DNN Power and Energy Consumption Analysis of Edge AI Devices

Henar Mike Canilang\textsuperscript{1}, Angela Caliwag\textsuperscript{2}, Jonghun Kwon, Wansu Lim\textsuperscript{3}

\textit{Department of Aeronautics, Mechanical and Electronic Convergence Engineering}
Kumoh National Institute of Technology, Gumi, South Korea
\textsuperscript{1}hmocanilang@kumoh.ac.kr, \textsuperscript{2}a_caliwag@kumoh.ac.kr, \textsuperscript{3}wansu.lim@kumoh.ac.kr

\textbf{Abstract}— The application of Deep Neural Networks for real-time data processing and AI applications on edge AI devices is developing rapidly. Deep learning models in real-time applications that yields high accuracy proves to consume more power and energy resources. This gives constraints in deploying computation-intensive DNN-based application to edge AI devices due to its limited computational resources. This paper focuses on power consumption and energy consumption analysis of different DNN models deployed to edge AI devices. To show emphasis with the overall power and energy consumption analysis results, a network optimization is applied for comparisons purposes. This bridges the gap between the deployment of complex DNN designs and its effect to the resource constrained edge AI devices.

\textbf{Keywords}— Edge AI Device, Deep Neural Network, Deep Learning Model, Power Consumption Analysis, Energy Consumption.

\section{I. INTRODUCTION}

Edge AI devices enables machine learning computations locally and serves as an expansion in the field of data processing applications. The number of edge devices grows rapidly throughout the years, adhering to the fact that data generated and processed at the edge also grows constantly. International Data Corporation predicted that 180 zettabytes of data and 70\% of the total data generated by IoT will be processed at the edge which leads to the need of capable edge AI devices. Processing the data at the edge through edge devices aims to overcome the shortcomings of current data processing such as latency, bandwidth, availability, energy, security and privacy. Deploying AI applications are progressing with the advancement in deep learning and the continuous development of hardware architectures such as edge AI devices. The forecasted demand for edge devices to process the billions of data bytes in the future requires data processing and structural optimization for hardware-software implementation. This leads to the integration of AI to edge AI devices \cite{1}. DNNs is a bridges the gap of enabling AI applications in this 5G era \cite{2} so as in application in edge AI devices for data processing.

The implementation of DNN to edge AI devices requires a design that has low-power consumption requirements. Efforts of using edge AI devices for low power real time deep learning application is increasing in demand. A deep learning model that requires high accuracy in real-time has high power consumption. It’s also mentioned that a deep learning model have high computational complexity and computing energy requirement which can affect its implementation to edge AI devices \cite{3-4}. In this matter, researchers’ focuses in optimizing the DNN applied in edge AI devices to cope up with it limited hardware resources such as power and computing capability. \cite{5} presented a method to estimate the energy consumption of DNN based on its architecture, sparsity, and bitwidth. This links the role of developing energy efficient-DNNs to address its application to the resource constrained edge AI devices.

This paper presents an analysis on deployed DNN to an edge AI device. The deployed DNN to the edge AI device is for real-time face detection application. A network optimization for DNN is applied to highlight its effect on the power and energy consumption analysis of the edge AI devices. For comparison and to support the results and claim of this paper, an energy estimation tool provided by \cite{5} is used as reference to show the effect of network optimization to power and energy consumption on existing DNN models.

\section{II. PROPOSED METHODOLOGY}

\textbf{A. Experimental Setup}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig1.png}
\caption{Actual Experimental Setup.}
\end{figure}

The power and energy consumption analysis of a DNN deployed in an edge AI device presented in this paper, presents the actual results from real-time face detection application. NVIDIA Jetson Nano is used as an edge AI device in this implementation. The input comes from a Raspberry Pi v2 camera module. Network optimization is used to show the results of an optimized DNN to its host Edge AI device. Network optimization is applied in the training process to reduce the complexity of the DNN model applied, particularly, pruning and quantization method. Reducing the complexity of the DNN model leads to the reduced layer energy thus reducing both computational and data movement energy. \begin{equation}
E_{\text{layer}} = E_{\text{comp}} + E_{\text{data}}
\end{equation}
(1) represents energy consumption model \cite{5}.

Reducing each layers’ energy consumption leads to optimized power consumption.

\section{Theoretical Analysis}

\section{Experimental Analysis}

\section{Conclusion}

\section{Acknowledgment}

\section{References}

\begin{thebibliography}{9}
\bibitem{1}...
\bibitem{2}...
\bibitem{3}...
\bibitem{4}...
\bibitem{5}...
\end{thebibliography}
III. RESULTS AND DISCUSSIONS

A. Power Consumption and Energy Consumption Analysis

Figure 1a shows the relationship of average power consumption for the NVIDIA Jetson Nano to the face detection algorithm. The compression ratio includes (40x, 60x and 100x) which yielded varying CPU and GPU power consumption. This is compared to the uncompressed version of the face detection algorithm which showed non-linear relationship. It is observed that the CPU power consumption, of the model with 40x and 60x compression ratio is reduced from 2603mW to 1837mW and 1733mW respectively. The model with 100 times compression ratio yielded 2379mW which is smaller than the uncompressed model but greater than the 40x and 60x compressed model. This is due to overfitting in the data. We included this result as part of the analysis to highlight the fact that applying compression or increasing compression ratio of the model will not all the time help the performance of the edge AI device. To overcome this situation, enough datasets should be available for the training of the compressed model to avoid overfitting. The GPU power consumption as observed in Fig.2, which shows that the GPU power is lowest in the 100x compressed model compared with the uncompressed model. Compared with the uncompressed model with a GPU power consumption of 1023mW, the 40x and 60x compressed model has higher GPU power consumption of 1688mW and 1783mW respectively. Considering this power consumption analysis done in NVIDIA Jetson Nano, the model with 60x compression ratio proves to be the most power-energy wise model in this real-time face detection application.

The network optimization applied in this model aimed in reducing the processing and memory transactions of each layers which adheres to computation and data movement energy. This leads to reducing each layers’ energy consumption which leads to the edge AI’ devices overall power consumption optimization. Refer to equation [5] for the energy consumption model.

B. DNN Energy Estimation

To support this claim, an energy estimation tool [6] for an existing DNN model and their own model shows that when network optimization such as pruning is applied, the overall energy consumption of the DNN model is reduced. Fig.1b shows an example of an unpruned DNN model while Fig.1c shows the pruned DNN model. Based on Fig.1b and Fig.1c, pruning.

IV. CONCLUSION

This research paper presented a power consumption and energy analysis for DNN deployed in Edge AI Device, in particular, a face detection was implemented to gather and implement power consumption analysis. Network optimization was applied for the face detection model. For the energy consumption analysis, equation [5] shows the relation of the processing and memory transactions of each layers adheres to computation and data movement energy. Network optimization minimizes the network model hence reducing each layers’ energy consumption which results to edge AI’ devices overall power consumption optimization.

ACKNOWLEDGMENT

This work was supported by the Ministry of SMEs and Start-ups, S. Korea (S2829065, S3010704), and by the National Research Foundation of Korea (2020R1A4A10177511).

REFERENCES


