# Attention analysis of a sign language recognition task on the AUTSL dataset

***Jeanne Coppin, Mathieu De Coster and Joni Dambre***

Attention has been a breakthrough in various NLP tasks and is now commonly used in neural networks. By weighting the components of the input, the attention mechanism introduces the notion of relevance of those components for a given task. Some researchers use attention analysis to interpret their model's outputs and understand its decision-making process. While attention interpretability in action recognition is generally accepted, its interpretability in NLP models is still a contentious issue [1] [2].

Sign Language Processing is no exception and many best-performing models in sign language recognition and translation use attention mechanisms. However, attention analysis in Sign Language Processing tasks has only been slightly covered. The interpretability of attention in this field is an open issue: the features extraction of sign languages relates to action recognition tasks due to their visual-manual modality. Yet, sign language processing is a subfield of NLP, where attention interpretability is disputed.

This work analyzes temporal attention (i.e., which frames have the highest attention weights) of the transformer presented in [3] and aims to explore whether the model attention correlates to human intuition. [3] reaches an accuracy of 92.92% on the AUTSL dataset for isolated sign language recognition. This model is fed with cropped images of the hands, passed through the convolutional neural network ResNet-34, pretrained on ImageNet, and movement of the body, using pose flow estimation.

The analysis of the attention is restricted to a subset of signs, chosen for their representativeness of the dataset diversity as described in [4]. We choose [5]'s framework of the frame-level phonetic representation of signs for our analysis. This framework was developed to capture the important phonetic details of how signs are produced by individuals. Each frame of the sign is labeled either as postural (no change in any of the SL parameters) or trans-forming. Different types of trans-forming movement are further detailed. We thereby aim to define which types of frames have the highest attention weights in the transformer and examine whether the model learns linguistic features that humans use to describe signs.

At first sight, it looks as though the network used in [3] learns interpretable attention patterns. However, a systematic analysis of a larger subset of samples shows that we cannot observe consistent attention patterns with the input used in [3]. Analyses in future work will determine whether more linguistically informed inputs, that is to say, aligned with the sign language parameters, lead to more explainable attention patterns.

References:

[1] Serrano, S., Smith, N. A., "Is attention interpretable?", ACL, 2019.
[2] Wiegreffe, S., & Pinter, Y., "Attention is not not explanation", EMNLP, 2019.
[3] De Coster, M., et al., "Isolated Sign Recognition from RGB Video using Pose Flow and Self-Attention." CVPR, 2021
[4] Sincan, O. M., & Keles, H. Y., "AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods", IEEE Access, Vol. 8, 2020.
[5] Johnson, R., & Liddell, S., "A Segmental Framework for Representing Signs Phonetically", Sign Language Studies, Vol. 11, 2011.