Rethinking the Design of Sequence-to-Sequence Models for Efficient Machine Translation

Maha Elbayad[†]
LIG - Université Grenoble Alpes, France
Inria - Grenoble, France
maha.elbayad@inria.fr

In recent years, deep learning has enabled impressive achievements in Machine Translation. Neural Machine Translation (NMT) relies on training deep neural networks with large number of parameters on vast amounts of parallel data to learn how to translate from one language to another. One crucial factor to the success of NMT is the design of new powerful and efficient architectures. State-of-the-art systems are encoder-decoder models (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) that first encode a source sequence into a set of feature vectors and then decode the target sequence conditioning on the source features. In this thesis we question the encoder-decoder paradigm and advocate for an intertwined encoding of the source and target so that the two sequences interact at increasing levels of abstraction. For this purpose, we introduce Pervasive Attention, an NMT model with a computational graph different from existing encoder-decoder models. In Pervasive attention, the source and the target communicate and interact throughout the encoding process towards abstract features. To this end, our NMT model uses two-dimensional convolutional neural networks to process a grid of features where every position represents an interaction between a target and a source tokens.

To tackle a different aspect of efficiency in NMT systems, we explore the challenging task of online (also called simultaneous) machine translation (Fügen et al., 2007; Mieno et al., 2015; Dalvi et al., 2018; Ma et al., 2019) where the source is read incrementally and the decoder is fed partial contexts so that the model can alternate between reading and writing. To improve the translation's delay in online NMT systems, we first setup a common framework for online sequence-to-sequence models that will allow us to train existing deterministic decoders that alternate between reading the source and writing the target in a predetermined fashion, and dynamic decoders that condition their decoding path on the current input. We first prove the effectiveness of the deterministic online decoders and their ability to perform well outside the delay range they were optimized for. We then adapt Pervasive Attention models for the task of online translation with both a deterministic and a dynamic decoding strategy.

We also address the resource-efficiency of encoder-decoder models, namely Transformer models (Vaswani et al., 2017), state-of-the-art in a wide range of NLP tasks (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Ng et al., 2019). Models based on the Transformer architecture can grow deep, accumulating billions of parameters. We posit that going deeper in a neural network is not required for all instances, and design depth-adaptive Transformer decoders. These decoders allow for anytime prediction and sample-adaptive halting mechanisms, to favor low cost predictions for low complexity instances, and save deeper predictions for complex scenarios.

Pervasive Attention models and our Online NMT framework are implemented on top of the Fairseq library (Ott et al., 2019) in our open-source code.[1]

Acknowledgements The author would like to thank her Ph.D. supervisors, Laurent Besacier and

[†] Now at Meta AI Research

[1] https://github.com/elbayadm/attn2d

Jakob Verbeeek, and her thesis examiners, Holger Schwenk and Hermann Ney. The work presented in this thesis has been supported by the grant ANR-11-LABX-0025-01 “LabEx PERSYVAL”. The work on Depth-adaptive Transformers was done during an internship at Facebook AI research hosted by Michael Auli.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. of EMNLP*.

Dalvi, Fahim, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. 2018. Incremental decoding and training methods for simultaneous translation in neural machine translation. In *Proc. of NAACL-HLT*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*.

Fügen, Christian, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine translation*.

Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proc. of ICML*.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *ArXiv preprint*.

Ma, Mingbo, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proc. of ACL*.

Mieno, Takashi, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Speed or accuracy? a study in evaluation of simultaneous speech translation. In *Proc. of INTERSPEECH*.

Ng, Nathan, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair’s wmt19 news translation task submission. In *Proc. of WMT*.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proc. of NAACL-HLT*.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proc. of NeurIPS*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. of NeurIPS*.