RecVis 21 Final Project Proposal

Paul-Emile Zafar

I have chosen the topic C: "**Detecting Temporal Boundaries in Sign Language Videos**".

1 Method

The project will be split in two major parts.

1.1 Reproducing results

The first objective will be to reproduce the results presented in [1]. More precisely, we will focus on the fully supervised training results, done on an annotated part of the BSL-Corpus dataset [2]. Within the given code (available at `https://github.com/RenzKa/sign-segmentation`), the authors of [1] provide:

1. A usable demo that extracts features using the I3D model (a video classificatin model pretrained on the Kinetic dataset), then apply a pre-trained Multi-Stage Temporal Convolutional Network (MS-TCN) giving segmentations.
2. A training run-file that trains the MS-TCN network on the I3D pre-extracted features.

I could therefore apply the pretrained MS-TCN on various examples (both from BSL-Corpus and other video segments) to have a qualitative overview of what the model is capable of. Then I will familiarize myself with the code, and try to re-obtain the paper mF1B score after retraining the model.

1.2 Using transformers

The second step of the project will involve replacing the MS-TCN architecture by a transformer-based one [3], as transformers have proven to be efficient on various sequence segmentation cases. The series of image features of a clip extracted from the I3D architecture model (already pre-computed) will be the input, and we aim at predicting a segmentation output.

I will experiment with basic transformers architectures first, train them and evaluate my results on the annotated test dataset provided, using the mF1B score for comparison with the original paper.

References

[1] K. Renz, N. C. Stache, S. Albanie, and G. Varol, "Sign language segmentation with temporal convolutional networks," in *ICASSP*, 2021.

[2] A. Schembri, J. Fenlon, R. Rentelis, and K. Cormier, "British sign language corpus project: A corpus of digital video data and annotations of british sign language 2008-2017 (third edition)," 2017.

[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*, pp. 5998–6008, 2017.