Multi-channel Transformers for Multi-articulatory Sign Language Translation: Supplementary Material

Necati Cihan Camgoz[1], Oscar Koller[2], Simon Hadfield[1], and Richard Bowden[1]

[1] CVSSP, University of Surrey, UK, {n.camgoz, s.hadfield, r.bowden}@surrey.ac.uk,
[2] Microsoft, Munich, Germany, oscar.koller@microsoft.com

In this supplementary material, we report our quantitative experiment results for finding the best channel feature and word embedding setup. We also share qualitative translation samples produced by our best performing Multi-channel transformer model.

1 Channel Feature and Word Embeddings

In this preliminary set of experiments we investigate different ways to embed Convolutional Neural Network (CNN) based channel features and one-hot word vectors. Machine translation orientated transformer implementations either use pretrained word embeddings or train a linear projection layer from scratch. Although not stated in the original paper [3], the official transformer implementation also utilizes *embedding scaling*, where the projected word representations are multiplied by a constant which is the square root of the hidden size[3].

Compared to the one-hot vectors which have a constant scale between $[0-1]$, CNN features can have an arbitrary scale. To see how important the input scale is and to examine the effects of the different embedding setups, we initially trained translation networks that only used hand channel features as input. Each network has two layers, with a hidden size of 64 and 128 position-wise feed forward units.

As can been in the first row of Table 1, the translation performance degrades drastically when we apply the commonly used embedding scaling on either of the embeddings. We have experimentally found that transformer networks are extremely sensitive to input scale and this is substantiated by these results. Thus, in our next set of experiments we investigate ways to normalize inputs to control their scale. To do so, we utilize batch normalization [1] and soft-sign activation [2]. While batch normalization scales the inputs between $[-3, 3]$ it has also been shown to improve convergence rate. On the other hand, soft-sign activation scales the inputs between $[-1, 1]$ while also enhancing the representation capability of the embedding due to its non-linear nature.

Individually, both batch norm and soft-sign significantly improve the translation performance when applied to the projected CNN features (see second and third rows of Table 1). We then investigate their combined use on both CNN features and word embeddings. Although there were several comparable setups, we

[3] github.com/tensorflow/models/blob/master/official/nlp/transformer/

concur that the joint application of batch norm and soft-sign only on the CNN features yield the most stable and balanced performance in terms of development and test set for BLEU-4 and ROUGE scores. We believe this is due to the already scaled nature of the word embeddings and the additional stability and non-linearity introduced by applying soft-sign and batch norm to the projected CNN features. Therefore, we utilize this embedding setup for our experiments in the main manuscript.

Table 1. Effects of using different embedding setups on hand channel features to spoken language translation performance.

						Dev Set				Test Set			
Feature Embedding			Word Embedding			BLEU-4		ROUGE		BLEU-4		ROUGE	
Scaling	BatchNorm	SoftSign	Scaling	Batch Norm	Soft-Sign	Best	mean ± std	-	mean ± std	-	mean ± std	-	mean ± std
✗	-	-	✗	-	-	**15.46**	**15.08** ± 0.25	**40.57**	**39.39** ± 0.55	**15.98**	**15.23** ± 0.54	**39.97**	**39.44** ± 0.80
✓	-	-	✗	-	-	13.96	13.39 ± 0.36	37.51	37.07 ± 0.33	14.29	13.77 ± 0.47	37.91	37.54 ± 0.60
✗	-	-	✓	-	-	14.54	14.20 ± 0.26	37.86	38.02 ± 0.70	14.80	14.58 ± 0.46	38.35	38.38 ± 0.72
✓	-	-	✓	-	-	13.52	12.88 ± 0.36	36.93	36.14 ± 0.66	13.99	13.29 ± 0.53	37.72	36.58 ± 0.81
✗	✓	-	✗	✗	-	**16.35**	**15.64** ± 0.46	**41.30**	**40.41** ± 0.55	**16.09**	**15.60** ± 0.50	**40.89**	**40.15** ± 0.70
✗	✗	-	✗	✓	-	14.49	14.07 ± 0.25	37.96	37.86 ± 0.58	14.72	14.33 ± 0.38	37.49	37.53 ± 0.43
✗	✓	-	✗	✓	-	15.17	14.69 ± 0.30	39.92	39.10 ± 0.56	15.50	14.99 ± 0.64	39.83	39.26 ± 0.81
✗	-	✓	✗	-	✗	**16.40**	**15.74** ± 0.55	**41.90**	**40.73** ± 0.63	14.95	**15.63** ± 0.48	39.31	39.93 ± 0.57
✗	-	✗	✗	-	✓	15.75	15.10 ± 0.35	39.72	39.47 ± 0.46	**16.33**	15.41 ± 0.55	40.74	39.76 ± 0.67
✗	-	✓	✗	-	✓	15.98	15.50 ± 0.35	41.02	40.24 ± 0.51	15.64	15.49 ± 0.48	**40.79**	**40.02** ± 0.59
✗	✓	✓	✗	✗	✗	**16.44**	**15.70** ± 0.41	40.79	40.45 ± 0.64	**16.18**	15.63 ± 0.65	**40.62**	40.07 ± 0.80
✗	✗	✗	✗	✓	✓	15.15	14.18 ± 0.42	39.30	37.78 ± 0.77	15.14	14.30 ± 0.53	38.41	37.53 ± 0.82
✗	✓	✓	✗	✓	✓	15.59	15.04 ± 0.33	39.83	39.40 ± 0.46	14.85	14.99 ± 0.51	39.26	39.30 ± 0.65
✗	✓	✓	✗	✗	✓	15.98	15.62 ± 0.21	**41.63**	**40.63** ± 0.62	15.97	**15.65** ± 0.49	40.54	**40.24** ± 0.74
✗	✗	✓	✗	✓	✓	15.09	14.70 ± 0.26	39.36	38.95 ± 0.25	14.70	14.80 ± 0.49	39.38	38.87 ± 0.67

2 Qualitative Examples

In this section we share translation examples generated by our best performing model. As the ground truth spoken language annotations of the RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) dataset are in German, we share both the original German translations and their equivalent word-by-word translations in English. As can be seen in Table 2, we also categorize the results into three categories, namely *Good*, *Mediocre* and *Poor* translations, to give further insight to the reader on the limitations of the current approach.

We categorize translations as *Good* or *Mediocre* when the produced sentences convey the same or similar information as the reference sentences. These examples follow the standard grammar with few exceptions. We classify translations as *Poor* when the model fails to understand and translate the conveyed information in sign videos. Most of these examples contain repetitions. In some cases, the model is not able distinguish some sign glosses from another, such as named entities like locations or numbers which occur in limited contexts in the training data. One way to address this issue might be to utilize pretrained spoken language models to improve the produced translations.

Table 2. Spoken language translations produced by our best Multi-Channel Transformer model.

Good Translations:	
Reference:	und nun die wettervorhersage für morgen dienstag den ersten februar .
	(and now the weather forecast for tomorrow tuesday the first of february .)
Ours:	und nun die wettervorhersage für morgen dienstag den ersten februar .
	(and now the weather forecast for tomorrow tuesday the first of february .)
Reference:	der sorgt wieder für wolken die regen im bergland auch schnee bringen .
	(it provides clouds again and the rain in the mountains also brings snow .)
Ours:	im übrigen land fällt gebietsweise regen im bergland auch schnee .
	(in the rest of the country there is rain in the mountains and snow in some areas.)
Reference:	die neue woche beginnt wechselhaft und kühler .
	(the next week starts variable and colder .)
Ours:	auch am montag wechselhaft und deutlich kühler .
	(also on monday variable and significantly colder .)
Mediocre Translations:	
Reference:	ab sonntag wird es wieder milder dabei gibt es viele wolken zeitweise fällt regen im nordwesten windig .
	(from sunday on it will be more mild with many clouds partly rain in the northwest windy .)
Ours:	am sonntag mehr wolken als sonne hier und da regen im westen ist es windig .
	(on sunday more clouds than sun from time to time rain in the west windy .)
Reference:	im osten und südosten auch schnee oder schneeregen .
	(in the east and south-east also snow or sleet .)
Ours:	im südosten schnee oder schnee .
	(snow or snow in the south-east .)
Reference:	westlich des rheins und im nordosten bleibt es meist trocken .
	(west of the rhine and in the northeast it remains mostly dry .)
Ours:	im westen und südwesten bleibt es noch im nordosten trocken .
	(in the west and southwest it remains still dry in the northeast .)
Poor Translations:	
Reference:	deutschland liegt morgen unter hochdruckeinfluss der die wolken weitgehend vertreibt .
	(germany will be under the influence of high pressure tomorrow which will largely dispel the clouds .)
Ours:	deutschland liegt morgen über deutschland nach deutschland .
	(germany is tomorrow over germany to germany .)
Reference:	am freitag insgesamt viele wolken die regen bringen .
	(on friday overall many clouds bringing rain .)
Ours:	am freitag gibt es am freitag viele wolken .
	(on friday there are on friday many clouds)
Reference:	dazu weht ein starker wind vor allen dingen wieder über vorpommern aus südost .
	(in addition a strong wind blows before all things again over vorpommern from southeast .)
Ours:	es weht ein kräftiger nordostwind .
	(a strong north-easterly wind is blowing .)

References

1. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: Proceedings of the International Conference on Machine Learning (ICML) (2015)
2. Nwankpa, C., Ijomah, W., Gachagan, A., Marshall, S.: Activation Functions: Comparison of trends in practice and research for deep learning. arXiv:1811.03378 (2018)
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is All You Need. In: Proceedings of the Advances in Neural Information Processing Systems (NIPS) (2017)