Spiking Online Transformer with for Fast Prosthetic Hand Control

Nathan Leroux[1], Jan Finkbeiner[1], Emre Neftci[1]
[1]Forschungszentrum Jülich, PGI-15, RWTA Aachen, Germany

Summary. In this work [1], we demonstrate an online implementation of Transformers with Leaky-Integrate and Fire neurons [2]. Our model shows above state-of-the-art results on a finger position estimation dataset, and successfully shows that the sparsity of Spiking Neural Networks can be integrated into Transformers. These results therefore pave the way toward the implementation of state-of-the-art models like Transformers for signal processing in Neuromorphic hardware.

Transformers are state-of-the-art models for signal processing [3]. Unlike Recurrent Neural Networks, Transformers do not suffer from the vanishing gradient problem, and as such are very performant to learn long-range dependencies. Moreover, these models do not have inductive biases made from assumptions about the data structure and they can be trained very fast on GPUs since they can process entire temporal sequences in parallel. However, conventional Transformers relies on the self-attention mechanism, an operation that compares all the elements of a sequence with each other. If this operation is very powerful for Natural Language Processing, it is not suited for online processing of continuous signals because it scales quadratically with sequence length and require waiting for the end of a sequence before computing, which prevents smooth online computation.

In contrast, Spiking Neural Networks [2] integrate the concept of time in their operating model, and are therefore more suited for continuous signals and event base processing. Moreover, the inherent temporal sparsity makes them very promising for low power neuromorphic hardware implementations. However, the performance of SNNs still remain to match the performances of transformer models. Since SNNs integrate time in their operating models and Transformers must compute many time steps in parallel, integrating spiking neurons in Transformers is challenging and can lead to overheads. In this work, we adapt the attention mechanism of Transformers to make them compatible with the implementation spiking Leaky-Integrate and Fire neurons, thus implementing a Spiking Transformer.

We use our model to predict finger positions with surface Electromyography signals (sEMG). sEMG is a technique that senses currents running through muscular fibers' membrane [4]. Since they are triggered by electrical stimuli from the central nervous system and they only require electrodes on the skin, this method is gaining a strong interest as a mean for non-invasive Human-Machine Interfacing and prosthetic hand control. Deep learning models like Temporal Convolutional Neural Networks or Transformers are powerful to process sEMG signals. However, these techniques require to make inference on time windows larger than 100 ms, which prevent a very fine grain processing. For wearable prosthetic hand control systems, it is essential to develop algorithms able to process sEMG signals with a low power consumption, a high accuracy, and at a fast rate.

To process sEMG signals online, in this work we propose an online transformer that makes use of a linearized sliding window attention mechanism. In conventional Transformers, the self-attention mechanism compares three projections of the input sequence, which are called queries, keys, and values. Future and past projections of a sequence are compared all together, which induce a delay. With the sliding window attention mechanism, our model makes attention on past and local information. The depth of the information of our attention mechanism is a memory with a tunable length. We show how this attention mechanism can be serialized: it performs inference for each sequence element projection each as they are generated. To leverage information from past inputs, we store information in the keys and the values of the attention mechanism, and we update this memory dynamically as the projections are generated.

We test our model on a finger position regression through sEMG signals using the Non-Invasive Adaptive Hand Prosthetics Database 8 (NinaProDB8) dataset [5]. First, we show that our online transformer allows users to process sEMG signals with high accuracy using solely very short time windows of 3.5 ms, which permits a very smooth prosthetic hand control. Secondly, we show that selecting the temporal depth of the attention improves the results of signal processing and makes our model outperform a self-attention-based transformer, as well as previous state-of-the- art models. Finally, we show how our

custom online attention mechanism allows us to SNNs inside different modules of our Transformer architecture to increase the network sparsity, which in turn results in a reduction of the required number of synaptic operations by a factor of ×5.3 without loss of accuracy. In conclusion, our work is promising for fast, precise, and low-power Human-Machine Interfacing, and is another step toward bridging powerful deep learning models like Transformers and energy-efficient Neuromorphic models like SNNs.

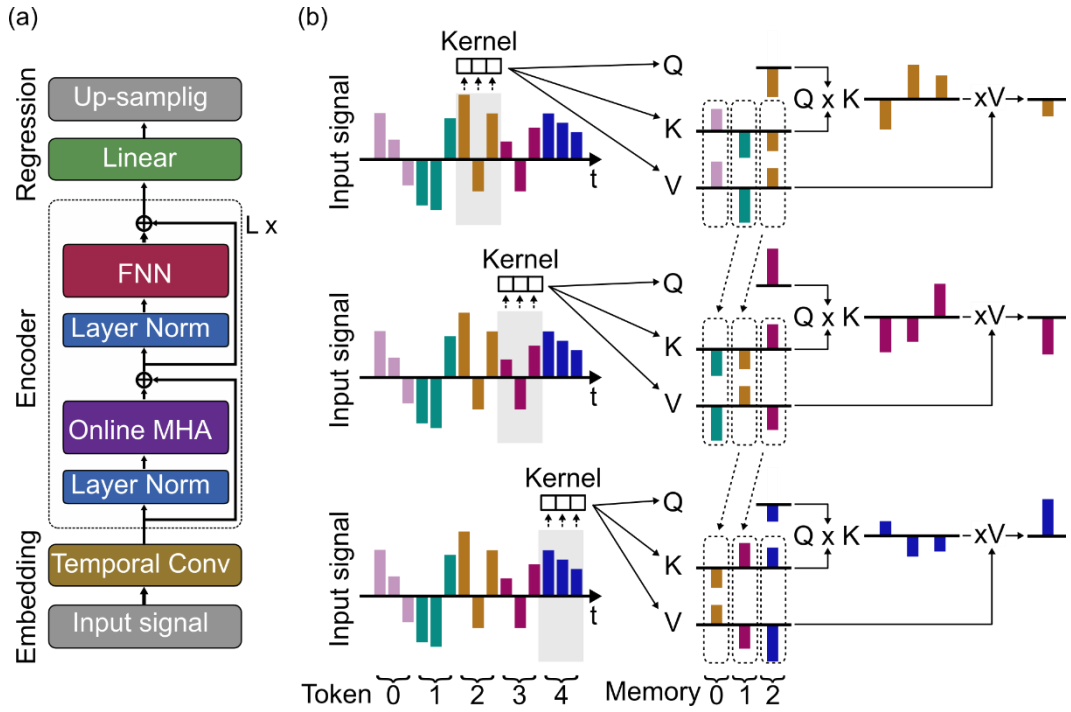

Figure 1: a) Online Transformer neural network architecture. In this architecture, we have replaced the Multi-Head Attention (MHA) and the Feedforward Neural Network (FNN) by Spiking Neural Networks (SNNs) to increase the model's sparsity. (b) Online Attention sketch: The different sequence elements are created by a temporal convolution (with a kernel size 3 and a stride 2 in this example). The sequence elements are linearly projected toward the queries, the keys and the values (Q, K, V). Q matches only the present sequence element whereas K and V store multiple previous tokens. The length M of this memory is 3 in this example. At each time step, K and V forget the projection of the oldest sequence element and store the projection of the new one.

References

[1] N. Leroux, J. Finkbeiner, and E. Neftci, *Online Transformers with Spiking Neurons for Fast Prosthetic Hand Control*, arXiv:2303.11860.

[2] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, *Deep Learning in Spiking Neural Networks*, Neural Networks **111**, 47 (2019).

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention Is All You Need*, in *Advances in Neural Information Processing Systems*, Vol. 30 (Curran Associates, Inc., 2017).

[4] M. Zheng, M. S. Crouch, and M. S. Eggleston, *Surface Electromyography as a Natural Human–Machine Interface: A Review*, IEEE Sensors Journal **22**, 9198 (2022).

[5] A. Krasoulis, S. Vijayakumar, and K. Nazarpour, *Effect of User Practice on Prosthetic Finger Control With an Intuitive Myoelectric Decoder*, Frontiers in Neuroscience **13**, (2019).