Predictive Model-Based Selection of Deep Neural Networks for Embedded Image Classification

Bussa Uday Kiran[1], Dr. Raghavendra D Kulkarni[2]
[1]*Research Scholor, Bir Tikendrajit University*
[2]*Research Supervisor, Bir Tikendrajit University*

***Abstract*—Efficient and accurate picture categorization in the realm of embedded systems is being pursued via the investigation of deep neural networks (DNNs) due to high demand. Nevertheless, the task of choosing the most suitable DNN structure for certain embedded applications continues to be difficult because of limitations on processing resources, power use, and memory.This research introduces an adaptive algorithm to ascertain the appropriate Deep Neural Network (DNN) model for a given input, taking into account the desired level of accuracy and inference time. Our methodology employs machine learning to create a prediction model that efficiently chooses a pre-trained deep neural network (DNN) based on a given input and optimisation constraint. We implement our methodology to the job of image classification and assess its performance on a Jetson TX2 embedded deep learning platform, using the ImageNet ILSVRC 2012 validation dataset for evaluation. We evaluate a variety of prominent deep neural network (DNN) models.**

***Index Terms*—Deep learning, Embedded, Predictive model, Inference, Accuracy**

I. INTRODUCTION

The advent of deep learning has ushered in a new era in computer vision, with deep neural networks (DNNs) proving to be exceptionally effective at image classification tasks. These models, characterized by their deep layers and complex structures, have consistently outperformed traditional machine learning algorithms, driving advancements in fields ranging from autonomous driving to medical imaging. However, the deployment of DNNs on embedded systems, such as smartphones, IoT devices, and drones, remains a formidable challenge. Embedded systems are typically limited in terms of computational power, memory, and energy resources, which makes it difficult to directly deploy state-of-the-art DNNs that are both resource-intensive and computationally demanding.Embedded image classification systems are increasingly ubiquitous, finding applications in diverse domains such as real-time surveillance, wearable health monitoring, and smart home devices. The effectiveness of these systems hinges on their ability to accurately classify images while operating within stringent resource constraints. Traditional approaches to deploying DNNs on embedded devices often involve manual tuning and model compression techniques such as pruning and quantization. While these methods can reduce the model size and computational load, they typically require extensive trial-and-error processes and may still fall short of fully optimizing the deployment for specific use cases.

Predictive models are trained to estimate various metrics of interest, such as inference time, energy consumption, and memory usage, based on the characteristics of the neural network architecture and the hardware specifications of the embedded system. These models enable a systematic and automated approach to selecting the most suitable DNN, thereby reducing the reliance on manual intervention and accelerating the deployment process.

One of the core components of our framework is the development of accurate and reliable predictive models. We explore a range of techniques for building these models, including regression analysis, machine learning-based predictors, and more sophisticated approaches like neural architecture search (NAS). The predictive models are trained on extensive datasets comprising performance metrics of various DNN architectures evaluated on different embedded platforms. By capturing the relationships between network characteristics and deployment outcomes, these models provide valuable insights into how different DNNs will perform under specific conditions.The framework also incorporates a

decision-making module that utilizes the predictions to select the optimal DNN for a given application. This module considers multiple objectives, balancing the need for high classification accuracy with constraints on computational resources and energy consumption. In practical terms, this means that the system can dynamically adjust its choice of neural network based on real-time requirements and available resources, ensuring efficient and effective operation.

II. REVIEW OF LITERATURE

Sanz Marco, Vicent et al. (2019). Deep neural networks, often known as DNNs, are increasingly becoming a crucial technology that enables many application fields. Nevertheless, doing on-device inference on embedding systems that are powered by batteries and have limited resources is often impractical because of the excessively lengthy time it takes to perform inference and the resource demands of several deep neural networks. Computation outsourcing to the cloud is often deemed unsuitable owing to privacy issues, significant latency, or insufficient connection. "Although compression methods often provide faster inferencing times, they also result in diminished accuracy." This research introduces a novel method to facilitate the effective implementation of Deep Neural Networks (DNNs) on embedded devices. Our methodology employs a dynamic process to ascertain the appropriate Deep Neural Network (DNN) to use based on the required accuracy and inference time. The system utilises machine learning techniques to create an inexpensive prediction model that efficiently chooses a pre-trained deep neural network (DNN) based on a given input and optimisation constraint. To do this, we begin by training a predictive model offline. Subsequently, we use the acquired model to choose a DNN model for processing fresh and unfamiliar inputs. We implement our methodology in two specific Deep Neural Network (DNN) fields: image classification and machine translation. We assess our methodology on a Jetson TX2 integrated deep learning platform and examine a variety of important DNN models, such as convolutional and recurrent neural networks. In the task of picture classification, we are able to reduce the time it takes to make predictions by 1.8 times, while also improving the accuracy by 7.52%, compared to the best single deep neural network model available. In the context of machine translation, we are able to decrease the time it takes to make predictions by 1.34 times compared to the most proficient individual model, while maintaining translation quality at a similar level.

Bharadiya, Jasmin. (2023). Deep learning has lately been used in several tasks such as scene labelling, object tracking, posture estimation, text detection and identification, visual saliency detection, and picture classification. Deep learning often employs models such as Auto Encoder, Sparse Coding, Restricted Boltzmann Machine, Deep Belief Networks, and Convolutional Neural Networks. Convolutional neural networks have shown superior performance in image classification compared to other models. "This research presents a simple and precise Convolutional Neural Network (CNN) designed for picture classification. This Convolutional neural network successfully completed the task of picture categorization. We explored several learning rate setting strategies and optimisation algorithms to determine the optimal parameters that have the most significant impact on picture classification, building upon the Convolutional neural network framework."

The authors of the study are Zhang, Dengqing, et al. (2021). Heart disease has become a significant concern to people's health in recent decades due to its widespread occurrence and high mortality rate. Consequently, the ability to forecast cardiac disease based on basic physical markers gathered during routine physical examinations in the early stages has become a significant topic of interest. From a clinical standpoint, it is crucial to be attentive to these symptoms associated with cardiac disease in order to make predictions and establish a dependable foundation for further diagnosis. Nevertheless, the substantial volume of data necessitates laborious and onerous human analysis and prediction. The objective of our study is to efficiently and precisely forecast cardiac disease by using many physiological signs. This research presents a unique approach for predicting cardiac disease. Our proposal entails the use of an algorithm that predicts heart disease by integrating the integrated feature selection approach with deep neural networks. "This integrated feature selection technique utilises the LinearSVC algorithm and employs the L1 norm as a penalty term to choose a subset of characteristics that are strongly correlated

with heart disease." The deep neural network we constructed is trained using these characteristics. The network's weight is initialised using the He initializer to mitigate the issues of gradient vanishing or explosion, hence enhancing the performance of the predictor. The heart disease dataset collected from Kaggle is used to test our model. The predictor is evaluated using indicators such as accuracy, recall, precision, and F1-score. The results indicate that our model achieves accuracy of 98.56%, recall of 99.35%, precision of 97.84%, and an F1-score of 0.983. Additionally, the average AUC score of the model is 0.983, confirming the efficiency and reliability of our proposed method for predicting heart disease.

Saito, Shota and colleagues (2018). Feature selection is a widely used strategy in machine learning to enhance the predicted accuracy and interpretability of a trained model. Feature selection strategies may be categorised under three categories: filter, wrapper, and embedding approaches. In general, the embedded technique effectively conducts feature selection during model training, resulting in a favourable trade-off between performance and computing cost. Our research introduces an innovative approach to feature selection, which involves embedding probabilistic model-based evolutionary optimisation. We present the multivariate Bernoulli distribution, which governs the choice of features, and we optimise its parameters throughout the training process. The update strategy for the distribution parameter is identical to that of population-based incremental learning (PBIL). However, we concurrently update the parameters of the machine learning model using a standard gradient descent technique. Non-linear models, such neural networks, may readily include this strategy. Furthermore, we include the penalty term in the goal function to regulate the quantity of chosen features. We use the suggested approach, using the neural network model, to perform feature selection for three classification challenges. The suggested technique demonstrates comparable performance and an acceptable computing cost when compared to traditional feature selection methods.

Bamidele, Awotunde, and colleagues (2021). The Industrial Internet of Things (IIoT) is a burgeoning field of study that connects digital devices and services to tangible systems. The Industrial Internet of Things (IIoT) has facilitated the collection of vast amounts of data from many sensors, leading to the identification of various device-related challenges. The Industrial Internet of Things (IIoT) has encountered many types of cyber assaults that endanger its ability to provide organisations with uninterrupted operations. These risks lead to financial and reputational harm for firms, as well as the unauthorised acquisition of sensitive information. Consequently, numerous Network Intrusion Detection Systems (NIDSs) have been created to combat and safeguard IIoT systems. However, gathering the necessary information for developing an intelligent NIDS is a challenging undertaking, resulting in significant difficulties in detecting both known and novel attacks. Hence, this research presents a sophisticated intrusion detection framework for Industrial Internet of Things (IIoT) using deep learning. The framework incorporates a hybrid rule-based feature selection method to effectively train and validate the information extracted from TCP/IP packets. The training method was executed with a hybrid approach that combined rule-based feature selection with a deep feed forward neural network model. The suggested technique was evaluated using two well-known network datasets, NSL-KDD and UNSW-NB15. According to the performance comparison, the proposed technique outperforms other related methods in terms of accuracy, detection rate, and FPR. "Specifically, for the NSL-KDD dataset, the suggested method achieves a 99.0% accuracy, a 99.0% detection rate, and a 1.0% FPR. For the UNSW-NB15 dataset, the suggested method achieves a 98.9% accuracy, a 99.9% detection rate, and a 1.1% FPR." Simulation trials using several assessment indicators ultimately demonstrated the suitability of the proposed strategy for classifying intrusion network attacks in the Industrial Internet of Things (IIoT).

Rawat, Waseem, and Wang, Zenghui. (2017). Convolutional neural networks (CNNs) have been used for visual applications since the late 1980s. However, despite being used in a few isolated cases, they remained inactive until the mid-2000s when advancements in computing power and the availability of large amounts of labelled data, along with improved algorithms, led to their progress and brought them to the forefront of a resurgence in neural networks that has been rapidly advancing since 2012. This article specifically examines the use

of Convolutional Neural Networks (CNNs) for the purpose of picture classification. We provide an overview of their evolution, starting with their ancestors and leading up to the most advanced deep learning systems now available. During our analysis, we examine (1) the first achievements of these models, (2) their significance in the resurgence of deep learning, (3) specific symbolic works that have played a part in their current appeal, and (4) various efforts to enhance them by evaluating the contributions and difficulties presented in more than 300 papers. In addition, we present some of their present-day patterns and persistent obstacles. Shen, Xin et al. (2019). Amazon production characteristics benefit from the use of fine-grained multi-label classification models, which can accurately predict labels based on visual information. "These models are versatile and can be used to many tasks, including fashion attribute identification and brand recognition." A major obstacle in achieving optimal performance for classification tasks in real-world scenarios is the presence of a complex visual background signal, which includes irrelevant pixels. This signal might confuse the model, making it difficult to concentrate on the area of interest and accurately forecast outcomes for that particular region. This research presents a novel semantic-embedding deep neural network that incorporates spatial awareness semantic features. The network utilises a channel-wise attention-based approach to enhance model performance for multi-label prediction by using localization guidance. We found a mean relative improvement of 15.27% in terms of AUC score across all labels when compared to the baseline strategy. The core experiment and ablation investigations focus on performing multi-label fashion attribute classification on images of fashion apparels from Instagram. We conducted a comparative analysis of the performance of our technique, the baseline approach, and three other ways in utilising semantic characteristics. The results indicate that our method performed well.

III. EXPERIMENTAL SETUP

We assess our methodology using the NVIDIA Jetson TX2 integrated deep learning platform. The system is equipped with a 64-bit dual-core Denver2 processor and a 64-bit quad-core ARM Cortex-A57 processor, both running at a speed of 2.0 GHz. Additionally, it has a 256-core NVIDIA Pascal GPU operating at a frequency of 1.3 GHz. The board is equipped with 8 gigabytes of LPDDR4 RAM and 96 gigabytes of storage, consisting of 32 gigabytes of eMMC storage and an additional 64 gigabytes via an SD card.The assessment platform we use operates on Ubuntu 16.04, using the Linux kernel version 4.4.15. We use Tensorflow version 1.0.1, cuDNN version 6.0, and CUDA version 8.0.64. The implementation of our prediction model use the Python scikit-learn library. The foundation of our feature extractor is based on the use of OpenCV and SimpleCV.We evaluate 14 pre-trained convolutional neural network (CNN) models for the purpose of image recognition. These models are sourced from the TensorFlow-Slim library. The models are constructed using TensorFlow and trained using the ImageNet ILSVRC 2012 training dataset.

We assess our methodology by using the criteria of (A) Inference time, (B) Energy usage, and (C) Accuracy. Accuracy (E) Retrieve (F) The F1 score

We calculate the geometric mean of the evaluation metrics described before for each cross-validation fold. In order to measure the time, it takes for a model to make inferences and the amount of energy it consumes, we continually run each model on each input until the 95% confidence bound per model per input is less than 5%. During the tests, we do not consider the loading time of the CNN models as they only need to be loaded once in practical applications. Nevertheless, we include the additional costs of our prediction model into all of our experimental data. In order to quantify energy use, we have created a streamlined programme that collects data from the energy sensors included into the device. Our work does not prioritise optimising for energy use. In our case, we discovered that there is little disparity between optimising for energy usage and optimising for time.

IV. RESULTS AND DISCUSSION

Inference Time

Figure 1 illustrates the disparity in inference time between individual DNN models and our technique.MobileNet is the most efficient model for performing inferences, with a speed that is 2.8 times quicker than Inception and 2 times faster than

ResNet. However, it has the lowest level of accuracy among these models. Our prediction model is three times quicker than MobileNet. The majority of the computational burden of our prediction model comes from the process of extracting features. Our technique has an average inference time of less than one second, which is somewhat more than the average time of 0.7 seconds for MobileNet. Our technique demonstrates 1.8 times increase in speed compared to Inception, which is the most precise inference model in our group of models. Considering the substantial increase in prediction accuracy that our technique brings to Mobilenet, we consider the reasonable cost of our predictive model to be appropriate.

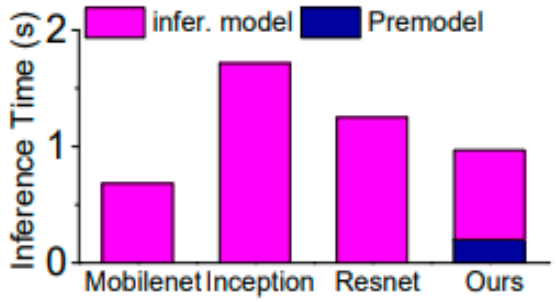

Figure 1: Inference Time

Energy Consumption

Figure 2 presents the data on energy use. The energy usage on the Jetson TX2 platform is directly proportional to the model inference time. By increasing the total inference speed, we are able to decrease the energy usage by nearly two times when compared to Inception and Resnet. Our predictive model has a tiny energy footprint, being 4 times lower than MobileNet and 24 times lower than ResNet. Therefore, it is appropriate for devices with limited power capabilities and may enhance the overall precision when using numerous inferencing models. Moreover, in situations when the predictive model anticipates that none of the deep neural network (DNN) models can accurately deduce an input, it has the capability to bypass the inference process in order to save power. It should be noted that our predictive model, which operates on the CPU, has a reduced energy footprint ratio compared to the runtime.

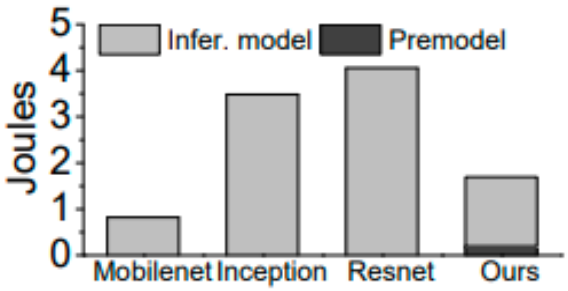

Figure 2: Energy Consumption

Accuracy

Figure 3 illustrates the comparison of the accuracy produced by each strategy, specifically in terms of top-1 and top-5 accuracy. We also provide the highest attainable accuracy provided by an ideal predictor for model selection, which we refer to as the Oracle. It should be noted that the Oracle does not provide complete accuracy since there are instances when all the Deep Neural Networks (DNNs) are unsuccessful. Nevertheless, it is important to note that different deep neural networks (DNNs) may not exhibit failure on the same pictures. For instance, whereas Inception may struggle to correctly identify some photos, ResNet is capable of effectively classifying them. Thus, by the efficient utilisation of numerous models, our technique surpasses the performance of each individual inference model.

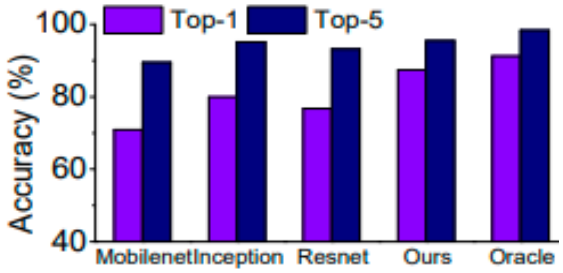

Figure 3: Accuracy

MobileNet's accuracy is enhanced by 16.6% and 6% for the top-1 and top-5 scores, respectively. Additionally, it enhances the top-1 accuracy of ResNet and Inception by 10.7% and 7.6% correspondingly. Although there is just a little increase of 0.34% in the top-5 score compared to Inception, our technique is twice as quick. Our methodology achieves a performance level of more than 96% compared to Oracle (87.4% for top-1 and 95.4% for top-5, as opposed to Oracle's 91.2% and

98.3% respectively). Furthermore, our technique consistently selects a model that never fails, even when others may succeed. This result demonstrates that our methodology may enhance the precision of inference in individual models.

Precision, Recall, F1 Score

Figure 4 demonstrates that our strategy surpasses individual DNN models in several assessment measures. Our technique specifically achieves the maximum overall accuracy, resulting in the best F1 score. Increased accuracy may effectively decrease the occurrence of false positive results, particularly in areas such as video surveillance. This reduction in false positives is crucial as it minimises the need for human intervention in verifying inaccurate predictions.

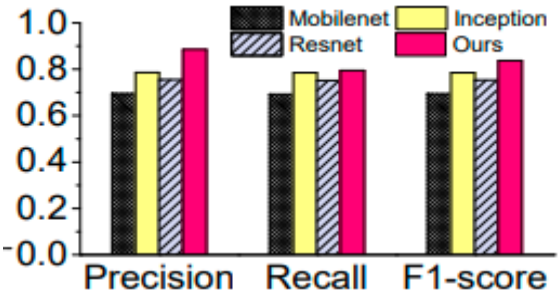

Figure 4: Precision, Recall & F1 score

V. CONCLUSION

The increasing ubiquity of embedded systems, coupled with the growing demand for intelligent image classification applications, necessitates a solution that balances the high accuracy of DNNs with the limited computational and energy resources characteristic of these systems. "Our approach leverages predictive modeling to guide the selection of the most appropriate neural network architecture, ensuring that the deployment is both efficient and effective." The predictive model-based selection framework represents a significant advancement in the deployment of deep neural networks for embedded image classification. By bridging the gap between the high accuracy of DNNs and the practical limitations of embedded systems, our approach paves the way for more efficient and effective use of deep learning in a wide range of embedded applications. The success of this framework underscores the importance of predictive modeling in the ongoing efforts to harness the full potential of artificial intelligence in resource-constrained environments.

REFERENCES

[1] Sanz Marco, Vicent & Taylor, Ben & Wang, Zheng & Elkhatib, Yehia. (2019). Optimizing Deep Learning Inference on Embedded Systems Through Adaptive Model Selection.

[2] Bharadiya, Jasmin. (2023). Convolutional Neural Networks for Image Classification. International Journal of Innovative Research in Science Engineering and Technology. 8. 673. 10.5281/zenodo.7952031.

[3] Zhang, Dengqing& Chen, Yunyi& Chen, Yuxuan & Ye, Shengyi& Cai, Wenyu & Jiang, Junxue& Xu, Yechuan& Zheng, Gongfeng& Chen, Ming. (2021). Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network. Journal of Healthcare Engineering. 2021. 1-9. 10.1155/2021/6260022.

[4] Saito, Shota & Shirakawa, Shinichi & Akimoto, Youhei. (2018). Embedded feature selection using probabilistic model-based optimization. 1922-1925. 10.1145/3205651.3208227.

[5] Bamidele, Awotunde& Chakraborty, Chinmay & Adeniyi, Emmanuel. (2021). Intrusion Detection in Industrial Internet of Things Network-Based on Deep Learning Model with Rule-Based Feature Selection. Wireless Communications and Mobile Computing. 2021. 1-17. 10.1155/2021/7154587.

[6] Rawat, Waseem & Wang, Zenghui. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Computation. 29. 1-98. 10.1162/NECO_a_00990.

[7] Shen, Xin & Zhao, Xiaonan& Luo, Rui. (2023). Semantic Embedded Deep Neural Network: A Generic Approach to Boost Multi-Label Image Classification Performance.

[8] Açıkgöz, Hakan. (2022). A novel approach based on integration of convolutional neural networks and deep feature selection for short-term solar radiation forecasting. Applied Energy. 305. 117912. 10.1016/j.apenergy.2021.117912.

[9] Sekaran, Karthik & Ramalingam, Srinivasa Perumal & P.V.S.S.R., Chandra Mouli. (2018). Breast Cancer Classification Using Deep Neural Networks. 10.1007/978-981-10-6680-1_12.
[10] Ajala, Sunday. (2021). Artificial Neural Network Implementation for Image Classification using CIFAR-10 Dataset. 10.13140/RG.2.2.12974.43844.
[11] Motzev, Mihail. (2023). Predictive Approach to Model Selection and Validation in Statistical Learning Networks. 10.13140/RG.2.2.36680.90888.
[12] Ashank, & Chakravarty, Soumen & Garg, Pranshu & Kumar, Ankit & Agrawal, Manish & Agnihotri, Prabhat. (2021). Deep neural networks based predictive-generative framework for designing composite materials.
[13] Sornam, Madasamy& Kavitha, Muthu Subash & Venkateswaran, Vanitha. (2017). A Survey on Image Classification and Activity Recognition using Deep Convolutional Neural Network Architecture. 10.1109/ICoAC.2017.8441512.
[14] Jha, Nandan Kumar. (2020). Hardware-Aware Co-Optimization of Deep Convolutional Neural Networks.
[15] Alpaydin, E. (2020) Introduction to Machine Learning. 4th ed. MIT, pp. xix.
[16] Brown, S. (2023) 'Machine Learning, Explained' Available at: https://mitsloan.mit.edu/ ideas-made-to-matter/machine-learning-explained (Accessed: 02 November 2023).
[17] Burns, Ed. (2017) 'Deep learning models hampered by black box functionality'. Available at: http://searchbusinessanalytics.techtarget. com/feature/Deep-learning-modelshampered-by-black-box-functionality (Accessed: 04 May 2017).
[18] Günnemann, St., Kremer, H., and Seidl, Th. (2011) 'An extension of the PMML standard to subspace clustering models. Proceedings of the 2011 workshop on Predictive markup language modeling, p. 48. doi:10.1145/2023598.2023605.
[19] Hastie, T., Tibshirani, R., and Friedman J. (2017) The Elements of Statistical Learning: data mining, inference, and prediction. New York: Springer.
[20] IBM newsletter. (2023) 'What is machine learning?' Available at: https://www.ibm.com/topics/machine-learning (Accessed: 02 November 2023).
[21] MicroStrategy, (2005) 'An Architecture for Enterprise Business Intelligence'. White Paper., pp. 162-173. Available at: http://www.microstrategy.com/Publications/Whitepapers (Accessed: 02 November 2023).
[22] Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2012) Foundations of Machine Learning. The MIT Press