Neural Sign Actors: A diffusion model for 3D sign language production from text Supplementary Material

Vasileios Baltatzis[1], Rolandos Alexandros Potamias[1], Evangelos Ververas[1], Guanxiong Sun[2], Jiankang Deng[1], Stefanos Zafeiriou[1]

[1]Imperial College London, [2]Queen's University Belfast

The supplementary material of this paper consists of a video file along with this document. The video file first contains examples of the fitting pipeline's outputs. Additionally, we include a qualitative comparison of generated signs between the proposed method and Stoll *et al*. [2], as well as a qualitative assessment of the ablations. The rest of this document contains details regarding the architecture and the implementation of the back-translation pipeline.

1. Back-translation

1.1. Architecture

The translation network is an encoder-decoder architecture. The encoder takes a series of 3d poses extracted from a video sequence as input and produces pose embeddings. Then, the decoder takes the pose embeddings as input and translates it into text.

The encoder consists of L_e transformer encoder layers [3]. L_e is set to 6 by default. Each transformer encoder layer contains two sub-modules: (1) a Multi-Head Self-Attention layer that allows each position in the input sequence to attend to all positions in the same input sequence; (2) a Feed-Forward Network to further fuse information of each position. The input pose tensor is a 2D tensor with a shape of $[T, c]$, where T denotes the number of frames and c denotes the dimension of pose vectors. Learnable queries Q are introduced to extract pose embeddings. Q contains N_q = 64 query tokens whose dimension is d = 768. Specifically, we pass the pose tensor through a linear layer which projects the pose dimension from c to d and then we concatenate Q and the projected pose tensor as the input of encoder layers. Finally, we keep the first N_q embeddings from the output of the last encoder layer as the pose embeddings.

We follow the general practice in the paper [3] to design the decoder. The decoder consists of L_d transformer decoder layers. L_d is set to 6 by default. Each transformer decoder layer has two Multi-Head Attention (MHA) layers: a masked self-MHA layer and a cross-MHA layer. The masked self-MHA layer attends to the target sequence and is masked to ensure that each position can only attend to previous positions, preventing information flow from future tokens during training. The cross-MHA layer attends to the encoder's output (pose embeddings) using the output of the previous self-MHA layer as the query and the encoder embeddings as the key and value.

1.2. Implementation Details

The training strategy employs the AdamW optimizer, featuring a learning rate of 1e-4 for parameter updates. A weight decay of 0.05 is applied to the model's weights for regularization, with an epsilon value of 1e-8 to enhance numerical stability. The betas parameter is set to (0.9, 0.999), dictating the exponential decay rates for gradient and squared gradient moving averages.

A Cosine Annealing learning rate scheduler is utilized. This scheduler operates on an epoch-by-epoch basis, starting with a learning rate at the beginning of training (epoch 0) and gradually reducing it to its minimum by epoch 50. This dynamic adjustment of the learning rate follows a cosine annealing schedule, contributing to the model's convergence.

We opted on learning relative joint features as we observed that they significantly increase the expressivity of the model to learn high frequency motions and achieve better performance [1]. For the text encoder, we use the pretrained CLIP model from Huggingface.

References

[1] Rolandos Alexandros Potamias, Alexandros Neofytou, Kyriaki Margarita Bintsi, and Stefanos Zafeiriou. Graphwalks: efficient shape agnostic geodesic shortest path estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2968–2977, 2022. 1

[2] Stephanie Stoll, Armin Mustafa, and Jean-Yves Guillemaut. There and back again: 3d sign language generation from text using back-translation. In *2022 International Conference on 3D Vision (3DV)*, pages 187–196. IEEE, 2022. 1

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1