ESC499 Thesis Interim Report

Automatic sleep staging transformer model and hardware accelerator

Tristan Robitaille (1006343397)
tristan.robitaille@mail.utoronto.ca

Supervisor: Professor Xilin Liu

January 12, 2024

Contents

List of Figures

List of Tables

1 Introduction[1]

As reported by Chaput *et al.* [1], insomnia impacts around 24% of Canadians adults. Detection and classification of sleep stages, known as *sleep staging*, followed by neuromodulation has been recently found by Yoon [2] to be a promising treatment against insomnia. The current stage-of-the-art for sleep staging involves the use of polysomnography to measure biosignals (at least 19 sensors are required, as explained by Levin and Chauvel [3]) and manual annotation by a sleep expert, which requires, on average, 2 hours of work [4]. This technique also does not provide neuromodulation. Thus, there is a need to develop an in-ear device performing electroencaphalogram (EEG) sensing, sleep staging and neuromodulation. To maximize treatment potential, the device should be as small and portable as possible such that it can be used at home.

This thesis focuses on the development of a deep learning model to perform sleep staging and on the design of an accelerator ASIC module to perform in-situ inference of said model. In the end, we aim to prove, by simulations, the merit of such an accelerator in order to potentially integrate it in the in-ear device. Multiple authors [5]–[7] have published high-accuracy results using a deep learning approach to sleep staging, and have done so with significantly fewer sensors than polysomnography. However, these AI models run on standard computers as software frameworks and are thus un-

[1] Adapted from the *Thesis Proposal*, submitted in October 2023.

suitable for a lightweight integrated solution. Google sells small custom AI-accelerators (such as the Coral Edge TPU) that could run these AI models, but they still consume too much power (2W, [8]) and do not readily integrate with custom neuromodulation hardware.

The proposed solution should match or exceed the accuracy of traditional polysomnography and published models in the literature with a power consumption low enough that the whole system can be powered for at least a full-night on a battery that fits in-ear.

This document serves to report the current state of literature in both sleep staging using deep learning and AI accelerator hardware in order to define a gap that is filled by this project. It also discusses the progress made to date and the work that is left.

2 Literature review: ML for sleep staging

Deep learning for sleep staging has been studied since around 2017. Broadly speaking, basic deep neural networks (DNN) cames first, followed by convolutional neural networks (CNN) and recurrent neural-networks (RNN) [9]. The transformer is a relatively new type of neural network based around the concept of "attention" and particularly suited for sequence inputs [10]. Since its introduction in 2017 [11], the transformer has been used for sleep staging tasks. Indeed, Dai *et al.* developed a transformer-like model without decoders which used three input EEG channels and achieved an impressive

87.2% accuracy on the popular SleepEDF-20 dataset [12]. Similarly, Phan *et al.* developed a model with a focus on outputting easily-interpretable confidence metrics for clinicians. Their model ingests multiple sleep "epochs" (small segments of EEG signals, typically 30s in length) for each inference, which allowed the team to achieve 84.9% accuracy on the SleepEDF-78 dataset. Eldele *et al.* managed an accuracy of 85.6% on SleepEDF-78 using a single-channel, single-epoch attention-based model [7].

In recent years, the accuracy of sleep staging by ML models has plateaued. In fact, Phan *et al.* claim that AI-based sleep-staging in heathly patients has been solved fully as the accuracy has reached the 'almost perfect' level of Cohen's kappa [4]. However, none of the models presented above meet our constraints. Indeed, we require a lightweight, single-channel, single-epoch model. Most models have more than 1M parameters [9]; even the smallest model by Eldele *et al.* has above 500k parameters. Furthermore, none have been optimized to run on custom hardware. Thus, there is a need to develop a novel lightweight transformer.

3 Literature review: AI accelerator hardware

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus vitae odio eget neque lacinia posuere. Donec et massa ut turpis interdum lobortis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Fusce eu aliquet quam, et sodales est. Duis vel elit nec odio ultricies ultrices. Nulla facilisi. Nullam eu ex eu odio volutpat efficitur.

4 Detailed design constraints and direction

Table 1 indicates precise design goals and their justification, which helps guide design decision and development effort. For example, to reach the target model size, time will be spent evaluating the impact of hyperparameters to find the combination that gives the lowest size while meeting the desired accuracy. Furthermore, quantization and pruning will be explored to reduce model size. For the AI accelerator, since inference power and frequency are inversely proportional, we must focus on reducing energy per inference. From first principles, this implies reducing the amount of charge that is displaced within the chip. Since the physical properties are locked for the target 65nm node, we focus on reducing the number of operations, simplifying operations, limiting data movement and reducing control logic.

Table 1: Design goals for AI model and ASIC accelerator

Type	Goals	Justification
Model	Size $< 200\,\mathrm{kB}$	Help reach ASIC area/power goals
	Accuracy $> 80\%$	Competitive with state-of-the-art
ASIC	$P_{\mathrm{avg}} < 1\,\mathrm{mW}$	System to function for whole night
	$T_{\mathrm{inference}} < 30\,\mathrm{s}$	Sleep epochs are 30s
	$A_{\mathrm{total}} < 1\,\mathrm{mm}^2$	Minimize cost (65nm node)

5 Progress to date

5.1 Transformer model and edge TPU

Since August 2023, I have completed Andrew Ng's Machine Learning Specialization course to build enough knowledge to tackle the transformer model. The model is based on a vision transformer [13], which accepts a 30s epoch and return the most likely sleep stage. The architecture is shown in Figure 1. Furthermore, I wrote feature-rich scripts to extract and preprocess PSG data, designed a complete vision transformer model with the required interface to run hyperparameter search and k-fold validation on the Compute Canada clusters. The model I developed contains only 63k parameters (pre-pruning), which can be quantized to 16-bit integers with a slight gain of accuracy. The accuracy on a 31-fold validation set is XYZ %. To help determine design priorities for the accelerator, the training script also exports the total number of different type of operations, as presented in Table 2.

In addition, I have ported to model to run on the Coral Edge TPU using TensorFlow Lite. This provided us with reference latency and power figures to size what I consider to be the most promising commercial alternative to a custom ASIC. This knowledge can be useful should we wish to develop a functional proof-of-concept prototype. In regular frequency mode (200MHz), the average inference time on the Edge TPU is 0.754ms. I have also researched initial ways the Edge TPU can be used with a microcontroller, which will be needed should we wish to make a prototype in the future.

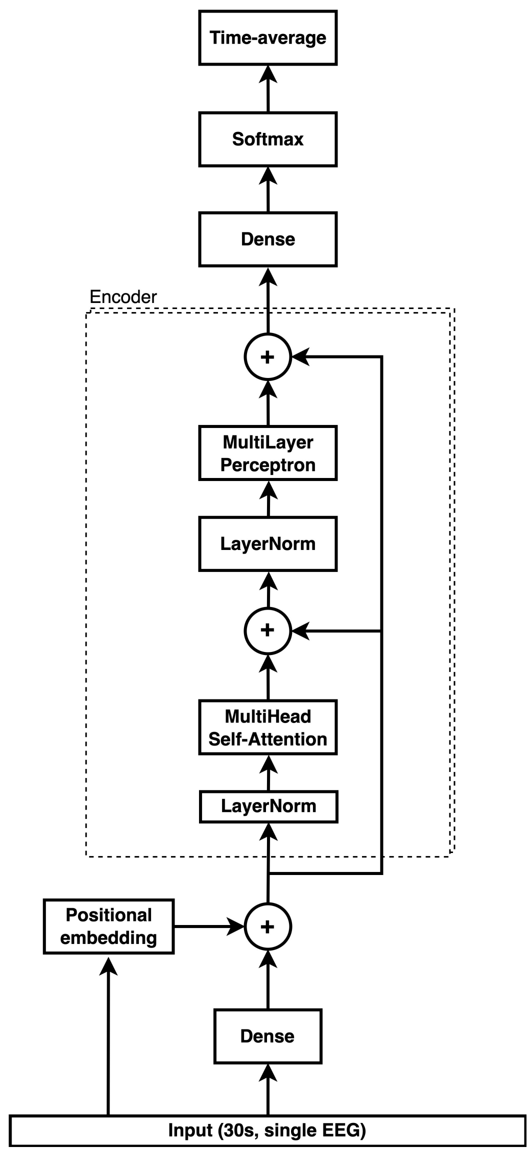

Figure 1: High-level vision transformer architecture for in-situ sleep staging

Table 2: Count of different types of operations in the transformer model

Operation	Total count	Percent of total
Addition	8581394	82.4%
Index increment*	1812531	N/A
Multiplication	1780528	17.1%
Division	34112	0.33%
Activation (SWISH)**	12064	0.12%
Substraction	9920	0.095%
Exponent	496	0.0048%
Square root	322	0.0031%

* Increments are excluded from total as they can be executed in parallel of any other element-wise operation.
** The count of elementary operations for activation depends on the approximation used in hardware.

5.2 ASIC accelerator

References

[1] Jean-Philippe Chaput, Jessica Yau, Deepa P. Rao, et al. "Prevalence of insomnia for Canadians aged 6 to 79". In: *Health Reports* 29.12 (2018).

[2] Ho-Kyoung Yoon. "Neuromodulation for Insomnia Management". In: *Sleep Medicine and Psychophysiology* 28.1 (2021), pp. 2–5.

[3] Jessica Vensel Rundo and Ralph Downey. "Chapter 25 - Polysomnography". In: *Clinical Neurophysiology: Basis and Technical Aspects*. Ed. by Kerry H. Levin and Patrick Chauvel. Vol. 160. Handbook of Clinical Neurology. Elsevier, 2019, pp. 381–392. DOI: `https://doi.org/10.1016/B978-0-444-64032-1.00025-4`. URL: `https://www.sciencedirect.com/science/article/pii/B9780444640321000254`.

[4] Huy Phan and Kaare Mikkelsen. "Automatic sleep staging of EEG signals: recent development, challenges, and future directions". In: *Physiological Measurement* 43.4 (2022), 04TR01.

[5] Micheal Dutt, Surender Redhu, Morten Goodwin, et al. "SleepXAI: An explainable deep learning approach for multi-class sleep stage identification". In: *Applied Intelligence* 53.13 (2023), pp. 16830–16843.

[6] Mingyu Fu, Yitian Wang, Zixin Chen, et al. "Deep learning in automatic sleep staging with a single channel electroencephalography". In: *Frontiers in Physiology* 12 (2021), p. 628502.

[7] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, et al. "An attention-based deep learning approach for sleep stage classification with single-channel EEG". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), pp. 809–818.

[8] *USB Accelerator datasheet*. Version 1.4. Coral. 2019. URL: https://coral.ai/docs/accelerator/datasheet/.

[9] Huy Phan, Kaare Mikkelsen, Oliver Y Chén, et al. "Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification". In: *IEEE Transactions on Biomedical Engineering* 69.8 (2022), pp. 2456–2467.

[10] Kai Han, Yunhe Wang, Hanting Chen, et al. "A survey on vision transformer". In: *IEEE transactions on pattern analysis and machine intelligence* 45.1 (2022), pp. 87–110.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[12] Yang Dai, Xiuli Li, Shanshan Liang, et al. "MultiChannelSleepNet: A Transformer-based Model for Automatic Sleep Stage Classification with PSG". In: *IEEE Journal of Biomedical and Health Informatics* (2023).

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. "An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020". In: *arXiv preprint arXiv:2010.11929* (2010).