Neural Sign Language Processing With Language Models

Ayush Zenith[1] and Samuel Ji[2]

Abstract— **This paper presents the research on neural sign language translation with transformer learning. We built upon the current state-of-the-art sign language translation architecture and utilize language models to get a robuster model.**

I. Justification

For those who are hard of hearing (or the deaf community in general), sign language is the primary means of communication. Sign language has all the characteristics of a natural language concept yet there doesn't seem to be any major effort from academics to model and study sign languages. With the advanced technology, we believe it is possible to build models that can capture and translate sign gestures into their counterparts in order to ease the daily lives of individuals of this community.

II. Introduction

In this paper, we are presenting our attempt to improve the Joint End-to-end Sign Language Recognition and Translation architecture introduced by Camgoz et.al.[2], by utilizing the concept of the transfer learning, which is first introduced by De Coster et.al.[4] We are going to show the result model with this approach is more robust compared to the one that Camgoz et.al. trained.

We are using the same data set that Camgoz et.al. was using, which is the PHOENIX14T. The data set is in German, and is consists of the signing video frames, their corresponding GLOSS representation and their verbal counterparts.

The high-level approach we are using for improving the transformer is to use retrained language models as the encoder. Specifically, we used BERT-base-German-cased and BERT-base-multilingual-uncased which includes German, as our goal is to translate from German sign to German. Our result shows that both the model with BERT-base-German-cased and the model with BERT-base-German-cased as encoder outperforms the model trained by Camgoz et.al., yet they fall behind the model trained by De Coster et.al. with the Bert-based-uncased. One interesting observation is the Bert-base-multilingual-uncased encoder results in a better training result than the one with the language model of the targeted language.

III. Related work

The sign Language Transformer introduced by Camgoz et.al.[2] is the current state-of-the-art sign language translation architecture. It solves the long term dependency issues other architectures face, as gloss usually has a shorter length than frames, and enable the model to jointly learn to recognize and translate sign video sequences into sing glosses and spoken language in an end-to-end manner, by taking account both the gloss loss and word loss. However the lack of labeled data may prevent the model has a better performance.

De Coster. et.al.[4] suggested to apply transfer learning to the original architecture by integrated the language model to the encoder (and the decoder) and evaluate the performance with BERT2RAND and BERT2BERT. The theoretical basis for this approach is that transfer learning from high-resource to low-resource language pairs result in better translation performance for low- resource language pairs [5], and the retrained models such as BERT can be adapted as encoders (and decoders) to improve machine translation performance[6].

De Coster. et.al. designed three experiment to improved the original model and concluded BERT2RND, which refers to the model that has the FPT in the encoder and the trained-from-scratch decoder performs the best. They believe it's because the decoder benefits more from having more degrees of freedom than the encoder. Among different set of fin-tuning value on hyper-parameters such as learning rate, number of layers, loss weights, decoding beam size and beam alpha, they derived the best BERT2RND model with the highest BLEU-4 score of 22.25. They suggested for the future work to investigate if a smaller translation models may obtain better performance. They also proposed to experiment on the use of the language models priors to regularize the SLT model.

IV. Data

While datasets like How2Sign Amanda et. al., 2021[3], with 16,000 signs spread over 35,000 sentences and over 79 hours of videos, exist; training translation models on them can be very computationally taxing and overwhelming. Which is why we opted to use the PHOENIX14T dataset which was introduced by Camgoz et al. 2020[2]. According to our knowledge it is the only dataset that has been used to study Continuous Sign Language translation amongst the Sign Language Processing community thus far. The dataset was constructed over a period of three years (2009 - 2011) with the daily news and weather forecast airings of the

[1]Ayush Zenith is an undergraduate student with the Khoury school of Computer Science, Northeastern University, Boston, MA zenith.a@northeastern.edu

[2]Samuel Ji is an undergraduate student with the Khoury college of Computer Science, Northeastern University, Boston, MA ji.xian@northeastern.edu

German public tv-station PHOENIX which featured sign language interpretation that have been recorded and the weather forecasts of a subset of 386 editions have been transcribed using gloss notation. Automatic speech recognition with manual cleaning was used to transcribe the original German speech in the newscast. The signing is recorded by a color camera placed in front of the sign language interpreters and the interpreters wear dark clothes in front of an artificial gray background with color transition. All recorded videos are at 25 frames per second and the size of the frames is 210 by 260 pixels with each frame showing only the interpreter. The dataset is divided into a training set, dev set, and test set. With each set having 7000+, 600+, and 500+ entries respectively.

The dataset is quite small for a translation task which was one of our primary inspirations that lead us to research using transfer learning to improve learning on small datasets. Another shortcoming of the dataset is that it has a very limited vocabulary limited to weather forecasting and since none of our pretrained models that we use have been fine-tuned with a weather limited vocabulary it might not perform very well together and models trained on this dataset might not perform well with data that isn't relevant to data.

V. Method

We first reproduced the baseline of Camgoz et al. 2020[2] using the code-base provided by the authors. We then reproduced a couple baselines of De Coster et al. 2021[4] using the code-base provided by the authors.

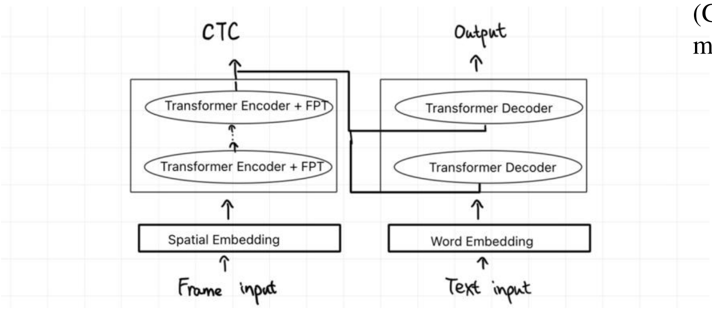

Fig. 1. Architecture used for FPT Models and only Transformer Models

In the article by De Coster et al. 2021 we find that they tune hyper-parameters including learning rate, number of layers, loss weights, decoding beam size, beam alpha, and study which layers are best frozen and which layers must be fine-tuned when using a pre-trained model. They also use a base-uncased-bert model as well as the mBart and mBart-50 model to help improve performance as the data-set being used is very tiny for the proposed translation task.

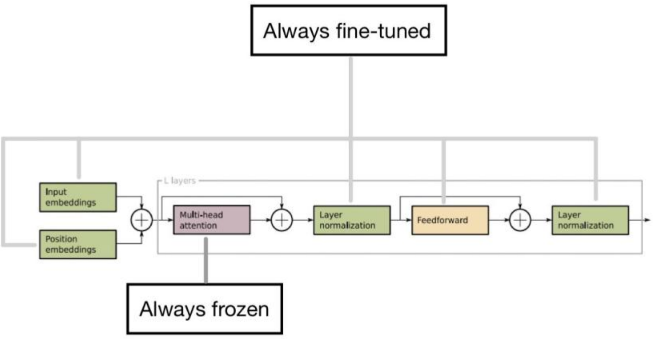

Fig. 2. Fine-tuned and frozen layers of our models (De Coster et al. 2021[4])

We use the findings from De Coster et al. 2021 to choose our hyper-parameters and to freeze the self-attention while fine-tuning other layers as the training continues on with our selection of pre-trained models. We built on their code-base to pre-train a continuous sign language translation model using a pre-trained bert-base-german-cased as well as a pre-trained bert-base-multilingual-uncased. We test by using the pre-trained models as encoders with a decoder trained from scratch (BERT2RND) and also by using the pre-trained models as encoders with a a cross-attention module added to the BERT model integrated in the decoder (BERT2BERT). We then evaluated the performance of the different architectures of BERT2RND and BERT2BERT and also compared how results may vary by using a base-bert model with the most parameters, a BERT model trained on the target language (German), and a BERT model trained on multiple languages making it a multilingual model.

VI. Evaluation

.

Model	BERT Variant	Test BLEU-4	Training Time*
Sign2(Gloss+Text) (Camgoz et al. 2020)[2]	N/A	18.32	2 hours and 41 mins
Bert2Rnd - Fine tuned	*bert-base-uncased (De Coster et al. 2021)*[4]	**19.44**	1 hour and 46 mins
	bert-base-german-cased	18.93	53 mins
	bert-base-multilingual-uncased	19.25	2 hours and 38 mins
Bert2Bert - Fine tuned	*bert-base-uncased (De Coster et al. 2021)*	18.65	1 hour and 9 mins
	bert-base-german-cased	17.64	**44 mins**

TABLE I
Results

We will be using BLEU-4 as the metric in order to validate the performance of the model. While we collect other metrics like WER for checking the performance on recognition, other BLEU metrics, ROUGE metrics, etc we have decided to dictate learning and final performance with best achieved BLEU-4 score.

As seen by our results table the BERT2RND models outperform the BERT2BERT models which outperformed the baseline of Camgoz et al. 2020 as we had expected. The best results were yielded when we used the bert-base-uncased model that was used by the De Coster et al. 2021 baseline. It seems to perform significantly better than the baseline of Camgoz et al. 2020 but also outperforms our models. We found it interesting that the bert-base-uncased model performed better than the bert-base-multilingual-uncased model which outperformed the bert-base-german-cased model.

This is interesting as the translation task is all in german yet the english model seems to outperform all after which the multilingual model follows and the worst one of the lot being the german model, which still outperforms the baseline of Camgoz et al. 2020 but seems to still be far from the english model. The cause of this might stem back to our dataset being very domain limited and since the bert-base-german-cased has the fewest parameters in comparison with the bert-base-uncased model it might not be very useful in this domain specific translation task. We also found it interesting that the bert-base-multilingual-uncased model performed quite similarly to the bert-base-uncased model as in De Coster et al. 2021's findings they found that multilingual models like the mBART and mBART-50 performed very poorly and often did even worse than the baseline of Camgoz et al. 2020. We believe that our multilingual model might have performed well in comparison because a large part of our multilingual model's training data resembled the bert-base-uncased model's training making it similar to the well performing bert-base-uncased model.

An interesting discovery we made was how computationally efficiently using Transfer Learning could be. It reached similar or better results within a fraction of the time when comparing training time with transformers that were trained from scratch. The training time of different models was also quite interesting, like how the multilingual model took a rather long time in comparison with the German model.

VII. Conclusion

In conclusion, while our results weren't not what we expected, we still do believe that if the translation were to take place in a non-domain specific challenge the use of the bert-base-german-cased model would outperform the bert-base-uncased model. In the future we would like to attempt to test this theory out by using a larger dataset that isn't domain specific. We would also convert the video to pose to eliminate noise and increase the recognition accuracy. Experimenting with a variety of other BERT variants like distilbert-base-uncased, distilbert-base-german-cased, and roberta-base or non BERT based models could also lead to interesting results that might lead to the development of better continuous sign language translation systems.

REFERENCES

[1] N. C. Camgoz, S. Hadfield, O. Koller and R. Bowden, ”SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition,” 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3075-3084, doi: 10.1109/ICCV.2017.332.

[2] Camgöz, N.C., Koller, O., Hadfield, S., Bowden, R. (2020). Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10020-10030.

[3] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i-Nieto, “How2Sign: A large-scale multimodal dataset for continuous American sign language,” arXiv.org, 01-Apr-2021. [Online]. Available: https://arxiv.org/abs/2008.08143. [Accessed: 13-Oct-2022].

[4] M. D. Coster, K. D’Oosterlinck, M. Pizurica, P. Rabaey, S. Verlinden, M. V. Herreweghe, and J. Dambre, “Frozen Pretrained Transformers for Neural Sign Language Translation,” 2021, Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), pages 88–97, Virtual.

[5] B. Zoph, D. Yuret, J. May and K. Knight, “Transfer Learning for Low-Resource Neural Machine Translation,” arXiv.org, 08-Apr-2016. [Online]. Available: https://arxiv.org/abs/1604.02201.

[6] S. Rothe, S. Narayan and A. Severyn, “Leveraging Pre-trained Checkpoints for Sequence Generation Tasks,” arXiv.org, 29-jul-2019. [Online]. Available: https://doi.org/10.1162%2Ftacl_a_00313.