FP101 - Topic F - Sign Language Translation from Video to Text

LIU Vincent
ENS Paris Saclay
liuvincent25@gmail.com

Abstract

Our project focuses on Sign Language Translation from Video to Text. We aim at reproducing and exploring the work of [2], the authors present a model based on Transformer architecture in order to perform jointly Continuous Sign Language Recognition (CSLR) and Translation (SLT). As suggested as future work, we study how modeling multi sign articulators such as faces, hands and body can improve the original method. To this end, we use Distillation Of Part Experts for whole-body 3D pose estimation in the wild (DOPE) [6] to obtain keypoints coordinates as well as hidden feature representation of the articulators. We explore, evaluate and compare some fusion strategies in order to incorporate the pose estimation to the existing model.

1. Introduction

Sign Language is the main medium of communication of the Deaf. It is a visual language using multiple channels to convey information such as manual features (hand shape, pose and movement) as well as non-manual features (facial expression, mouth and movement of the head) [2].

In Sign Language Translation (SLT), we are given a video of someone performing sign and we want to produce a model able to predict the corresponding translation in a spoken language sentence. Such system aims to help improving the Deaf daily life communication. It is a challenging task since it involves a huge amount of video processing, data annotation with expert knowledge and facing technical difficulties: we need to model each body part as well as its dependencies with each other to interpret the meaning of a sign sentence.

Deep Learning methods to solve this spatio-temporal machine translation task has shown great advances through the work of [2] and [5], which leverages the powerful transformers architecture.

2. Related work

Sign Language Transformers

In this paper, the authors propose a transformer based architecture that jointly learns Continuous Sign Language Recognition (CSLR) and Translation (SLT) while being trainable in an end-to-end manner, according to [2]. CSLR refers to the task of recognizing sign gloss representation while SLT is the task of generating corresponding spoken language sentences.

CSLR is an auxiliary task to provide the network additional supervision. It has been shown that it can improve significantly the SLT performances. It allows the network to learn meaningful spatiotemporal representations guided by gloss representations, to improve the end SLT objective.

Multi-channel Transformers for Multi-articulatory

The authors proposed an another transformer based architecture. It tries to learn the relationships between different sign articulators to be modelled within the transformer network itself [5]. The originality of this approach is to preserve channel specific information by introducing an anchor loss. In this paper, they want to reduce the need of gloss supervision, since gloss-level annotations can be expansive in practice as it needs expert knowledge and it is a repetitive task. The authors evaluate their approach against multiple standard configurations to feed the multi-articulatory inputs such as early fusion and late fusion ensembling.

DOPE

Distillation Of Part Experts for whole-body 3D pose estimation in the wild (DOPE) is a model that detect and estimate whole-body 3D human poses, including bodies, hands and faces [6]. While most of the methods focus on modeling separately body, hands, face poses individually, the authors proposed a method that could perform the three tasks within the same neural network.
To this end, they train the model in two stages. In the first step, they train separately three neural networks for the three tasks independently, which are called the experts. The second step is the procedure of teaching a single neural

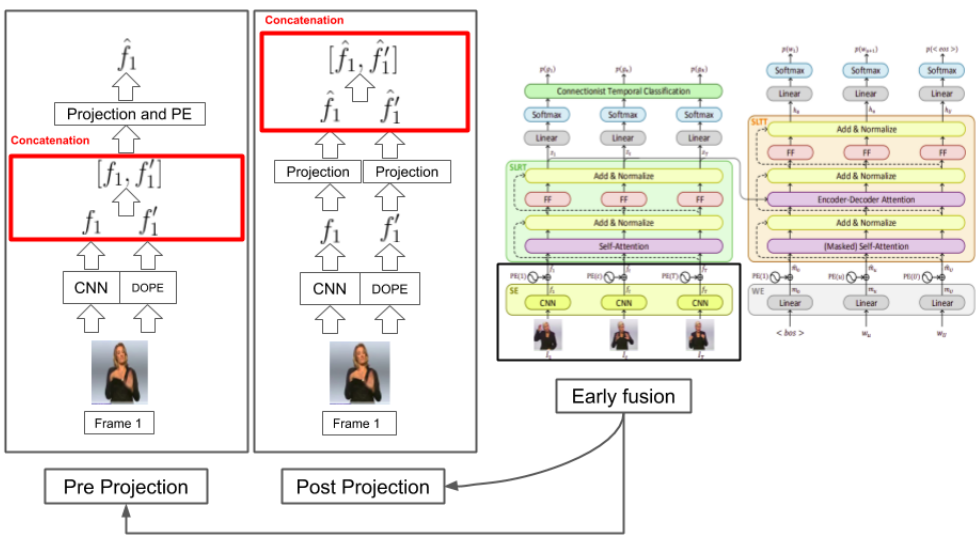

Figure 1. Scheme of early fusion before projection and post projection (left) and a detailed overview of a single layered Sign Language Transformer (right) taken from [2]. The image example is from PHOENIX14T datatset [1].

network with the output of the three experts. This process is similar to Knowledge distillation [3], in which a student model is trained to match the teacher models. Knowledge is transferred from the teacher models to the student by minimizing a loss function between the student and teacher outputs. According to the authors, a model using knowledge distillation outperforms the experts trained without distillation for the whole-body pose estimation.

3. Methodology

In this section, we present our methodology. First, we try to reproduce the results from the original paper [2] and set it as our baseline. Afterwards, we apply DOPE [6] to the compressed PHOENIX14T dataset [1] to obtain the whole-body pose estimation. We investigate several fusion strategies in order to incorporate the DOPE outputs as additional input to the Sign Language Transformer [2]. To evaluate the performances, we use the BLEU-4 score similarly to [2].

Extract whole-body pose estimation with DOPE

We extract 2D and 3D keypoints coordinates of the hands, face and body pose estimation as well as the hidden feature representation through DOPE [6].

A hidden feature representation corresponds to the feature maps before ROI Pooling, which are 1064 x 62 x 50 dimensional vectors. We use Global Average Pooling to map it to a 1064 dimensional vector. We can interpret this representation as a global encoding of a video frame.

Since DOPE can have many failure cases, especially when the hands are blurry (10% success on hand detection for some cases), we input the missing values to the most recent frame's valid pose estimation. The advantage of using hidden feature representations is that it does not meet failure cases. The drawback is that it is simply a global representation while the keypoints coordinates are more fine grained.

Fusion strategies

We explore several fusion strategies. The most straightforward way is to perform early fusion. We simply concatenate the spatial feature representations provided by the authors with those obtained with DOPE. The concatenated features are projected into a denser space where similar vectors are close to each other as shown in Figure 1.

A slight modification of early fusion is to perform concatenation at the embedding level, after applying independent spatial projection to each input type, as shown in Figure 1. We can call this approach early fusion post concatenation.

Another way to incorporate the whole body pose estimation into the model is to perform late fusion, which consists of ensembling different networks at the decoder level. We train some "weak" learners, each of them is trained on different input type. At testing time, we predict a new word by taking a weight sum of the output probabilities of each small model:

$$w_i^* = argmax \qquad \lambda_1 P_1(w_i) + \cdots + \lambda_N P_N(w_i)$$

with $\sum_i \lambda_i = 1$

3.1. Implementation Details

We use directly the original code from https://github.com/neccam/slt. For early fusion with post projection and late fusion, we change the minimal amount to make it work. We ran the code on Google Colab Pro with Tesla V100 GPUs. The history of our experiments and our code can be find at https://github.com/liuvince/mva-slt.

4. Evaluation

In this section, we present our experimental results.

4.1. Baseline

Table 1 shows the results when attempting to reproduce the Table 4 from the original paper [2]. We remark that the performances are overall lower the the published paper results. The reason might be that we do not use exactly the same hyper parameters.

We notice that we are able to verify the authors statement about Sign2(Gloss+Text) ($\lambda_R > 0$) outperforming Sign2Text ($\lambda_R = 0$): adding gloss recognition as an auxiliary task improve the performance. The baseline, with equal loss weights on recognition and translation is a good start, but it overfits the dev set a little (21.08 BLEU-4 on dev set compared to 20.41 BLEU-4 on test set). Moreover, we can check that increasing the recognition loss weight can also improve both the recognition and the translation task.

Loss Weights		DEV		TEST	
λ_R	λ_T	WER	BLEU-4	WER	BLEU-4
0.0	1.0	-	18.75	-	18.92
1.0	1.0	41.53	**21.08**	41.11	20.41
2.5	1.0	31.71	20.40	31.32	19.79
5.0	1.0	**28.82**	19.39	**28.55**	20.49
10.0	1.0	30.48	20.18	29.96	19.86
20.0	1.0	30.61	19.80	31.30	20.52
40.0	1.0	32.53	20.32	32.50	**20.64**

Table 1. **B**aseline: Results trying to reproduce Table 4 from [2]. Training SLT to jointly learn recognition and translation with different weight on recognition loss with baseline features obtained from [4].

4.2. Early fusion

We study the impact of adding each articulatory individually on Table 2. We see that the performances are also impacted by the loss recognition weight. We are able to beat the baseline slightly for same specific cases (**B**aseline + body $\lambda = 40.0$, **B**aseline + hands $\lambda = 20.0$).

Table 3 shows the impact when provided multiple articulately cues. We concatenate the **K**eypoints coordinates altogether and use **H**idden features. Overall, we don't have a meaningful improvement. However, we have one run with **B** + **K** + **H** as input in which we have a strong BLEU-4 score on both DEV (20.62) and TEST set (20.81).

Input	λ_R	2D	3D
Baseline	40.0	**20.64**	20.64
B + hands	1.0	19.23	20.66
	20.0	20.60	19.40
	40.0	19.20	**20.72**
B + body	1.0	20.13	18.83
	20.0	19.40	**20.87**
	40.0	19.43	20.51
B + face	1.0	18.50	**20.54**
	20.0	20.45	20.03
	40.0	19.63	20.29

Table 2. Early Fusion: Training SLT [2] by adding single articulatory keypoint coordinates to the **B**aseline features [4], with different combination of loss weights. Measurement metric is BLEU-4 on TEST set.

		DEV		TEST	
Input	λ_R	WER	BLEU-4	WER	BLEU-4
Baseline	40.0	32.53	20.32	32.50	20.64
B + **K**	1.0	49.64	20.08	48.09	19.58
	20.0	32.05	19.72	32.43	20.05
	40.0	28.32	**20.76**	29.21	20.19
B + **H**	1.0	46.41	20.16	45.50	19.67
	20.0	**27.89**	19.89	**28.39**	19.79
	40.0	29.44	20.20	29.21	20.55
B + **K** + **H**	1.0	43.88	19.84	43.51	19.48
	20.0	28.82	20.62	29.04	**20.81**
	40.0	30.85	19.91	30.66	19.44

Table 3. Early Fusion: Training SLT [2] by adding multiple articulately cues (**K**eypoints coordinates, **H**idden features from [6]) to the **B**aseline features [4] for different combination of loss weights.

		DEV		TEST	
Input	λ_R	WER	BLEU-4	WER	BLEU-4
Baseline	40.0	**32.53**	**20.32**	**32.50**	**20.64**
B + **K**	1.0	37.02	18.70	36.60	19.17
	20.0	34.67	18.98	34.89	18.59
	40.0	39.34	19.42	39.14	19.29
B + **H**	1.0	39.34	19.62	39.73	18.68
	20.0	33.87	19.61	34.28	19.42
	40.0	47.32	18.51	46.04	18.48
B + **K** + **H**	1.0	39.93	19.56	39.77	19.26
	20.0	33.33	19.75	33.74	18.96
	40.0	38.00	18.18	37.47	18.96

Table 4. Early Fusion post projection: Training SLT [2] by adding multiple articulately cues (**K**eypoints coordinates, **H**idden features from [6]) to the **B**aseline features [4] for different combination of loss weights.

4.3. Early fusion with post projection

Table 4 shows experiments where the concatenation is done after the projection. This method is not promising and leads to poor results. We were not able to reach 20 BLEU-4 score on TEST set.

4.4. Late fusion

For late fusion, we first train some weak learners. Table 5 shows the results for model having as input either only the **K**eypoints coordinates or the **H**idden features. We can notice that the pose estimation gives little cues compared to the baseline features as the performances are far lower. We observe that setting higher the recognition loss weight has more impact on this type of features. It makes sense since DOPE has never seen sign video before, contrarily to the baseline features obtained through [4].

Input	λ_R	DEV	TEST
Baseline	40.0	**20.64**	**20.64**
K	1.0	10.21	10.21
	20.0	11.82	11.49
	40.0	11.84	**12.04**
H	1.0	13.50	13.11
	20.0	16.20	15.17
	40.0	15.51	**15.58**

Table 5. Baseline for late fusion: Training SLT [2] with only dope outputs [6] as input, with different combination of loss weights.

Table 6 shows the resulting when ensembling the weak learners with the baseline model at the decoder level. We see that in the specific setup with **B** + **K** as input feature, $\lambda_R = 40$, we achieve our best result with 21.51 BLEU-4 score on TEST set.

		DEV		TEST	
Input	λ_R	WER	BLEU-4	WER	BLEU-4
Baseline	40.0	**32.53**	20.32	**32.50**	20.64
B + **K**	1.0	54.68	19.49	53.60	19.01
	20.0	46.25	19.36	46.58	19.23
	40.0	47.56	19.63	47.45	19.81
B + **H**	1.0	52.60	21.08	51.02	20.05
	20.0	51.19	20.42	50.72	20.15
	40.0	48.76	**20.84**	47.90	**21.51**
B + **K** + **H**	1.0	61.01	20.30	59.29	19.73
	20.0	58.47	20.35	57.38	20.11
	40.0	59.33	20.61	58.23	21.24

Table 6. Late Fusion: Predicting SLT [2] by averaging the class probabilities between different small models.

5. Conclusion

We notice that the baseline model described in [2] provide already good performances. Overall, adding DOPE [6] features does not give much improvement except in some specific cases.

As future work, we could test it on the original dataset [1] (not compressed) to have better DOPE keypoints detections, to improve the subsequent SLT task.

References

[1] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 4

[2] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation, 2020. 1, 2, 3, 4

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 2

[4] Oscar Koller, Necati Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 04 2019. 3, 4

[5] Necati Cihan Camgoz and Oscar Koller and Simon Hadfield and Richard Bowden. Multi-channel Transformers for Multi-articulatory Sign Language Translation, 2020. 1

[6] Weinzaepfel, Philippe and Bregier, Romain and Combaluzier, Hadrien and Leroy, Vincent and Rogez, Gregory. DOPE: Distillation Of Part Experts for whole-body 3D pose estimation in the wild. In *ECCV*, 2020. 1, 2, 3, 4