Cross+Self-Attention for Transformer Models

Stephan Peitz Sarthak Garg Udhay Nallasamy Matthias Paulik
Apple Inc.
{sarthak_garg, speitz, udhay, mpaulik}@apple.com

Recently proposed simplifications of the Transformer model (Vaswani et al., 2017) suggest merging the encoder and the decoder into a single joint model (He et al., 2018; Fonollosa et al., 2019). This type of model works without cross-attention and solely relies on the self-attention mechanism which spans the input and output sequences. Furthermore, this approach reduces the amount of trainable parameters by up to 50% which speeds up training and results in a smaller model size. However, this reduction seems to lead to a degradation of translation quality which can be equalized by increasing the depth of the Transformer model (He et al., 2018).

In this work, we argue that the degradation is not solely caused by the reduction of parameters. The approach introduced by He et al. (2018); Fonollosa et al. (2019) takes the intermediate representation of the input sequence rather than the final one as input to the joint self-attention. Furthermore, while the proposed joint model shares all encoder and decoder parameters, we show that some parameters are shareable and others should not be shared. We propose keeping the encoder-decoder architecture and merging the cross-attention and the decoder self-attention module into a cross+self-attention module. This module is basically a multi-head self-attention taking the concatenation of encoded representation of the input and output sequences as key-value pair and query.

Our experimental results (Table 1) with Transformer `big` on the WMT18 English→German news translation task show that replacing cross-attention and decoder self-attention by cross+self-attention does not impact the translation quality while reducing the amount of parameters and improving inference time (on a single GPU) by 12%. We further observed that taking the final encoder state as input for the cross+self-attention module (final) is crucial. This approach performs better than attending to the encoder state from the same layer (layer-wise) which simulates the joint model suggested in (Fonollosa et al., 2019). In addition, we show that sharing the (cross+) self-attention between the encoder and decoder further reduces the amount of parameters while slightly hurting the translation quality. Sharing the fully-connected feed-forward network and the layer normalization module, however, severely impacts the translation quality. Finally, we compare our approach with a Transformer model using an average attention network (Zhang et al., 2018) which provides a similar speed-up but degrades the translation quality.

Table 1: Results on the WMT18 English→German task. Bleu (Papineni et al., 2002) is computed with `sacreBLEU` (Post, 2018). †`https://github.com/pytorch/fairseq/issues/506#issuecomment-464411433`

	Cross+Self-			newstest2014	
Model	Attention	Shared Layers	Parameters (M)	Bleu [%]	tokens/sec.
(Edunov et al., 2018)†	-	all embeddings	213	29.0	-
Our baseline	-	all embeddings	213	29.0	113
+average attention			200	28.6	126
w/o cross-attention	layer-wise	all embeddings	188	28.2	**129**
		+ self-attention	162	28.0	
		+ feed-forward	112	27.5	
		+ layer-norm	112	27.5	
	final	all embeddings	**188**	**29.0**	
		+ self-attention	162	28.5	
		+ feed-forward	112	27.4	
		+ layer-norm	112	27.7	

References

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium.

José A. R. Fonollosa, Noe Casas, and Marta R. Costa-jussà. 2019. Joint source-target self attention with locality constraints. *arXiv e-prints*.

Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. In *Advances in Neural Information Processing Systems*, pages 7944–7954, Montréal, Canada.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Conference on Statistical Machine Translation*, pages 186–191, Belgium, Brussels.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 1–11, Long Beach, CA, USA.

Biao Zhang, Deyi Xiong, and Jinsong Su. 2018. Accelerating neural transformer via an average attention network. In *Annual Meeting of the Association for Computational Linguistics*, pages 1789–1798, Melbourne, Australia.