UNIVERSITY OF CALIFORNIA
RIVERSIDE

Statistical Analysis of WCET Estimation on DNNs

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Electrical Engineering

by

Ankith Jain Rakesh Kumar

March 2018

Thesis Committee:

Dr. Hyoseung Kim, Chairperson
Dr. Daniel Wong
Dr. Salman Asif

The Thesis of Ankith Jain Rakesh Kumar is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to express my sincere gratitude to my advisor Professor Hyoseung Kim, for his support in study and research. His guidance and experience helped me in understanding the research topics and enabled me to shape my Thesis.

I also thank my friends and lab-mates for helping me out in understanding various road-blocks and motivating me to solve them through discussions.

To my parents for all the support.

ABSTRACT OF THE THESIS

Statistical Analysis of WCET Estimation on DNNs

by

Ankith Jain Rakesh Kumar

Master of Science, Graduate Program in Electrical Engineering
University of California, Riverside, March 2018
Dr. Hyoseung Kim, Chairperson

The current research work on determining the worst-case execution time (WCET) focuses mainly on real-time systems since this is a key parameter in evaluating the reliability of a time-critical entity. There is a real dearth of research in estimating WCET measurements in the area of deep neural networks (DNN). This work proposes a novel approach that predicts the probabilistic WCET (pWCET) of DNN based image classification models such as GoogleNet and CaffeNet. The proposed approach uses actual measurement of the DNNs total inference time that considers any variations in the input size and employs Extreme Value Theory (EVT) to estimate the pWCET.The work also discusses a unique approach to predict the pWCET of image resizing given the variations in the input sizes of the images by estimating the pWCET of the single pixel and multiplying it with the actual image size. In addition to this, it achieves a confidence level of 99% for its pWCET estimates.

Contents

List of Figures

Chapter 1

Introduction

In the field of neural networks, the relevance of inference time analysis in the prediction process is very important. Its significance is to comprehend and evaluate the time taken for the neural networks to predict the various classes of its applications such as image classification and recognition, text mining, speech recognition and many more as early as possible.

The timing performance of many hardware and software components seen in isolation can be modeled with deterministic worst-case methods such as state-of-the-art static timing analysis (STA) and the probabilistic timing analysis (PTA) [16].The STA method is based on analyzing the source code or binary files. On the other hand, PTA is based on probability distributions of the execution time and computational complexity. In recent times the hardware and software architectures have grown in complexity. The interaction between various components in the hardware have drastically increased, as a result deriving the deterministic limits on the execution time of the task are probable to produce more pessimistic results if STA is employed. This motivates us to use probabilistic timing analysis that has emerged as a reliable and accurate methodology in the recent past. PTA considers the worst-case execution time(WCET) bounds for arbitrarily low probabilities and addresses the violations. This makes the system reliable by meeting the system safety requirements.

Another key concept in determining the worst-case execution time(WCET) is that of Extreme Value Theory (EVT). Extreme value theory (EVT) is a statistical framework dealing with inferring the extreme (rare) behavior. This gives us a way to reason about the tail behavior of the sample by using a small amount of data. EVT has been used in many applications such as hydrology, finance & insurance, Geo-spatial issues and real-time systems.

In this work, we focus on a variant of PTA, known as measurement-based probabilistic timing analysis (MBPTA). MBPTA uses Extreme Value Theory (EVT) as its building block. Its timing analysis is performed using EVT, that relies on the samples of inference time observed during the analysis. This nature of modeling provides a probable worst-case execution time that can be deemed as a safe estimate of WCET.

EVT does not directly pertain to representativeness of the inference time. Here, representativeness deals with the timing analysis of the samples which capture the impact of both software and hardware events (such as cache misses, number of layers in DNN and many more) that affects the execution time of the application during operations [1] . Therefore, EVT assumes the computing platform (neural network model in our case) as a black box and does not deal with the representativeness of the execution conditions under which the input data are collected. But, representativeness is very important in order to apply the EVT approach accurately, and to ensure that analysis-time observations can be applied and used to derive the pWCET estimation during system operation.

Contributions:

In particular my contributions are as follows:

- Our proposed approach estimates WCET for the total inference time of a DNN based image classification model. Here the total inference time is the sum of the layer processing time and the image resizing time.In addition to this, our work also supports variations in input sizes and takes care of various hardware and software stacks.

- A novel approach that estimates the pWCET of the image resizing process by considering a single pixel. The approach calculates the pWCET of the single pixels and multiplies this value with the actual image size to obtain the pWCET of the image resizing.

Chapter 2

Background

EVT is a statistical based approach used to predict the extreme or rare events. EVT uses a set of input data to determine the parameters required to produce the output curve. This curve describes the tail (right or left) of the distribution based on the respective application. The necessary and important condition for applying EVT is the independent and identically distributed condition (IID) that will be discussed later in the work. Furthermore, EVT does not make any assumptions about the condition of the system while considering the execution time measurements of samples. As a result, we can take EVT as a method that can be used to predict the probabilities of the overall events during the timing analysis of the measured execution time of the sample. There are certain limitations to applying EVT, one such limitation is: Difficulty in applying EVT to the unseen or unobserved data samples. The reason for this could be due to extreme execution times of the samples ie either too large or too small when compared to the measured data sets [1]. However, we can use the MBPTA approach that ensures events such as the ones mentioned above are included in the analysis tests by structuring it, using the representativeness argument that imposes requirements on the timing performance of the platform. Platforms that fulfill all these requirements is called MBPTA-complaint platform [10]. Hence, MBPTA considers and identifies those platform components (hardware and software) with the jittery timing behavior that affects the execution time measurements [4].

Sources of jitter (SoJ) are usually upper bounded by either deterministic or probabilistical methods during the validation period [10]. Deterministically upper bounding at the validation phase can be achieved by enforcing the sources of jitter to have maximum latency. Similarly, probabilistically upper bounding is enforced by applying time-randomized resources. Time-randomized resources can be achieved in both hardware and software concepts.

MBPTA compliance properties can be deployed and achieved at both hardware and software level. Therefore, at the hardware level both deterministic and probabilistic upper bounding can be achieved. For instance, at deterministic upper bounding, the sources of jitter can be forced to work on their worst latency such as the floating-point units [13]. For probabilistic upper bounding at the hardware level can be achieved by bus arbitration requests using the random policy of the resources such as the random replacement policies of the cache [10]. For software level, randomization can be achieved and deployed by using random allocation of the resources during the operation and validation time for the calculation of program execution time. These features of MBPTA-compliance properties have benefits in controlling the Source of Jitter (SoJ).

2.1 Extreme Value Theory Approaches

There are two main approaches in EVT: Block Maxima approach (BM) and Peak over Threshold (POT) approach

2.1.1 Block Maxima Approach

Block Maxima approach [5] takes the input parameter as the block size b, that divides the measured observations into equally sized blocks. For instance, consider a sample of N measured observations X_1, X_2,..., X_N, given a sample of S = 2000 observations and block size of b = 5 elements, Block maxima approach splits the data into nb (number of blocks) i.e. nb = 100 blocks (S/b) with 5 observations in each block. Given that the observations collected meet the independent and identically distributed tests, distribution of samples into blocks is valid and blocks are created with consecutive observations in them. Typically, the maximum value from each block be denoted as M_N, that are selected to create a smaller group of samples (of only maximum values) with as many observations as number of blocks present. Here M_N = max $\{X_1,....., X_N\}$ for each block. Then those nb number of samples are used to fit the Generalized Extreme Value (GEV) distributions.

The Fisher-Tippett-Gnedenko theorem states that, if there exist sequences of normalizing constants $a_n > 0$ [5], and b_n. $M_n{}^* = M_n$ - b_n / a_n , where the $M_n{}^*$ (maximum values of the distribution). The values of a_n and b_n can be selected such that the location and scale parameter of $M_n{}^*$ can be stabilized as value of n increases, evading the difficulties which occur with the variable M_n.

In the Block Maxima approach there are many appropriate continuous probability distribution families such as Gumbel distribution, Weibull distribution and Frchet distribution.

Extremal Types Theorem

Theorem: [5]If there exist any sequences of constants $a_n > 0$ *and* b_n *such that*

Pr$\{(M_n$ - $b_n)/a_n \leq$ z $\}$ $\rightarrow$ G(z) as n $\rightarrow \infty$,

where G is a non-degenerate distribution function, then G belongs to one of the following family of distribution:

$$Type1 : G(z) = \exp\{-\exp[-\frac{z-b}{a}]\}, -\infty < \infty; \tag{2.1}$$

$$Type2 : G(z) = \begin{cases} \exp\{-(\frac{z-b}{a})^{-\alpha}\} & \text{if } z > b \\ 0, & \text{if } z \leq b \end{cases} \tag{2.2}$$

$$Type3 : G(z) = \begin{cases} \exp\{-(-\frac{z-b}{a})^{\alpha}\} & \text{if } z < b \\ 0, & \text{if } z \geq b \end{cases} \tag{2.3}$$

In other words, the theorem states that the sample maxima converge in distribution to a variable having a distribution within one of the families mentioned such as Gumbel, Weibull and Frchet based on the tail behavior of the distribution (shape parameter (ξ)). These family of distribution can be generalized, and they are widely known as the extreme value distributions. Each family of distribution have a shape parameter and scale parameter along with these two parameters the Frchet and Weibull distribution have the shape parameter.

2.1.2 Generalized Extreme Value (GEV) distribution

The family of distribution mentioned in the theorem have a distinct form of behavior, corresponding to the different tail behavior for the distribution. The three distributions give different representation based on the shape of the tail. The problem in selecting the distribution based on this behavior can be overcome by using GEV distribution. The Generalized Extreme Value (GEV) distribution uses the parameters to integrate the three families of distribution into one.

A better approach to this is integrating the three families into one by reformulating the theorem. Therefore in doing so the GEV distribution restates the theorem.

2.1.3 Unified Extremal Types Theorem (UETT)

Theorem : [5]If there exist any sequences of constants $a_n > 0$ *and* b_n *such that*

Pr$\{(M_n$ - $b_n)/a_n \leq$ z $\} \rightarrow$ G(z) as n $\rightarrow \infty$,

where G is a non-degenerate distribution function, then G belongs to the member of the GEV family:

$$G(z) = \exp\{-[1 + \xi(\frac{z-\mu}{\sigma})^{-\frac{1}{\xi}}]\}, \tag{2.4}$$

Thus, the GEV distribution with the shape parameter (ξ), scale parameter (σ) and the location parameter (μ) is defined by z : 1 + ξ ($\frac{z-\mu}{\sigma}$) > 0, where - $\infty < \mu < \infty$, σ > 0 and - $\infty < \xi < \infty$.

The family of distribution is selected based on the tail behavior or the shape parameter (ξ). If ($\xi \leq 0$) then it is characterized as Weibull distribution, Gumbel when

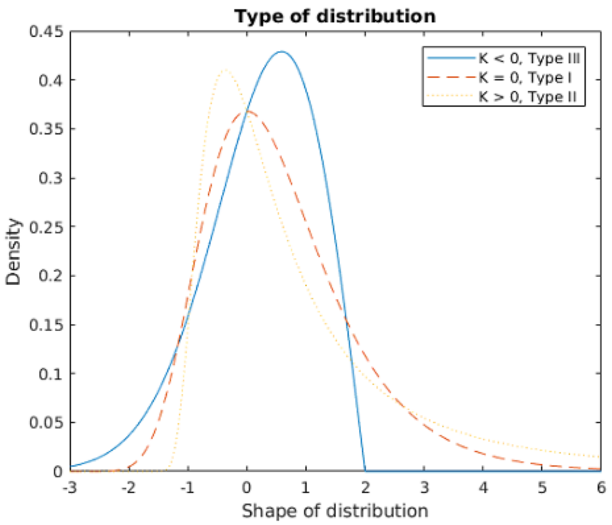

Figure 2.1: Type of distributions based on the shape parameter (ξ).

(ξ = 0) and Frchet ($\xi \geq 0$), which are also known as light tail distribution, exponential distribution and the heavy tail distribution respectively.

The shape of the tail for each GEV distribution can be illustrated from the figure 2.1. As shown, the Weibull distribution has sharp slope with ξ = -0.5 and exponential distribution $\xi = 0$ also has a relatively sharp slope. The exceedance probability plot for the Frchet distribution with ξ = 0.5 decreases only polynomially.

2.1.4 Peak over Threshold (PoT) Approach

The classical approach of block maxima to model the extreme events using the blocks is not efficient when there is large variation of samples within the same block or when one set of blocks have more extreme events than the other set. PoT approach [5] is one of the methods used to estimate the probabilistic worst-case execution time (pWCET). It basically relies on the samples that are higher than the threshold value u. This leads to having a smaller but extreme set of samples that corresponds to the tail behavior. The tail behavior obtained from the probability distribution are described by the Generalized Pareto Distribution (GPD) family.

2.1.5 Generalized Pareto Distribution (GPD)

Let X_1,.....,$X_N X_1$,.....,X_N. be a sequence of independent and identically distributed samples with F as the marginal distribution function. Consider samples of extreme events ξ which exceed high threshold value u. The value of u can be selected by using mean residual life plot and parameter stability plot.

The Generalized Pareto distribution is a family of distribution based on a integrated equation of all the distributions such as Gumbel, Weibull and Frechet distributions. The theorem states the GPD approach:

Theorem: Let X_1,.....,X_N be a sequence, and for large enough u, the distribution function of (X - u), conditional on X > , is approximately

$$H(y) = \{[1 - (1 + \frac{\xi y}{\sigma})^{-\frac{1}{\xi}}]\}, \tag{2.5}$$

denoted on y : y > 0 and (1+ $\frac{\xi y}{\sigma}$) > 0,

The family of GPD distribution is based on the shape parameter (ξ). The figure 2.2. describes the distributions with different values of shape which represents tail behavior of the samples.

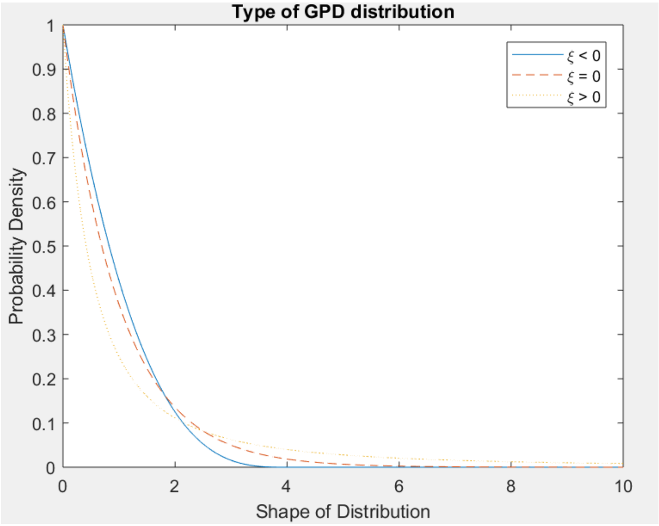

Figure 2.2: Type of GPD distributions based on the shape parameter (ξ).

2.1.6 Measurement based Probabilistic Timing Analysis - Co-efficient of Variation) (MBPTA-CV)

MBPTA-CV approach is a variation of the Generalized Pareto distribution that helps in obtaining the pWCET estimates. It is built upon certain customized applications of EVT to estimate the pWCET problem. This approach uses the new findings of CV plot for the calculation of pWCET.

The co-efficient of variation (CV) of a distribution is defined as the ratio between its standard deviation and its mean, which makes the CV independent of the scale of distribution. For any given distribution the residual CV is used to determine the behavior of the tail: if CV $= 1$ the distribution is said to be exponential(gumbel distribution), similarly if the CV ≥ 1 it is said to be heavy tail (Frchet distribution), if CV ≤ 1 then light tail (Weibull distribution) [8].

MBPT-CV: Parameter Selection

In general parameter set are selected based on the experience of the statistical models. As we know that many hypothesis tests are built upon the statistical significance level that is typically 0.01, 0.02 and 0.05 which indicates and helps in rejecting the hypothesis under consideration and determines how the samples are biased towards the hypothesis. In this scenario 0.05 is used so that we can satisfy many statistical tests and properties such as independent and identically distribution.

The number of extreme measurements required to obtain the actual distributions of the tail is based on the selection of parameters. According to the [3] it is mentioned that the minimum number of samples to obtain the real behavior of the tail can be done using 10 to 50 extreme samples. According to [2] for the tail measurements they have considered no fewer than the 50 tail or extreme samples for accuracy and safety purposes.

Rules for MBPTA-CV approach

The CV-plot is based on the rules to select the number of tail values (N_{+th}).

Rule 1: Samples with fewer than 10 values are not considered due to the unreliability caused with fewer samples in the approximations of tail measurements [3].

Rule 2: The cv(th) test for the exponential tail behavior is performed for all the values in the range of $50 > N_{+th} \geq 10$. If the test fails, for the N_{+th} values in that range then it is considered that some of the observations in that range do not belong to the tail observations. It means that the values can be above the exponential behavior indicating that they might be in the heavy tail region. In this scenario, there is a need to collect more samples in such a way that the 50 samples belong to the exponential tail and do not have high values.

But if the test passes i.e., the tail values belong to either exponential or light region for all those values in the range of $50 > N_{+th} \geq 10$, in such a scenario any suitable value of N_{+th} can be considered.

Rule 3: Once the cv(th) test is passed for $50 > N_{+th} \geq 10$, the final N_{+th} value is the one which satisfies the following conditions: (1) $N_{+th} \geq 50$, (2) the value of cv(th) is closer to 1, (3) the cv(th) test is passed for the value of N_{+th} such that $N_{+th} \geq N_{+th} \geq 10$.

Rule 4: For MBPTA-CV plot, half of the samples are not considered as they do not belong to the tail values. Therefore, only the other half of the samples ($N_{+th} >$ N/2) are considered for the tail fitting plot i.e., MBPTA-CV

Rule 5: In this MBPTA-CV approach, light tails are also considered for the pWCET estimates instead of requesting for the more samples. This is due to exponential tail upper bounds as a result light tails can be used for the estimation of pWCET.

MBPTA-CV: Fitting a distribution for the tail

The tail of the distribution depends on the type of distributions obtained. In this case the tail of the distribution is fitted using the exponential distribution. The value of the pWCET is calculated and estimated using the probability of exceedance at 10^{-15} per runs.

2.2 Deep Neural Networks (DNN)

Deep Neural Network is also known as deep structured learning. It is a study of artificial neural networks and related machine learning algorithms. DNN can model very complex nonlinear relationships. Their architectures create many compositional models where the given image or object is expressed as a layered composition of its primitives. Deep neural networks basic structure has three parts, such as input nodes, hidden nodes and the output nodes. Here, hidden nodes can have multiple hidden layers between the input and output nodes. If the network has more than two hidden nodes the model is called as Deep Neural Network. Each layer in the hidden nodes has many channels inside it. The successive layers use the output from the previous layers. Deep nets use a cascade of many layers of nonlinear processing units for feature extraction and transformation. As the number of channels inside the layers increases, the accuracy and computation both increase which leads to better accuracy and worse execution time. Channels contribute to the feature pools for that layer, each channel comes at a cost in terms of weight storage, communication and computation in a forward pass.

The structure of a typical DNN is shown in figure 2.3. Each hidden layer consists of Convolution layer + RELU, Pooling layer and Fully Connected layer which can be seen in figure 2.3. Convolution layer is the main and core building block of a DNN, consisting of a set of learnable filters, it consists of small receptive field, that we slide over the image spatially, computing dot products between the entries of the filter and the input image. The filter will extend to the full depth of the input image. Pooling layer is a nonlinear down sampling part of the network. Here we use the max pooling nonlinear function. Basically,

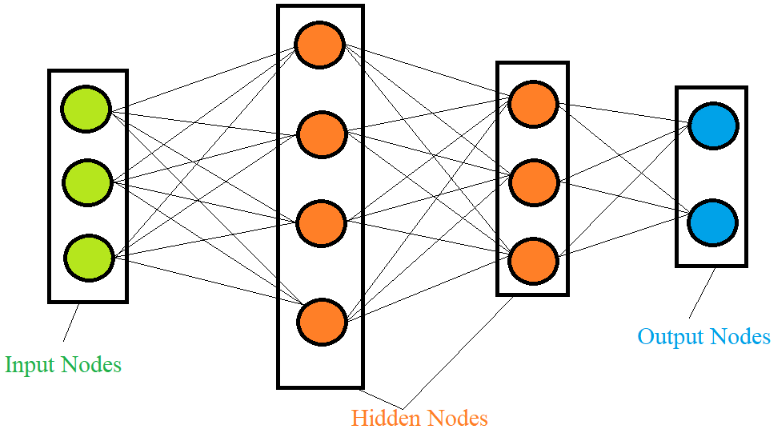

Figure 2.3: Structure of a typical DNN.

the main goal of this layer is to reduce the spatial size of the representation and prevent the over-fitting. RELU layer is rectified linear units. It helps in improving the nonlinear properties of the decision function and the overall network without affecting the receptive fields of the convolution layer. Fully connected layer is used to perform the high-level reasoning of the model after several convolutional layers, max pooling layer. It means they have full connections to all activations in the previous layer

Chapter 3

Requirements on the use of EVT for the estimation of WCET

The safe upper bound of pWCET estimations for given input samples can be achieved by applying EVT. To achieve this it is necessary for the samples to satisfy the independence and identically distributed (i.i.d) conditions.The independent condition states that two random variables are said to be independent if the occurrence of one event does not impact the occurrence of other events. Furthermore, samples are said to be identically distributed if and only if they hold the following conditions such as all random items taken from the samples follow the same probability distribution and they do not fluctuate. If both the conditions are met, then the sample of measured data can be considered as safe for the calculation of pWCET. There are many other tests which makes sure that the data samples are safe for calculation such as stationarity of the sample set, extremal dependence of the samples.

The independence is not a necessary hypothesis [5], where they developed the EVT model based on the stationary weakly dependent time series. In [14] they consider and support stationarity and extremal dependences as necessary conditions for EVT to be applied.

Stationarity can be defined as a process in which the mean, variance and autocovariance structure of the samples do not change over a given amount of time. It can be verified from the given execution time of data sets, by performing the autocorrelation which can be computed with lag plots.

3.1 Dependence of the extreme samples

According to [14] stationarity is a necessary and sufficient condition for applying EVT, but they still find the independence of the extreme samples of measured data that are concerned. Indeed, EVT can be applied on the time series data samples with the extreme samples separated in time, which states that the extreme samples are independent. To estimate the dependence level of extreme samples, extremogram test of the samples is performed.

An extremogram is a measure of extremal dependence for time series measurement [7] where this test considers only the extremal values for the dependence check. The theoretical definition of the extremogram of a stationary time series is defined by:

$$\rho(h) = \lim_{n \to \infty} \left(\frac{P[\chi_o > a_n, \chi_h > a_n]}{P[\chi > a_n]} \right), \tag{3.1}$$

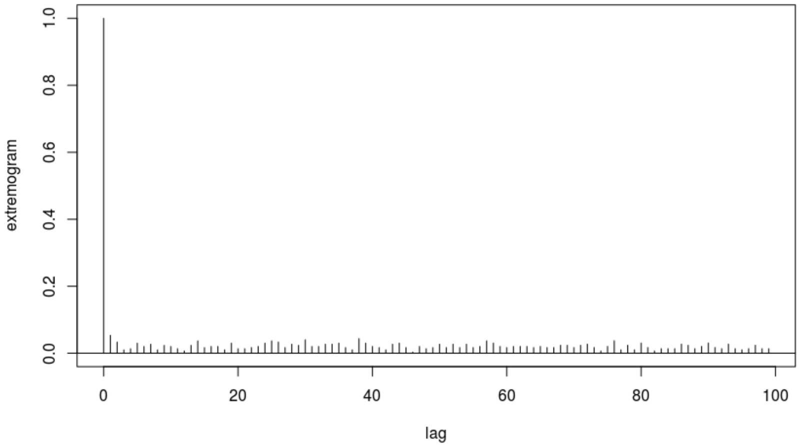

Figure 3.1: Extremogram plot for extremal dependency check.

with a_n a sequence such that P$[|\chi| > a_n] \approx n^{-1}$. The variable h can be seen as a correlation length

When ρ(h) $\rightarrow$ 1, the extremal samples are highly correlated. If ρ(h) $<$ 0.1 or 0.2 the samples are not correlated. Hence, the samples do not need any de-clustering.

The figure 3.1 shows the extremal dependence of the samples of data, where the ρ(h) $<$ 0.2. Therefore, we can conclude based on the figure that the samples do not have any extremal dependency.

Chapter 4

Related Work and Motivation

4.1 Related Work

Based on Timing Analysis Techniques : There are different types of timing analysis techniques and they can be categorized into two methods such as deterministic timing analysis (DTA) method and probabilistic timing analysis (PTA). Deterministic methods considers only the single WCET value whereas the probabilistic ones estimates the probabilistic WCET value for its distribution. Each techniques consists of static and measurement based approaches. Here DTA considers the working environment as time-deterministic platforms and where as the PTA considers the time-randomized platforms.

In this paper [9] the author uses the probabilistic timing analysis technique to obtain the estimates of pWCET of the data samples where in he assumes the data samples follow the exponential tail behavior and neglects the light and heavy tail data samples. In this work [12] they mainly concentrate on probabilistic timing analysis approach and they apply EVT on two different hardware platforms considering the Generalized Extreme value approach. MBPTA [6], performed the measurement based approach to estimate the pWCET. Most of the research is based on time-randomized platforms as they find that this technique gives better results of pWCET as the system does not have a significant interference between the cores and the the cache hierarchy.

Based on Representativeness : This concept is important in determining if the obtained pWCET estimations are accurate and also have the required considerations of the hardware and software concepts while performing EVT. Also, the timing analysis measurement results might lack in evidence of the samples that cannot be captured during the operation.

In these works [9] [14] WCET estimates have been performed on the non-MBPTA complaint platform [2]. In this work [2] they perform the EVT and estimate the pWCET on the MBPTA- complaint platforms.

Based on EVT requirements: A lot of research is available in this field that considers different conditions such as independent and identically distributed, samples which do not obey i.i.d conditions, applying stationarity and the extremal dependence tests.

In a few research studies they show that it is not necessary to have independent samples if the dependency is weak it is enough to perform EVT [5]. In this paper [14] they consider the stationarity and extremal dependence as the necessary conditions to apply EVT to estimate the pWCET and not the i.i.d. condition.

4.2 Motivation

Neural networks are mainly known for classifications of data patterns and prediction of objects such as object detection and pattern recognition in various fields. Currently, there is no research in the area of timing analysis pertaining to DNNs. In addition to this the research community in the field of WCET measurement focuses only on traditional embedded real time benchmarks that does not account for any variations in the input size [2] [12]. Hence, this opens up various opportunities to conduct research in estimating the WCET of neural networks.The idea behind analyzing pWCET of DNN is to create opportunities to use DNN in real time systems.

Chapter 5

Proposed Work

In recent research studies, we come across the estimation of pWCET of the traditional embedded real time benchmarks. In these studies, to estimate pWCET of the system [12]they do not consider the variations in the input size of the samples. As a result, in our proposed approach we consider the variations in the input sizes while estimating the pWCET along with the variations in the hardware and software stacks.

The DNN models are segmented into two parts. The first part deals with the layer processing time and the second part deals with the resizing time of an image. Here, the layer processing time is the time taken for the layers to process and classify the image to a particular class and resizing time is the time taken to resize the input image. The layer processing time is insensitive to the variations in the input size whereas the opposite is true for the resizing time i.e. it is sensitive to the input size. The WCET of the layer processing time can be predicted using the traditional EVT approach. On the other hand, for the estimation of pWCET of resizing part is considered in our approach. The following sections discusses the proposed approach used to estimate the pWCET for the total inference time and also estimates pWCET for the image resizing using single pixel. Here the total inference time is a sum of the layer processing time and the resizing time.

5.1 Estimation of pWCET for the total inference time

In this section, our proposed approach discusses WCET estimation for the total inference time of a DNN model that supports variations in input sizes and also considers variations in hardware and software stacks such as cache. The block diagram of my proposed implementation is as shown in figure 5.1.

The following points summarize my approach to find the pWCET estimate:

- The input images/datasets that are fed to the image resizing block wherein all the datasets are resized to a certain standard value.

- The above process is timed and it's execution period is stored in the system. This is nothing but the resizing time of an image that is used to calculate the total inference time.

- The resized image is later fed to the image classifier that is essentially a Deep Neural Network. The DNNs used in the experiments are CaffeNet and GoogleNet that have their respective layer processing times.

- The layer processing time and the resizing time are summed together to obtain the main variable for the estimation of WCET i.e. total inference time.

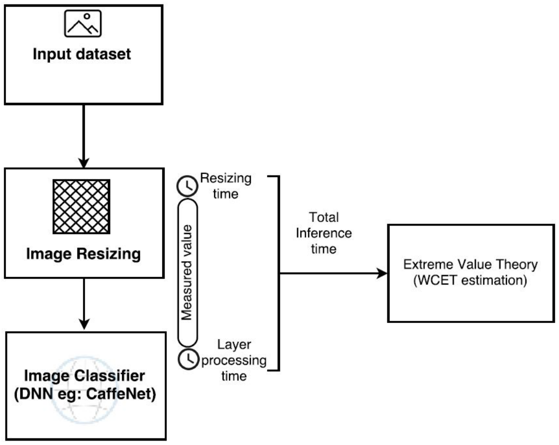

Figure 5.1: Block diagram for the Total Inference time calculation and estimating pWCET.

- The main block of this implementation is the Extreme value theory block that is used to estimate the pWCET. This block takes in the various inference times from multiple inputs. With the help of methodologies such as GEV, GPD and MBPTA-CV these inference times along with other parameters extracted such as(shape parameter, location, scale) help estimate pWCET with a confidence level of 99%.

5.2 Image Resizing Time using a Single Pixel

In this section of proposed work, describes a novel approach to estimate the pWCET for the image resizing part of DNN models. The image resizing in a normal image classification application is a pre-processing step, where the images are resized to a certain value. As this is a key component of the image classifiers it is essential to consider this process in determining the worst case execution time. Although this is a very important parameter most of the research studies do not consider the pre processing parts in their WCET estimation. This work uses a standard 227x227 image resizing algorithm to find

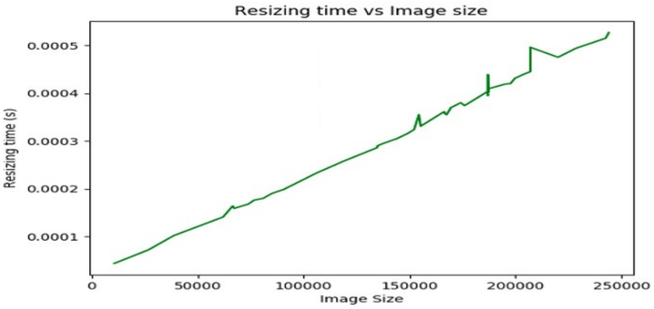

Figure 5.2: Linear relationship between image size and resizing time.

out the resizing time of a single pixel. The main goal behind finding the resizing time of a single pixel is to use this metric to calculate the resizing times for varying sizes of input images.

Let us consider a sample experiment: A 100x100 image is to be resized to a 10x10 image before classification. The resizing time for this process is 10s. Now we know that there exists a linear relationship between the resizing size and its resizing time i.e. if an image resizing of 10x10 takes 10s then the image resizing of 1x1 of the same image takes 0.1s. After determining the resizing time for the single pixel we can apply the same methodology used in the above section (Extreme Value Theory) to estimate the pWCET for a single pixel. The above logic can be applied to varying sizes of input images and given a pWCET for the single pixel value, we can easily calculate the pWCET of the desirable scale. Simply put, the single pixel pWCET value can be multiplied with the desirable resizing scale and it's pWCET can be conveniently derived.

The linear relationship of the size of an image with respect to its resizing time can be shown in figure 5.2.

Chapter 6

Evaluation

6.1 Experimental Setup

The training part of the neural network is performed on the GPU, where as the validation of the model is performed on the CPU (single core).The execution time of the neural network are obtained using the Caffe framework. The platform is instantiated and configured by performing the validation part of the neural network in a particular hardware configurations such as: (1) disabling the hyper-threading so that there are no interference between the core while calculating the execution time of the neural network, (2) disabling the Intel turbo-boost, (3) switching off the DVFS so that there are no fluctuations in the frequency and voltage during measurements of the execution time, (4) run the inference of neural networks on a single cpu core to avoid the interference between the cores, (5) the validation part and training part of the network is performed using the CLI(command line interface) mode to avoid any delay in timing analysis of the DNN models. These arrangements help in designing a platform using the hardware configurations.

The MBPTA-compliant platform is achieved using the software configurations such as randomization of the inputs so that random replacement policy is achieved in the cache. Cache randomization helps and initiates cache placement random across each runs so that in every new run, samples are assigned randomly across the cache. As a result, the cache layouts and memory mapping is being captured during the validation part.

6.2 Information about the Data

In order to perform and estimate the pWCET in the DNN models such as CaffeNet and GoogleNet, the data samples incorporating the training and validation parts for these networks are about 15,000 and 25,000 image samples during training and the 50,000 and 100,000 samples during validation part. The images used for image classification have the input size ranging from 400x500 to 4000x4200.

6.3 Results

This section consists of the estimation of the pWCET using Generalized Extreme Value (GEV) approach, Generalized Pareto Distributions (GPD) and MBPTA-CV approach on two deep neural network models such as CaffeNet and GoogleNet. The neural network comprises of two different processes such as image resizing and prediction(image classification). The sum of layer processing time and resizing time of images is the Total inference time. In each network model the pWCET is calculated in two ways: (1) the calculation is made by isolating each of the above mentioned processes, (2) an aggregate calculation of both the processes is made.

GoogleNet and CaffeNet

GoogleNet [15] is a 22 layer deep neural network. It has 12x less parameters than the AlexNet model and much more accurate model than the AlexNet. Inception layer is included in this model which adds efficiency and accuracy to the model.

CaffeNet [11] an image classifier built on the AlexNet architecture. It has 5 convolutional layers and three fully connected layers as its structure. The images are resized to 227 * 227 in the network.

The training part of this network model is performed on the GPU for 50,000 iterations. The Validation of the data is collected by the single cpu core to avoid any interference which can be caused by the other cores. There are different steps being used while obtaining the estimation of pWCET of the models considering different methods. Each and every method has their own approach in estimating the pWCET. Now, let us consider the approaches individually and find out the pWCET based on their approaches.

6.3.1 Statistical tests on the data

In this section we perform the required and necessary tests on the data samples collected (unit image resizing time and layer processing time) as explained in the chapter 3 such as the independent and identically distributed test, extremal dependency check etc.

The figures 6.1 and 6.2 shown below for the extremogram tests shows that the ρ(h) values are less than 0.1. Therefore, the data samples are independent at their extremes.

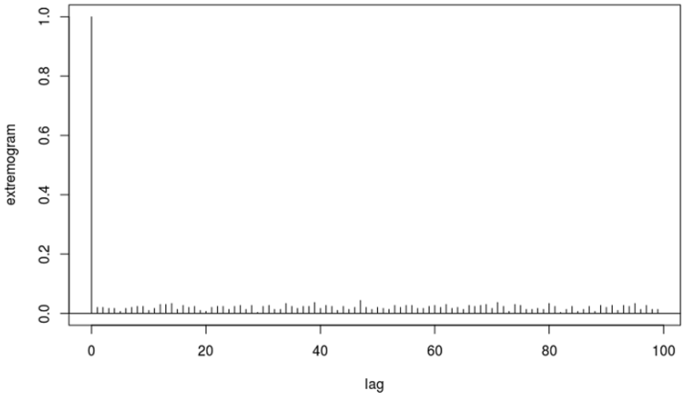

Figure 6.1: Extremogram plot for extremal dependency tests for GoogleNet layer processing time.

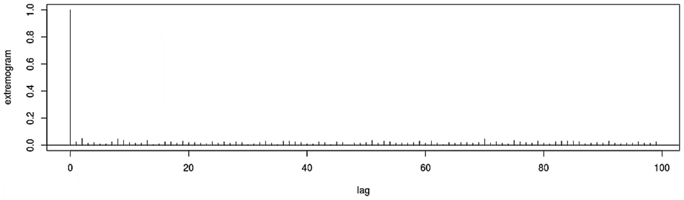

Figure 6.2: Extremogram plot for extremal dependency tests for CaffeNet layer processing time.

6.3.2 Generalized Extreme Value Distribution

As mentioned in section 2.1.2 is a block maxima approach where the maximum data from each block is considered and to this set of data EVT is applied to obtain the estimation of pWCET of the model. The following steps are being performed to do so:

- Set the initial block size b to 5.

- If the number of blocks [N/b] where N is the total number of execution time samples is less than 30 , then stop and collect more samples.

- For each of the [N/b] blocks find the maximum values y_1,....,$y_{N/b}$ where y_i = max($x_{(i-1)b+1}$, $x_{(i-1)b+2}$,...,x_{ib}).

- Estimate the best-fit GEV parameters ξ, σ and μ to the block maximum values y_1,...,$y_{N/b}$ using maximum likelihood estimation approach.

- Based on the GEV parameters the qq- plot is plotted to check if the empirical quantile values of the sample data is same as the quantile of the standard form of a target distribution.

- Chi-square tests is used to verify the goodness of fit between the block maximum values and the estimated GEV parameters based on this test.

- Estimate the pWCET time from the GEV parameters and probability of exceedance (p_e) for the validation data samples..

Based on the above mentioned steps these are the results obtained from the Generalized Extreme Value approach for unit resizing time, layer processing time and the total inference time(resizing time and layer processing time) for both the models.

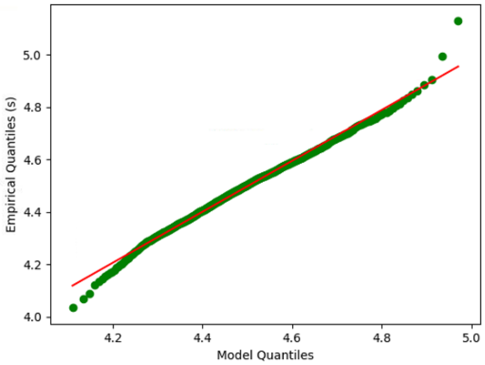

Figure 6.3: QQ-Plot of Block Maxima values with GEV quantiles.

GoogleNet-Layer Processing time

The pWCET for the layer processing time of the GoogleNet model is estimated based on the steps mentioned above in section 6.3.2. The figure 6.3 refers to the QQ-Plot. QQ-Plot or quantile-quantile plot is defined as a plot of the empirical quantile values of the sample data and the samples of the standard quantile. This plot helps in understanding whether the two distributions are identical or not. We can observe from the figure 6.3 that the dot represents the block maximum values and the line represents the GEV fit to the data. It is clear and evident that we get a good linear fit of the data.

The figure 6.4 refers to the chi-square test. The chi-square tests is performed to verify the goodness of fit between the Block maximum values and the estimated values obtained based on the GEV parameters. From the figure 6.4 it is evident that the observed and measured samples are almost the same and they satisfy the chi-square test with 95% confidence level.

The figure 6.5 refers to the estimated pWCET of the data samples vs measured

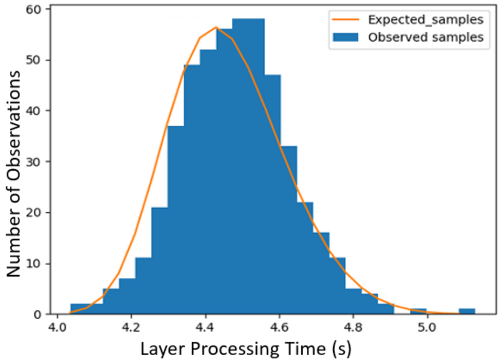

Figure 6.4: Fit of Task Block Maxima Values for GoogleNet model.

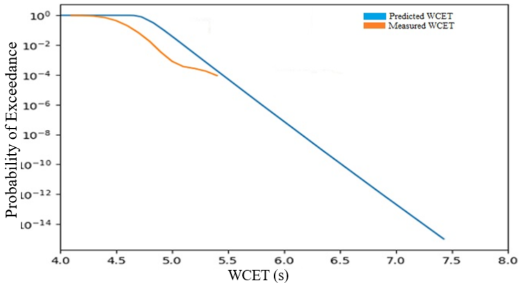

Figure 6.5: Predicted vs Measured Exceedance Probability for GoogleNet layer processing time

probability of exceedance (p_e). From the graph, it is clear that the predicted pWCET upper bounds the measured values from 10^{-1} to 10^{-15}. Here probability of exceedance(p_e) refers to the number of samples which are above that WCET for a given run from 10^{-1} to 10^{-15}.

CaffeNet-Layer Processing time

The pWCET for the layer processing time of the CaffeNet model is estimated based on the steps mentioned above in subsection: 6.3.2 distribution. The figure 6.6, 6.7 and 6.8 depicts the qq-plot, chi-square test and exceedance probability plot.From figure 6.8 its seen that predicted WCET upper bounds the measured WCET from 10^{-1} to 10^{-15}

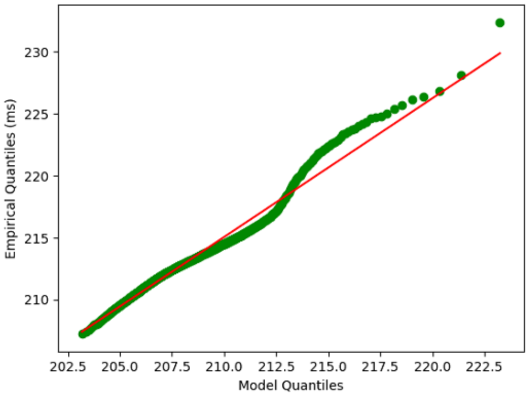

Figure 6.6: QQ-Plot of Block Maxima values with GEV quantiles for CaffeNet layer processing time.

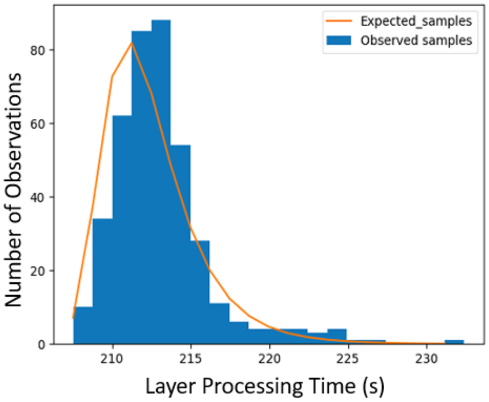

Figure 6.7: Fit of Task Block Maxima Values for CaffeNet model.

run.Here probability of exceedance(p_e) refers to the number of samples which are above that WCET for a given run from 10^{-1} to 10^{-15}

CaffeNet single pixel resizing time

In this section, we use the single pixel for the resizing time of an image to estimate the pWCET of the data samples. The procedure is same as mentioned in the subsection

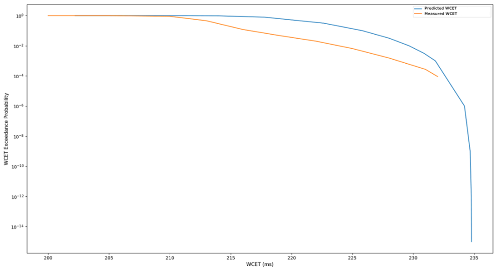

Figure 6.8: Predicted vs Measured Exceedance Probability for CaffeNet layer processing time

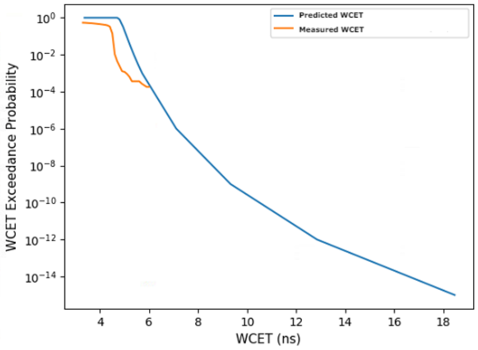

Figure 6.9: Predicted vs Measured Exceedance Probability for CaffeNet single pixel resizing time

6.3.2. Following are the results obtained for the estimation of pWCET for single pixel resizing time. The figures 6.9 depicts the result of exceedance probability plot. The 6.9 shows that the predicted exceedance probability upper bounds the measured p_e for a given runs from 10^{-1} to 10^{-15}.

GoogleNet Total Inference Time

In this section, we use the total inference time of a network to estimate the pWCET of the data samples. The procedure is same as explained in the section 6.3.2 using the steps mentioned in the section: 6.3.2. The figure 6.10 depicts the result for exceedance probability

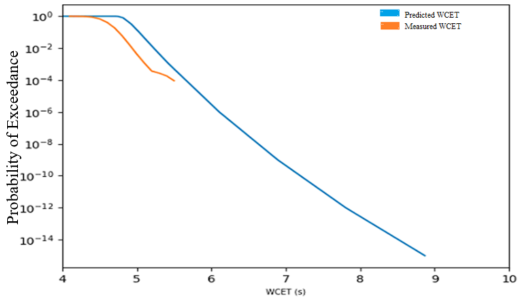

Figure 6.10: Predicted vs Measured Exceedance Probability for GoogleNet (Total Inference time)

plot. The 6.10 shows that the predicted exceedance probability upper bounds the measured p_e for a given runs from 10^{-1} to 10^{-15}.

6.3.3 Generalized Pareto Distribution

As mentioned in section 2.1.5 this is an approach based on the given samples that are above a certain threshold value u, where the data above a threshold value are considered. This set of data is applied as input to EVT in order to obtain the estimate of pWCET. of the model. The following steps are being performed to do so:

• Find the threshold value for a given sample of data using the mean residual life plot or the mean excess plot.

• Once the threshold value is calculated from the plot then the data above threshold are used to find the GPD parameters using the maximum likelihood estimation or the l moments approach.

• Using the parameters, the pWCET execution estimation is made using the exceedance probability plot.

Based on the above-mentioned steps these are the results obtained from the Gen-

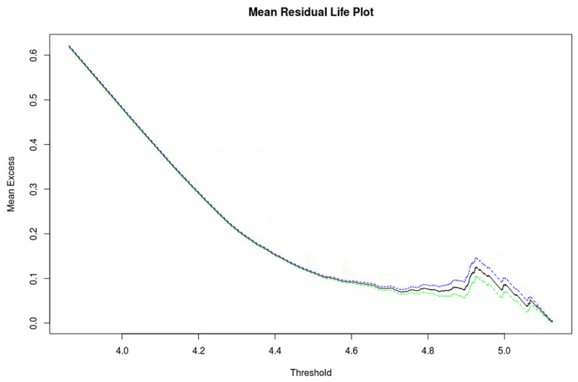

Figure 6.11: Threshold selection for the GoogleNet layer processing time

eralized Pareto distribution approach for single pixel resizing time,layer processing time and the total inference time(resizing time and layer processing time) for both the models.

Threshold selection

In this section, given a data samples we find out the threshold value u such that all the data samples selected are above the threshold value and are extreme values that depicts the tail behavior of the distributions. The threshold value is selected using mean residual life plot and mean excess plot. In this case, we use the mean residual life plot to find the threshold value for a given data sample.

The figure 6.11 refers to the threshold selection for the GoogleNet layer processing time. In this figure we observe that there is evidence of linearity in the graph at u = 4.85. As a result threshold value is selected at u $\approx$ 4.85. Similarly, the threshold value is calculated in a similar way for other models.

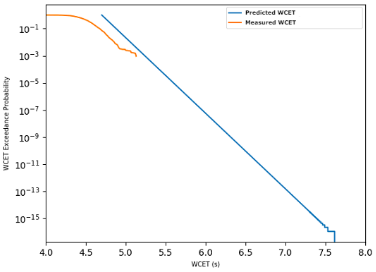

Figure 6.12: Predicted vs Measured Exceedance Probability for GoogleNet (layer processing time)

Googlenet layer processing time

The pWCET for the layer processing time of the GoogleNet model is estimated based on the steps mentioned above in section 6.3.3. As the threshold value is obtained using the mean residual life plot where u = 4.85, now the data above the threshold value is considered for parameter estimation using L moments approach. Furthermore, using the parameters the estimation of pWCET is predicted and exceedance probability plot is plotted for the 10^{-1} runs to 10^{-15} for particular WCET values.

The figures 6.12 shows that the expected probability of exceedance always upper bounds the measured exceedance probability at all runs.

CaffeNet layer processing time

The pWCET for the layer processing time of the CaffeeNet model is estimated based on the steps mentioned in section 6.3.3. The threshold value is obtained using the mean residual life plot,the data above the threshold value is considered for parameter esti-

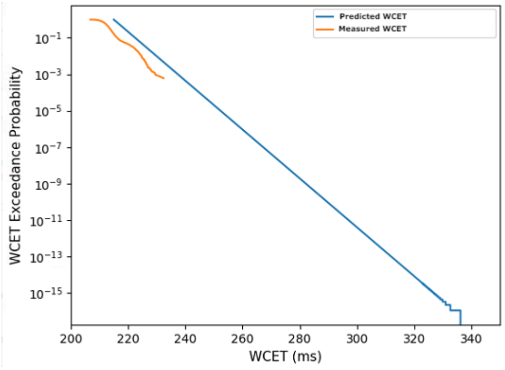

Figure 6.13: Predicted vs Measured Exceedance Probability for CaffeNet (layer processing time)

mation using L moments approach. Furthermore, using the parameters the estimation of pWCET is predicted and exceedance probability plot is plotted for the 10^{-1} runs to 10^{-15}.

The figures 6.13 shows that the expected probability of exceedance always upper bounds the measured exceedance probability at all runs.

CaffeNet single pixel image resizing time

In this section, the single pixel resizing time of an image is used to estimate the pWCET of the data samples. The procedure is same as mentioned in the subsection 6.3.3.The figures 6.14, depicts the results for exceedance probability plot. The 6.14 shows that the expected exceedance probability upper bounds the measured p_e for maximum time but there is an exception at WCET = 7.5 ns.

GoogleNet Total Inference Time

In this section, we use the total inference time of a network model to estimate the pWCET of the data samples. The procedure is same as explained in the section 6.3.3 to

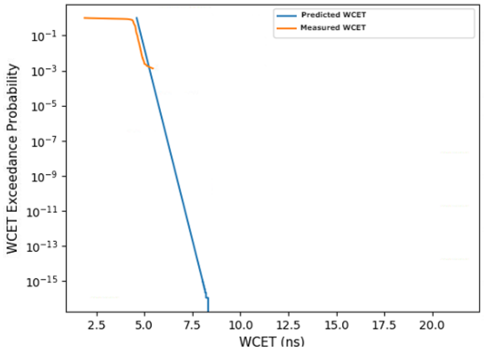

Figure 6.14: Predicted vs Measured Exceedance Probability for unit resizing time of an image

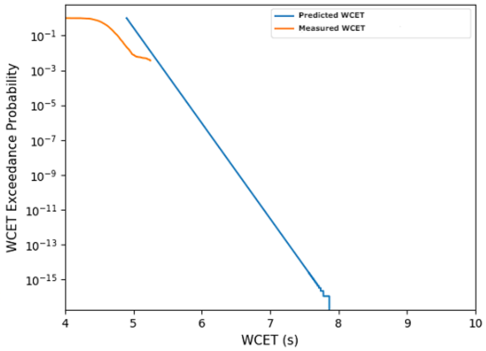

Figure 6.15: Predicted vs Measured Exceedance Probability for GoogeleNet (total inference time)

the other methods using the steps mentioned in the section: 6.3.3. The figures 6.15 depicts the result for exceedance probability plot. The 6.15 shows that the predicted exceedance probability upper bounds the measured p_e for a given run from 10^{-1} to 10^{-15}.

6.3.4 MBPTA-CV

In this section 2.1.6 the pWCET is estimated based on the co-efficient of variation plot. This plot is used to remove the samples that are not in the exponential region of the cv plot i.e., CV = 1 and considers only the samples that are in this region. The rules are mentioned in the section 2.1.6 where it describes how to find the pWCET using the MBPTA-CV approach.Using the rules mentioned in section 2.1.6 following are the results for the estimation of pWCET for the given models.

Googlenet and CaffeNet layer processing time - MBPTA-CV approach

The pWCET for the layer processing time of the GoogleNet and CaffeNet model is estimated based on the steps mentioned above in section 2.1.6.Furthermore, the samples that lie in the exponential region or the light tail are only considered for the estimation of pWCET. The minimum number of samples used for the estimation of pWCET are 50.

The figure 6.16 helps in rejecting the samples that do not lie within the exponential region. The red lines indicate the exponential region. The region above both the red lines is heavy tail and below the red lines is light tail. The samples which are within the exponential and light tail regions are considered for the probability of exceedance plot. In the figure 6.16, we see that the values left of the brown line are all rejected and do not contribute to estimate the pWCET for the GoogleNet model layer processing time. In the figure 6.17 there are about 2055 values that determine the tail behavior of the distribution which are used to estimate the pWCET for the CaffeNet model layer processing time.

In the figure 6.18 and 6.19 we see that in both the graph the expected pWCET upper bounds the measured values.

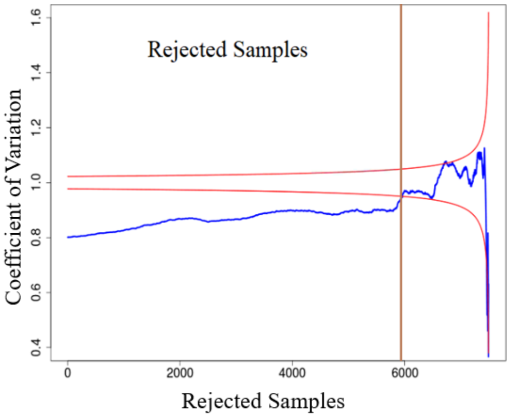

Figure 6.16: CV plot to determine the number of samples to be considered for the estimation of pWCET for GoogleNet layer processing time

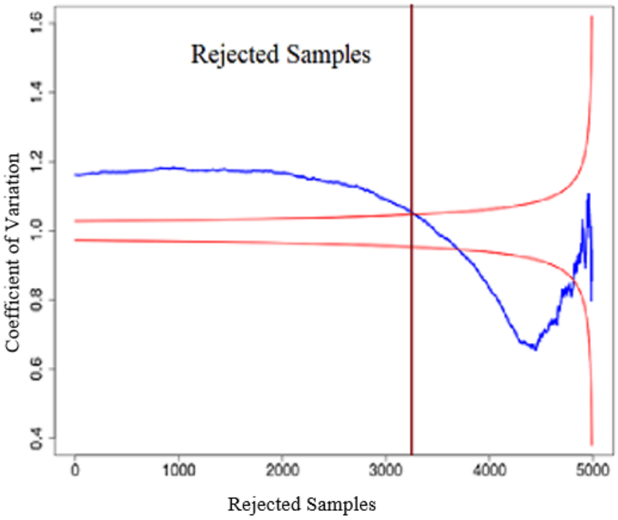

Figure 6.17: CV plot to determine the number of samples to be considered for the estimation of pWCET for CaffeeNet layer processing time

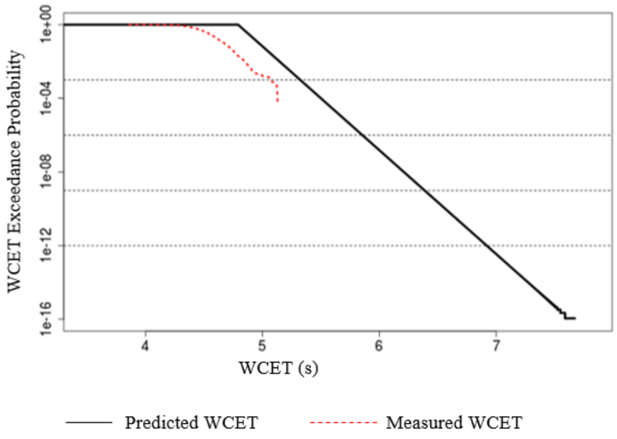

Figure 6.18: Predicted vs Measured Exceedance Probability for GoogleNet layer processing time

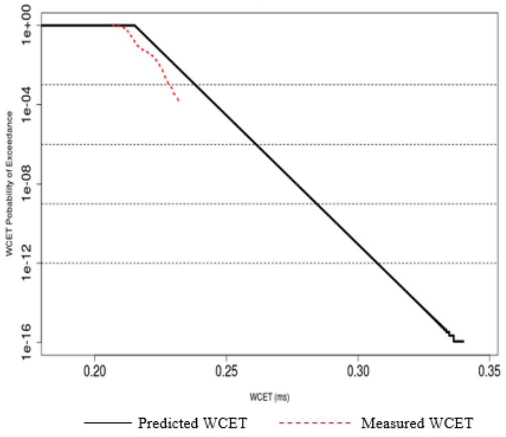

Figure 6.19: Predicted vs Measured Exceedance Probability for CaffeeNet layer processing time

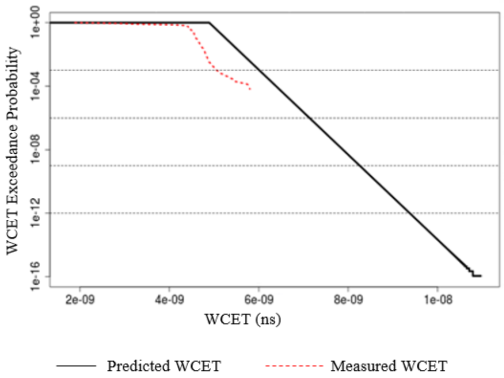

Figure 6.20: Predicted vs Measured Exceedance Probability for unit resizing of an image

pCWET estimate for the Single pixel of an image

The pWCET is estimated for a single pixel resizing of an image using the MBPTA-CV approach, where it follows the rules mentioned in the section: 2.1.6. In figure6.20 depict MBPTA-CV pWCET estimation, which shows the predicted pWCET upper bounds the measured data for all the runs from 10^{-1} to 10^{-15}.

GoogleNet Total Inference Time

In this section, pWCET estimation is predicted for the total inference time of a model. The figures 6.21 depict the result of exceedance probability plot, where the predicted exceedance probability upper bounds the measured p_e from 10^{-1} to 10^{-15} runs.

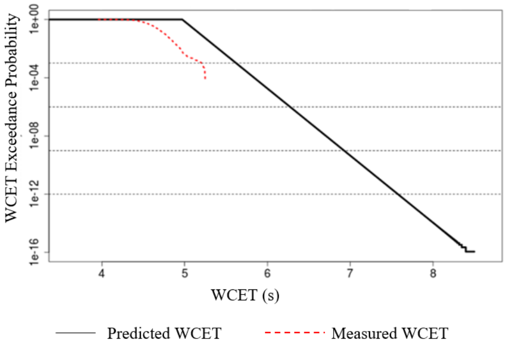

Figure 6.21: Predicted vs Measured Exceedance Probability for Total Inference time

Chapter 7

Conclusions

In this work, we discuss the implementation of probabilistic models that determine the probabilistic worst-case execution time (pWCET) of Deep Neural Networks (DNN). This is a novel approach in determining pWCET for the total inference time of DNN models considering the variation in the input size of the images. The DNN in consideration here are CaffeNet and GoogleNet. The proposed approach to estimate pWCET for the total inference time of the DNNs can be approached using any of the EVT approaches (1) Generalized Extreme Value, (2) Generalized Pareto Distributions and (3) MBPTA-CV approach (Variation of GPD). After careful evaluation of pWCET estimates based on our approach, we achieve a confidence level of 99%.

In the current work, we estimate pWCET of the DNNs using offset data, wherein the data samples are collected after performing the execution. In the future, we will concentrate on the pWCET estimation during the runtime of a program.

Bibliography

[1] J. Abella, E. Quiones, F. Wartel, T. Vardanega, and F. J. Cazorla. Heart of gold: Making the improbable happen to increase confidence in mbpta. In *2014 26th Euromicro Conference on Real-Time Systems*, pages 255–265, July 2014.

[2] Jaume Abella, Maria Padilla, Joan Del Castillo, and Francisco J. Cazorla. Measurement-based worst-case execution time estimation using the coefficient of variation. *ACM Trans. Des. Autom. Electron. Syst.*, 22(4):72:1–72:29, June 2017.

[3] Yuzhi Cai and Dominic Hames. Minimum sample size determination for generalized extreme value distribution. *Communications in Statistics - Simulation and Computation*, 40(1):87–98, 2010.

[4] Francisco J. Cazorla, Tullio Vardanega, Eduardo Quiñones, and Jaume Abella. Upper-bounding Program Execution Time with Extreme Value Theory. In Claire Maiza, editor, *13th International Workshop on Worst-Case Execution Time Analysis*, volume 30 of *OpenAccess Series in Informatics (OASIcs)*, pages 64–76, Dagstuhl, Germany, 2013. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[5] S. Coles. *An Introduction to Statistical Modeling of Extreme Values.* Lecture Notes in Control and Information Sciences. Springer, 2001.

[6] L. Cucu-Grosjean, L. Santinelli, M. Houston, C. Lo, T. Vardanega, L. Kosmidis, J. Abella, E. Mezzetti, E. Quiones, and F. J. Cazorla. Measurement-based probabilistic timing analysis for multi-path programs. In *2012 24th Euromicro Conference on Real-Time Systems*, pages 91–101, July 2012.

[7] Richard A. Davis and Thomas Mikosch. The extremogram: A correlogram for extreme events. *Bernoulli*, 15(4):977–1009, 11 2009.

[8] Joan del Castillo, Jalila Daoudi, and Richard Lockhart. Methods to distinguish between polynomial and exponential tails. 41, 12 2011.

[9] Jeffery P. Hansen, Scott A. Hissam, and Gabriel A. Moreno. Statistical-based WCET estimation and validation. In *9th Intl. Workshop on Worst-Case Execution Time Analysis, WCET 2009, Dublin, Ireland, July 1-3, 2009*, 2009.

[10] L. Kosmidis, E. Quiones, J. Abella, T. Vardanega, I. Broster, and F. J. Cazorla. Measurement-based probabilistic timing analysis and its impact on processor architecture. In *2014 17th Euromicro Conference on Digital System Design*, pages 401–410, Aug 2014.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[12] G. Lima, D. Dias, and E. Barros. Extreme value theory for estimating task execution time bounds: A careful look. In *2016 28th Euromicro Conference on Real-Time Systems (ECRTS)*, pages 200–211, July 2016.

[13] Marco Paolieri, Eduardo Quiñones, Francisco J. Cazorla, Guillem Bernat, and Mateo Valero. Hardware support for wcet analysis of hard real-time multicore systems. *SIGARCH Comput. Archit. News*, 37(3):57–68, June 2009.

[14] Luca Santinelli, Jérôme Morio, Guillaume Dufour, and Damien Jacquemart. On the Sustainability of the Extreme Value Theory for WCET Estimation. 39:21–30, 2014.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[16] Franck Wartel, Leonidas Kosmidis, Adriana Geanina Gogonel, Andrea Baldovin, Zoe Stephenson, Benoit Triquet, Eduardo Quinones, Code Lo, Enrico Mezzetti, Broster Ian, Jaume Abella, Liliana Cucu-Grosjean, Tullio Vardanega, and Francisco J. Cazorla. Timing analysis of an avionics case study on complex hardware/software platforms. In *DATE 2015 - Design, Automation and Test in Europe*, pages 397–402, Grenoble, France, March 2015.