

CUNI Non-Autoregressive System for the WMT 22 Efficient Translation Shared Task

Jindřich Helcl

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 11800 Prague, Czech Republic
helcl@ufal.mff.cuni.cz

Abstract

We present a non-autoregressive system submission to the WMT 22 Efficient Translation Shared Task. Our system was used by Helcl et al. (2022) in an attempt to provide fair comparison between non-autoregressive and autoregressive models. This submission is an effort to establish solid baselines along with sound evaluation methodology, particularly in terms of measuring the decoding speed. The model itself is a 12-layer Transformer model trained with connectionist temporal classification on knowledge-distilled dataset by a strong autoregressive teacher model.

1 Introduction

In the past few years, non-autoregressive (NAR) models for neural machine translation (NMT) attracted interest from the research community (Gu et al., 2018; Lee et al., 2018). Given the conditional independence between the output states, the decoding process can be parallelized across time steps. In theory, this leads to higher decoding speeds.

Since efficient decoding is claimed to be the main motivation of non-autoregressive models, the Efficient Translation Shared Task seems to be the appropriate venue to provide fair comparison between these models and their autoregressive counterparts. However, all submissions to this task were autoregressive so far (Birch et al., 2018; Hayashi et al., 2019; Heafield et al., 2020, 2021).

Recently, Helcl et al. (2022) pointed out common flaws in the evaluation methodology of NAR models. We found that optimized autoregressive models still achieve superior performance over NAR models. The only scenario where NAR models showed some potential is GPU decoding with batch size of 1 (latency). Nevertheless, optimized autoregressive models were still both faster and better in terms of translation quality. The main purpose of this submission is to provide a reasonable baseline to future non-autoregressive submissions.

2 Model

In our experiments, we use the non-autoregressive model proposed by Libovický and Helcl (2018) based on Connectionist Temporal Classification (CTC; Graves et al., 2006). We submit models that have been trained as a part of Helcl (2022).

Architecture. The architecture is a 6-layer Transformer encoder, followed by a state-splitting layer and another stack of 6 Transformer layers. The state-splitting layer takes the encoder states, project them into k -times wider states using an affine transformation, and then split the states into k -times longer sequence while retaining the original model dimension. In the submitted model, we set $k = 3$. The latter 6 layers cross-attend to the states immediately after state-splitting. We use Transformer model dimension of 1,024, 16 attention heads and a dimension of 4,096 in the feed-forward sublayer.

The defining property of non-autoregressive models is that the decoding process treats output states as conditionally independent. In this architecture, we set the output sequence length to $k \times T_x$ where T_x is the length of the source sentence. To allow for shorter output sequences, the any output state can produce an empty token. The training loss is then computed using a dynamic programming algorithm as a sum of losses of all possible empty token alignments which lead to the same output sentence. The schema of the architecture is shown in Figure 1.

Training. We train our model on the knowledge-distilled data generated by the provided teacher (Chen et al., 2021). We use learning rate of 0.0001 in a inverse square-root decay scheme with 8,000 warm-up and decay steps.

Implementation. We implement and train our model in the Marian toolkit (Junczys-Dowmunt et al., 2018). For the CTC implementation,

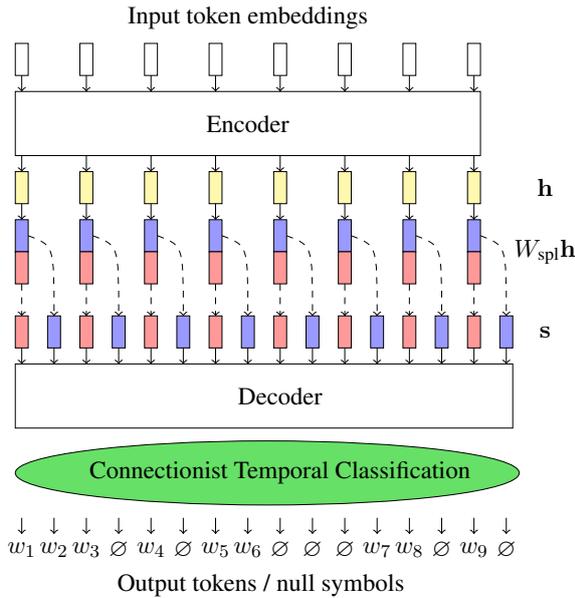


Figure 1: The CTC-based model architecture. We show the original image from Libovický and Helcl (2018).

we use the warp-ctc library¹. We release our code at <https://github.com/jindrahelcl/marian-dev>. The trained model (and a number of different variants including models in opposite translation direction) can be downloaded at <https://data.statmt.org/nar>.

3 Results

We refer the reader to the original paper for more details about the evaluation and its results. The model we submitted is denoted in the paper as “large”. A summary of the results follows.

Translation Quality. To summarize the main findings, the model achieves a competitive BLEU score (Papineni et al., 2002) on the WMT 14 news test set (Bojar et al., 2014), which serves as a comparison to other non-autoregressive models that use this test set as the de facto standard benchmark. When evaluated on the WMT 19 news test set, our model obtains BLEU of 47.8, and a COMET score (Rei et al., 2020) of 0.1485. Compared to a similarly-sized autoregressive teacher model with 50.5 BLEU and COMET of 0.4110, we see a somewhat surprising gap between the COMET scores while BLEU scores are relatively close. We hypothesize that the errors that the non-autoregressive model makes are out of the training domain of the COMET models, which makes them more sensitive towards this kind of errors.

¹<https://github.com/baidu-research/warp-ctc>

Decoding Time. We evaluated our models on the one million sentences benchmark used in the previous editions of this task (Heafield et al., 2021), and we tried to reproduce the official hardware setup to large extent. For CPU decoding, we measured time to translate the test set on an Intel Xeon 6354 server from Oracle Cloud, with 36 cores. We run the evaluation only in the batch decoding mode, as the models were too slow to decode with a single sentence in batch. With the submitted model, the translation on CPU took 7,434 seconds (using batch of 16 sentences).

We used a single Nvidia A100 GPU for GPU decoding. In the latency setup, the translation took 7,020 seconds, and the batched decoding ($b = 128$) took 782 seconds. When compared with other submissions to this task, we find that the smallest difference is indeed found in the GPU decoding latency setting. However, the optimized models submitted to last year’s round still achieved significantly better decoding times.

4 Conclusions

We submit a non-autoregressive system to the Efficient Translation Shared Task to the WMT 22. The model is trained with connectionist temporal classification, which allows the generation of empty tokens and thus making generation of sentences of various length possible while retaining the conditional independence among output tokens without explicit length estimation.

The main motivation of this submission is to provide a reasonable baseline system for future research. We believe that the sub-field of non-autoregressive NMT cannot progress without controlled decoding speed evaluation, which is exactly what the shared task organizers provide.

Acknowledgements

This project received funding from the European Union’s Horizon Europe Innovation programme under grant agreement no. 101070350 (HPLT). Our work has been using data provided by the LINDAT/CLARIAH-CZ Research Infrastructure, supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

References

- Alexandra Birch, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Yusuke Oda. 2018. [Findings of the second workshop on neural machine translation and generation](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.
- Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. [Findings of the 2014 workshop on statistical machine translation](#). In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Pinzhen Chen, Jindřich Helcl, Ulrich Germann, Laurie Burchell, Nikolay Bogoychev, Antonio Valerio Miceli Barone, Jonas Waldendorf, Alexandra Birch, and Kenneth Heafield. 2021. [The University of Edinburgh’s English-German and English-Hausa submissions to the WMT21 news translation task](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 104–109, Online. Association for Computational Linguistics.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 369–376, Pittsburgh, PA, USA. JMLR.org.
- Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. [Non-autoregressive neural machine translation](#). In *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada.
- Hiroaki Hayashi, Yusuke Oda, Alexandra Birch, Ioannis Konstas, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Katsuhito Sudoh. 2019. [Findings of the third workshop on neural generation and translation](#). In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 1–14, Hong Kong. Association for Computational Linguistics.
- Kenneth Heafield, Hiroaki Hayashi, Yusuke Oda, Ioannis Konstas, Andrew Finch, Graham Neubig, Xian Li, and Alexandra Birch. 2020. [Findings of the fourth workshop on neural generation and translation](#). In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 1–9, Online. Association for Computational Linguistics.
- Kenneth Heafield, Qianqian Zhu, and Roman Grundkiewicz. 2021. [Findings of the WMT 2021 shared task on efficient translation](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 639–651, Online. Association for Computational Linguistics.
- Jindřich Helcl, Barry Haddow, and Alexandra Birch. 2022. [Non-autoregressive machine translation: It’s not as fast as it seems](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1780–1790, Seattle, United States. Association for Computational Linguistics.
- Jindřich Helcl. 2022. *Non-Autoregressive Neural Machine Translation*. Ph.D. thesis, Charles University.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. [Marian: Fast neural machine translation in C++](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.
- Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. [Deterministic non-autoregressive neural sequence modeling by iterative refinement](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Brussels, Belgium. Association for Computational Linguistics.
- Jindřich Libovický and Jindřich Helcl. 2018. [End-to-end non-autoregressive neural machine translation with connectionist temporal classification](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3016–3021, Brussels, Belgium. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.