CS5052 – Data Intensive Systems

220031985

University of
St Andrews

29th March 2023

“Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge”

Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski,
Trevor Mudge, Jason Mars, Lingjia Tang

INTRODUCTION

Intelligent personal assistants such as Siri, Alexa, Ok Google, and Cortana are now run in the cloud. The cloud-only method necessitates sending large amounts of data over the network to the cloud, putting enormous computing demand on the data centre. This method is *effective*, but it requires a dependable and fast internet connection, which is not *always available.*

Deep learning has become one of the most prominent topics in artificial intelligence due to its capacity to autonomously learn representations from raw data. It has been employed in a variety of applications, including *image identification*, *natural language processing*, and *autonomous driving.*

Deep learning models, on the other hand, are computationally costly and require a huge amount of processing resources, making them challenging to execute on mobile devices. This is owing to mobile devices' *limited computational capability*, which cannot manage the computational needs of deep learning models.

In the current period, mobile device computational capacity has become increasingly powerful and energy efficient, raising issues about whether cloud-only processing is desirable in the future, and what the ramifications are of transferring some or all of this computing to mobile devices at the edge.

The authors propose a collaborative intelligence framework called Neurosurgeon for optimising **deep neural network (DNN) inference** on mobile devices. In the paper, the authors address the problem of the limitations of mobile devices in performing complex DNN computations due to their limited computational power and memory capacity.

BACKGROUND

The Neurosurgeon framework is based on the concept of cloud-to-mobile edge intelligence collaboration. It optimizes the deep learning model's performance by dynamically changing the computational effort between the device and the server.

The app is in charge of executing the deep learning model on the mobile device. It continuously checks the model's performance and delivers performance metrics to the application's server. The performance metrics are the *current model accuracy*, the *computational complexity of the model*, and *the mobile device's processing capacity.*

These metrics are received by the server and analysed to establish the ideal computing workload for the model. It then employs a **reinforcement learning algorithm** to figure out the best workload distribution for the model.

The approach considers the current model accuracy, the computational complexity of the model, and the mobile device's available resources. Once the optimal workload distribution is determined, the server sends the workload partition instructions back to the device.

When the mobile device receives this data, it modifies the computational workload accordingly. If the device has enough resources, it will perform the computation locally; otherwise, it will offload the computation to a cloud server.

The model may *achieve high accuracy* while running effectively on the mobile device because of this dynamic workload adjustment.

NOVELTY

The work is **novel** as it introduces a *collaborative intelligence approach* for optimising DNN interference on mobile devices. Its framework uses a collaborative approach between the cloud and mobile devices to optimise deep learning inference on mobile devices.

Previous studies have attempted various methods for reducing the computational complexity of deep learning models on mobile devices, but the Neurosurgeon framework addresses these limitations by dynamically adjusting the computational workload between the mobile device and the server, allowing for efficient use of computational resources while maintaining high efficiency.

The use of a reinforcement learning algorithm to learn the optimal workload distribution for the model sets this *dynamic and adaptive approach* apart from other approaches and makes it *a contribution to the field of deep learning inference* on mobile devices.

Ultimately, the framework is a novel method for optimising deep learning inference on mobile devices, with the potential to improve the performance of deep learning models on a wide range of mobile devices, enabling the development of more efficient and accurate mobile applications.

EVALUATION

The authors evaluate Neurosurgeon as doing 8 DNNs as their benchmarks across Wi-Fi, LTE and 3G wireless connections with both CPU-only and GPU mobile platforms. They also demonstrate that it achieves significant *end-to-end latency* and *mobile energy* improvements over the status quo cloud-only approach.

The authors pit Neurosurgeon against **MAUI a well-known computation offloading framework** and then also evaluate its robustness to variations in wireless network connections and server load demonstrating the need for such a dynamic run time system.

The final evaluation is the data datacentre throughput improvement the Neurosurgeon achieves by pushing computing out of the cloud to the mobile device.

To improve energy efficiency, Neurosurgeon spends **24.2 per cent less energy** on average than the status quo method for suboptimal selections, and when optimising for best energy consumption, Neurosurgeon achieves a **59.5 per cent reduction** in mobile energy and up to **94.7 per cent reduction** over the status quo.

When compared to MAUI, both the Neurosurgeon and MAUI properly identify that local processing on the mobile device is optimal for NLP applications. The neurosurgeon correctly determines that in this particular scenario, it is best to execute the DNN entirely in the cloud, achieving comparable performance to the status quo and a **20.5 speedup over MAUI**.

Regardless of the server load, Neurosurgeon keeps the *end-to-end latency of executing image classification below 380ms.* By considering server load and its impact on the server performance, Neurosurgeon consistently delivers the *best latency* regardless of the variation in server load.

Overall, the authors report that it improves end-to-end latency by 3.1× on average and up to 40.7×, reduces mobile energy consumption by 59.5% on average and up to 94.7%, and improves data-center throughput by 1.5× on average and up to 6.7×.

CONCLUSION

To conclude, the Neurosurgeon system is a **promising solution** for collaborative intelligence between cloud and mobile edge devices, with the potential to improve the efficiency and effectiveness of intelligent applications in a variety of domains, including healthcare, autonomous driving, and smart homes.

Total Words: 985