Hand Gesture Recognition Using Temporal Convolutions and Attention Mechanism

Elahe Rahimian†, Soheil Zabihi‡, Amir Asif ‡, Dario Farina†† S. Farokh Atashzar‡‡, and Arash Mohammadi†

†Concordia Institute for Information System Engineering, Montreal, Canada

‡ Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada

††Department of Bioengineering, Imperial College London, London, UK

‡‡Electrical & Computer Engineering, Mechanical & Aerospace Engineering, New York University, USA

Motivations/Contributions

Motivation:

- Providing a novel and innovative architecture (named as TC-HGR which refers to "Temporal Convolutions-based Hand Gesture Recognition architecture") for hand-gesture recognition.
- Data-driven models have been challenged by their need for a large number of trainable parameters and their structural complexity.

Contribution:

- The TC-HGR framework is proposed based on self-attention mechanism and temporal convolution to address the aforementioned challenges with the recurrent architectures.
- The TC-HGR reduces the number of parameters, which is a key step forward to embed the DNN models into prostheses controllers.
- The TC-HGR divides the sEMG signals into patches, which reduces the computational burden of the system.

Structure of the TC-HGR

- After pre-processing, we segment the sEMG signals based on a window of size $W \in \{200\,ms, 300\,ms\}$, resulting in the dataset $\mathcal{D} = \{(\boldsymbol{X}_i, y_i)\}_{i=1}^{M}$. More specifically, $\boldsymbol{X}_i \in \mathbb{R}^{C\times L}$ is the i^{th} segment with label y_i, for $(1 \leq i \leq M)$.
- Embedded Patches: The input segment $\boldsymbol{X}_i$ is divided into N non-overlapping patches. Here, $N = L/P$, where P shows the size of each patch. This patching mechanism helps reduce memory and computation requirements.
- Temporal Convolution Block: Inspired by the performance of Temporal Convolutions (TCs) for the sequential data, we used "Temporal Convolution Block" instead of recurrent networks for the sEMG-based HGR.
- Self-Attention Module with Residual Connection: In the proposed TC-HGR architecture, we used the "Temporal Convolution Block" along with the "Attention" mechanism. Attention mechanism allows a model to present important information in a given input sequence.

2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'22).

The proposed TC-HGR Architecture

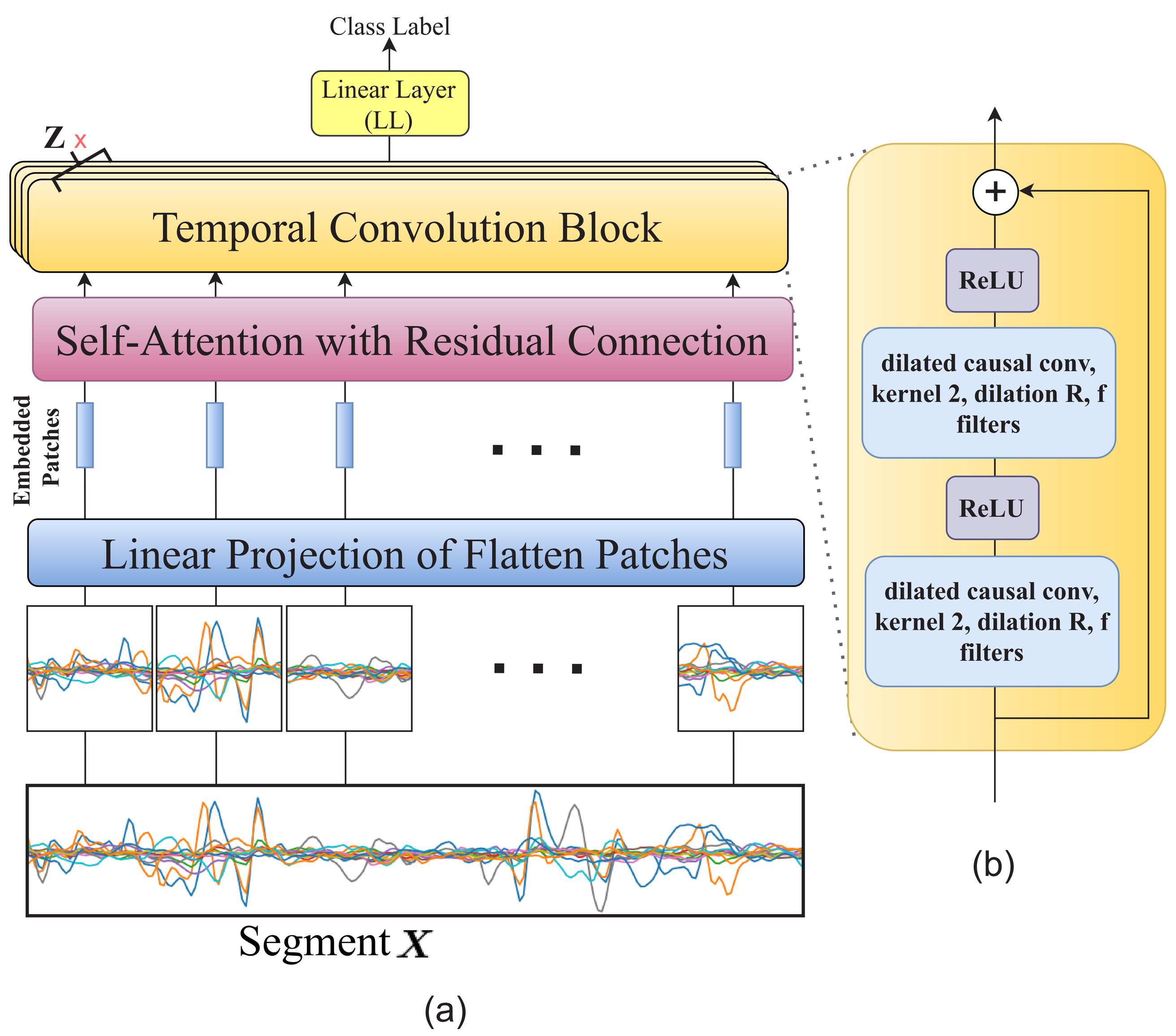


Figure 1: The proposed TC-HGR architecture: (a) Each input segment $\boldsymbol{X}$ (for simplicity, we dropped the index i) is divided into N non-overlapping patches. Then, each patch is flattened and mapped to model dimension D (blue block). We refer to the output of this process as Embedded Patches. The sequence of the Embedded Patches is passed into the Self-Attention module, which includes the residual connection (purple block). Afterward, we used $\boldsymbol{Z}$ number of Temporal Convolution Blocks to access a long history (orange block). (b) Each Temporal Convolution Block consists of two dilated causal convolutions, each followed by a ReLU activation function. Again, we used residual connections to concatenate the output and input. Finally, a Linear Layer (LL) is adopted to output the class label.

Conclusion

- The proposed model showed strong capability in addressing several existing challenges of gesture recognition based on the temporal convolutions and attention mechanism.
- We showed that by proper design of convolution-based architectures, we can extract temporal information of the sEMG signal and improve the performance.
- the proposed architecture can reduce the required number of trainable parameters with respect to the state-of-the-art, which is a key enabling factor to reduce the complexity and embed DNN-based models into prostheses controllers.

References

[1] E. Rahimian, S. Zabihi, F. Atashzar, A. Asif, A. Mohammadi, "XceptionTime: Independent Time-Window XceptionTime Architecture for Hand Gesture Classification," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1304-1308, 2020.

[2] E. Rahimian, S. Zabihi, A. Asif, S.F. Atashzar, and A. Mohammadi, "Few-Shot Learning for Decoding Surface Electromyography for Hand Gesture Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1300-1304.

[3] E. Rahimian, S. Zabihi, A. Asif, D. Farina, S.F. Atashzar, and A. Mohammadi, "FS-HGR: Few-shot Learning for Hand Gesture Recognition via ElectroMyography," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2021.

Experimental Results

Experiment 1: Effect of the Model's Dimension D::

Table 1: Descriptions of TC-HGR architecture variants.

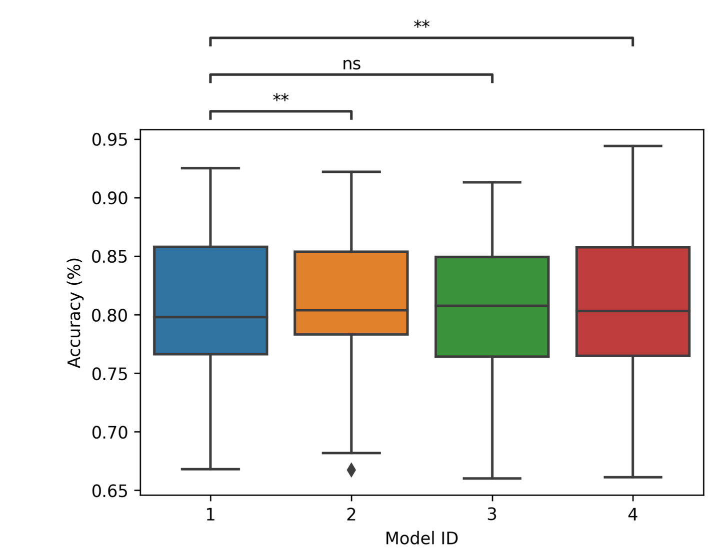
Window size W	Model ID	Number of Patches N	Model dimension D	Params
200 ms	1	10	12	49,186
	2	10	16	68,445
	3	16	12	69,076
	4	16	16	94,965
300 ms	1	10	12	52,066
	2	10	16	72,285
	3	15	12	67,651
	4	15	16	92,945

Experiments 2: Effect of the Number of Patches N and Window size W:

Table 2: Classification accuracies for TC-HGR architectures variants for different Window Size (W).

	Model ID	1	2	3	4
$W=$ 150 ms	Accuracy (%)	79.78	80.22	80.00	**80.72**
	STD (%)	6.5	6.5	6.5	**6.4**
$W=$ 200 ms	Accuracy (%)	80.29	80.63	80.51	**80.83**
	STD (%)	6.7	6.8	6.7	**6.6**
$W=$ 250 ms	Accuracy (%)	80.34	80.99	80.73	**81.35**
	STD (%)	6.4	6.5	6.6	**6.6**
$W=$ 300 ms	Accuracy (%)	80.84	81.59	80.95	**81.65**
	STD (%)	6.4	6.5	6.5	**6.7**

Statistical Comparisons of the Different TC-HGR Variants for Window Size 300 ms::



Comparison with the State-of-the-art Research [10]:

Table 3: Comparison between the proposed TC-HGR methodology and previous works [10], which used recurrent architectures (LSTM).

		200 ms		300 ms	
		Params	Accuracy (%)	Params	Accuracy (%)
Reference [10]	4-layer 3rd Order Dilation	1, 102, 801	79.0	1, 102, 801	82.4
	4-layer 3rd Order Dilation (pure LSTM)	-	-	466, 944	79.7
	SVM	-	26.9	-	30.7
Our Method	Model 1	**49,186**	80.29	**52,066**	80.84
	Model 4	94, 965	**80.72**	92, 945	**81.65**