Assignment 3

Michael Resplandy michael.resplandy@polytechnique.edu
Gabriel Faivre gabriel.faivre@polytechnique.edu

Abstract

In this project, we will study the recent approaches [1] and [2] for the task of sign language translation on the PHOENIX14T dataset. We will answer the question of whether there is any benefit from explicitly modeling the body pose (e.g. hands, face, upper body). Those body-pose keypoints will be computed using [3] and added as an additional input to the model of [1].

1. Introduction

Neural Machine Translation has made significant advances to translate from one spoken language to another. Sign languages received less attention; however, they are more complex due to being visual signals. Those visual languages use manual features (hand shape, movement and pose), but also facial expression and movement of the whole upper body (mouth, head, shoulders and torso). Those non-manual features are often neglected, like in [1], that focuses on the video input (Fig. 1). To take those additional features into account, we will use the DOPE, a method described in [3], that allows to detect and estimate whole-body 3D human poses, including bodies, hands and faces, in the wild. We will add those human poses as an additional input to the model in [1] and compare our performance on the PHOENIX14T dataset.

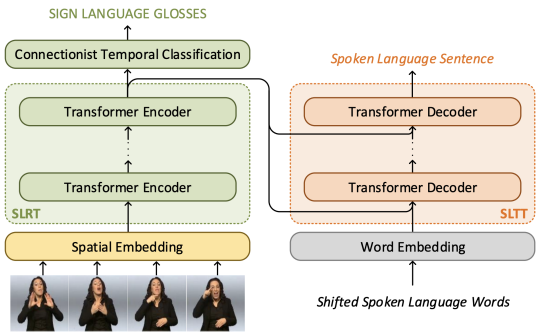

Figure 1. An overview of an end-to-end Sign Language Recognition and Translation approach using transformers implemented (taken from [1])

2. Team Work

Given the choice of the architecture of [1], their models can be used on four different learning tasks: Recognition from sign to gloss annotations, from sign to spoken language sentence, translation from gloss annotations to spoken language sentence and the combined recognition-translation task. Whereas, the architecture of the model of [2] allow us to evaluate only on the recognition task from sign to spoken language sentence.

- In the first part of our project, we will reproduce the results of [1] for all the metrics on the four learning tasks described above. (**Gabriel Faivre**)
- Secondly, we will use DOPE [3], to obtain 2D and 3D body pose keypoints (face, hands, body) from Phoenix-2014T. (**Michaël Resplandy**)
- Finally, we will investigate different strategies to fuse the information extracted by DOPE with the Sign Language Transformer model of [1]. We will compare the obtained results with the [1] and [2] benchmark. (**Gabriel and Michaël**)

This organization seems optimal as it allows the two of us to work in parallel on two different parts and still have a global vision on the project.

A natural extension of our work would be to substitute the encoding first part of the [2] model and replace it by DOPE - or a part of DOPE.

References

[1] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10023–10033, 2020. 1

[2] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Multi-channel transformers for multi-articulatory sign language translation," *arXiv preprint arXiv:2009.00299*, 2020. 1

[3] P. Weinzaepfel, R. Brégier, H. Combaluzier, V. Leroy, and G. Rogez, "Dope: Distillation of part experts for whole-body 3d pose estimation in the wild," in *European Conference on Computer Vision*, pp. 380–397, Springer, 2020. 1