MATHÉMATIQUES VISION APPRENTISSAGE

école normale supérieure paris-saclay

université PARIS-SACLAY

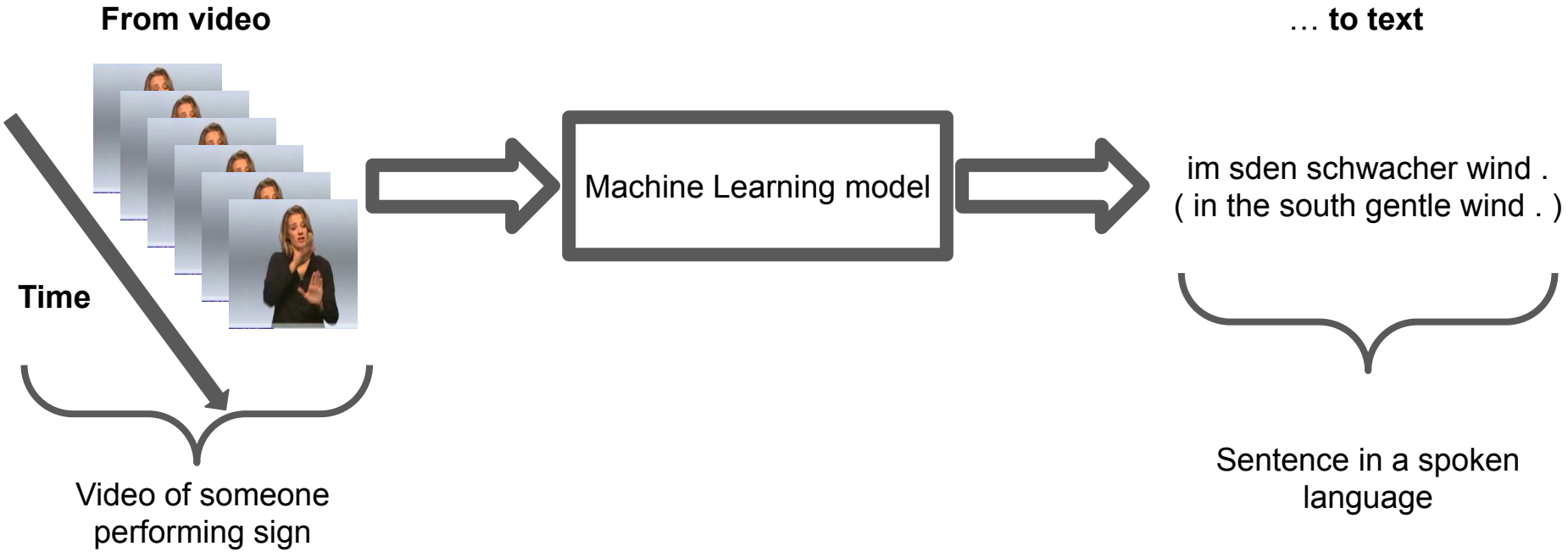
Final Project Presentation

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

Vincent LIU liuvincent25@gmail.com

Supervised by Gül Varol

Sign Language Translation (SLT)



Goal: Produce a model able to help the Deaf daily life communication.

Video taken from [5] and output sentence taken from [1].

Why is it difficult ?

Numerous challenges, including:

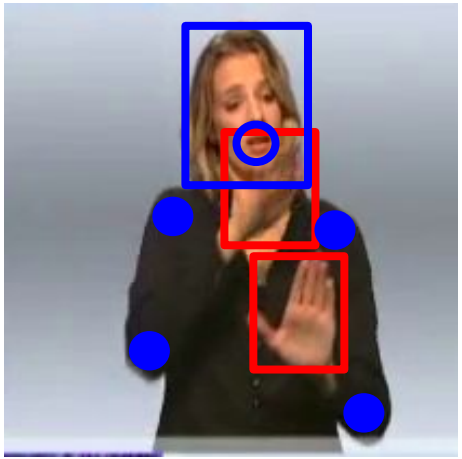
- Memory challenges (videos are big).
- Annotations are expensive (it requires expert knowledge).
- Task is also difficult: Sign2Text is not an one to one mapping [1].
- Multiple channels to convey information [3].

Example of:

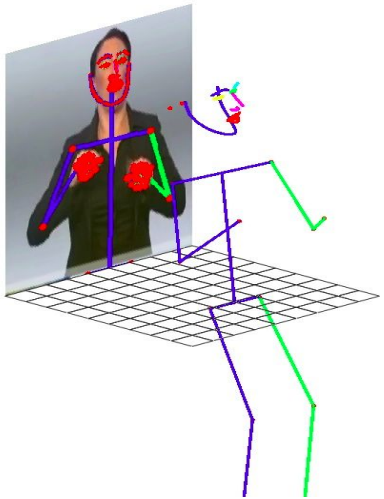
Manual channels

Non-manual channels

Also temporal dependencies



Methodology: DOPE [2] on compressed PHOENIX 14T dataset [5]



Extract for each frame:

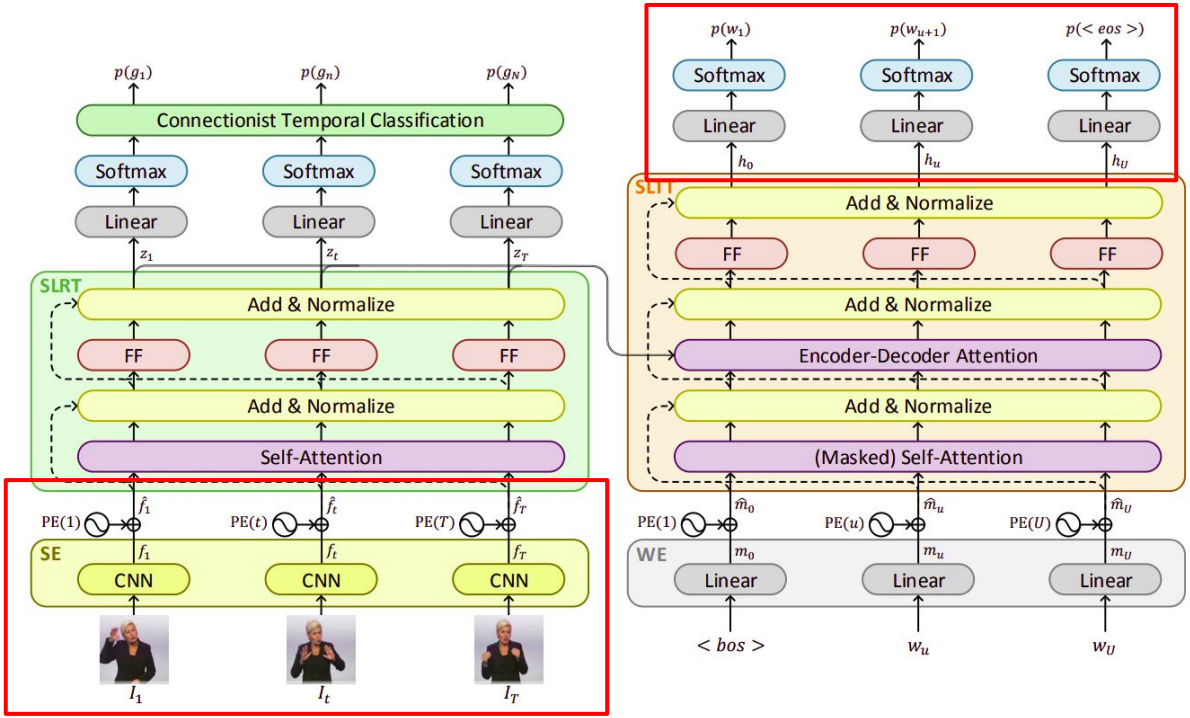
- 2D and 3D keypoints coordinates of face, hands and body.
- Hidden feature representation (before ROI Pooling), and apply global average pooling to obtain 1024 dimensional vectors.

The process took ~10 hours on Tesla V100 GPU.

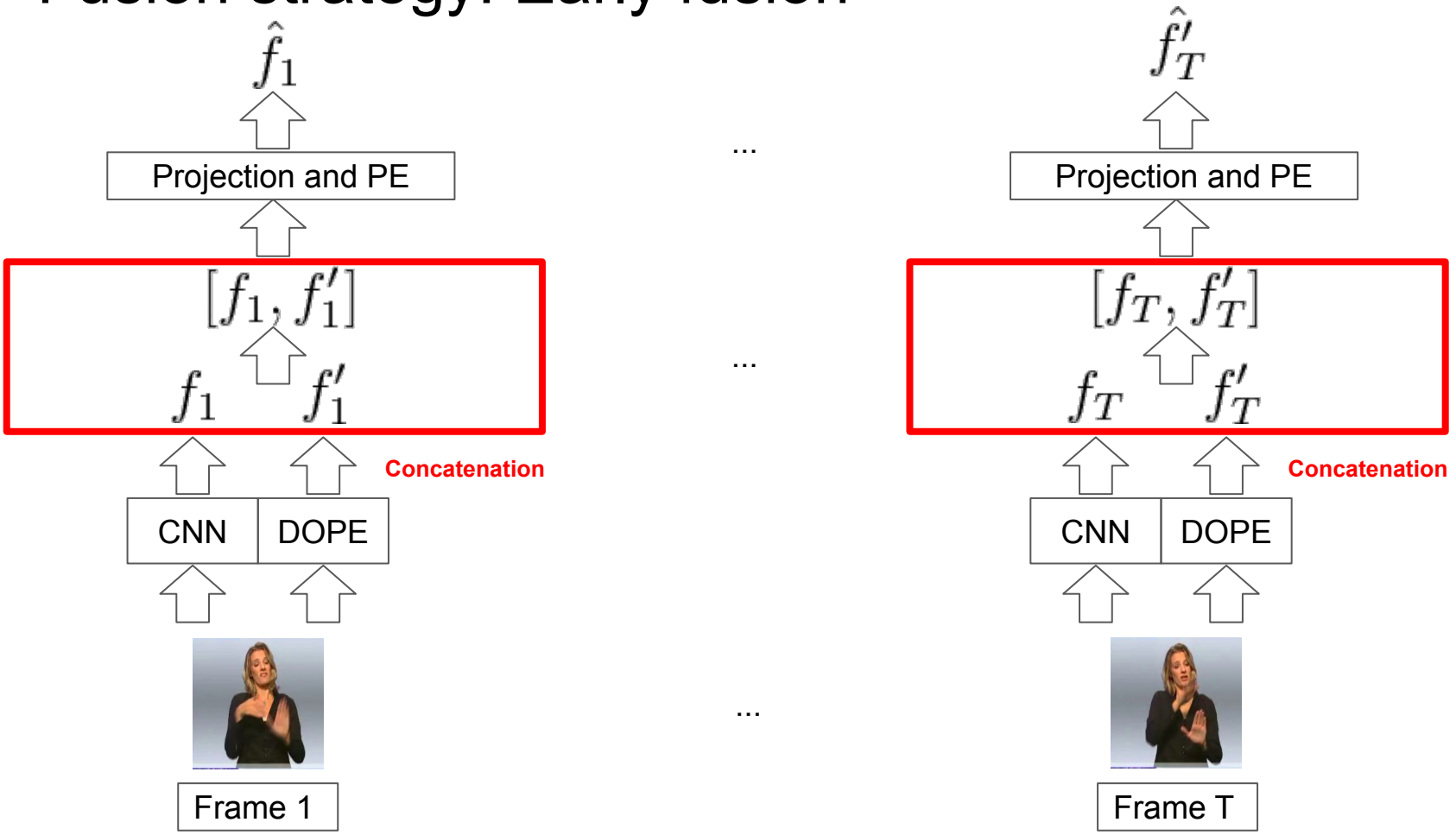
High error rate on hands detection (for some videos <10% success).

Fusion strategy on Sign Language Transformers [1]

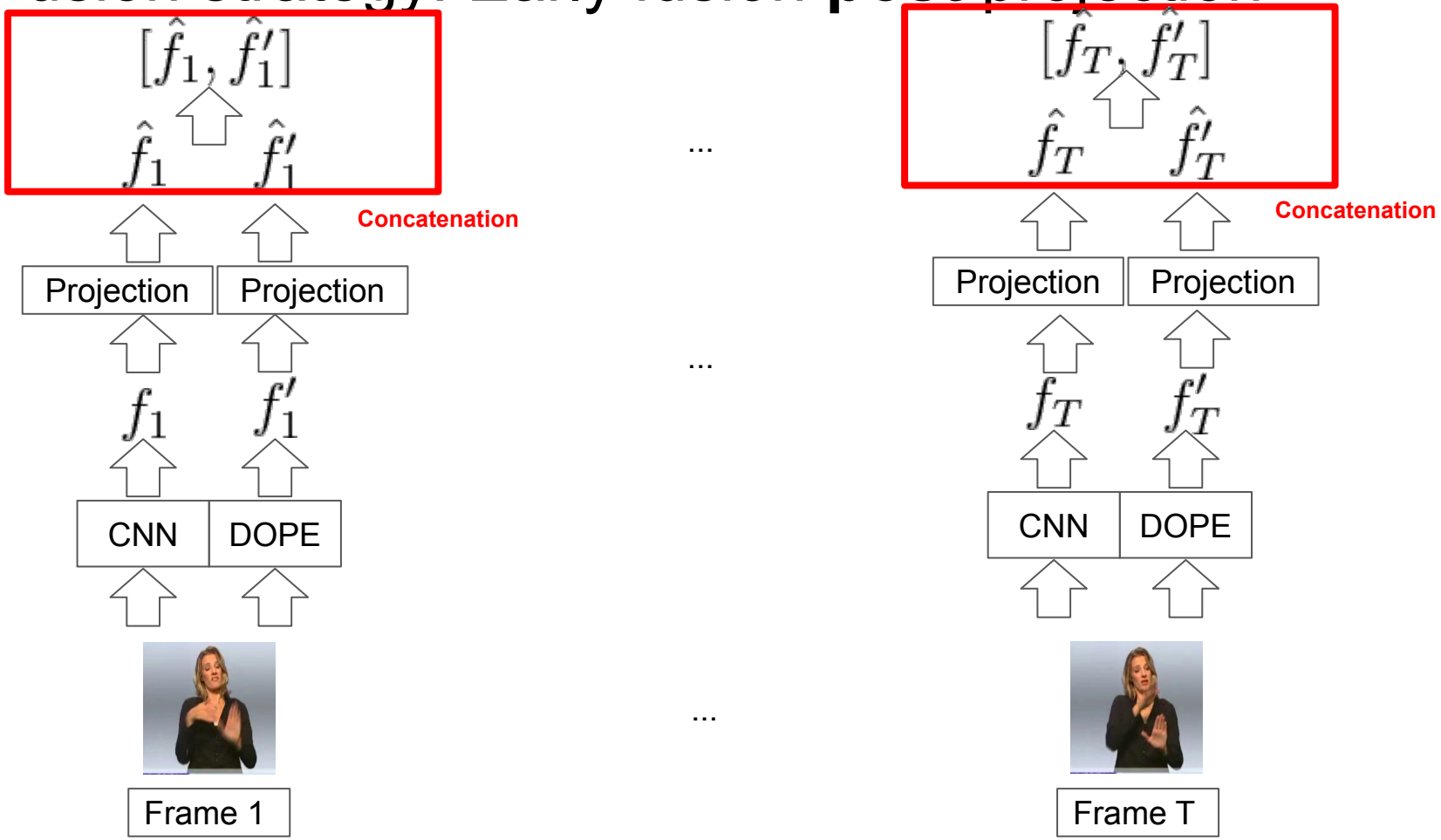
Figure taken from [1].



Fusion strategy: Early fusion



Fusion strategy: Early fusion **post** projection



Fusion strategy: Late fusion

Train N models independently for each input type.

Average output probabilities at test time.

$$w_i^* = argmax \quad \lambda_1 P_1(w_i) + \lambda_2 P_2(w_i) + \cdots + \lambda_N P_N(w_i)$$

$$\text{with } \lambda_1 + \ldots \lambda_N = 1$$

Quantitative results: reproduce [1]

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 [1]. Training SLT to jointly learn recognition and translation with different weight on recognition loss with baseline features obtained from [4].

Observations

- Sign2(Gloss+Text) is better than Sign2Text (First row has lower result).
- High λ_R can help but baseline $\lambda_R = 1$ and $\lambda_T = 1$ is a good start because the 2D CNN has already seen sign videos according to [1].
- Results lower than the original paper results.

Early Fusion with single cue

Input	λ_R	2D	3D
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 [1] 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.

Observations

- Body pose estimation may help the model.
- 3D scaled features was better (A little bit unfair since 2D raw features were not scaled).

Early Fusion with multiples cues

Input	λ_R	DEV WER	DEV BLEU-4	TEST WER	TEST BLEU-4
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	19.84
	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 by adding multiple articulately cues (**K**eypoints coordinates, **H**idden features from [4]) to the **B**aseline features [2] for different combination of loss weights.

Observations

- B + K + H with $\lambda_R = 20$ good results on **both** dev and test set.
- K = 3D Keypoints concatenated, H = Hidden features

Early Fusion **post** projection

Input	λ_R	DEV WER	DEV BLEU-4	TEST WER	TEST BLEU-4
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 [1] by adding multiple articulately cues (**K**eypoints coordinates, **H**idden features from [2]) to the **B**aseline features [4] for different combination of loss weights.

Observations

It did not work, it could not exceed 20 BLEU-4 Score.

Late Fusion: train weak models

First, train a single model for each input type.

Input	λ_R	DEV	TEST
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 [1] with only dope outputs [2] as input, with different combination of loss weights.

Observations

- Since DOPE has not seen sign video before, it makes sense to add **more loss recognition** weight when training with only DOPE input features.

Late Fusion: combine models

At test time, combine the small models.

Input	λ_R	DEV WER	DEV BLEU-4	TEST WER	TEST BLEU-4
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 [1] by averaging the class probabilities between different small models.

Observations

- Slight better result with $\lambda_R = 40$ and 0.8 B and 0.2 H weight sum on output.
- We didn't have time to tune the weight sum coefficients.

Conclusion

Baseline gives already strong results.

Adding DOPE features does not give much improvement overall except in specific cases.

Further possible work:

- Need to test it on original dataset (not compressed) to have better DOPE keypoints detections, to improve subsequent SLT task.

References

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

[2] 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.

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

[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.

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

Thank you!

Table of contents

Sign Language Transformers

Reported 21.80 BLEU-4 score on test set.

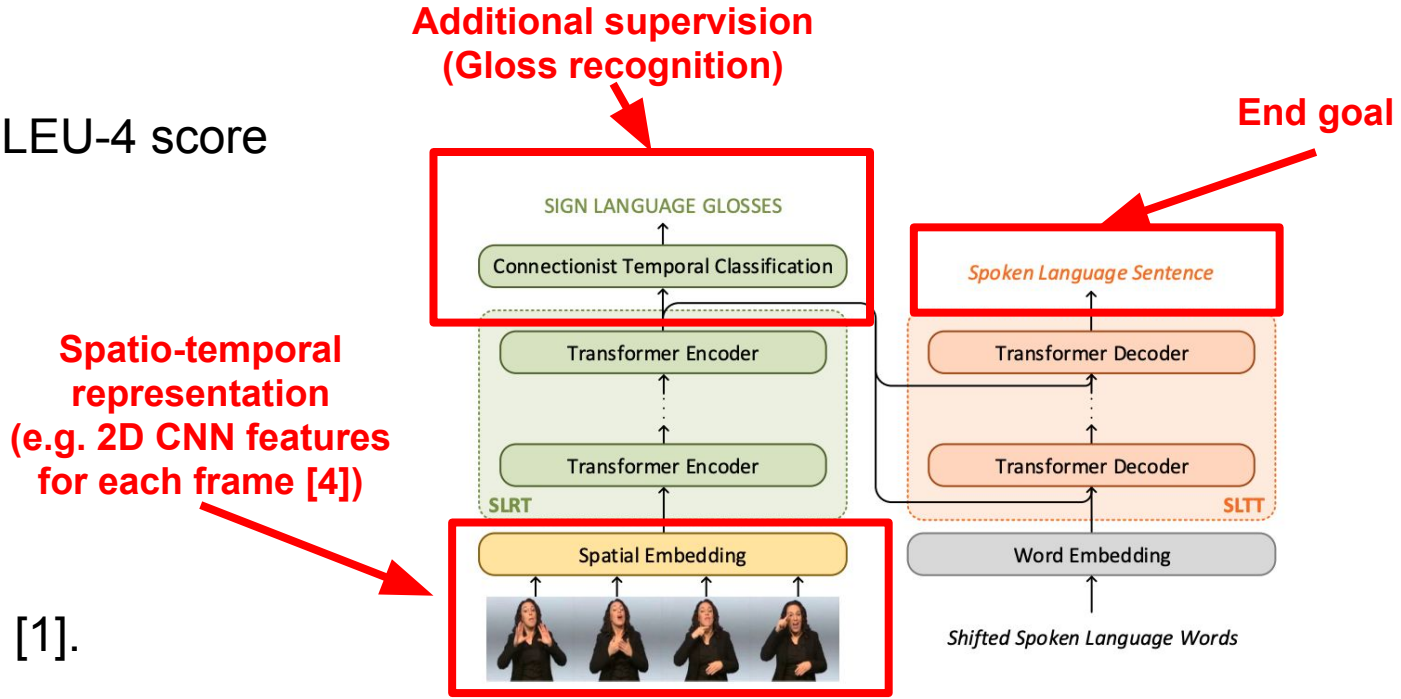


Figure taken from [1].