which combines bounding box, scene graph, and textual explanations. The study was conducted with 90 participants and a total number of more than 10,000 trials. Table 1.1 shows the number of participants and the number of trials in each group.

Table 1.1 User study statistics for multi-modal explanations

	Group	Subjects	Trials
NE	Control group	15	4124
SP	Spatial attention	15	1826
SA	Active attention	15	1021
SE	Semantic	15	1261
OA	Object attention	15	1435
AL	All explanations	15	846
	Total:	90	10513

A total of 3969 image-question pairs were randomly selected from the overlap of the VG dataset (Krishna et al., 2017) and VQA dataset (Goyal et al., 2017) to be used in the trials. The

questions asked on each trial are selected from the VQA dataset, and the annotations used in generating the explanations are extracted from the VG dataset. In the selection, all yes-no and counting questions were excluded to draw the focus of the test to non-trivial questions and less obvious answers with higher levels of detail in explanations.

1.7 Results

After assigning different groups of participants to specific combinations of explanations (including a control group that received no explanations) and having them perform the VQA prediction task, we evaluated different hypotheses about the explanations' impact on various aspects of human-machine task performance. The results are either based on the average of all trials within certain groups or the progress throughout the tests. Since each group and trial's task can be different from other groups and trials, the number of trials finished by subjects varies between groups and even within groups.

1.7.1 Impact on User-machine Task Performance

The first metric we used to assess user-machine task performance is the user's accuracy for predicting the machine's correctness and whether explanations affect this. We tested for any effect (positive or negative) between accuracy and the presence of explanations using a chi-squared test. The results from different groups show an overall accuracy increase in all explanation groups compared to the control group; however, this is statistically significant only for cases where the system's answer is wrong (Figure 1.11).

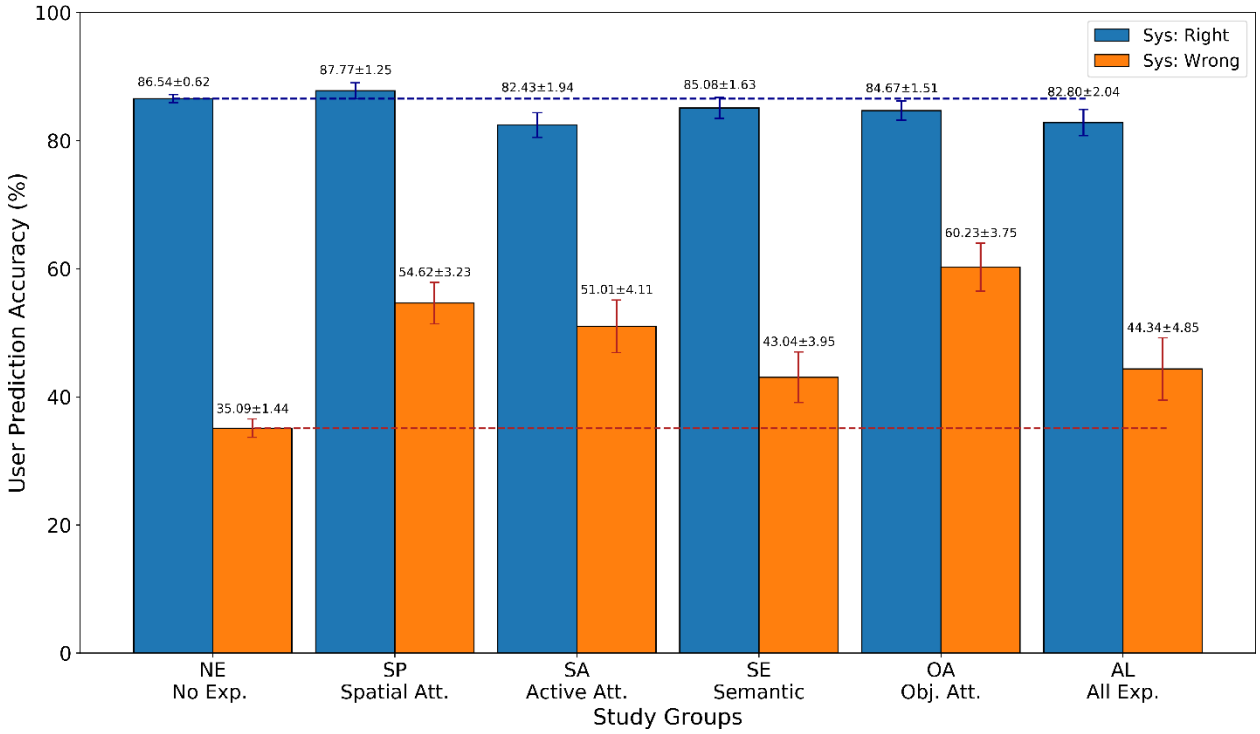


Figure 1.11 The average values of user's prediction accuracy (user performance) compared between different groups (system: right $p = 0.061$, system: wrong $p < 0.0001$, overall $p = 0.0001$).

The progress of prediction accuracy is also another metric to quantify subjects' progress in understanding and predicting systems behavior. As subjects go through trials in different groups, we compare the improvement of their mental model based on their prediction accuracy (Figure 1.12-left). Whether the system is right or wrong, the subjects in explanation groups show a steadier improvement in their prediction accuracy.

1.7.2 User Explanation Helpfulness Ratings

Before predicting the VQA model's answer, users rate each explanation mode based on how much it helped them in the prediction task, a rating that we call "explanation helpfulness." Comparing these helpfulness ratings with the users' prediction accuracy reveals a positive correlation with accuracy improvement (accuracy after minus accuracy before) and helpfulness of explanations, but only in cases where the system is *right*. Figure 1.13(right) implies that when

users find explanations helpful, they do better on the prediction task. On the other hand, a higher rating for explanations when the system is wrong (Figure 1.13-left) has lower human prediction accuracy. This observation shows the influential role of explanations in the process of decision-making for users.

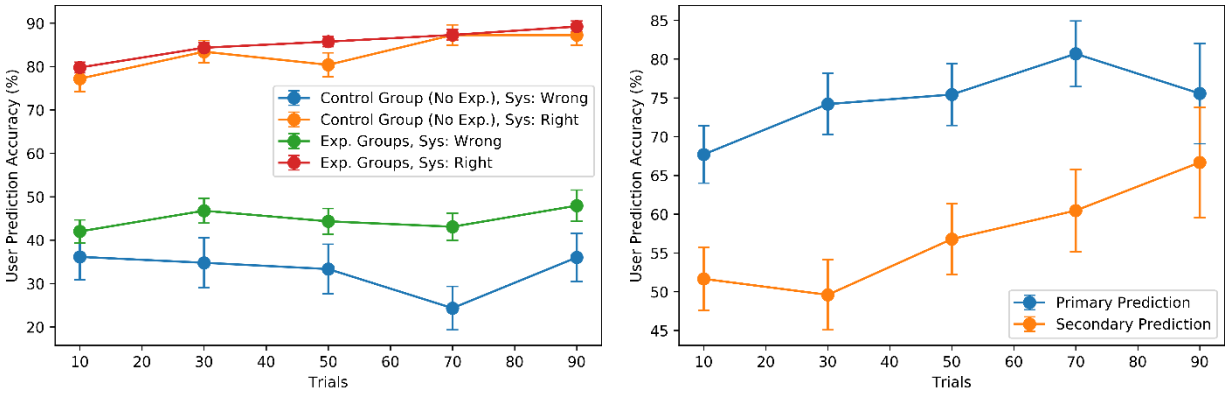


Figure 1.12 **Left:** The user's prediction accuracy progress compared between the control group and all explanation groups. The results are separated based on the accuracy of the system. **Right:** Prediction accuracy progress in the Active Attention explanation group. Primary prediction is made based on the model's original attention map, and secondary prediction is based on the modified attention map that the subject provides.

1.7.3 Active Attention Explanation

Within-group SA, subjects view and interact with active attention explanations before making their prediction. Like spatial attention, users first make a prediction based on the attention map created by the VQA model. In the second step, subjects draw a new attention map for the model to change AI's answers. Subjects can compare their attention with the model's attention and answer that is created based on each of them.

Figure 1.12(right) illustrates the trend of prediction accuracy progress as subjects interact with active attention maps. While the explanation helps subjects improve their primary prediction of AI correctness, they also substantially improve predicting AI when working with their modified attention (secondary prediction).

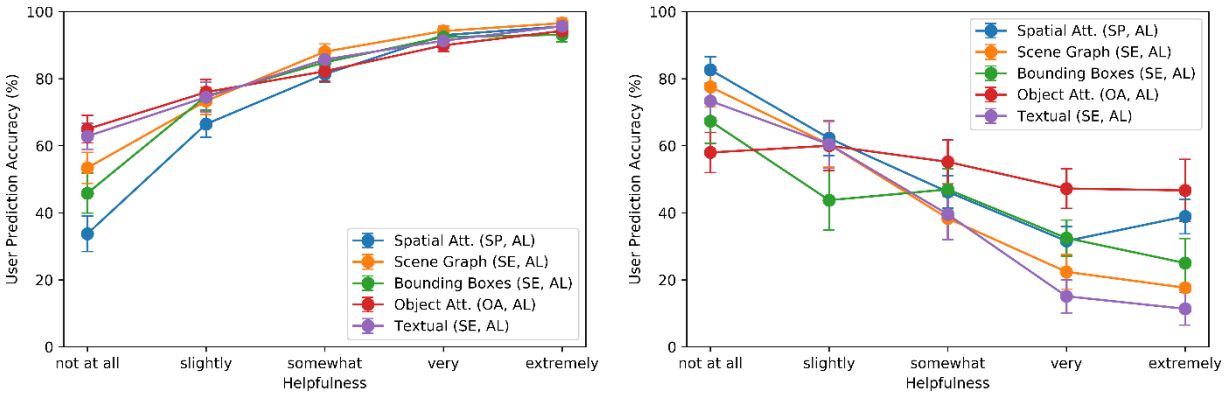


Figure 1.13 The average values of user's prediction accuracy for each explanation mode vs. user's ratings on explanations' helpfulness for cases where the system is right (**left**) and where the system is wrong (**right**). Overall $p < 0.0001$.

1.7.4 Impact of Active Attention on User Confidence

Active attention explanation provides users with a feedback loop to modify the system's attention and see the result of attention changes in the model's answer. In trials with active attention explanation, users make two predictions: one based on the original spatial attention provided by the user and a secondary prediction after they modify the attention map. We consider the accuracy of the primary prediction as an indicator of the user's mental model state. The secondary prediction is more specifically dependent on users' general mental model of the attention map.

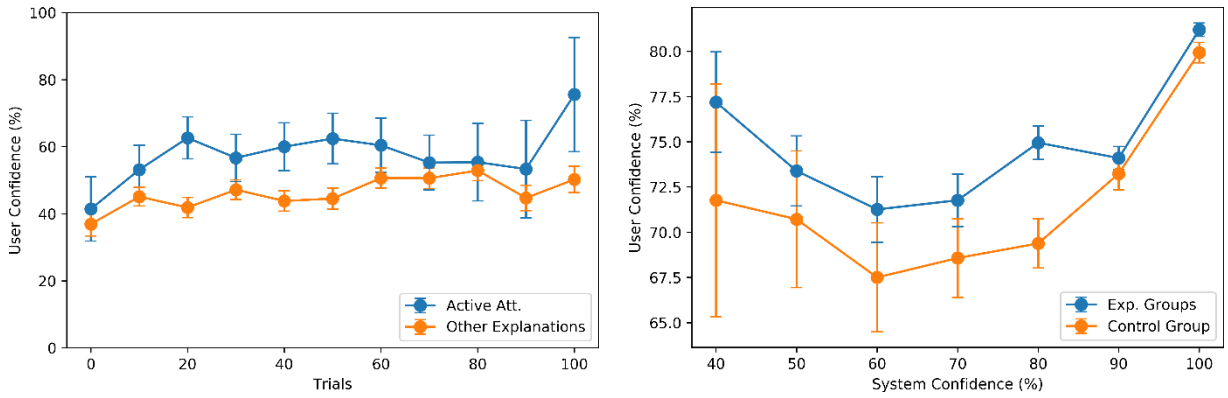


Figure 1.14 **Left:** User confidence progression comparison between the active attention group and other groups. **Right:** User confidence in prediction vs. system confidence in the answer.

Comparing results from different explanation groups with the active attention group shows that users in the active attention group have higher average confidence in their primary predictions than other explanation groups (Figure 1.14 left).

While the increase in user confidence points out the confidence and trust built by the active attention explanation, the average prediction accuracy in this group of participants is lower than in other groups. These results suggest a higher potential for this technique to produce real insight into the model if used in multiple feedback loops instead of just one.

1.7.5 Impact on Trust and Reliance

Another essential purpose of explanation systems is to create user trust in AI machines so that they can rely on the outcome of the system.

Our user study asks users about their level of reliance (in the Likert scale) on the explanation section while predicting AI performance. Comparing users' reliance concerning their performance indicates a correlation between the reliance and users' accuracy in those cases when the system is wrong (Figure 1.15 Left).

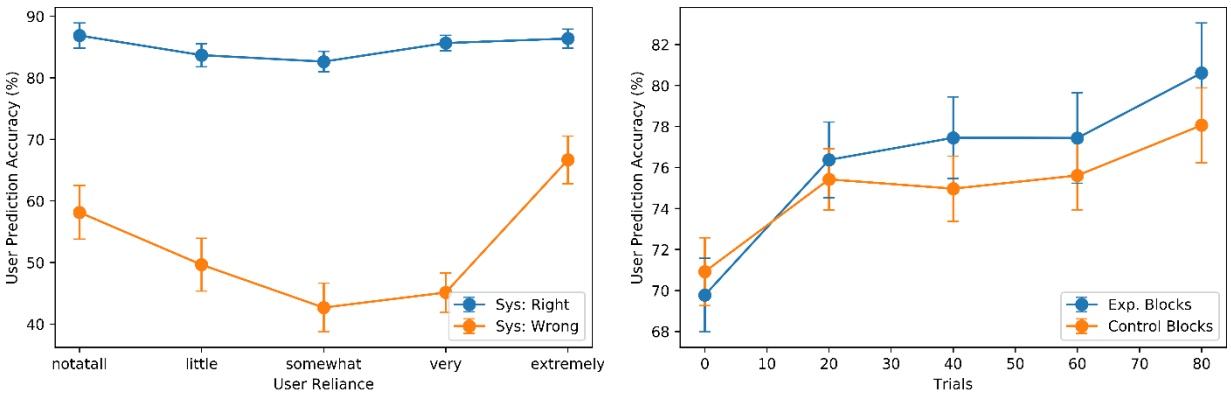


Figure 1.15 **Left:** Users' prediction accuracy vs. their reliance on explanation divided by the system's accuracy. **Right:** Prediction accuracy growth in explanation groups compared between exp. Blocks and no-exp. Blocks.

Moreover, users declare their confidence level in their prediction on a Likert scale. Generally, we can assume the users' confidence level in their prediction as a function of user

confidence in the system and the system's confidence in its answer. In the control group with no explanations, confidence mainly stems from system performance in previous trials (mental model). In contrast, in other groups, the explanations directly affect the level of confidence.

Figure 1.14 (right) shows average user confidence compared with system confidence (provided to users after they make their predictions) in those cases when the user's prediction is correct. The results indicate a consistent increase in user confidence when exposed to explanations against the control group with no explanations.

1.7.6 Impact of Explanation Goodness

As mentioned earlier, users go through blocks of trials in explanation groups. To assess the goodness of explanations in helping users predict systems to answer, each block of trials with explanations is followed by a block without explanations. Comparing the user prediction accuracy between these blocks illustrates the progress of users' mental models in the presence of explanations (Figure 1.15 Right). Results indicate that users have built a better mental model to predict the system within explanation blocks and made progress in understanding system answers.

1.8 Discussion

The overall assessment of user performance reveals a substantial improvement in prediction accuracy in the presence of explanations while the system is not correct. Users also provide higher ratings for the explanations when they perform better and vice versa. This direct correlation in all explanation modes strongly suggests the effectiveness of these explanations within the prediction task.

Among different modes, object attention (group OA) helped the users the most by a more precise segmentation of images and labeling of the relevant entities. Similarly, providing the labels

and bounding boxes of objects and semantic explanations (SE) yet did not offer the same level of insight as Object attention (OA). The difference in helpfulness between these two groups indicates how the coherency within the presented explanations can over-shadow the amount of information.

Although the subjects viewed all explanation modes in group AL, we do not see a higher accuracy level than other groups. The feedback from the post-study interviews pointed out two possible reasons for such observation: 1) the overwhelming amount of information in the group decreased the performance level of the subjects; 2) those cases where explanation modes conflicted with each other confused some of the subjects.

In cases where the system is right, the explanation groups do not reach the same accuracy level in prediction as to the control group. One reason can be the strong mental model shaped in the final trials of the control groups as the subject can traverse more trials compared to other groups. A closer look at the early trials in all groups reveals a higher accuracy for explanation groups in all cases (Figure 1.12 Left).

The prediction accuracy progress under active attention explanations shows early improvement in primary and secondary predictions. On the other hand, users show higher levels of confidence when exposed to active attention within explanation groups. Nonetheless, the active attention group (SA) does not yet exceed the spatial attention group (SP). This drawback can be that active attention's limited to only visual features and not question features. Possibly, multiple feedback loops can help users better understand the role of image features as only one of (and not all) the contributors in the final answer.

In cases where the system is wrong, users' accuracy correlates with users' reliance. The subjects seem to do well either when they are significantly relying on the explanations or completely ignoring them. For those cases where the users ignore the explanations, post-study

interviews suggest that the subjects decide based on their mental model of the system and previous similar trials.

1.9 Conclusion

We designed an interactive experiment to probe explanation effectiveness in improving user prediction accuracy, confidence, and reliance in the context of a VQA task. The results of our study show that the explanations help improve VQA accuracy, and explanation ratings approve the effectiveness of explanations in human-machine AI collaboration tasks.

To evaluate various modes of explanations, we conducted a user study with 90 participants. Users interactively rated different explanation modes and used them for predicting AI system behavior. The user-machine task performance results indicate improvements when users were exposed to the explanations. User confidence in predictions also improved when they viewed explanations that display the potential of our multi-modal explanation system in building user trust.

The strong correlation between the users' rating on explanation helpfulness and their performance in the prediction tasks shows the effectiveness of explanations in the user-machine task performance. Those explanations identified as more helpful helped users in cases where the system was accurate. On the other hand, in cases where the system was inaccurate, those explanations ranked as more helpful became more misleading.

We also introduced an interactive explanation mode (Active attention) where users could directly alter the system's attention and receive feedback. Comparing the user confidence growth between Active attention and other explanation groups shows a higher level of trust from the users, which shows the effectiveness of interactive explanations in building a better mental model of the AI system.

As a future direction, we may investigate other interactive explanation modes to maximize the performance in human-machine tasks. On the other hand, user feedback and ratings for the different modes explored in this study can guide us towards more effective explanation models in XAI systems.

1.10 Acknowledgments

Chapter 1, in part, is a reprint of the material as it appears in A study on multimodal and interactive explanations for visual question answering. Alipour, Kamran; Schulze, Jurgen P.; Yao, Yi; Ziskind, Avi; Burachas, Giedrius. In SafeAI workshop at AAAI conference (2020). The dissertation author was the primary investigator and author of this paper.

CHAPTER 2 EXPLAINING AI COMPETENCY

2.1 Abstract

Explainability is one of the critical elements for building trust in AI systems. Among numerous attempts to make AI explainable, quantifying the effect of explanations remains a challenge in conducting human-AI collaborative tasks. Aside from the ability to predict the overall behavior of AI, in many applications, users need to understand an AI agent's competency in different aspects of the task domain. In this chapter, we evaluate the impact of explanations on the user's mental model of AI agent competency within the task of visual question answering (VQA). We quantify users' understanding of competency, based on the correlation between the actual system performance and user rankings. We introduce an explainable VQA system that uses spatial and object features and is inspired by the transformer technology utilized in the BERT language model. Each group of users sees only one kind of explanation to rank the competencies of the VQA model. The proposed model is evaluated through between-subject experiments to probe explanations' impact on the user's perception of competency. Comparing two VQA models shows transformer-based explanations, and the use of object features improves the user's prediction of the model's competencies.

2.2 Introduction

Recent developments in AI, specifically deep neural networks (DNN), have brought them into various applications. DNNs have automated a wide range of human activities resulting in reduced complexity of many tasks. Users of AI systems, though, need to maintain at least a

minimal level of understanding and trust in the system, i.e., they need a proper mental model of the system's internal operations for anticipating success and failure modes.

While accuracy is well-known as the primary metric for AI efficiency, it cannot guarantee a collaborative human-machine interaction in the absence of trust. If the users do not trust a model or a prediction, they will not use it (Ribeiro et al., 2016). This mistrust escalates in the presence of adversarial attacks where imperceptible changes to the input lead to wrong outputs and the susceptibility of DNNs to non-intuitive errors.

Explainable AI aims to gain users' trust in two significant steps: interpretability and explainability. Interpretable models provide a basic comprehension of their inner processes through visual or textual cues. On a higher level, explainable models attempt to give reason and causality behind their decisions (Gilpin et al., 2018).

The appearance of various methods of explanations calls for a parallel effort to evaluate and quantify their efficiency. While previous works introduce nominal visualizations and textual justifications of the inner features of DNN models, they do not assess the impact of explanations on various aspects of users' understanding and trust.

Evaluation techniques for explanations include automatic and human-based methods. Automated approaches provide quantifiable measures over relevant benchmarks, e.g., alignment with human attention datasets (Das et al., 2017); however, they still cannot propose a straightforward metric for trust in actual human-machine tasks.

Furthermore, human-based approaches attempt to quantify explanation effectiveness by collecting user ratings (Chandrasekaran et al., 2017; Lu et al., 2016). Despite their insightful results, these methods do not measure the user's perception of AI competency in the entire domain.

Users can benefit from AI systems more if they are familiar with the AI agent's competency in the operational domain. The biases can impact the competency of AI in the training data or the limited representation of crucial features. An explanation system that provides case-by-case reasoning for AI behavior does not automatically produce a higher view of competency. Notably, deep learning models are notoriously opaque and difficult to interpret and often have unexpected failure modes, making it hard to build trust.

As our prior work showed, explanations improve user prediction of system accuracy (Alipour, Schulze, et al., 2020). This chapter focuses on the user's mental model of an AI system in terms of competency understanding. Specifically, we evaluate the importance of explanations to help users interpret how a VQA system performs with different types of questions. We model the users' learning process under two explanation systems to identify the role of the attention-based explanations in the users' prediction of competency. For this purpose, we evaluate the impact of explanations on user learning rate and their ultimate score on the task of competency prediction.

2.3 Related Work

Visual question answering (VQA). Initially introduced by Antol *et al.*(Antol et al., 2015), the VQA problem involves answering questions about the visual content of an image. The VQA task is explicitly challenging due to the complex interplay between the language and visual modalities (Zhang et al., 2019). Limited labeled data and the complex feature space complicate developing VQA models. These challenges result in inconsistent outputs and profound logical contradictions (Ray, Sikka, et al., 2019). In such an environment, the choice of hyper-parameters

and architectural designs can drastically impact the performance of VQA models (Teney et al., 2017).

A common approach to VQA is to use DNNs with attention layers that select specific regions of the image, guided by the question for inferring an answer (Fukui et al., 2016; Teney et al., 2018; Xu & Saenko, 2016). Herein, we also study two attention-based VQA models with different attention structures. As a baseline, we use a model based on the approaches of Kazemi and Elqursh (Kazemi & Elqursh, 2017) and Teney *et al.(Teney et al., 2018).* We propose a new VQA architecture by replacing the attention mechanism with a transformer similar to the BERT model (Devlin et al., 2018) in the baseline VQA model.

Previous work in VQA includes various attempts to optimize the attention mechanism. To improve the attention to the question, Lu *et al.*(Lu et al., 2016) utilize a co-attention model to reason about images and questions on hierarchical levels jointly. Anderson *et al.*(Anderson et al., 2018) propose a combined bottom-up and top-down attention mechanism to calculate attention at the level of objects. The model is further upgraded and fine-tuned to win the VQA Challenge 2018(Jiang et al., 2018).

Despite all the advancements in the overall accuracy of VQA models, their unbalanced performance in different aspects of the task is overtly noticeable. Some prior approaches address this issue by focusing on tasks such as reading text in images (Singh et al., 2019) or counting objects (Zhang et al., 2018). Other studies introduce new datasets to reduce bias (Goyal et al., 2017) or to enforce the logical consistency of models through *visual commonsense reasoning (VCR)* for challenging questions (Zellers et al., 2019).

Explainable AI (XAI). The ever-increasing complexity of the modern AI machine demands a trustable source of explanation for all AI users. Generating automated reasoning and

explanations dates to very early work in the AI field with direct applications from medicine (Shortliffe & Buchanan, 1984) and education (Lane et al., 2005; Van Lent et al., 2004) to robotics (Lomas et al., 2012). In computer vision, several explanation systems focus on the importance of image features in the decision-making process (Hendricks et al., 2016; Jiang et al., 2018; Jiang et al., 2017; Zeiler & Fergus, 2014).

AI explanations for the task of visual question answering usually include image and language attention (Kazemi & Elqursh, 2017; Lu et al., 2016). Besides saliency/attention maps, other studies investigated different explanation modes like layered attention (Yang et al., 2016), bounding boxes around important regions (Anne Hendricks et al., 2018), textual justifications (Huk Park et al., 2018; Shortliffe & Buchanan, 1984), or a combination of these modalities (Alipour, Schulze, et al., 2020).

We propose an explainable VQA system that produces justifications for system answers in the form of an attention map. Unlike previous post-hoc saliency approaches such as GradCAM (Selvaraju et al., 2017), our method seeks causal explanations by providing attention as an inherent step of answer inference. Our proposed model uses visual features on both the spatial and object level. For better performance in a VQA task, our proposed model utilizes the transformer mechanism to process question features and visual features.

Explanation evaluation. As AI enters our daily lives, a new interest has surged among the AI community to make AI algorithms more understandable to regular people without knowledge of AI (Drozdal et al., 2020). In this study, we choose the subjects for explanation evaluation from a group of individuals with minimum knowledge about AI and deep neural networks.

The XAI literature widely discusses the impact of explanations on user mental models and human-machine performance. Some of the earlier work quantifies the efficacy of explanations

through user studies to assess the role of explanations in building a better mental model of AI systems for their human users.

Some of the previous studies introduce metrics to measure trust with users (Cosley et al., 2003; Ribeiro et al., 2016) or the role of explanations in achieving a goal (Kulesza et al., 2012; Narayanan et al., 2018; Ray, Burachas, et al., 2019). Dodge *et al.*(Dodge et al., 2019) investigated the fairness aspect of explanations through empirical studies. Lai and Tan (Lai & Tan, 2019) assessed the role of explanations in user success within a spectrum from human agency to full machine agency. Lage *et al.*(Lage et al., 2019) proposed a method to evaluate and optimize the human interpretability of explanations based on measures such as size and repeated terms in explanations. Other approaches measured the effectiveness of explanations in improving the predictability of a VQA model (Alipour, Schulze, et al., 2020; Chandrasekaran et al., 2018).

Unlike prior approaches, our work is focused on evaluating human agents' knowledge of AI competency. Specifically, we are interested in the user's mental model of AI performance in different aspects of the VQA task. We conduct a user study to investigate the impact of explanations on the user's mental model of system competency. In our study, subjects rank system performance among different input questions. The results indicate a positive influence on the accuracy of the user's mental model in the presence of explanations. We show the overall and temporal effect of the explanations on the user's interpretation in two explainable VQA models.

2.4 Method

Our approach aims at evaluating the role of attentional explanations in the user's mental model of AI competency. We compare two explainable VQA models and test them through user studies to accomplish this task. This section covers the architecture details for these VQA models

and the differences in their attention mechanisms. The following section continues with sample cases from explanation models and their differences.

2.4.1 Explainable VQA (XVQA) Models

Our work compares two VQA agents: spatial attention VQA (SVQA) and spatial-object attention BERT VQA (SOBERT). Both agents are trained with the VQA 2.0 dataset. SVQA is based on a 2017 SOTA VQA model with a ResNet (Szegedy et al., 2017) image encoder (Figure 2.1). The agent uses an attention mechanism to select visual features generated by an image encoder and an answer classifier that predicts an answer from 3000 candidates.

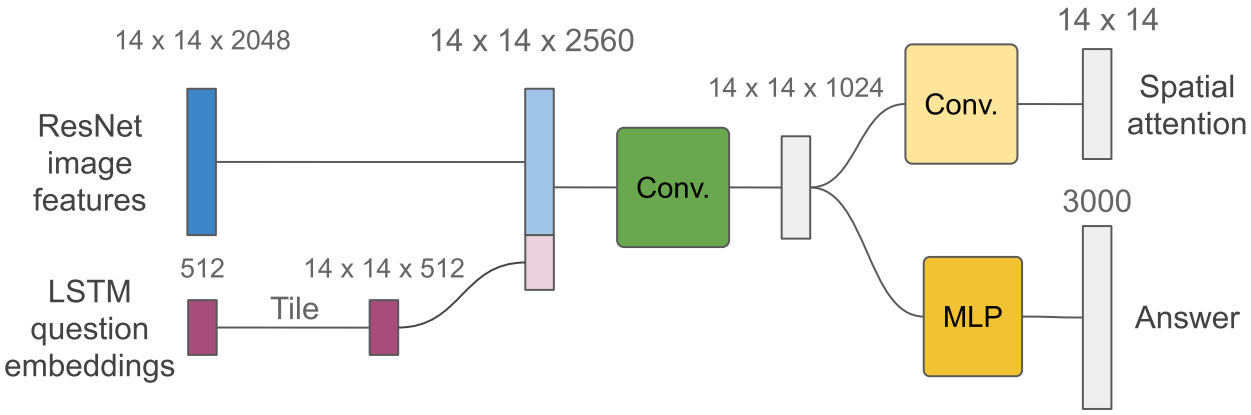


Figure 2.1 The architecture of our explainable SVQA model.

As shown in Figure 2.1, SVQA takes as input a 224×224 RGB image and questions with at most 15 words. A ResNet subnet encodes the image into a $14 \times 14 \times 2048$ feature representation. An LSTM model, GloVe (Pennington et al., 2014), encodes the input question word embeddings into a feature vector of 512 dimensions.

The attention layer in the SVQA model transfers the question and image features to a set of attention weights on the image features. The model convolves the concatenation of weighted image features and question features to produce the attention layer with $14 \times 14 \times 1024$ dimensions. The model predicts the probability of the final answer from a set of 3000 answer

choices using a multilayer perceptron (MLP). The attention layer also uses a convolution block to generate the spatial attention map.

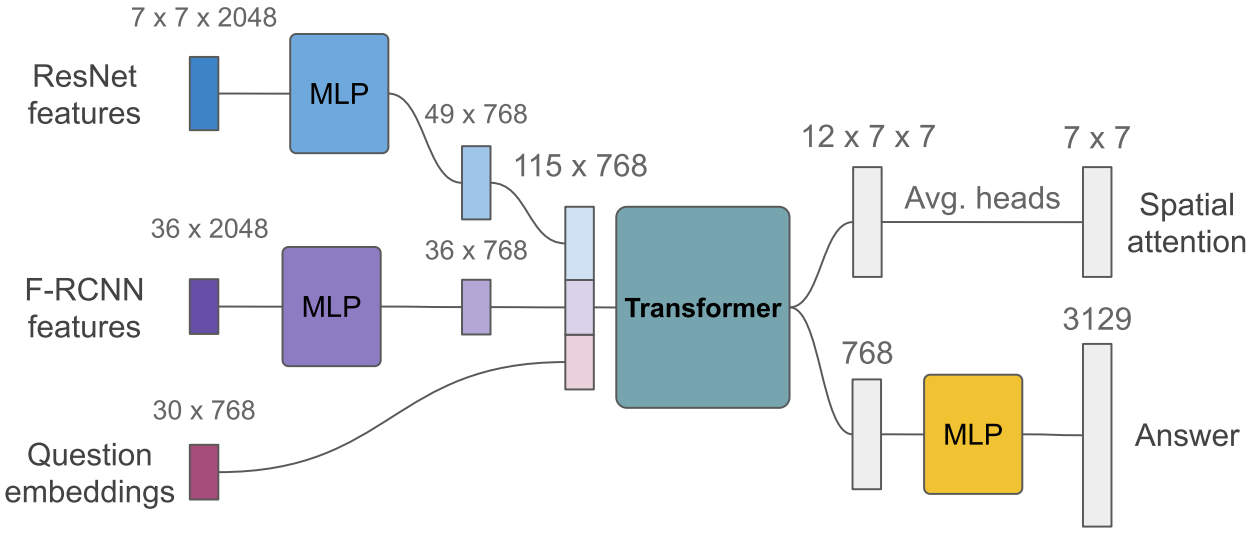


Figure 2.2 The architecture of the explainable SOBERT model. This model combines visual features from ResNet and FRCNN and question embeddings into a BERT model to produce answers and spatial attention.

On the other hand, the SOBERT agent uses a combination of visual embeddings of the image from ResNet and Faster RCNN (FRCNN)(Ren et al., 2015) alongside question embeddings (Figure 2.2). SOBERT accepts questions with a maximum length of 30 words. The input question embeddings contain location and token information of words. The location features are encoded in both ResNet and question embeddings.

The SOBERT agent employs a four-layer transformer module with 12 attention heads. This transformer converts hidden features (115×768) into spatial attention heads ($12 \times 7 \times 7$) and an output layer. An MLP translates the output layer to the final answer prediction of 3129 candidates.

Based on their training process and characteristics, VQA agents can reach certain levels of accuracy in each type of question. We limit the cases into a subset of a VQA 2.0 validation set with questions about action, attribute, object, and count for our tests. We classify the questions using automated methods, including word matching in questions and their answers. Questions about activity inside an image are labeled as "Action." Questions about objects inside the image

are labeled as "Object." Questions that are specific about attributes of entities in the image (e.g., color) are labeled as "Attribute."

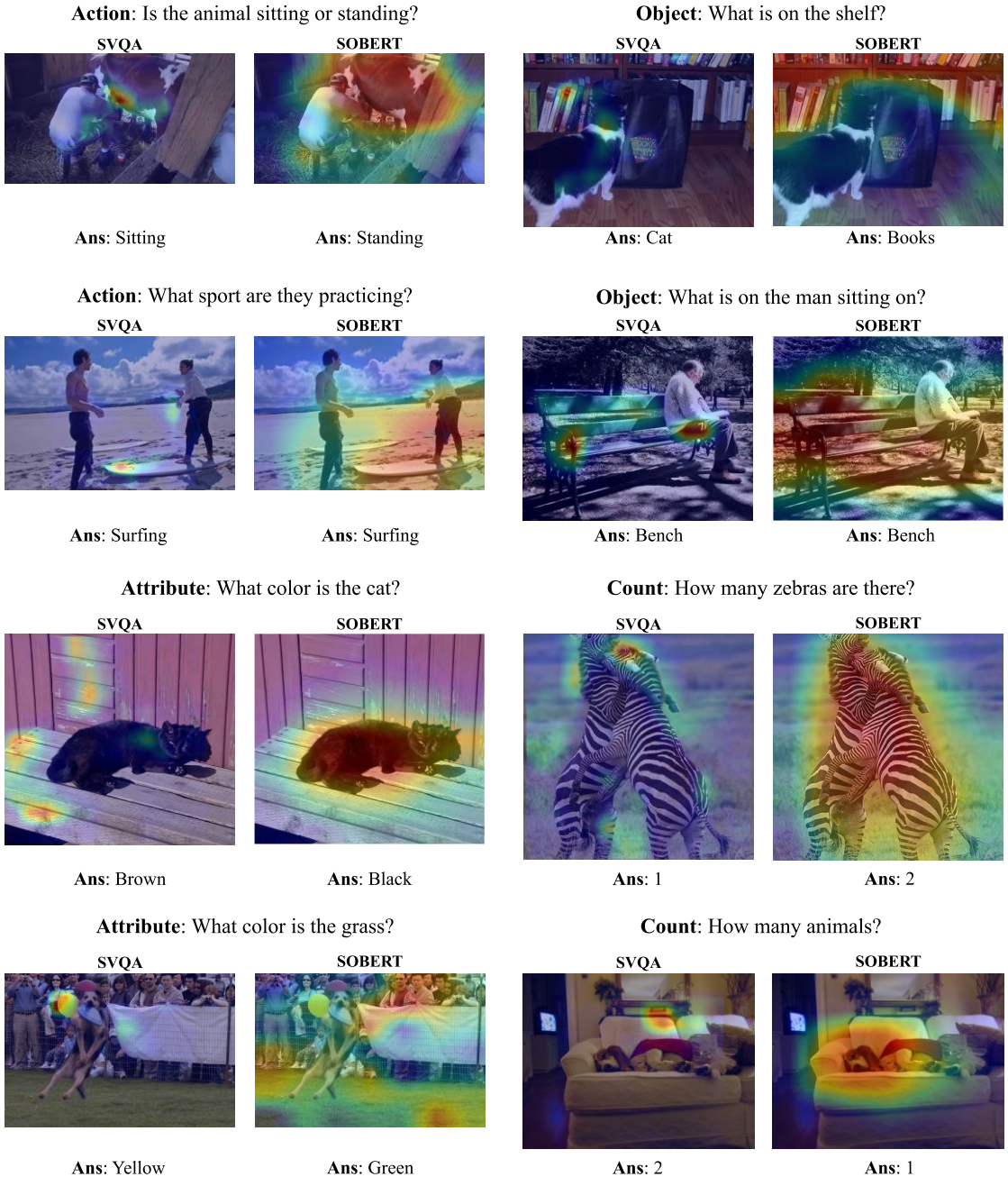


Figure 2.3 Attention maps generated by the AI agents for questions in different question type categories. As illustrated by the results, the SOBERT model produces attention maps with more focus on the areas related to the question.

Finally, questions about counting entities in the image are categorized as "Count." Table 2.1 shows the accuracy of SVQA and SOBERT agents in these four categories. The models'

accuracy is calculated across the four categories of the VQA validation dataset. As Table 2.1 shows, the two models produce a similar ranking between the four types of questions, while the SOBERT model can reach a higher accuracy than the SVQA model.

Table 2.1 The accuracy of VQA agents in four selected types of questions.

Model	Action	Attribute	Object	Count
SVQA	81.21%	70.83%	64.46%	45.78%
SOBERT	**88.35%**	**86.63%**	**71.84%**	**60.14%**

2.4.2 Explanations

The VQA agents can produce a spatial attention map to visualize focus areas while making the answer. The SVQA model convolves the attention tensor into a 14×14 spatial map. In the SOBERT model, the attention tensor is averaged over the 12 attention heads into a 7×7 spatial attention map.

The attention maps generated by the VQA agents provide a causal explanation to the users as they illustrate AI spatial/object attentions as an inherent step in answer inference. Both models use spatial features from the images while gaining a general representation of the images' content. The SOBERT model also incorporates object-level F-RCNN features into the process. One significant impact of including object-level attention emerges in the attention map outputs of the model. As shown in Figure 2.3, the attention from the SOBERT model covers broader areas associated with objects in the scene. Also, the averaging layer that generates attention produces smoother attention distributions in the SOBERT model compared to more localized and scattered attention in SVQA.

Table 2.2 The maximum learning rate of the users and the final correlation value in the competency ranking task. Both explanation models show an improvement in early learning rates. While explanations from the SOBERT model increase the learning rate as much as SVQA, SOBERT reaches a higher final learning rate.

Model	Condition	Final ranking corr.	Max. user learning rate (corr. / blocks)
SVQA	Baseline	0.757	0.0105
	Explanation	**0.805**	**0.0769**
SOBERT	Baseline	0.611	0.0253
	Explanation	**0.921**	**0.0468**

2.5 Experiments

We designed an interface for an in-person user study to evaluate the impact of explanations on the user's understanding of AI agent competency among different question types. At the introductory section of each study session, subjects are reminded that the model competency and accuracy of the AI model are unknown to minimize their prior knowledge and judgment of the AI agent competency.

In this user study, subjects go through a set of trial blocks where the AI agent answers questions about images. Each block consists of four trials with one image-question of each type: object, attribute, action, and count. In each trial, subjects first see the input image and question, and then they proceed to see the outputs of the AI agents.

The study is divided into two groups for each model: baseline and explanation. Each study group contains ten subjects, and each subject goes through 100 trials (25 blocks). In all groups, users see the agent's top five answers, their probabilities, and agents' Shannon confidence in each trial. In the explanation group, subjects first view the attention map from the model and then see

the top answers and confidence value. The subjects are asked to rank the helpfulness of the attention maps for understanding the AI's performance in that trial.

At the end of each block, subjects rank the trials within the block based on system performance for each question type. These rankings reflect the subject's opinion of AI competency. Comparing rankings between baseline and explanation groups measures the impact of explanations on subjects' mental models (Figure 2.4).

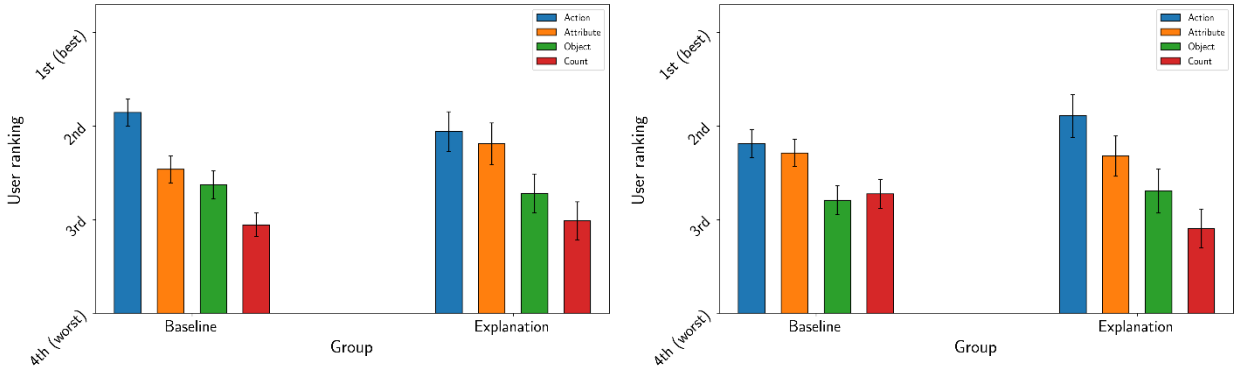


Figure 2.4 The average of all rankings entered by the subjects at the end of every block of trials (**Left**: SVQA model, **Right**: SOBERT model).

In each block of trials, four question-images show up in randomized order. The AI agent's success ratio in each block is also randomized. In the baseline group, users can only rely on the top answers and their probabilities to understand system performance for that question and image. On the other hand, subjects from the explanation groups have the extra information provided by the attention maps (Figure 2.5).

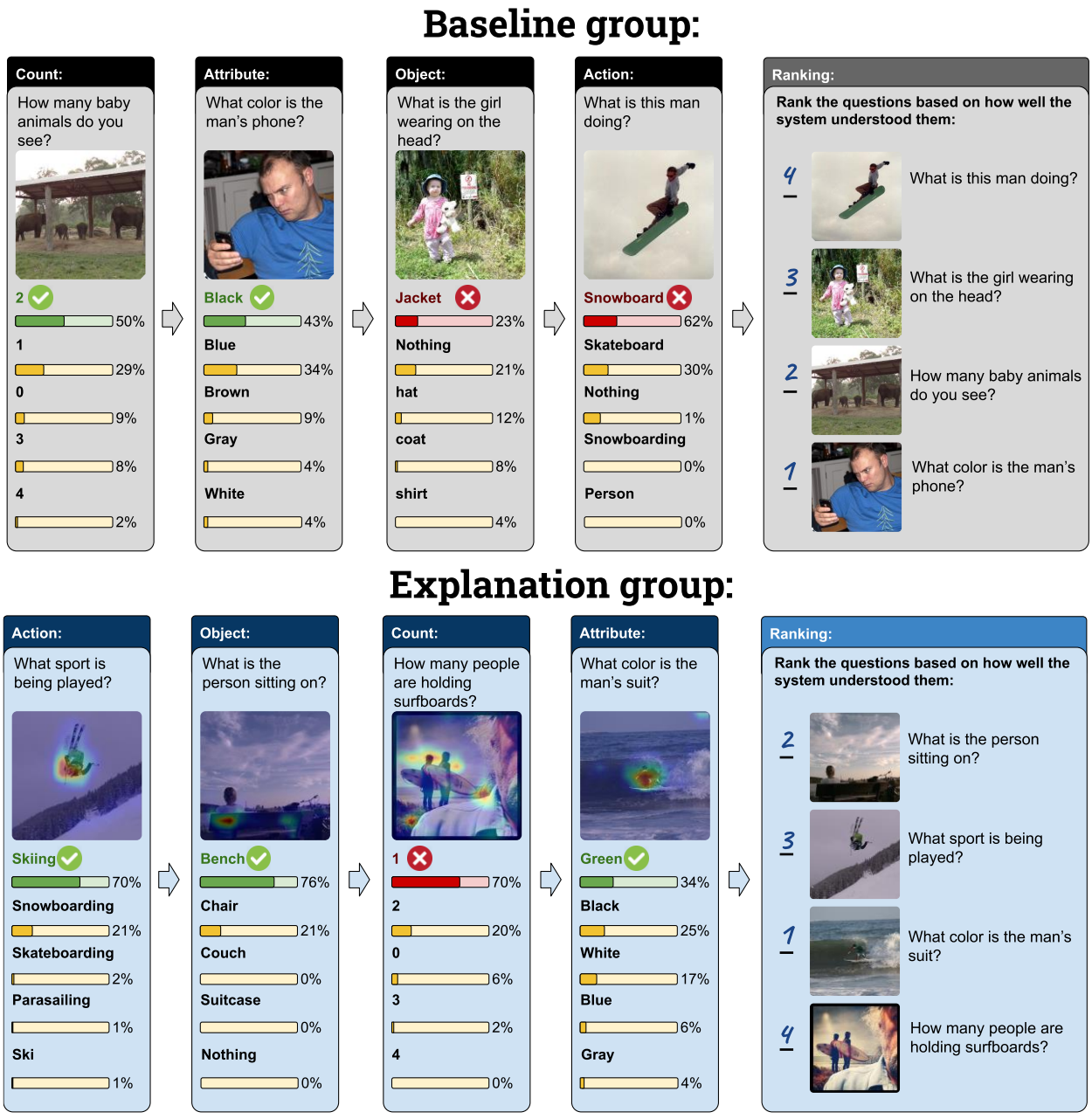


Figure 2.5 The workflow for user study groups: on the left is the baseline group, where the users only view the top five answers from the model and the probability of the answers. Shown on the right: users inside the explanation group also view the attention maps generated by the model. Each group considers blocks of trials. Users are requested to score the question-images at the end of each block based on how well they appear to be comprehended by the model.

2.5.1 Explanation Helpfulness

In the explanation group, the subjects view the attention explanations before they see AI's final answers and accuracy. At this stage, the subjects rate the explanations based on their helpfulness in understanding the AI's performance.

The helpfulness rankings are specifically interesting for action and count question types within which the VQA agents show their highest and lowest competencies. The helpfulness rankings within these categories on SOBERT explanations show an increase compared to SVQA (Figure 2.7). While subjects ranked 17% of SVQA explanations as "not helpful" in count questions, this number is reduced to 7% when using SOBERT explanations. SOBERT also decreased unhelpful explanations from 8% to 3% in action questions.

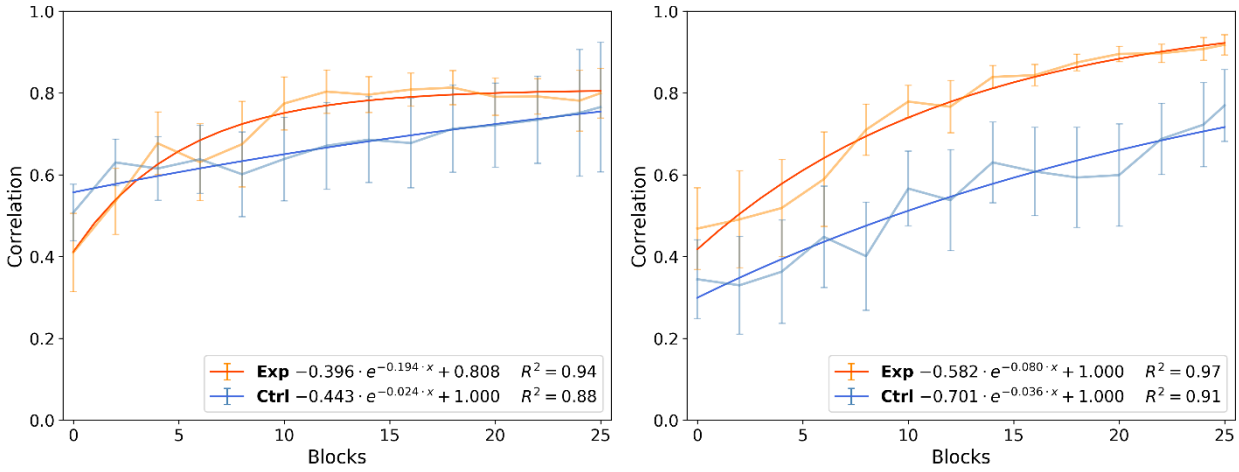


Figure 2.6 Temporal impact of attention maps on user rankings. Left: the growth of correlation in baseline and explanation groups is compared between baseline (blue) and explanation (orange) groups for two models, SVQA (left) and SOBERT (right). T-test p-values for SVQA and SOBERT data are 0.07 and $3.7E-8$ respectively.

2.5.2 Competency Ranking

We assess the accuracy of the subject's rankings by measuring the correlation between that and the ground truth competency ranking of AI agents (Figure 2.5) and the collected rankings at the end of each block. Figure 2.7 illustrates this correlation between each study group's starting

and finishing blocks. The start and finish values of correlation are the average of 1-5 and 20-25 blocks, respectively.

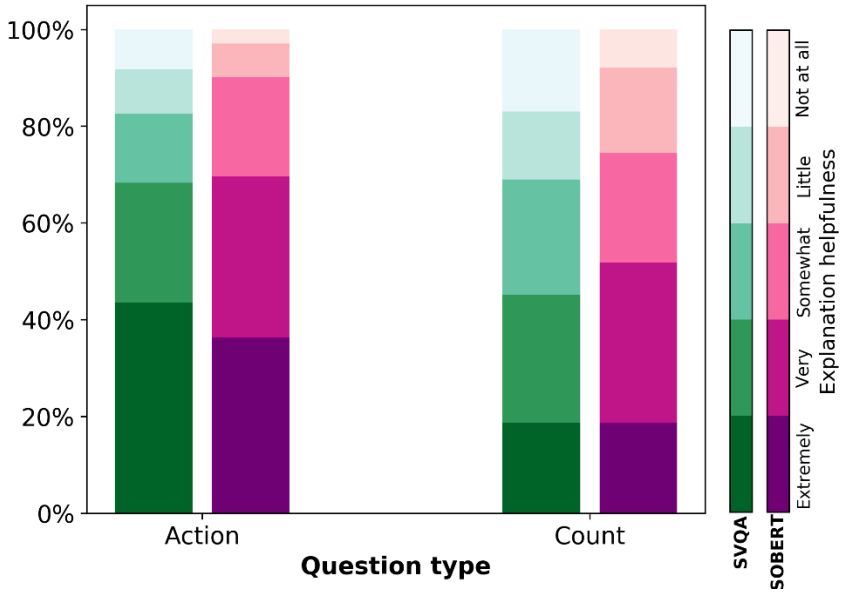


Figure 2.7 Histogram of ratings of how "helpful" explanations are for the subjects. The subjects give these helpfulness ratings as they view the explanations before seeing the system's top 5 answers. Consequently, these ratings are not confounded by the accuracy of the AI.

Overall, the ranking correlation shows an increase in both models with a slightly higher slope in the presence of explanations (Figure 2.8). To better picture the temporal impact of explanations on the users' mental models, Figure 2.6 presents the progress of ranking correlation throughout the study. In the early blocks of both models, the explanation groups increase their ranking correlation at a higher rate than baseline.

2.5.3 Competency Learning Curves

We also investigate the temporal pattern of temporal ranking correlation by fitting curves into the data in baseline and explanation groups. This problem, in general, can be viewed as modeling a user learning curve for a particular task.

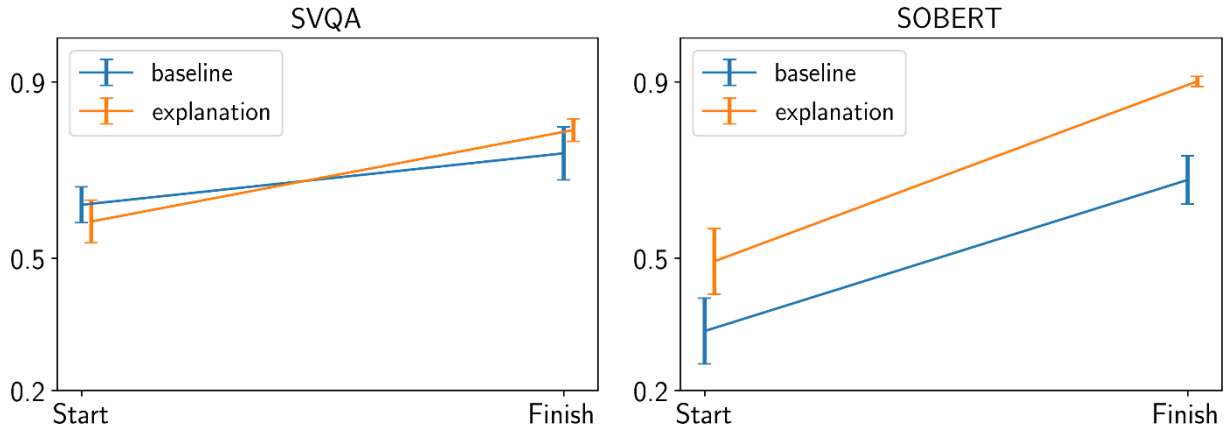


Figure 2.8 The overall correlation between the users' rankings and the system's actual competencies. Comparing the results from the SVQA model (left) and our SOBERT model (right) suggests a better improvement of correlations in the presence of SOBERT attention maps.

The modeling of a user learning curve is widely discussed in cognitive science. In previous work, researchers analyzed exponential learning equations to describe user improvement in the task (Heathcote et al., 2000; Ritter & Schooler, 2001). The assumption of a monotonically decreasing improvement is the main foundation beneath the exponential learning curves.

Here, in the context of learning AI competency rankings, subjects start the study with no prior knowledge of the AI agent's rankings. Also, the correlation metric cannot exceed a value of 1.0. Considering these similarities to the general learning model, we also assumed an exponential curve with an upper bound as blocks grow to infinity. With this analogy, we considered the following curve to fit the ranking correlation trends:

$$c = \alpha \cdot e^{-\beta \cdot b} + \delta \qquad 2.1$$

where b and c are the block count and ranking correlation, respectively. In this setting, the ranking correlation approaches δ as the subjects continue the study. The value of δ is penalized for curves fitting to satisfy the condition $\delta \leq 1.0$.

The slope of the curves in Figure 2.6 represents the growth rate of ranking correlations with respect to the number of blocks. Higher rates of correlation growth show faster learning by

the subjects. To compare the learning rates, we consider the maximum slope of each curve (Table 2.2).

The results indicate a higher rate of learning for users in the presence of an explanation. The explanation from the SVQA agent causes a higher increase in the learning rate compared to SOBERT. However, the ultimate value of ranking correlation in the SVQA model is bound to $\delta = 0.808$, while the SOBERT model approaches the maximum correlation at $\delta = 1.0$ (Figure 2.6).

2.6 Discussion

In our user studies, the overall progress of ranking correlations is measured as a metric to evaluate the users' mental model of system competency. We tested the users' mental model after seeing only 100 instances (trials) of the AI agent's performance. However, the results strongly suggest that even with this limited view of system performance, the subjects learn the overall competency of AI agents throughout these tests.

Adding the attentional explanations for both models results in a significant improvement over competency rankings. Comparing the early learning rates between baseline and explanation groups suggests a considerable improvement by attention map explanations, especially for the SVQA model. However, the SVQA learning curve suggests an upper bound to the correlation in the presence of explanations. On the other hand, the SOBERT model shows a higher learning rate with explanations than the baseline while still reaching the maximum correlation value. These results highlight the effect of input features on the information that the explanations can carry. The SOBERT model uses object and spatial features vs. the spatial features in the SVQA model. The SOBERT model also uses a transformer to construct the attention maps based on the

features. These changes compared to the SVQA have raised the upper bound on the maximum reachable competency prediction by the subjects.

In our study, we control two factors: the presence or absence of explanation and the source of explanation. The comparison between control groups and explanation groups shows the impact of explanations in general. On the other hand, the comparison between the two models investigates the role of explanation quality on user predictions. The quality of explanations can impact the user's mental model. For instance, the scattered/smooth attention maps may be construed as a sign of lower/higher competency in the eyes of lay users. To restrict the impact of such assumptions and comparisons, each person in our study can only participate in the study once and is only exposed to one AI agent. However, discovering the chain-of-thought and decision process can be investigated in future research to identify the explanation qualities that play a role in the subject's prediction.

2.7 Conclusion

This chapter evaluates the role of attention map explanations on the user's mental model of AI competency. We designed an experiment where subjects rank the performance of the VQA model among four different types of questions. To quantify the subjects' mental model, we compute the correlation between user rankings and the AI's actual ranking among the question types.

We propose a new XVQA model that produces answers and attention maps from spatial and object features of the image. This explainable model uses a transformer attention module to better process the visual and textual embeddings of the input. The proposed model is compared with a baseline model to show the effect of input object features and the attention module.

Overall results from the experiment suggest an improvement in the user's mental model when exposed to the attention map explanations. The progress of the user's mental model (ranking correlations) throughout the experiments indicates a higher learning rate in the presence of explanations. Furthermore, the subject group interacting with the newly proposed model shows a higher ranking correlation rate than the baseline model. This improvement positively impacts the explanations by including the object feature and the transformer model.

2.8 Acknowledgments

Chapter 2, in full, is a reprint of the material as it appears in The impact of explanations on AI competency prediction in VQA. Alipour, Kamran; Ray, Arijit; Lin, Xiao; Schulze, Jurgen P.; Yao, Yi; Burachas, Giedrius. In 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI) (pp. 25-32) (2020). IEEE. The dissertation author was the primary investigator and author of this paper.

CHAPTER 3 ATTENTION-BASED EDITS

3.1 Abstract

In the domain of Visual Question Answering (VQA), studies have shown improvement in users' mental model of the VQA system when they are exposed to examples of how these systems answer certain Image-Question (IQ) pairs. This chapter shows that showing controlled counterfactual image-question examples is more effective at improving users' mental models than simply showing random examples. We compare a generative approach and a retrieval-based approach to show counterfactual examples. We employ recent breakthroughs in generative adversarial networks (GANs) to construct counterfactual images by removing and repainting certain regions of interest in the image. We then expose users to changes in the VQA system's answer on those altered images. To select the area of interest for inpainting, we experiment with using both human-annotated attention maps and a fully automatic method that uses the VQA system's attention values. Finally, we test the user's mental model by asking them to predict the model's performance on a counterfactual test image. We note an overall improvement in users' accuracy to predict answer change when shown counterfactual explanations. While realistic retrieved counterfactuals are the most effective at improving the mental model, we show that a generative approach can also be equally effective.

3.2 Introduction

With the growing application of AI in high-risk domains, it is crucial to understand the extent and limits of AI system competencies to ensure such systems' efficient and safe deployment. While deep neural networks have made impressive strides, they are notorious for being

unpredictable to a human user when they succeed or fail in producing correct outputs. Hence, we need practical approaches to improve the end user's mental model of the deep neural network-based AI systems.

There have been studies in the literature (Chandrasekaran et al., 2018) that show humans can improve their mental models by mere exposure to the system predictions for various inputs. A mental model is a person's internal representation of the AI system she is interacting with and ideally builds a correct understanding of how that system works (Rutjes et al., 2019).

This chapter explores the various ways we can present such explanatory additional input-output examples to a user to maximize their mental model improvement. We ask the question: are specific examples of how the machine behaves better than other examples to teach humans when to trust the model and when not to? We examine the effect of exposing the users to explanatory examples where the inputs are controlled to observe better how the machine output changes to controlled changes in the input. We call these controlled changes in input "counterfactuals." We hypothesize that such controlled changes in the examples shown are better for mental model improvement than showing random examples.

Many approaches (Alipour, Ray, et al., 2020; Chandrasekaran et al., 2018; Ray, Burachas, et al., 2019) to improving mental models also focus on using explanations that aid the user in understanding how a deep network arrived at a particular conclusion. While many existing explanation approaches, such as attention maps, attempt to provide insights into the inner working of AI machines, they don't necessarily convey the causal chain of inference that happens in the algorithm (Ray et al., 2021). As a result, the research community actively seeks novel explanation modalities to probe the causality of AI as this form of explanation can resonate better with human logic. Humans learn better from explanations that easily convey when a machine is about to be

correct and when not (Ray et al., 2021). Among different techniques, showing counterfactual examples is considered a *human-friendly* explanation because they are contrastive and selective when showing the feature changes (Molnar, 2020). Counterfactuals allow the user to explore the range of responses from AI as they manipulate certain features of the inputs and the conditions.

In this chapter, we focus on improving users' mental models by generating counterfactual explanations for the task of visual question answering (VQA) - answering natural language questions asked about images. Specifically, we compare various methods of generating counterfactual examples to maximize a user's accuracy in predicting when a model will fail or succeed.

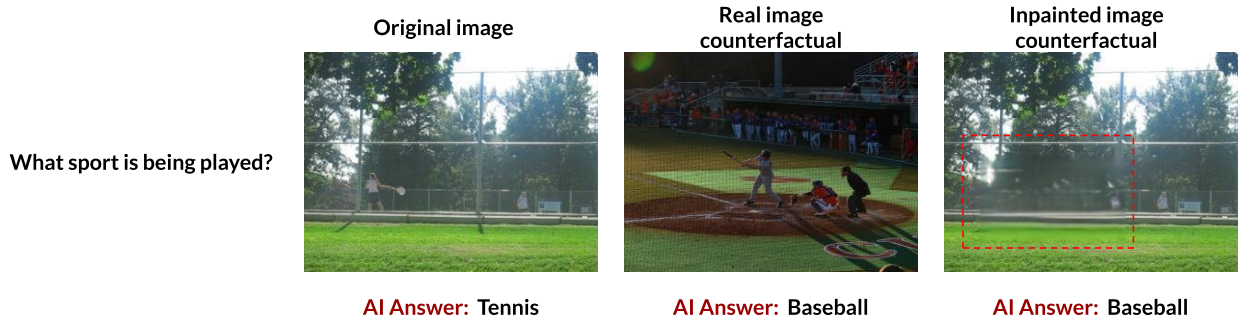


Figure 3.1 While alternative real images may present a convincing counterfactual case for a visual question answering (VQA) model; they are expensive to harvest and often incapable of selecting specific features. In this sample, while the real-image counterfactual may suggest that the AI agent is correctly capturing the type of sport, the in-painted counterfactual indicates that the change in the answer is not necessarily correlated to the changes in the input.

In this setting, given an image-question pair, a counterfactual explanation shows the model's output for the same question but on a different image where the answer should be different. For example, as shown in Figure 3.1, for the question "what sport is being played?" on the original image of playing tennis, the counterfactual examples could be showing the answer of the model on an image where someone is playing baseball (middle image), or where a tennis racket is absent (rightmost image).

Specifically, we compare a retrieval-based approach and a GAN-based approach to generate counterfactual images for a given question. For example, as shown in Figure 3.1, we can generate a counterfactual image (an image where the answer may be different from the original image) by either retrieving an image where the answer is different (middle image) or by removing the tennis racket using a GAN network (rightmost image). Our automated approach using a GAN provides the opportunity to produce counterfactuals at scale and evaluate their effectiveness on a large population of AMT workers.

One major challenge of automatically generating counterfactual images is that the capability of current GAN models limits us. We chose to use an in-filling network (Chang et al., 2018) to remove parts of the image since we observed that current networks could achieve this with a good performance. Limited by the capability of only being able to remove parts of images, we need to decide the most effective parts of images to remove to generate counterfactual examples that help users to learn the idiosyncrasies of the model to improve their mental model. In this regard, we experiment by using attention maps (heatmaps that point to where a machine looks while answering the question) to decide on relevant and irrelevant parts of the image to remove to generate counterfactual ideas.

In summary, our contributions include:

Proposing effective ways to generate counterfactual examples: We outline several ways of generating counterfactual images. Specifically, we compare a retrieval-based method and an automated GAN-based method to generate counterfactual images.

We evaluate the effectiveness of counterfactual examples relative to providing random examples empirically. We improved users' mental model when offering controlled counterfactual examples compared to simply showing random or no examples at all.

In the following sections, we first look at the related work. We then discuss the methodology behind our approach. We then cover the details of our hypotheses and experimental designs. Finally, we provide the results from our studies and discuss our interpretations.

3.3 Related Work

VQA/Explanations. Our approach is based on interactions with a visual question answering (VQA)(Antol et al., 2015) machine. The use of attention-based layers and explanations in VQA has been a highly popular approach (Alipour, Ray, et al., 2020; Fukui et al., 2016; Teney et al., 2018; Xu & Saenko, 2016). Previous work in the attention-based VQA includes attempts to improve the attention mechanism through co-attention between image and question (Lu et al., 2016) or a combined bottom-up and top-down to compute object-level attention (Anderson et al., 2018). In recent work, Peng *et al.*(Peng et al., 2020) propose a Multi-modal Relation Attention Network (MRA-Net) model with textual and visual relation attention for higher performance and interpretability. Patro *et al.*(Patro et al., 2020) utilize adversarial training of the attention regions as a two-player game between attention and explanation. We adopted a VQA model similar to what was proposed by Alipour *et al.*(Alipour, Ray, et al., 2020), where the attention is derived from a transformer model (Devlin et al., 2018).

Counterfactuals. Counterfactual examples have also been used to explain image classifiers (Chang et al., 2018; Goyal, Wu et al., 2019). They have also been used in an optimization process where Wachter *et al.*(Wachter et al., 2017) proposed a loss function to find the minimum changes in the input that results in a shift in the classifier's output. Using counterfactual images as explanations can also be thought of as the visual equivalent of observing VQA behavior by rephrasing the question and checking if the model responds consistently

(Agarwal et al., 2020; Ray, Sikka, et al., 2019; Selvaraju et al., 2020). Hence, such counterfactual images hint at how consistent these models are to users, which aids in their mental model improvement.

Mental model evaluation. Some of the previous studies introduce metrics to measure trust with users (Cosley et al., 2003; Ribeiro et al., 2016) or the role of explanations in achieving a goal (Kulesza et al., 2012; Narayanan et al., 2018; Ray, Burachas, et al., 2019). Dodge *et al.* investigated the fairness aspect of explanations through empirical studies (Dodge et al., 2019). Lai and Tan (Lai & Tan, 2019) examined the role of explanations in user success within a spectrum from human agency to full machine agency. Lage *et al.* proposed a method to evaluate and optimize the human-interpretability of explanations based on measures such as size and repeated terms in explanations (Lage et al., 2019). Other approaches measured the effectiveness of explanations in improving the predictability of a VQA model (Alipour, Schulze, et al., 2020; Chandrasekaran et al., 2018).

In this chapter, we develop a series of user studies with a subject population of lay users with minimum knowledge about AI. The experiments are designed to investigate effective methods to produce counterfactuals that can improve the user's mental model of a VQA system.

3.4 Method

This section describes our VQA model and then explains how we generate counterfactual images using a GAN.

3.4.1 VQA Model

Our VQA model is trained based on the VQA 2.0 dataset (Goyal et al., 2017) and can answer questions about images in textual format. The model is a transformer-based neural network that can parse a combination of visual and textual embeddings from an image and question. The

model encodes the image into a 49×512 feature map with the help of ResNet152(He et al., 2016). The objects in the image are also encoded separately into a 36×512 feature map using a Region Proposal Network (He et al., 2017). The model accepts questions with a maximum length of 30 words, and all questions below this limit are padded with 0s. The question array is also embedded into a 30×512 vector of features.

The model employs transformer-based attention layers that receive all the visual, object, and textual features in the concatenated shape of 115 $(30 + 36 + 49)$ tokens. The transformer is comprised of four layers with 12 heads in each layer. Consequently, the model can provide an attention tensor between these tokens with a $4 \times 12 \times 115 \times 115$ dimension. The model offers its prediction as a SoftMax probability distribution over 3129 answer choices from the attention-weighted feature values.

We use a subset of the VQA 2.0 validation dataset for our experiments. We first show the VQA model's answer to this subset's original images and questions. For each example, we also present the answer to two counterfactual images for the question to the user. We finally test the user's mental model by asking the user to predict the correctness of the answer on a test image for the same question. We will now describe how we generate counterfactual images.

3.4.2 Generating Counterfactual Images

We generate counterfactual images to serve as examples of VQA behavior under differing inputs to improve users' mental models. For example, a user who sees a VQA not counting oranges correctly when changing the number of oranges in a picture and asking, "How many oranges?" will learn that the VQA model has a low accuracy for counting oranges. This sort of mental model improvement might not have been noticed if we presented the user with only one counting example and other random examples of images and questions. Our study focuses on altering objects in the

image for a particular question. Specifically, we use a GAN trained to in-paint areas of the image such that it looks natural (Chang et al., 2018). When asked to in-paint an area of a specific object in an image, such a GAN would usually omit the object and in-paint its content that matches the background/surrounding scenery. We use such an approach to remove objects from the scene. However, such methods are currently noisy, and we often note artifacts in the image that make it seem unnatural. Hence, we limit the size of all in-painting bounding boxes to 10% to 20% of the whole image area.

3.4.3 How to Choose Objects to In-paint

The goal of this algorithm is to generate counterfactual explanations that are helpful to the users in predicting AI's response. Given the diverse combinations and interactions between objects in a real scene, it is not obvious how to define an algorithm to select and alter the objects from images to maximize mental model improvement. In our approach, we use attention maps - heatmaps that convey the essential regions of the image for answering the question - to decide on objects to remove from the image. We use two different sources of attention to identify the in-paint candidates and then produce the counterfactual images based on them:

- Human annotated attentions for the image-question pair, which come from the human attention dataset (Das et al., 2017).
- The attention layers from the AI system. As described in Section VQA Model, our VQA system has multiple layers of attention that weigh the image and question features. We select the weights from the last layer (averaging over the transformer heads) to display the attention values over the image regions. The attention values over the image regions are also computed as the average weight over question tokens.

Based on the above-computed attention maps, we generate two counterfactual images- 1) we remove a box that falls in a region of high attention, and 2) we remove a box that falls in a region of low attention. This ensures we remove a relevant and irrelevant object in the image to introspect how the VQA model's answer changes. Based on this observation, a user can hypothetically learn whether the VQA model is behaving rationally or not. To select the low and high attention boxes, we employ a threshold that first segments the attention map's high and low attention regions. The bounding boxes surrounding these regions provide the in-paint area. In cases where the bounding boxes are outside the limits (10% - 20% of the image area), the proposed box is scaled to a size within the range.

3.5 Experimental Settings

We conduct experiments to quantify the improvement in the mental model for users after being exposed to counterfactual explanations. We measure the user mental model by asking them to predict the answer-change or the correctness of the answer for a given image-question (IQ) pair, similar to concurrent studies on user mental model evaluation (Alipour, Schulze, et al., 2020; Chandrasekaran et al., 2018). We use the Amazon Mechanical Turk platform to recruit users for our study. We recruit workers located in the United States (due to IRB regulations) who exceed 98% approval rating on at least 50 such human-intelligence tasks (HITs).

In our study, each user goes through 1 HIT, consisting of 20 episodes of IQ pairs. In each episode, the users first see the VQA model's response to the original Image-Question (IQ) pair. Based on their group configuration, then they may or may not see two counterfactual forms of the original image and AI's response to the original question for those counterfactual images. In the evaluation section of each episode, users attempt to predict AI's response to a test image for the

same question. We quantify users' mental model states based on their accuracy in predicting AI's response.

We use two tasks to measure a user's mental model - a) answer-change prediction to see if users can predict if the answer will change when a certain object is removed, and b) answer correctness prediction on a real test image based on the lessons learned from counterfactual examples.

3.5.1 Answer Change Prediction

In this setting, we show an IQ pair to a user, the VQA model's answer to the original IQ. One group of users - Counterfactual Group (CF Group) - sees two examples of objects being removed from the image and the VQA model's answer on these two altered images. The Control Group of users sees no such altered examples. Finally, both the groups of users are presented with another new object removed from the same image and are asked to predict if the VQA model's answer for that image will change from the original image or not (Figure 3.3).

In the experiment, the counterfactuals were generated based on human attention (Das et al., 2017) annotations from the VQA-HAT dataset. Note that the object removed in the test image is always different from the objects removed in the counterfactual examples shown. We do this by choosing separate regions of minimum (min), maximum (max), or medium (mid) attention based on the human-attention values on the image. While the Min and Max regions are determined by extracting the areas from the two sides of the attention spectrum (Figure 3.2), the Mid area is identified by avoiding the overlap with Min and Max and maintaining the minimum attention possible. This procedure of in-painting assures the minimum overlap among the sample and the test in-paintings and therefore minimizes the chance of overlap between counterfactual samples

and test images. While testing the users, we randomly choose to show two of min, mid, and max as counterfactual examples and test on the unseen third.

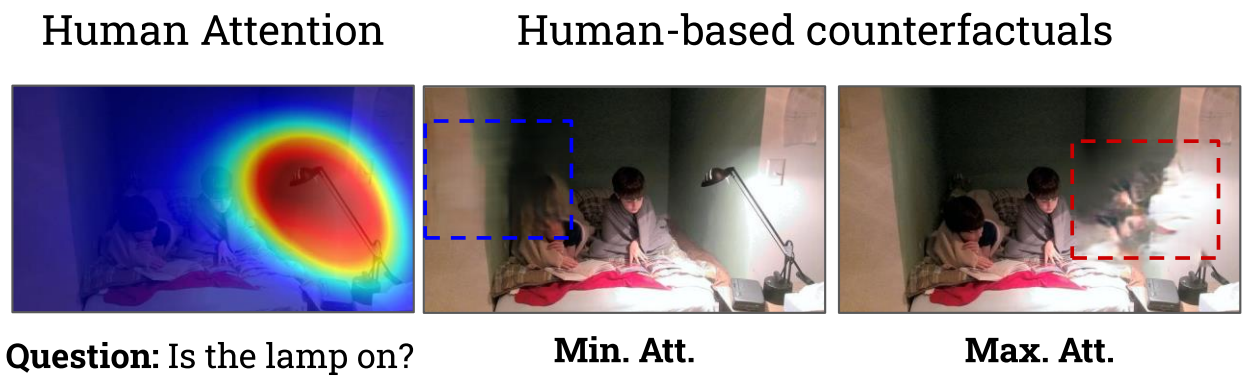


Figure 3.2 Generating counterfactual images based on human annotation attentions. The algorithm first identifies the most attended and least attended bounding boxes in the image and then applies the generative adversarial networks (GAN) to in-paint those bounding boxes and produce the counterfactual images.

The group **CF-MinAtt** shows mid and max attention as samples and tests on the image with the min attention region removed. Similarly, **CF-MidAtt** tests on the image with the mid attention region removed, and **CF-MaxAtt** tests on the image with the max attention region removed.

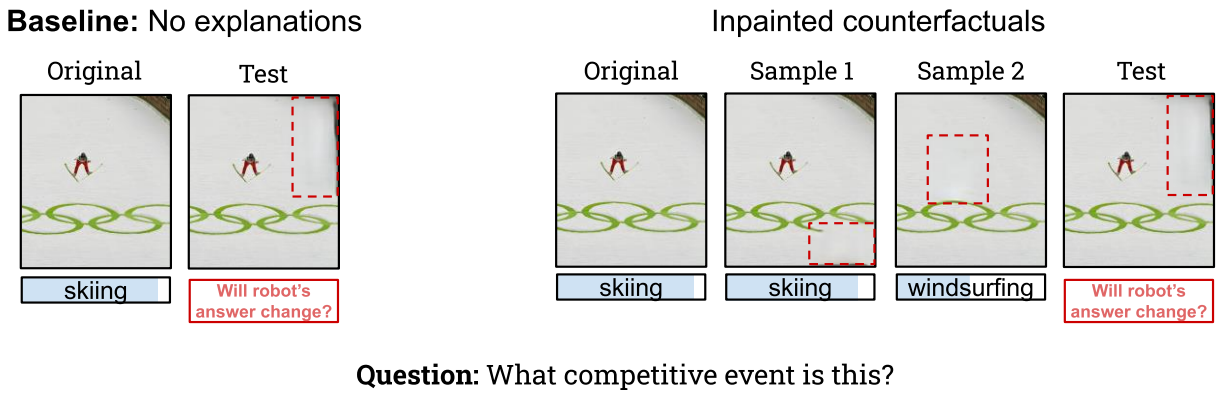


Figure 3.3 The interfaces for the experiments that evaluate the impact of in-painted counterfactuals for the task of answer-change prediction. Users in both groups are evaluated based on the same in-painting patterns. While the users in the Counterfactual Groups can utilize the counterfactual samples in their prediction, the baseline group attempts to predict the answer-change merely based on the original image-question (IQ) response. For the input and sample images, users see AI's top answer along with its probability (blue bar beneath the answers).

3.5.2 Answer Correctness Prediction

In the second set of experiments, we provide a more realistic setting to evaluate the user's mental model. Instead of predicting an answer change for a counterfactual test image, the users now attempt to predict whether the model will answer the same question correctly for a different test image. Since the test images are also selected from the IQ pairs in the VQA dataset, that guarantees that the test image is relevant for the question asked.

We define four groups (shown in Table 3.1) to check whether counterfactual examples improve users' mental models to be able to predict the model's correctness on an unseen test image:

- Control Group (**CG-NoExp**) sees no explanations and just the VQA model's answer on an IQ pair.
- The counterfactual group is either based on counterfactual images generated using human-annotated attention (**CF-HAT**) or the VQA model's attention (**CF-AIAtt**). These groups are to examine how the process of generating counterfactual images affects the mental model.
- A group that sees retrieved real counterfactual images (**CF-AltImg**). In the VQA dataset, each question is correlated with multiple images that can be relevant to those questions. In an automated process, we randomly retrieve images that are relevant to the question but have a different answer. We can think of these as ideal counterfactual examples. The performance of this group compared to the CF-HAT and CF-AIAtt groups would tell us if generated counterfactuals (CF-HAT and CF-AIAtt) can be used in place of real counterfactual data to reduce dataset collection costs.
- A group that sees random IQ pairs instead of counterfactuals (**CG-RandIQ**). This group

is to understand how much we gain from simply presenting two samples of random IQ pairs instead of two counterfactual IQ examples.

Table 3.1 User study groups for answer correctness prediction.

Group	Samples		Attention source
	Images	**Questions**	
CG-NoExp	-	-	-
CG-RandIQ	Random images	Random questions	-
CF-HAT	Original image in-painted over least/most attended areas	Original question	Human annotations
CF-AIAtt	Original image in-painted over least/most attended areas	Original question	AI attention
CF-AltImg	Alternative real images	Original question	-

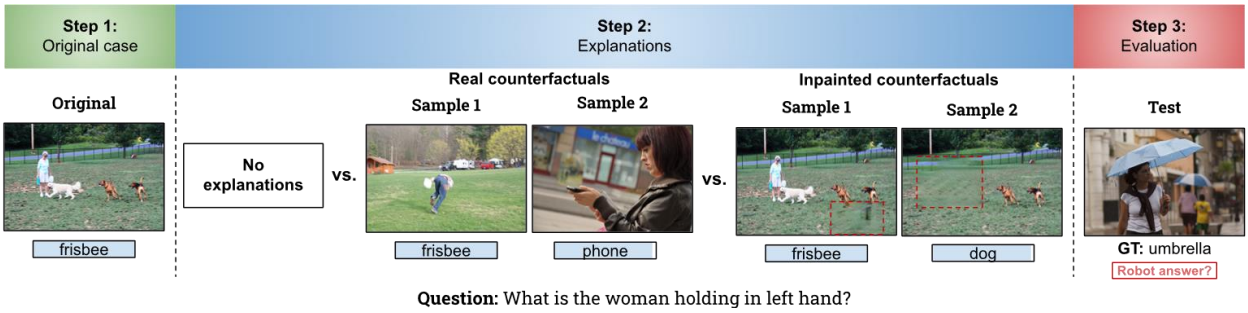


Figure 3.4 The workflow for different groups of the study. While steps 1 and 3 are shared among groups, the explanation step differentiates between them. In in-painted counterfactuals, samples 1 and 2 are in-painted over the least attended and most attended areas, respectively. The real counterfactual images are sampled from the VQA dataset.

Note that in all cases, we make sure all images are relevant to the question asked since VQA models are not trained to answer irrelevant questions about images (Ray et al., 2016). Figure 3.4 visualizes the different interfaces used for the CG-NoExp group and the CF groups.

3.6 Results and Discussion

In this section, we cover the results from the user studies conducted for the two major tasks described previously: answer change prediction and answer correctness prediction.

3.6.1 Answer Change Prediction

Table 3.2 provides detailed numbers on the user accuracy in all groups. We show the accuracy of users correctly predicting the system would be INCORRECT for the cases when the VQA model is INCORRECT and similarly for when the VQA model is CORRECT. In the last column, we finally present the normalized accuracy, which is the average of the accuracy for the CORRECT and INCORRECT cases. Since the number of correct cases is more than the number of incorrect cases for a VQA model, a normalized accuracy score mitigates potential spurious increases in the accuracy simply because a user always predicted a model would be correct. If a user always predicted a model would be correct, the recall for CORRECT cases would be 100% and 0% for the INCORRECT cases, resulting in a normalized accuracy of only 50%.

Table 3.2 Normalized user accuracies in answer change prediction task.

Group	**Baseline (%)**	**Counterfactual (%)**
CF-MinAtt	54.23	62.52*
CF-MidAtt	56.29	66.38**
CF-MaxAtt	62.91	66.01**
All	61.30	66.98**

Note: Bold values mark higher accuracies in each group and asterisks reflect statistical significance based on p-value (* < 5%, ** < 1%, and *** < 0.1%).

Users exposed to the counterfactual samples can predict the answer change better than the users in the baseline group. Moreover, we observe a consistent improvement in all subgroups

regardless of the in-painting patterns. Even in CF-MinAtt and CF-MidAtt, users tend to do better when exposed to the counterfactuals, although predicting an answer change for those cases can be inherently harder. These findings suggest a positive impact of the counterfactual samples on the mental model independent of the testing scenario.

3.6.2 Answer Correctness Prediction

Here, we evaluate the user's accuracy in predicting whether a VQA model will be correct or not on an unseen test image. We conduct most of our experiments on this task since this task is more realistic and challenging. Our goal is to see if users can learn from counterfactual explanations to predict the model's performance on unseen images. We divide the correctness prediction results into two subgroups based on cases where the VQA model is CORRECT and INCORRECT, as shown in Table 3.3.

Table 3.3 User accuracy in answer correctness prediction task.

	AI correct		AI incorrect		
Group	N	User Acc. (%)	N	User Acc. (%)	Norm. Acc. (%)
(a) CG-NoExp	2995	70.42	2605	45.60	58.01
(b) CG-RandIQ	2921	78.26	2659	46.30	62.28
(c) CF-HAT	3005	73.14	2625	**54.48**	63.81
(d) CF-AIAtt	2942	77.16	2588	42.81	59.98
(e) CF-AltImg	1643	**85.88**	917	43.84	**64.86**

Note: Bold values mark the highest user accuracies among groups categorized by AI correctness.

Users tend to have an initial optimistic bias toward AI accuracy, and as a result, they are more inclined to predict that the AI machine would be correct. As described previously, to prevent

a spurious accuracy increase simply due to a user predicting a model will be correct more often, we compute the normalized user accuracy as an average between AI correct and AI incorrect cases.

The data for the Answer Correctness Prediction can be found in the following directory: https://drive.google.com/drive/folders/1O4LKCJ4qA5FRRaquBiN-XoEsuX4sxf6_?usp=sharing the directory contains the input data used for the AMT studies as well as the feedback data which is the collection of responses from the study subjects. Each input datasheet contains the series of data for 20 episodes of trials. For each episode, the information such as image link, question, top answers and their probabilities can be found.

In the following, we summarize our findings from the Answer Change and Answer Correctness prediction experiments:

Counterfactual examples help over showing no examples: All CF groups - CF-AltImg, CF-HAT, CF-AIAtt - show improvement over the control group where no explanations are shown for users' mental model as shown in Table 3.3 row a vs. rows c,d,e. This is hardly a surprising result since counterfactual examples provide more information.

Counterfactual examples help over showing random examples: To check how much we gain in the mental model from simply providing more information, we check the performance of users when we show two random examples to the users. We see that the counterfactual groups CF-AltImg and CF-HAT both improve the mental model by simply showing random examples, as shown in Table 3.3, row b vs. rows c and e. This shows that counterfactuals are indeed an effective form of showing examples of how a model behaves to users.

Generated counterfactual images can be a close substitute for realistic counterfactual images: We see that a generated counterfactual image using an in-painting network (Chang et al., 2018) based on human-annotated attention (row c of Table 3.3) can be almost as effective as a real

retrieved counterfactual image from the VQA dataset (row e). While human-attention annotation is still currently needed, it is a step towards automating the counterfactual generation process.

Fully automating the generation process for counterfactual images can be tricky and currently does not seem to help mental model improvement: As seen from row *d* of Table 3.3, if we use the model's attention values to decide on objects to remove, the counterfactual images generated do not improve the user's mental model significantly over no example cases or when random cases are shown. This suggests that further research is needed to automate the counterfactual generation process effectively.

Overall, the results indicate that counterfactual explanations have a positive impact on the user's mental model. While studies on case-based explanations (Keane & Smyth, 2020; Kenny et al., 2021) have shown random examples can improve users' mental models, our results indicate that controlled counterfactuals can better improve the mental model with the same number of examples shown. In certain application fields such as medicine, data is expensive, and hence, counterfactuals can help achieve an increase in mental models with fewer data points than showing random examples. We also see that GAN-generated counterfactual examples show comparable efficacy when evaluated against real retrieved counterfactual examples. However, our best-performing GAN-generated counterfactual relies on human-annotated attention maps being available. Further research needs to be conducted to explore effective ways of generating GAN counterfactuals without the need for human attention for the GAN-based method to be scalable.

3.7 Conclusion

In this chapter, we demonstrated that showing counterfactual images is helpful for the mental model improvement of users in predicting a VQA model's performance. We showed that

counterfactual examples are more effective than showing random examples or not showing any examples at all. We also showed that a generative approach to generating counterfactual images could also be effective at improving the mental model of users. Investigating different image editing methods and the impact of the counterfactual quality on users' mental models can serve as interesting topics for the next steps of this work. We hope these results can serve as a foundation to improve generative models for producing effective counterfactual explanations to improve user mental models for the safe and effective deployment of AI systems in the wild.

3.8 Acknowledgments

Chapter 3, in full, is a reprint of the material as it appears in Improving users' mental model with attention-directed counterfactual edits. Alipour, Kamran; Ray, Arijit; Lin, Xiao; Cogswell, Michael; Schulze, Jurgen P.; Yao, Yi; Burachas, Giedrius. Applied AI Letters, e47 (2021). The dissertation author was the primary investigator and author of this paper.

CHAPTER 4 CAUSAL COUNTERFACTUALS

4.1 Abstract

Despite their high accuracies, modern complex image classifiers cannot be trusted for sensitive tasks due to their unknown decision-making process and potential biases. Counterfactual explanations are very effective in providing transparency for these black-box algorithms. Nevertheless, generating counterfactuals that can consistently impact classifier outputs and yet expose interpretable feature changes is a very challenging task. We introduce a novel method to create causal and yet interpretable counterfactual explanations for image classifiers using pretrained generative models without any re-training or conditioning. The generative models in this technique are not bound to be trained on the same data as the target classifier. We use this framework to obtain contrastive and causal sufficiency and necessity scores as global explanations for black-box classifiers. On the task of face attribute classification, we show how different attributes influence the classifier output by providing both causal and contrastive feature attributions and the corresponding counterfactual images.

4.2 Introduction

Regardless of their accuracy, AI algorithms have yet to provide a level of interpretability to be accepted as trustable assets by their lay users in real-world applications. In recent years, XAI has made an outstanding effort to bring transparency to AI, focusing on fairness and bias. Counterfactual explanations have received much attention among different methods in this field due to their scalable, intuitive, and logical approach (Goyal, Wu et al., 2019; Mothilal et al., 2020; Verma et al., 2020; Wachter et al., 2017).

A counterfactual explanation for a black-box AI should provide transparency to the inner functionality of the algorithm through causal arguments and yet be interpretable to human users. Some methods specifically focus on generating counterfactuals with a very high chance of changing AI's output, such as adversarial examples (Goodfellow et al., 2014; Guo et al., 2019). On the other hand, studying the impact of interpretable attributes in the input on AI output usually goes as far as minimal correlations with model output. It fails to meet the following two steps in Pearl's ladder of causation (Pearl, 2019): intervention and counterfactuals.

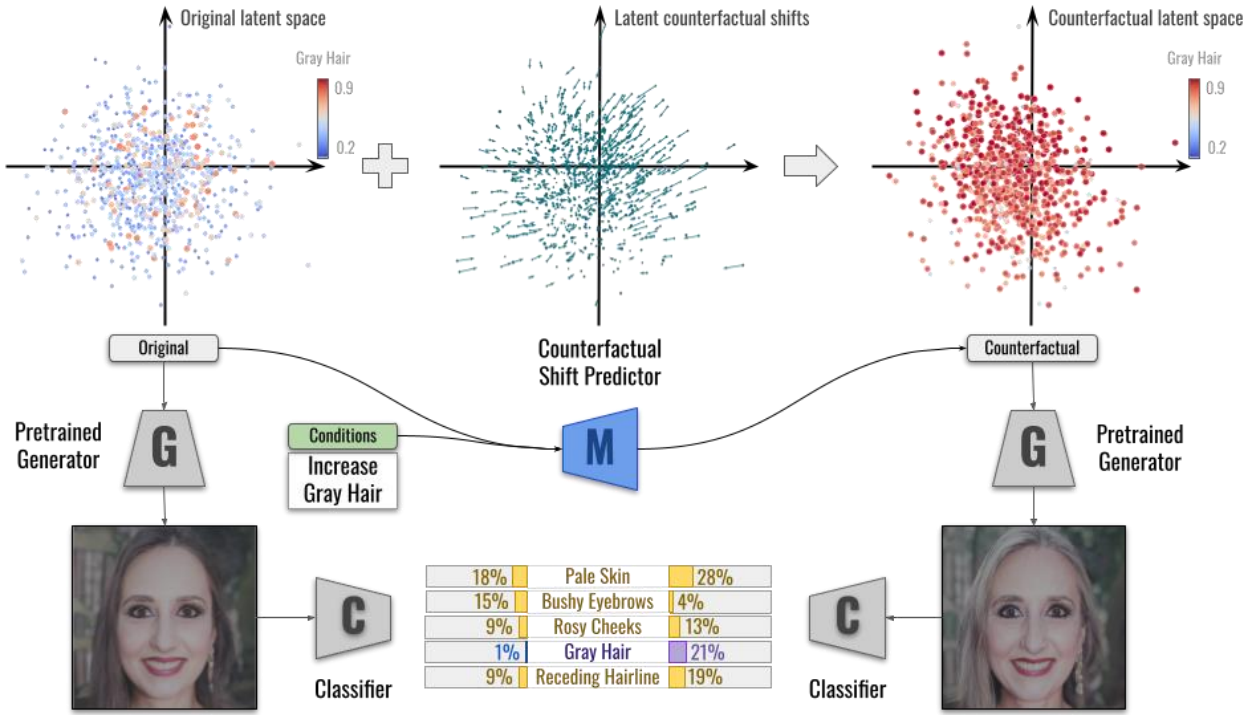


Figure 4.1 We introduce a method to learn counterfactual generation for a black-box classifier C. In this process, a shift-predictor M is trained to predict contrastive counterfactuals for a classifier in the latent space of a generative model G. The shift-predictor can produce shift directions for any combination of attributes that are predicted by the classifier. Sampling from these probabilistic contrastive counterfactuals provides a framework to explain the biases and interactions across different predicted attributes.

While we acknowledge that AI machines are not necessarily trained to follow causal reasoning based on interpretable features, we are interested in attributes that can provide the best of both worlds: to be as interpretable and as causal as possible. Learning such attributes can bridge

the gap between causality and interpretability and generate counterfactual explanations by changing meaningful attributes and still have a strong causal influence on AI outcomes.

Learning the interactions between causality and interpretability in feature space can bring us closer to the true definition of the explainability of AI. Such a framework can also provide us with means to measure whether an AI machine is following any human-understandable pattern to produce its output or not.

Comparing and contrasting target points by observing their differences along a fixed set of understandable dimensions has been one of the primary ways humans have always explained and understood concepts (Gerstenberg et al., 2015; Morton, 2013). These notions are also a natural way of explaining image classifiers. We can see and understand the difference between a pair of images distinctly different from each other in specific known attributes. This enables us to reason how those differences may cause them to obtain different outcomes in some downstream tasks performed on them. Therefore, we aim to generate explanations of the following general form: "For images with attributes equal to <value> for which the algorithm made decision <outcome>, the decision would have been <foil-outcome> with probability <score> had the attribute been <counterfactual-value>" (Galhotra et al., 2021). This chapter seeks to bring this contrastive framework for counterfactual explanations for image classification.

Notions of sufficiency and necessity build on this general form and allow us to reason about the necessary and sufficient conditions for a specific outcome. For an individual who received a positive outcome, necessity captures the importance of the current value of an attribute in obtaining this outcome. On the other hand, for individuals who received a negative outcome, sufficiency reflects the ability of an attribute to flip the negative result into a positive one by modifying its existing value to some new value.

Using a probabilistic interpretation of contrastive counterfactuals, we quantify the sufficiency and necessity of attributes to compute their causal responsibility towards the classifier's output. Obtaining these probabilities of sufficiency and necessity is a challenging task. They require generating counterfactual images that reflect the exact changes we desire in a set of human-understandable attributes. In our work, we formalize the definition of these scores in the context of images and traverse the latent space of generative models to obtain these counterfactual images, which correspond to user-defined values of the set of these interpretable attributes. This enables us to compute these scores efficiently.

An added benefit of our method is that instead of going through the expensive process of creating new datasets, it allows to sample many inputs through pre-trained generative models and estimate contrastive counterfactuals over this sub-population for explaining any black-box image classification model (Figure 4.1).

In summary, the contributions of this work are as follows:

1) We introduce a method to produce contrastive counterfactuals for an image classifier.
2) Our method can use generative models pre-trained on any dataset independent of the classifier training dataset.
3) We propose contextual, contrastive, and causal explanations in the form of sufficiency and necessity scores to explain the black-box model.
4) We use our method to provide global explanations for a black-box classifier trained on the CelebA dataset (Liu et al., 2015).

4.3 Related Work

Previous work in this area generally approaches the problem from several different perspectives. Some of the prior work take a fundamental approach and revolve around exposing the causal roots and achieving causal models.

Recent studies take a more rigorous approach by implementing contrastive counterfactuals for various applications. On the other hand, some of the methods in the vision area are centered around the use of generative models such as auto-encoders or GANs to produce interpretable counterfactuals. These generative approaches are divided into supervised and unsupervised techniques based on whether they involve annotations or classifiers.

Causality. When estimating the causal effect of annotations, it's essential to consider the confoundedness between these attributes for any intervention. In that regard, most previous work attempts to learn a form of a structural causal model (SCM)(Pawlowski et al., 2020; Sani et al., 2020).

Parafita *et al.*(Parafita & Vitrià, 2019) use causal counterfactuals to provide explanations by obtaining attributions for known latent factors. Dash *et al.*(Dash et al., 2022) use a conditional GAN to generate counterfactuals. Bahadori *et al.*(Bahadori & Heckerman, 2020) use a prior causal graph and existing annotations to explain the predictions.

Khademi and Honavar (Khademi & Honavar, 2020) compare different methods of causal effect estimation, such as CBPS and NPCBPS, to interpret predictive models and explain their prediction based on input average causal effect (ACE). In a different approach, Zaeem and Komeili (Zaeem & Komeili, 2021) introduce interpretable attributes as “concepts” and propose learning the presence of each “concept” in different layers of the classifier. Ghorbani *et al.*(Ghorbani et al. 2019) also developed a systemic framework to automatically identify higher-level concepts which

are both human-interpretable and important for the ML model. However, these concepts are neither necessarily causal nor require any prior interpretable attributes as input. While these techniques provide comprehensive explanations of the causal effect of attributes on model output, the causality is often quantified over a population and in correlation metrics. Such correlations are represented as global explanations and satisfy the first step of Pearl's ladder of causation (Pearl, 2019); however, they cannot guarantee a causal impact on a case-by-case intervention (local explanation). This chapter aims to go beyond correlations and provide a framework for a complete implementation of the causation ladder.

Contrastive counterfactuals. Contrastive counterfactuals have been the building blocks of ideas in philosophy and cognition that guide people's understanding and dictate how we explain things to one another (De Graaf & Malle, 2017; Morton, 2013) and have been argued to be central to explainable AI (Miller, 2019). To quantify these notions, probabilistic measures have been formalized and applied to various fields, including AI, epistemology, and legal reasoning (Greenland & Robins, 1999; Russell et al., 2017). Recent work has also focused on using them in the field of Explainable AI (Galhotra et al., 2021; Kommiya Mothilal et al., 2021; Watson et al., 2021).

Our work is following a trend of ongoing research into generating counterfactual explanations for AI algorithms (Bertossi et al., 2020; Karimi et al., 2020; Laugel et al., 2017; Mothilal et al., 2020; Stepin et al., 2021; Ustun et al., 2019; Verma et al., 2020; Wachter et al., 2017). We are specifically interested in implementing this framework in the image classification problem. This topic is inherently challenging as it demands a probabilistic causal model based on the algorithm's output.

Explainable Autoencoders. Due to their strong abilities in representation learning, auto-encoders have received a lot of attention in the XAI community. Variational auto-encoders (VAEs) have shown promising results in learning causal (O'Shaughnessy et al., 2020) and interpretable (An et al., 2021) representations or interception of interpretable attributes (Goyal, Feder, et al., 2019). Similarly, Castro *et al.* (Castro et al., 2019) propose a framework to measure how much the latent features in a VAE represent the morphometric attributes in the MNIST images. Some studies propose a feature importance estimation based on Granger causality (Thiagarajan et al., 2020) (Schwab & Karlen, 2019). While auto-encoders are very strong in representation learning, they tend to lose detail in regenerating highly complex data. We focus our approach on using pre-trained generative models that can produce high-resolution counterfactuals with accurate shifts in interpretable features.

Unsupervised disentanglement in GANs. Within recent related work in this area, several of them are dedicated to the unsupervised disentanglement of latent space of GAN models to interpretable feature spaces. Some approaches use fundamental techniques such as projection (Shen et al., 2020), PCA (Härkönen et al., 2020), and orthogonal regularization (Liu et al., 2022), while others use self-supervised techniques to learn interpretable representations (Nitzan et al., 2020). Another popular approach is the unsupervised discovery of linear (Lu et al., 2020; Voynov & Babenko, 2020) or nonlinear (Tzelepis et al., 2021) directions that correlate with interpretable features. Moreover, adversarial methods (Yang et al., 2021) alongside contrastive learning-based and intervention-based approaches (Gat et al., 2021; Yüksel et al., 2021) have also been studied for interpretable direction discovery in GANs.

Despite their impressive results, these techniques may combine multiple distinguishable attributes in one detected direction. On the other hand, the detected interpretable directions are not

always easy to label and hence cannot be used for any label correction or model improvement. In this study, we propose learning latent directions that correspond to actual labels so the explanation results can be used in improving datasets and training procedures.

Supervised direction discovery. On the other hand, many contributions in this area are focused on the supervised discovery of interpretable directions in the latent space of generative models. The majority of these models are implemented for GANs such as StyleGAN (Karras et al., 2020) due to their resounding success and high quality. A classic solution in this area is finding the class boundary hyperplane in the latent space of GAN (Li & Xu, 2021). Some approaches attempt to find counterfactuals with the use of gradient descent in a GAN's latent space (Liu et al., 2019). Some existing techniques train GANs to either apply residuals (Nemirovsky et al., 2020) or masked transformations (Samangouei et al., 2018) on images to generate counterfactuals.

Moreover, some of the prior work experiments incorporated the classifier (Lang et al., 2021) or contrastive language-image models (Patashnik et al., 2021) into GAN to accommodate attributes in the latent space. Another novel approach uses energy-based models (EBMs) for a controllable generation with GAN; however, this technique requires manual labeling of the latent samples (Nie et al., 2021). Styleflow (Abdal et al., 2021) introduces counterfactuals with conditional continuous normalizing flows in the latent space; however, their solution is tailored explicitly for the extended latent space of StyleGAN. In our proposed approach, we seek a minimal training process by utilizing only pretrained GANs. Our methodology follows a simple and scalable implementation to be compatible with different generative models.

4.4 Method

In this section, we first discuss the kind of explanations obtained using contrastive counterfactuals and provide some necessary background. We then formalize the probability of sufficiency and necessity, which are at the core of our explanations. We also describe how we can compute those scores in the setting of image classifiers.

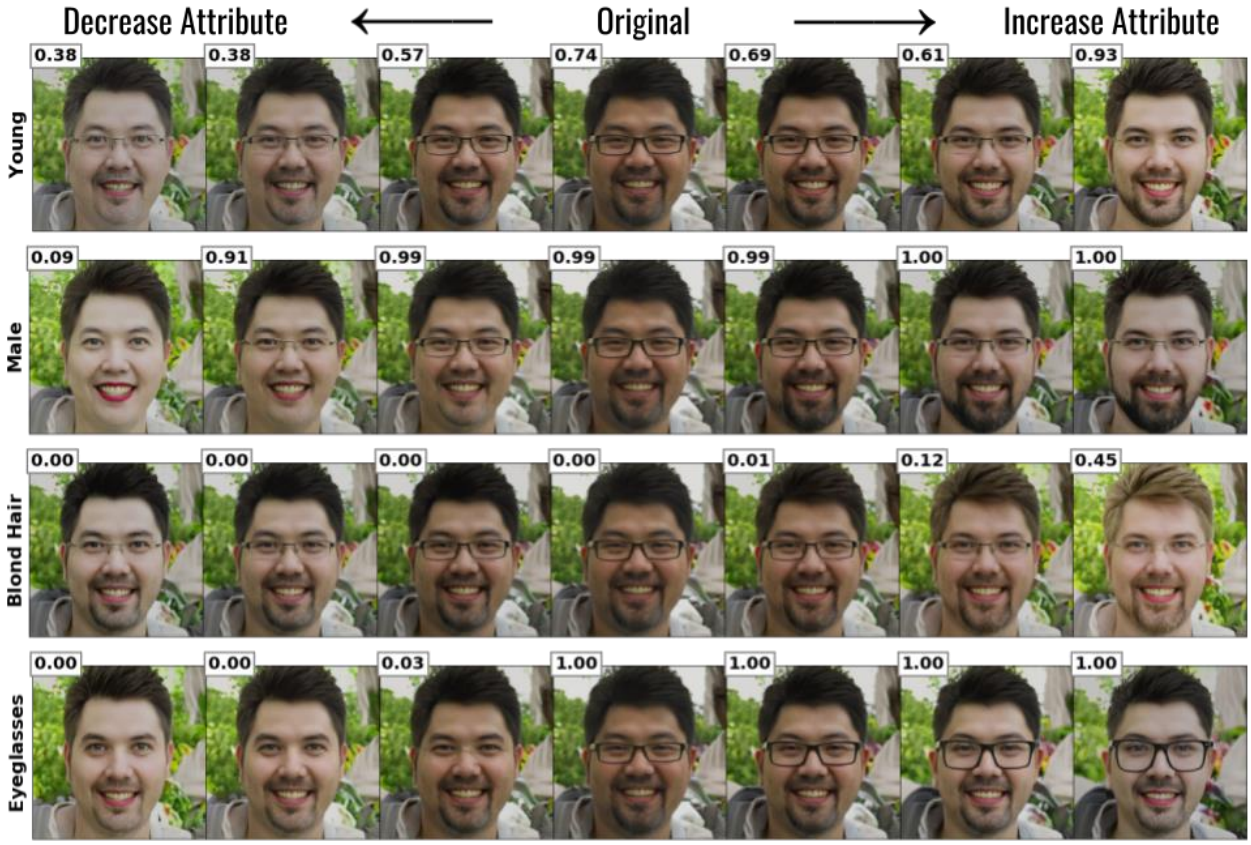


Figure 4.2 A shift predictor learns to predict optimum shifts in the proximity of any input and manipulate the input for different attributes and different directions. Aside from the intended attribute that has changed in each direction, some other attributes also change, which shows the shift predictor's ability to sample from a more realistic distribution and be mindful of potential confoundedness across different attributes.

We later explain the algorithm pipeline, which consists of a pretrained generative model (G) to produce realistic images and introduce a shift predictor to achieve counterfactual latent vectors for (G) (Figure 4.2). These allow us to compute the necessity and sufficiency scores for explaining black-box image classifiers.

4.5 Contrastive Counterfactual Explanations

Contrastive counterfactuals play an essential role in the explanations generated and understood by humans. Our goal is to use probabilistic contrastive counterfactuals and develop a feature attribution method that generates an explanation for an image classifier. These feature attributions quantify the causal contribution of a set of interpretable attributes to the outcome of the classifier. Specifically, for an image classifier that predicts the output Y for input images with an interpretable attribute A, our framework generates explanations of the following form: "For an input image with an attribute a for which the classifier outcome is y, the classifier outcome would be $\hat{y}$ with probability s, had the input attribute been $\hat{a}$ instead of a".

4.5.1 Sufficiency and Necessity

To comprehend the function of distinct conjunctive causes in the framework of causal reasoning, it is required to define all causes as necessary or sufficient. Necessity is a more complex concept than sufficiency. Sufficiency is a more straightforward notion than necessity. In the case of sufficiency, one can simply test if the cause is always followed by the effect; however, in the case of necessity, two alternatives may be tested: does the cause always precede the effect, and can the effect exist without the cause? More importantly, these definitions have distinct structures: necessity is seen as an all-or-nothing trait, whereas sufficiency is regarded as a more permissive attribute (Verschueren et al., 2004).

In the task of attractiveness classification for face images, these explanations pertain to images that had the positive outcome of being classified as attractive. For those cases, such explanations measure the probability that increasing an interpretable attribute such as baldness could lead to a negative outcome instead. Therefore, they measure the extent to which the original

value of the attribute is necessary for positive outcomes, hence called *probability of necessity*. Moreover, we provide sufficiency scores for input images that receive a negative output. Such explanations compute how changing an attribute is sufficient to flip a negative outcome to a positive one, hence called *probability of sufficiency*.

In this study, we rely on Pearl's probabilistic causal models (Pearl, 2009) to formalize and evaluate the notions of the probability of sufficiency and necessity. Next, we briefly review probabilistic causal models and then build on that to mathematically define *Necessity* and *Sufficiency* scores.

4.5.2 Causal Models and Counterfactuals

A probabilistic causal model (PCM) consists of (1) a set of observable *endogenous* attributes $\mathcal{A}$, (2) a set of latent *background (exogenous)* variables $\mathcal{U}$, (3) a set of structural equations $\mathcal{F}$ that capture the causal dependencies between the attributes by associating a function $F_A \in \mathcal{F}$ to each endogenous attribute $A \in \mathcal{A}$ that expresses the values of each endogenous attribute in terms of $\mathcal{U}$ and $\mathcal{A}$, and (4) a probability distribution $P(\boldsymbol{u})$ over the exogenous variables U. Given a probabilistic causal model, an intervention on an endogenous attribute $A \subseteq \mathcal{A}$, denoted $A \rightarrow a$, is an operation that modifies the underlying causal model by replacing F_A, the structural equations associated with A, with a constant a. The *potential outcome* of an attribute Y after the intervention $A \leftarrow a$ in a context of exogenous variables u, denoted $Y_{A \leftarrow a}(u)$, is the solution to Y in the modified set of structural equations.

The distribution $P(\boldsymbol{u})$ induces a probability distribution over endogenous attributes and potential outcomes. Considering proper PCMs, one can express counterfactual queries of the form

$P(\hat{y}_{A\leftarrow a} = \hat{y})$, or simply $P(\hat{y}_{A\leftarrow a})$; this reads as "What is the probability that we would observe $Y = \hat{y}$ had A been a?" and is given by the following expression:

$$P(\hat{y}_{A\leftarrow a}) = \sum_{u} P(\hat{y}_{A\leftarrow a}(u))P(u).$$

4.5.3 Probability of Necessity and Sufficiency

We are given a binary image classifier with the output $Y = \{y, \hat{y}\}$, where $y = 1.0$ and $\hat{y} = 0.0$ denote the positive (favorable) and negative (unfavorable) outputs, respectively. We also have a binary attribute $A = \{a, \hat{a}\}$ associated with the input images, where $a = 1.0$ and $\hat{a} = 0.0$ respectively denotes the presence or absence of the attribute.

The probability of SUFficiency and NECessity of A for Y measures are as follows:

$$NEC = P(\hat{y}_{A\leftarrow\hat{a}} | a, y). \quad 4.1$$

$$SUF = P(y_{A\leftarrow a} | \hat{a}, \hat{y}). \quad 4.2$$

Given a sub-population of input images with attributes a, for which the classifier returns the positive output, the notion of probability of necessity (Eq. 4.1) captures the probability that on changing the attribute A from its default value of a to the intervened value of $\hat{a}$, the classifier will return a negative outcome instead. In other words, it

measures the extent of positive classifications that are attributable to the *original state* of the attribute $A = a$. The probability of sufficiency (Eq. 4.2) is the dual of the probability of necessity. It applies to the sub-population of input images with the default attribute value $\hat{a}$, for which the classifier produced a negative output. It measures the effect of changing the attribute by intervention to a from its default state of $\hat{a}$. It computes the probability that this change could cause the classifier to return a positive outcome for these cases, which were initially handed out a

negative outcome. Hence, it measures the *capacity* of setting A to a to *flip* the negative outcome from the classifier.

We can choose to *change* the attribute from its default state by moving in its direction of increase or decrease. This would allow us to measure the sufficiency and necessity of changing the attributes in both directions and give a more in-depth understanding of how features are influencing the outcome of the black-box classifier. We denote necessity scores of increasing the attribute with NEC^{+} and the necessity scores of decreasing the attribute with NEC^{-}. We follow a similar notation for sufficiency scores.

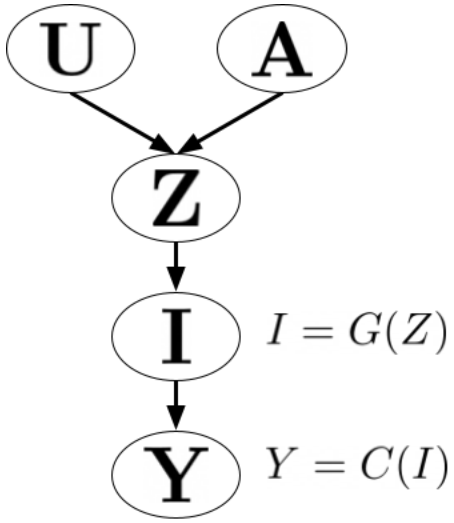


Figure 4.3 Causal model.

4.5.4 Computing Necessity and Sufficiency

As shown in Figure 4.3, we assume a causal model that has the following components. *a*) Unobserved attributes ($\mathcal{U}$) - These are all the attributes for which we have no observations and cannot account. They constitute the exogenous variable in our causal model. *b*) A set of interpretable attributes ($\mathcal{A}$) - This is a set of attributes for which we know the values. They are our endogenous variables in the causal model. *c*) A latent space ($\boldsymbol{Z}$) - The unobserved (exogenous) attributes and observed set of interpretable attributes (endogenous) together directly affect the

value of the latent space. *d*) A generative model (G) - It takes as input the above-mentioned latent space $\mathbf{Z}$ and transforms it into an image I. *e*) The classifier to be explained (C) - It takes as input the image I and produces the target label, Y.

We generate counterfactuals and pass the produced images (both originals and counterfactuals) to the black-box classifier to obtain the classifier's output. We use this information to compute the sufficiency and necessity scores. We generate counterfactuals by following Pearl's three-step procedure (Pearl, 2009):

- **Abduction** Given the prior distribution of the latent variable of a generative model G and the set of attributes $\mathcal{A}$, train a model M that estimates the updated probability of latent variable conditioned on any subset of attributes:

$$P(\boldsymbol{z} \mid \bar{\mathcal{A}} = \boldsymbol{a}), \bar{\mathcal{A}} \subseteq \mathcal{A}. \quad 4.3$$

- **Action** Take a set of interpretable attributes $\bar{\mathcal{A}}$. Based on the causal model, perform an intervention by setting a subset of attributes $\bar{\mathcal{A}} \subseteq \mathcal{A}$ to their determined values $\bar{\mathcal{A}} \leftarrow \hat{\boldsymbol{a}}$.
- **Prediction** Given the model M, obtain the modified latent vector probability corresponding to the new value of the attribute (s). Pass this modified latent vector into the generative model to obtain the corresponding image. Finally, run black-box classifier inference for this counterfactual image to obtain the target output:

$$\hat{I} = G(\hat{\boldsymbol{z}}), \hat{\boldsymbol{z}} = M(\boldsymbol{z}, \bar{\mathcal{A}} = \hat{\boldsymbol{a}}), \hat{Y} = C(\hat{I}). \quad 4.4$$

4.6 Counterfactual Generation

Conducting the three steps of counterfactual generation is a non-trivial task due to the complication that arises when setting the attributes in the Abduction and Prediction steps. Specifically, generative models tend to have very complex latent spaces, where finding a path from

$\boldsymbol{z}$ to $\hat{\boldsymbol{z}}$ for attribute change is intractable. To reconcile it, we propose training an MLP model M that serves as the shift predictor in our pipeline and can provide a prediction for $\hat{\boldsymbol{z}} = M(\boldsymbol{z}, \bar{\mathcal{A}} = \boldsymbol{a})$. With the use of M, we can now update the probability of latent variable $P(\boldsymbol{z})$ to the probability of counterfactual latent variable $P(\boldsymbol{z} \mid \bar{\mathcal{A}} = \boldsymbol{a})$ in the prediction step and follow Pearl's procedure. In the following, we provide the details on the generative model and shift predictor algorithm.

Generative model. Generative models are vastly popular in different fields of AI, and their recent advancements in creating realistic images have made them a viable approach to producing a latent representation of an image dataset. In our experiments, we utilize StyleGAN2(Karras et al., 2020) as a state-of-the-art generative model which can be used to generate high resolution and realistic images in different domains. StyleGAN feeds the latent variable into a mapping network that transforms it into an intermediate latent variable. Aside from its ability to produce styles, this transformation also provides the intermediary latent space as a more regulated domain to learn and traverse through interpretable attributes.

Shift predictor. A shift predictor model is an MLP model that can take the latent variable of an image from a generative model G and generates the latent variable for its counterfactual based on the attributes produced by a classifier (Figure 4.4). For a generative model, $G: \mathcal{R}^d \rightarrow \mathcal{R}^n$ that has a latent space with dimension d and a classifier $C: \mathcal{R}^n \rightarrow \mathcal{R}^m$ that predicts m attributes, we define our shift predictor as $M(\boldsymbol{z}, \hat{\boldsymbol{y}}): \mathcal{R}^d \times \mathcal{R}^m \rightarrow \mathcal{R}^d$, where $\boldsymbol{z} \boldsymbol{\epsilon} \mathcal{R}^d$ is the latent variable for the input image and $\hat{\boldsymbol{y}}$ denotes the attributes for the intended counterfactuals.

Algorithm 4.1 Training a shift predictor for binary attributes

Data: Classifier C. Generative model G. Counterfactual faithfulness ratio γ.

Result: Parameters θ_M for the shift predictor model M.

init:

$\theta_M \leftarrow$ Random initialization

For *number of iterations* **do**

Sample a batch of b noise variables and target outputs:

$\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(b)}\} \leftarrow p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\{\hat{\boldsymbol{y}}_i^{(1)}, \ldots, \hat{\boldsymbol{y}}_i^{(b)}\} \leftarrow Bern.(p = 0.5), \qquad \forall i = 1..m$

Predict counterfactual latent codes:

$\hat{\mathbf{z}}^{(j)} \leftarrow M(\mathbf{z}^{(j)}, \hat{\boldsymbol{y}}^{(j)}), \forall i = 1..b$

Generate images from the noise variables and predict attributes by the classifier:

$I^{(j)} \leftarrow G(\hat{\mathbf{z}}^{(j)}), \forall i = 1..b$

$\mathbf{y}^{(j)} \leftarrow G(I^{(j)}), \forall i = 1..b$

Compute attribute conditioning and faithfulness loss:

$$\mathcal{L}_a \leftarrow \frac{1}{b}\sum_{j=1}^{b}\sum_{i=1}^{m} -\hat{\boldsymbol{y}}_i^{(j)} \boldsymbol{log}(y_i^{(j)})$$

$$\mathcal{L}_f \leftarrow \frac{1}{b}\sum_{j=1}^{b} ||\hat{\mathbf{z}}^{(j)} - \mathbf{z}^{(j)}||$$

Update the shift predictor parameters: $M \leftarrow \nabla\theta_M(\mathcal{L}_a + \gamma\mathcal{L}_f)$

end

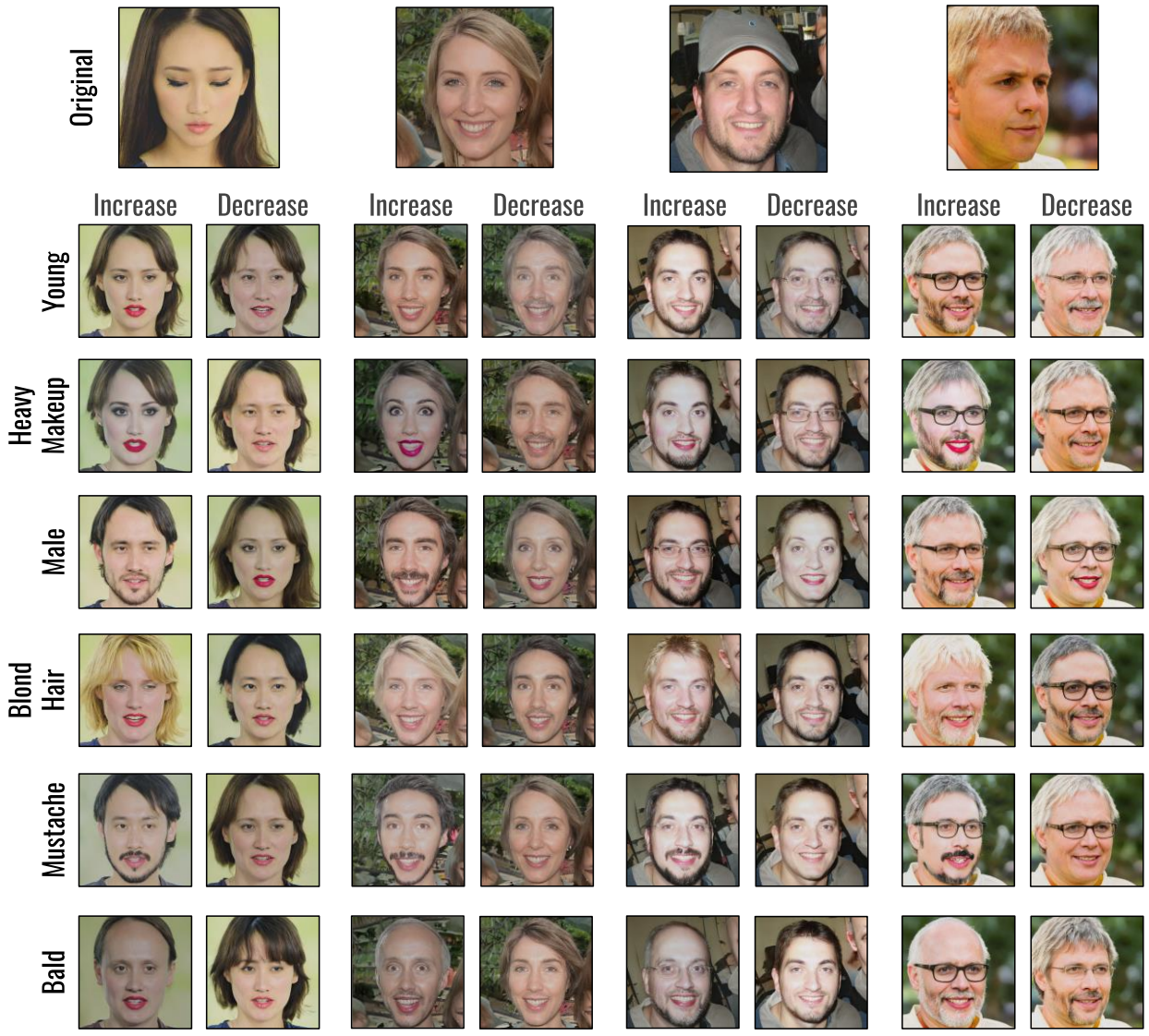


Figure 4.4 Examples of counterfactual images generated during the computation of the explanations scores. The first row shows the original images. Each subsequent row shows the original image modified to move towards the direction of increase and decrease of the labeled attribute. Along with necessity and sufficiency scores, these representative images are provided to the user as global explanations for the classifier.

In the training process, the shift predictor learns the directions in the latent space of G that correspond to changes in the attributes predicted by the classifier. Without any manual labeling, the training procedure only requires the latent variables of images from G to input the shift predictor and supervise it with the labels generated by the classifier (see Algorithm 4.1).

During the training, the shift predictor learns to produce a counterfactual latent variable that satisfies any combination of attributes defined by $\hat{\boldsymbol{y}}$. In other words, if the classifier predicts a

set of attributes $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$, shift predictor can provide a counterfactual latent variable compatible with any selected subset of attributes $\bar{\mathcal{A}}$:

$$\hat{\boldsymbol{z}} = M(\boldsymbol{z}, \{A_i = \hat{a}_i \mid A_i \in \bar{\mathcal{A}}\}). \quad 4.5$$

Under the assumption of proper training, the shift predictor approximates latent variable distribution conditioned by the subset of attributes $\bar{\mathcal{A}}$:

$$\hat{\boldsymbol{z}} \sim P(\boldsymbol{z} \mid \{A_i = \hat{a}_i \mid A_i \in \bar{\mathcal{A}}\}), \bar{\mathcal{A}} \subseteq \mathcal{A}. \quad 4.6$$

The loss function in the training process pursues two objectives: 1) minimizing the error in the prediction of attributes $\bar{\mathcal{A}}$ for the counterfactual image, 2) assuring a level of faithfulness and similarity between the original input image and its newly generated counterfactual. The attribute loss $\mathcal{L}_a$ is defined as a cross-entropy between the conditioned attributes and the attributes predicted by the classifier. In the training process, the conditioned attributes $\bar{\mathcal{A}}$ are distinguished from unset attributes so that the loss will be only calculated for them. On the other hand, the faithfulness loss $\mathcal{L}_f$ is calculated as the normal distance between the original latent variables and their counterfactuals. The overall loss in the training process is defined as a combination of these two losses with a faithfulness factor γ which establishes a balance between attribute accuracy of counterfactuals and their faithfulness to the original input:

$$\mathcal{L} = \mathcal{L}_a + \gamma \mathcal{L}_f = \sum_{A_i \in \bar{\mathcal{A}}} -\hat{\mathrm{y}}_i log(\mathrm{y}_i) + \gamma ||\hat{\boldsymbol{z}} - \boldsymbol{z}||. \quad 4.7$$

4.7 Experiments and Results

We run our experiments to explain black-box classifiers that are trained on the task of classifying face images. We have annotations for the set of interpretable attributes A that we will choose to use to explain the model's behavior. We train a multi-task classifier built on top of a pretrained VGG backbone to predict the set of interpretable attributes A for new unseen images.

The CelebA dataset is used as the training set and provides 39 binary attributes, including attractiveness which we use as the target output Y for any black-box classifier of choice. As the set of interpretable explanatory attributes (A), we choose six other labels: blonde hair, heavy makeup, baldness, mustache, youngness, and maleness. We make a simplifying assumption to use an underlying causal model in which the explanatory attributes are independent of each other. We model attractiveness as the positive class $(y = 1.0)$ and unattractiveness as the negative class $(\hat{y} = 0.0)$ for the black-box classifier to predict. In the set of interpretable attributes (A), an attribute has its default value as (a) when it is not explicitly set. Otherwise, during an intervention, it is set to value $(\hat{a})$ which can be $+1$ if we want to move in the direction of its increase and -1 if we want to move in the direction of its decrease. We intervene and set $\hat{a}$ to 0 if we do not wish to modify the attribute. Our initial dataset consists of 200 images randomly produced by the generative model. We pass the images through the multi-task classifier to obtain the attribute values and the black-box classifier to obtain the target label. We seek to explain the behavior of the target output Y using the attributes A on this dataset by performing the following experiments:

- **Linear baseline**, we first consider an interpretable linear approximation of target label behavior w.r.t. the attributes as the underlying black-box model. We use this approximation as ground truth and assess the validity of necessity and sufficiency scores in capturing this ground truth linear behavior.

- **Black-box explanations**, we consider a complex black-box classifier built on a pretrained VGG backbone and generate sufficiency and necessity scores to explain it. In conjunction with generated counterfactual images, we use these scores to analyze how increasing and decreasing the attributes affects the classification into attractive and not-attractive labels by the classifier.

4.7.1 Linear Baseline

Our pipeline of generating counterfactual images and explanation scores is agnostic to the type of black-box model being explained. This implies that it is independent of the kind of machine learning or deep learning model used. However, one way to test the quality of our explanations is by generating explanations for a case where we have access to the decision-making rationale of the underlying black-box model. To this end, we choose a logistic regression classifier as the model whose decision we seek to explain. This classifier takes actual values corresponding to the feature attributes obtained from the multi-task classifier and predicts the target label of attractiveness using only these values. The coefficients of the logistic regressor corresponding to the different features indicate how the model is making its decisions. We compare this to sufficiency and necessity scores generated by our method, which explains this logistic regression model.

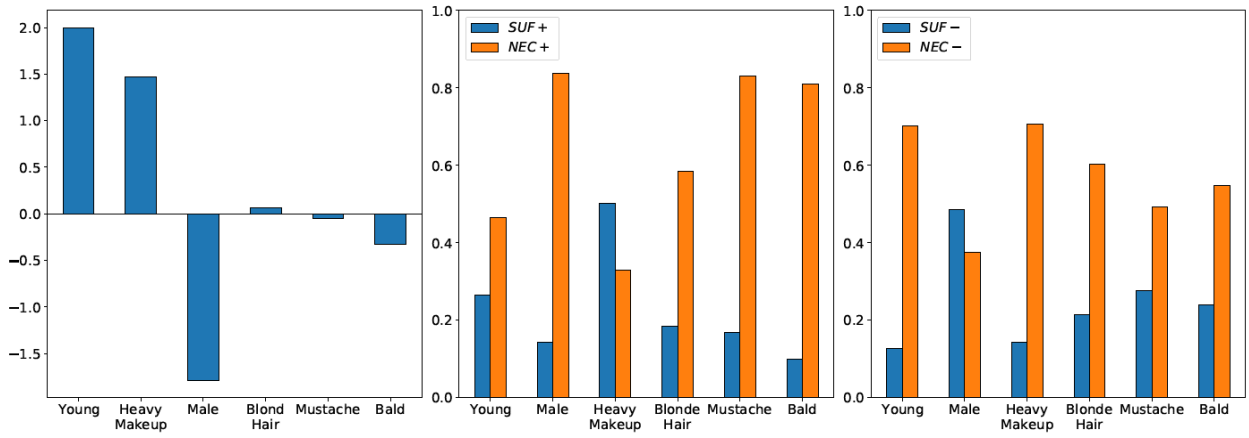


Figure 4.5 **Left**: The coefficients of the known black-box logistic regressor. The sufficiency and necessity scores explain the logistic regressor behavior when attributes are increased (**Center**) and decreased (**Right)**.

We observe from Figure 4.5 that features that have negative attributions, such as those for male, mustache, and bald, in the logistic regressor model carry a significant value of necessity when we move in the directions of their increase. This high necessity score indicates that leaving them unset compared to increasing them is most important to allow attractive individuals to keep

their attractiveness score high. Similarly, for people classified as unattractive, the features that had high positive attributions in the logistic regressor, such as Young and Heavy Makeup, carry the highest values of sufficiency when we increase their values. This indicates that increasing heavy makeup and youngness are the two most important factors to flip the outcome from not attractive to attractive. We can interpret the SUF/NEC^{-} scores in a similar way. This contrastive and counterfactual analysis is not possible through simple coefficients obtained from the logistic regressor. This makes it essential to use these notions of sufficiency and necessity over standard coefficient-based attributions that have been observed to have multiple shortcomings due to their overly simplistic nature. These include issues like their dependence on feature pre-processing methods and instability due to different feature selections.

4.7.2 Black-box Explanations

We use our pipeline to generate global explanations for the black-box attractiveness classifier. Here, the black-box classifier is built on a pretrained VGG backbone. Our explanations are two-fold. First, we provide sufficiency and necessity scores on a population level for the six different attributes. In addition to this, we also give the users the counterfactual images that were generated by our shift predictor during the computation of the scores. The ability to have both feature attributions and the images that led to the calculation of those attributions allows the user to understand model behavior at a deeper level.

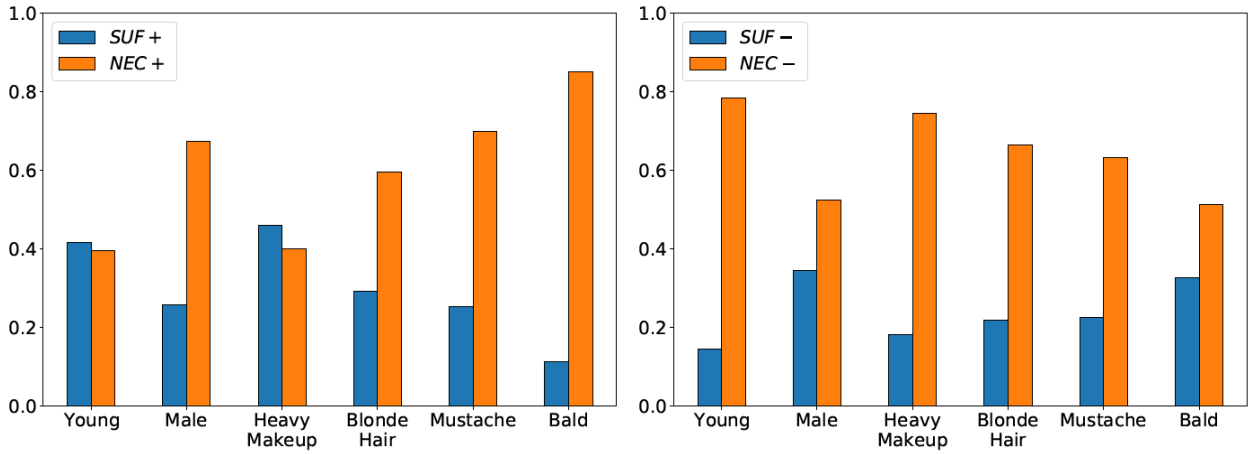


Figure 4.6 Sufficiency and necessity scores as global explanations. **Left**: Not increasing baldness and mustache are most necessary to remain classified as attractive. Increasing heavy makeup and youngness is most sufficient to flip an unattractive outcome to attractive. **Right**: Not decreasing youngness and heavy makeup is necessary to stay classified as attractive. Reducing maleness and baldness are most sufficient to reverse the outcome from unattractive to attractive.

Figure 4.4 contains a set of representative images. We can see how the original image changes in the direction of decrease and increase of the explanatory attributes. Figure 4.6 shows the overall sufficiency and necessity scores of attributes in both positive and negative increase directions. This gives us a detailed analysis of how the features affect the classifier output. For instance, a high SUF^{+} value of the attributes Heavy Makeup, Blond Hair, and Young implies that moving in the direction of increase of these attributes is most sufficient to flip an outcome of unattractiveness to attractiveness. Similarly, a high SUF^{-} value of the attributes Male, and Bald reflects that when moving in the direction of a decrease for these attributes, we are most likely to be able to flip our outcome from not attractive to attractive. The necessity scores inform us about the most important attributes to be left "unset" in their default state compared to increasing or decreasing them for a person classified as attractive to maintain that classification. A high NEC^{+} score of Baldness, Moustache, and Male is indicative of the fact that one should avoid increasing these attributes if they wish to remain classified as attractive by the classifier. Similarly, the high NEC^{-} scores of Young, Heavy Makeup, and Blond Hair are indicative of the fact that one should

avoid decreasing these attributes if they wish to remain classified as attractive by the classifier. With the generated counterfactual images as evidence, the necessity and sufficiency scores provide a holistic understanding of the black-box classifier to the end-user.

4.8 Conclusion

This chapter proposed an end-to-end pipeline that generated counterfactuals from a pretrained generative model and used that to help compute probabilistic causal counterfactual scores. These scores, along with the generated images, served as explanations for any underlying black-box image classifier. Our work also highlighted the need and advantages of these contrastive explanations over simple feature attributions. However, one of the drawbacks of our current method is that it does not effectively disentangle the effects between attributes. We would want to improve on that aspect by learning a structural causal model that can model the impact that attributes have on one another. This would also allow us to extend our analysis to compute the direct and indirect effects (Pearl, 2022) of attributes on the target label. Furthermore, we would like to apply this pipeline to detect and mitigate bias in image classification systems.

4.9 Acknowledgments

Chapter 4, in part, has been submitted for publication of the material as it may appear in the European Conference on Computer Vision (ECCV), 2022, Alipour, Kamran; Lahiri, Aditya; Adeli, Ehsan; Salimi, Babak; Pazzani, Michael. The dissertation author was the primary researcher and author of this paper.

CONCLUSION AND FUTURE WORK

In this dissertation, we introduce methods to generate explanations for AI machines and mechanisms to evaluate them effectively for complex computer vision tasks such as Visual Question Answering (VQA) and face image classification.

Explanation Evaluation

In Chapter 1, we create an interactive experiment to test explanations' effectiveness in boosting user prediction accuracy, confidence, and reliance in a VQA task. Our findings suggest that explanations aid in improving VQA accuracy, and explanation evaluations support the effectiveness of explanations in human-machine AI collaboration tasks.

We performed a user survey with 90 participants to evaluate various modalities of explanation. Users rated different explanation modalities interactively and used them to predict AI system behavior. In the cases of inaccurate AI response, users had improved performance on user-machine tasks when exposed to the explanations. When users read explanations that demonstrated the potential of our multi-modal explanation system in gaining user trust, their confidence in predictions increased.

One of the critical aspects of explanations that can be further evaluated is the scalability of the explanation. Our study focused on the specific task of visual question answering and conducted extensive user studies to assess different modalities of explanation. While such an approach provides valuable lessons about the interpretability of that particular task, it cannot guarantee the scalability of results in other AI applications. We can conduct similar studies for a broader range of applications in future work or choose primary AI tasks (e.g., image classification) that can be arguably generalizable to different applications.

On the other hand, we inherently focused on the AI side during our human-AI collaboration experiments. However, another crucial control here would be the human factors. We applied soft controls to the recruitment process as well as the study guide to limit the impact of prior knowledge/biases. In future versions of the experiment, we can apply a more specific control in these areas and also extend the evaluations into qualitative analysis conducted by professionals.

Attention-based Explanations

Chapter 2 extends the explanation evaluation experiment to include competency measurements. We evaluate the role of attention map explanations on the user's mental model of AI competency. We designed an experiment where subjects rank the performance of the VQA model among four different types of questions. To quantify the subjects' mental model, we compute the correlation between user rankings and the AI's actual ranking among the question types.

We propose a new XVQA model that produces answers and attention maps from spatial and object features of the image. This explainable model uses a transformer attention module to better process the visual and textual embeddings of the input. The proposed model is compared with a baseline model to show the effect of input object features and the attention module.

We also include an interactive explanation mode (Active Attention), in which users may guide the system's attention and get responses from it. When the user confidence development between active attention and other explanation groups is compared, the active attention explanation group exhibits a greater degree of trust from the users, demonstrating the efficiency of interactive explanations in constructing a better mental model of the AI system. For future improvements, we can look at more effective methods to intervene with the attention mechanisms of AI models so

that such interventions become more intuitive to the lay users and guarantee a causal effect on AI behavior.

The experiment's overall findings indicate that the user's mental model improves when exposed to attention-based explanations. The strong correlation between users' evaluations of explanations and their performance in the prediction tasks demonstrates the efficacy of explanations in user-machine task performance. The evolution of the user's mental model throughout the studies suggests that the learning rate is faster in the presence of explanations. Furthermore, our experiments show that adding the object feature and the transformer model positively influences the explanations. While evaluating the explanations, we primarily relied on Likert scales and user input; we can improve this process with more enriched data acquisition techniques such as asking users to provide textual feedback on explanation modalities. Such feedback can be processed automatically into a more descriptive evaluation system.

In our studies, we focused on models with attention mechanisms trained only based on accuracy objectives. In this setting, we attempted to improve the attention-based explanations either by upgrading the attention mechanism or by combining attention with interpretable annotations. Another direction worth investigating is modifying the training objectives for these models and considering human interpretability in their attention alongside accuracy. Analyzing the tradeoff between interpretability and accuracy in such models would be an interesting research topic.

In Chapter 3, we demonstrate that showing counterfactual images is helpful for the mental model improvement of users in predicting a VQA model’s performance. We offer that attention-based edits are more effective than showing random examples or not showing any examples at all. We also exhibit a generative approach to generating counterfactual images that can also be

effective at improving users' mental models. Investigating different image editing methods and the impact of the counterfactual quality on users' mental models can serve as interesting topics for the next steps of this work. These results can serve as a foundation to improve generative models for producing effective counterfactual explanations to improve user mental models for the safe and effective deployment of AI systems in the wild.

In our current implementation, we limit the counterfactual edits to a specific size to avoid noise in the counterfactuals. In future versions of this technique, we can take a closer look at improving the algorithms that edit the images. Such algorithms need to minimize artifacts in their edits to avoid unwanted impact on AI's performance and output. Also, with the newer generation of attention-based AI machines, such as vision transformers, we can define a more enriched attention mechanism to choose the counterfactual edit areas.

Causal Counterfactuals

In Chapter 4, we take a more fundamental look at the process of creating counterfactuals for explanation purposes. We introduce a method to produce contrastive counterfactuals for an image classifier. Our approach can use generative models pre-trained on any dataset independent of the classifier training dataset. We propose contextual, contrastive, and causal explanations in the form of sufficiency and necessity scores to explain the black-box model. We use our method to provide global and local explanations for a black-box classifier trained on the CelebA dataset.

We present an end-to-end pipeline that utilizes a pretrained generative model and computes probabilistic causal counterfactual scores to produce contrastive counterfactuals. Together with the produced images, these scores are used to explain the underlying black-box image classifier. Our research also demonstrates the importance and benefits of contrastive explanations over

fundamental feature attributions. However, one disadvantage of our current strategy is that it does not adequately separate the impacts of qualities. We'd want to build on that by developing a structural causal model that can represent the impact these qualities have on one another. This would also allow us to broaden our study to calculate the direct and indirect effects of qualities on the target label. Furthermore, we would like to apply this pipeline to detect and mitigate bias in image classification.

We propose an end-to-end pipeline that generated counterfactuals from a pretrained generative model and used that to help compute probabilistic causal counterfactual scores. These scores can be viewed as global explanations, while the generated images can serve as explanations for any underlying black-box image classifier. Our work also highlighted the need and advantages of these contrastive explanations over simple feature attributions.

One of the main downsides of the proposed approach is a lack of disentanglement between explainer attributes. Learning a structural causal model that can model the potential confoundedness across these attributes can provide a more straightforward explanation. This perspective would also extend our analysis to compute attributes' direct and indirect effects on the target label.

On the other hand, our method relies on existing labels to generate contrastive counterfactuals, while such annotations may not be readily available in so many applications. In future implementations of this approach, we can investigate techniques to learn and predict interpretable labels and use them in causal modeling. However, this objective is an open research question under the unsupervised causal interpretation that can impose its own challenges.

REFERENCES

Abdal, R., Zhu, P., Mitra, N. J., & Wonka, P. (2021). Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG), 40*(3), 1-21.

Agarwal, V., Shetty, R., & Fritz, M. (2020). Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Alipour, K., Ray, A., Lin, X., Schulze, J. P., Yao, Y., & Burachas, G. T. (2020). The Impact of Explanations on AI Competency Prediction in VQA. 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI),

Alipour, K., Schulze, J. P., Yao, Y., Ziskind, A., & Burachas, G. (2020). A study on multimodal and interactive explanations for visual question answering. *arXiv preprint arXiv:2003.00431*.

An, S., Choi, H., & Jeon, J.-J. (2021). EXoN: EXplainable encoder Network. *arXiv preprint arXiv:2105.10867*.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the IEEE conference on computer vision and pattern recognition,

Angwin, J., & Larson, J. (2016). *Machine Bias*. Retrieved 2022 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Anne Hendricks, L., Hu, R., Darrell, T., & Akata, Z. (2018). Grounding visual explanations. Proceedings of the European Conference on Computer Vision (ECCV),

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. Proceedings of the IEEE international conference on computer vision,

Bahadori, M. T., & Heckerman, D. E. (2020). Debiasing concept-based explanations with causal analysis. *arXiv preprint arXiv:2007.11500*.

Bertossi, L., Li, J., Schleich, M., Suciu, D., & Vagena, Z. (2020). Causality-based explanation of classification outcomes. Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning,

Castro, D. C., Tan, J., Kainz, B., Konukoglu, E., & Glocker, B. (2019). Morpho-MNIST: quantitative assessment and diagnostics for representation learning. *Journal of Machine Learning Research*, *20*(178), 1-29.

Chandrasekaran, A., Prabhu, V., Yadav, D., Chattopadhyay, P., & Parikh, D. (2018). Do explanations make VQA models more predictable to a human? *arXiv preprint arXiv:1810.12366*.

Chandrasekaran, A., Yadav, D., Chattopadhyay, P., Prabhu, V., & Parikh, D. (2017). It Takes Two to Tango: Towards Theory of AI?s Mind. *arXiv preprint arXiv:1704.00717*.

Chang, C.-H., Creager, E., Goldenberg, A., & Duvenaud, D. (2018). Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024*.

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. (2003). Is seeing believing?: how recommender system interfaces affect users? opinions. Proceedings of the SIGCHI conference on Human factors in computing systems,

Das, A., Agrawal, H., Zitnick, L., Parikh, D., & Batra, D. (2017). Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, *163*, 90-100.

Dash, S., Balasubramanian, V. N., & Sharma, A. (2022). Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,

De Graaf, M. M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). 2017 AAAI Fall Symposium Series,

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment.*IUI â€ ™19* Proceedings of the 24th International Conference on Intelligent User Interfaces, New York, NY, USA.

Drozdal, J., Weisz, J., Wang, D., Dass, G., Yao, B., Zhao, C., . . . Su, H. (2020). Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems.*IUI â€ ™20* Proceedings of the 25th International Conference on Intelligent User Interfaces, New York, NY, USA.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Galhotra, S., Pradhan, R., & Salimi, B. (2021). Explaining black-box algorithms using probabilistic contrastive counterfactuals. Proceedings of the 2021 International Conference on Management of Data,

Gat, I., Lorberbom, G., Schwartz, I., & Hazan, T. (2021). Latent Space Explanation by Intervention. *arXiv preprint arXiv:2112.04895*.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. CogSci,

Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. (2019). Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems, 32*.

Ghosh, S., Burachas, G., Ray, A., & Ziskind, A. (2019). Generating natural language explanations for visual question answering using scene graphs and visual attention. *arXiv preprint arXiv:1902.05715*.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA),

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Goyal, Y., Feder, A., Shalit, U., & Kim, B. (2019). Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. Conference on Computer Vision and Pattern Recognition (CVPR),

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual visual explanations. International Conference on Machine Learning,

Greenland, S., & Robins, J. M. (1999). Epidemiology, justice, and the probability of causation. *Jurimetrics*, *40*, 321.

Guo, C., Gardner, J., You, Y., Wilson, A. G., & Weinberger, K. (2019). Simple black-box adversarial attacks. International Conference on Machine Learning,

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. Proceedings of the IEEE international conference on computer vision,

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition,

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review*, *7*(2), 185-207.

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. European Conference on Computer Vision,

Huk Park, D., Anne Hendricks, L., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Härkönen, E., Hertzmann, A., Lehtinen, J., & Paris, S. (2020). Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, *33*, 9841-9850.

Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., & Parikh, D. (2018). Pythia v0.1: the Winning Entry to the VQA Challenge 2018. *CoRR*, *abs/1807.09956*.

Jiang, Z., Wang, Y., Davis, L., Andrews, W., & Rozgic, V. (2017). Learning discriminative features via label consistent neural network. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV),

Jonassen, D. H., & Ionas, I. G. (2008). Designing effective supports for causal reasoning. *Educational Technology Research and Development*, *56*(3), 287-308.

Karimi, A.-H., Barthe, G., Balle, B., & Valera, I. (2020). Model-agnostic counterfactual explanations for consequential decisions. International Conference on Artificial Intelligence and Statistics,

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Kazemi, V., & Elqursh, A. (2017). Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*.

Keane, M. T., & Smyth, B. (2020). Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). International Conference on Case-Based Reasoning,

Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, *294*, 103459.

Khademi, A., & Honavar, V. (2020). A Causal Lens for Peeking into Black Box Predictive Models: Predictive Model Interpretation via Causal Attribution. *arXiv preprint arXiv:2008.00357*.

Kommiya Mothilal, R., Mahajan, D., Tan, C., & Sharma, A. (2021). Towards unifying feature attribution and counterfactual explanations: Different means to the same end. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society,

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., . . . Fei-Fei, L. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, *123*(1), 32-73.

Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012). Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., & Doshi-Velez, F. (2019). An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*.

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. Proceedings of the Conference on Fairness, Accountability, and Transparency,

Lane, H. C., Core, M. G., Van Lent, M., Solomon, S., & Gomboc, D. (2005). *Explainable artificial intelligence for training and tutoring*.

Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., . . . Irani, M. (2021). Explaining in Style: Training a GAN to explain a classifier in StyleSpace. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., & Detyniecki, M. (2017). Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*.

Li, Z., & Xu, C. (2021). Discover the Unknown Biased Attribute of an Image Classifier. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Liu, K., Cao, G., Zhou, F., Liu, B., Duan, J., & Qiu, G. (2022). Towards Disentangling Latent Space for Unsupervised Semantic Face Editing. *IEEE Transactions on Image Processing*.

Liu, S., Kailkhura, B., Loveland, D., & Han, Y. (2019). Generative counterfactual introspection for explainable deep learning. 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP),

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. Proceedings of the IEEE international conference on computer vision,

Lomas, M., Chevalier, R., Cross Ii, E. V., Garrett, R. C., Hoare, J., & Kopack, M. (2012). Explaining robot actions. Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction,

Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. Advances In Neural Information Processing Systems,

Lu, Y.-D., Lee, H.-Y., Tseng, H.-Y., & Yang, M.-H. (2020). Unsupervised discovery of disentangled manifolds in gans. *arXiv preprint arXiv:2011.11842*.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence, 267*, 1-38.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Morton, A. (2013). Contrastive knowledge. *Contrastivism in philosophy*, 101-115.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems?: an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*.

Nemirovsky, D., Thiebaut, N., Xu, Y., & Gupta, A. (2020). Countergan: Generating realistic counterfactuals with residual generative adversarial nets. *arXiv preprint arXiv:2009.05199*.

Nie, W., Vahdat, A., & Anandkumar, A. (2021). Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems, 34*.

Nitzan, Y., Bermano, A., Li, Y., & Cohen-Or, D. (2020). Face identity disentanglement via latent space mapping. *arXiv preprint arXiv:2005.07728*.

O'Shaughnessy, M., Canal, G., Connor, M., Rozell, C., & Davenport, M. (2020). Generative causal explanations of black-box classifiers. *Advances in Neural Information Processing Systems, 33*, 5453-5467.

Parafita, Á., & Vitrià, J. (2019). Explaining visual models by causal attribution. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW),

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., & Lischinski, D. (2021). Styleclip: Text-driven manipulation of stylegan imagery. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Patro, B., Namboodiri, V., & others. (2020). Explanation vs attention: A two-player game to obtain attention for VQA. Proceedings of the AAAI Conference on Artificial Intelligence,

Pawlowski, N., Coelho de Castro, D., & Glocker, B. (2020). Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems, 33*, 857-869.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM, 62*(3), 54-60.

Pearl, J. (2022). Direct and indirect effects. In *Probabilistic and Causal Inference: The Works of Judea Pearl* (pp. 373-392).

Peng, L., Yang, Y., Wang, Z., Huang, Z., & Shen, H. T. (2020). MRA-Net: Improving VQA via Multi-modal Relation Attention Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),

Ray, A., Burachas, G., Yao, Y., & Divakaran, A. (2019). Lucid Explanations Help: Using a Human-AI Image-Guessing Game to Evaluate Machine Explanation Helpfulness. *arXiv preprint arXiv:1904.03285*.

Ray, A., Christie, G., Bansal, M., Batra, D., & Parikh, D. (2016). Question relevance in VQA: identifying non-visual and false-premise questions. *arXiv preprint arXiv:1606.06622*.

Ray, A., Cogswell, M., Lin, X., Alipour, K., Divakaran, A., Yao, Y., & Burachas, G. (2021). Knowing What VQA Does Not: Pointing to Error-Inducing Regions to Improve Explanation Helpfulness. *arXiv preprint arXiv:2103.14712*.

Ray, A., Sikka, K., Divakaran, A., Lee, S., & Burachas, G. (2019). Sunny and Dark Outside?! Improving Answer Consistency in VQA through Entailed Question Generation. *arXiv preprint arXiv:1909.04696*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems,

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,

Ritter, F. E., & Schooler, L. J. (2001). The learning curve. *International encyclopedia of the social and behavioral sciences*, *13*, 8602-8605.

Russell, C., Kusner, M. J., Loftus, J., & Silva, R. (2017). When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in neural information processing systems*, *30*.

Rutjes, H., Willemsen, M., & Ijsselsteijn, W. (2019). Considerations on explainable AI and usersâ€ ™mental models. CHI 2019 Workshop: Where is the Human? Bridging the Gap Between AI and HCI,

Samangouei, P., Saeedi, A., Nakagawa, L., & Silberman, N. (2018). Explaingan: Model explanation via decision boundary crossing transformations. Proceedings of the European Conference on Computer Vision (ECCV),

Sani, N., Malinsky, D., & Shpitser, I. (2020). Explaining the behavior of black-box prediction algorithms with causal learning. *arXiv preprint arXiv:2006.02482*.

Schwab, P., & Karlen, W. (2019). Cxplain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems*, *32*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision,

Selvaraju, R. R., Tendulkar, P., Parikh, D., Horvitz, E., Ribeiro, M. T., Nushi, B., & Kamar, E. (2020). SQuINTing at VQA Models: Introspecting VQA Models With Sub-Questions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Shen, Y., Gu, J., Tang, X., & Zhou, B. (2020). Interpreting the latent space of gans for semantic face editing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Shortliffe, E. H., & Buchanan, B. G. (1984). A model of inexact reasoning in medicine. *Rule-based expert systems*, 233-262.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., . . . Rohrbach, M. (2019). Towards VQA Models That Can Read. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, *9*, 11974-12001.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-First AAAI Conference on Artificial Intelligence,

Teney, D., Anderson, P., He, X., & Hengel, A. v. d. (2017). Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge. *CoRR, abs/1708.02711*.

Teney, D., Anderson, P., He, X., & Van Den Hengel, A. (2018). Tips and tricks for visual question answering: Learnings from the 2017 challenge. Proceedings of the IEEE conference on computer vision and pattern recognition,

Thiagarajan, J. J., Narayanaswamy, V., Anirudh, R., Bremer, P.-T., & Spanias, A. (2020). Accurate and robust feature importance estimation under distribution shifts. *arXiv preprint arXiv:2009.14454*.

Tzelepis, C., Tzimiropoulos, G., & Patras, I. (2021). WarpedGANSpace: Finding non-linear RBF paths in GAN latent space. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. Proceedings of the conference on fairness, accountability, and transparency,

Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. Proceedings of the national conference on artificial intelligence,

Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

Verschueren, N., Schroyens, W., Schaeken, W., & d'Ydewalle, G. (2004). The interpretation of the concepts 'necessity'and 'sufficiency'in forward unicausal relations. *Current psychology letters. Behaviour, brain & cognition*(14, Vol. 3, 2004).

Voynov, A., & Babenko, A. (2020). Unsupervised discovery of interpretable directions in the gan latent space. International conference on machine learning,

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.

Watson, D. S., Gultchin, L., Taly, A., & Floridi, L. (2021). Local explanations via necessity and sufficiency: unifying theory and practice. Uncertainty in Artificial Intelligence,

Xu, H., & Saenko, K. (2016). Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. European Conference on Computer Vision,

Yang, H., Chai, L., Wen, Q., Zhao, S., Sun, Z., & He, S. (2021). Discovering Interpretable Latent Space Directions of GANs Beyond Binary Attributes. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. Proceedings of the IEEE conference on computer vision and pattern recognition,

Yüksel, O. K., Simsar, E., Er, E. G., & Yanardag, P. (2021). Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Zaeem, M. N., & Komeili, M. (2021). Cause and effect: Concept-based explanation of neural networks. 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC),

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. European conference on computer vision,

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Zhang, Y., Hare, J., & Prügel-Bennett, A. (2018). Learning to count objects in natural images for visual question answering. *arXiv preprint arXiv:1802.05766*.

Zhang, Y., Niebles, J. C., & Soto, A. (2019). Interpretable visual question answering by visual grounding from attention supervision mining. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV),

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*.