Neural Architecture Search for Efficient Brain-Machine Interfaces

Noah Marti[(1)] and Federico Corradi [(1)]

[(1)] Dep. of Electrical Engineering, Eindhoven University of Technology, The Netherlands

Humans naturally communicate through language; losing this ability through disease or injury can be devastating [1]. Recent progress in the area of speech neuroprostheses enable the restoration of low-latency and high throughput communication in disabled patients by decoding signals directly from the brain [3].
Despite those advancements, these models use a significant amount of compute resources and have a large memory footprint. In short, they are not yet suitable for embedded, wearable devices. Our goal is to solve this issue by developing neural architecture search methods for designing efficient, low-power neural networks tailored to neuromorphic hardware and brain decoding tasks.

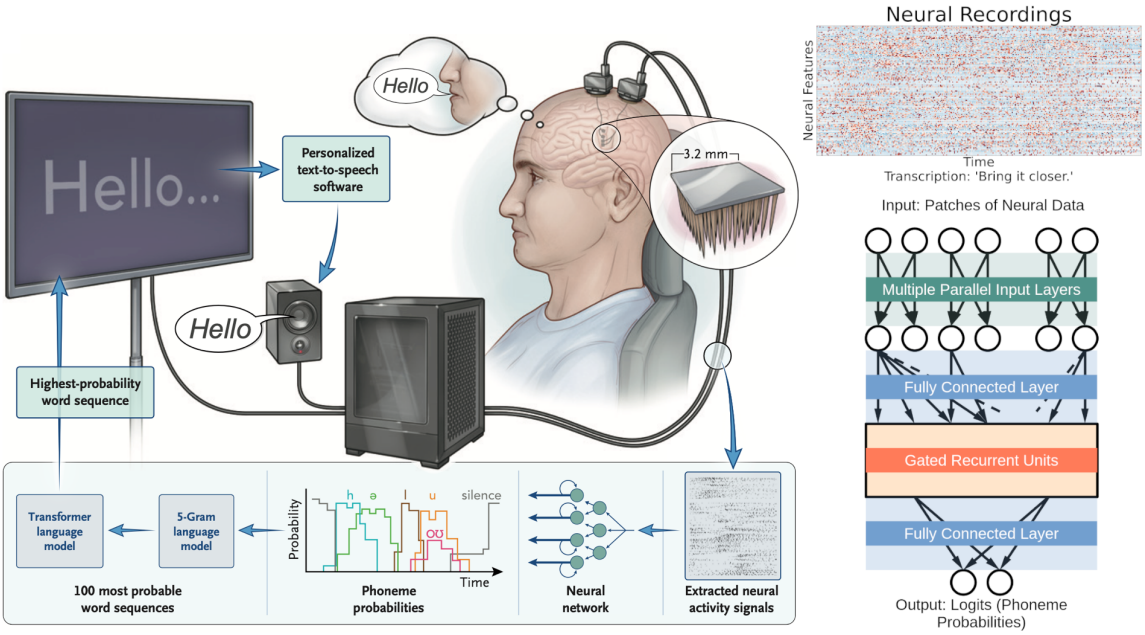

Figure 1: Methods adapted from N.S. Card et al. [2] extracting 512 from microelectrode arrays in the brain using multilayer perceptrons, gated recurrent units and language models to decode sentences.

We aim to reduce compute and memory needed for inference by introducing a hardware aware neural architecture search. With methods such as grid-search, bayesian inference and reinforcement learning we optimize an event-driven neural network for low memory, energy-efficiency and sparsity.
An initial approach is to optimize input encoding, layer types, number of layers and their sizes. Preliminary results show that we can reduce the number of parameters by 68.5% and FLOPs by 57.3%, without decreasing the raw phoneme accuracy. Achieving this result was possible by reducing parallel input layers and the number of GRU units through neural architecture exploration.

[1] M. Wairagkar et al., Nature (644.8075), 145-152, 2025.

[2] N.S. Card et al., New England Journal of Medicine (391.7), 609-618, 2024.

[3] A.B. Silva et al., Nature Reviews Neuroscience (25.7), 473-492, 2024.