Non-Autoregressive Sequence Generation

Jiatao Gu
Facebook AI Research
jgu@fb.com

Xu Tan
Microsoft Research Asia
xuta@microsoft.com

1 Tutorial Description

State-of-the-art sequence generation models are mostly autoregressive (AR, Vaswani et al., 2017; Brown et al., 2020) where each generation step depends on the previously generated tokens. However, such models are inherently sequential, leading to high latency at inference time and suffering label bias (Lafferty et al., 2001) problem due to the locally normalized searching steps and exposure bias (Bengio et al., 2015) problem due to mismatch between training and inference.

Recently, increasing attention has been paid to modeling sequence generation in a non- or semi-autoregressive manner, which attempts to generate the entire or partial output sequences in parallel to speed up the decoding process and avoid potential issues (e.g., label bias, exposure bias) in autoregressive generation. In this tutorial, for simplicity, we summarize both approaches as *non-autoregressive* (NAR) sequence generation models. NAR models have been explored in many sequence generation tasks for text (e.g., neural machine translation (Gu et al., 2018), text summarization (Gu et al., 2019), text error correction (Awasthi et al., 2019; Leng et al., 2021b)), speech (e.g., speech recognition (Chen et al., 2019) and speech synthesis (Ren et al., 2019)). However, naive NAR models still face many challenges to close the performance gap between state-of-the-art autoregressive models because of a lack of modeling power. This tutorial will provide a thorough introduction and review of the basics of non-autoregressive sequence generation, including the background, the capabilities, and limits, popular methods that improve NAR models, and their applications on text and speech generation.

Introduction The tutorial will start with a brief discussion on the motivation of NAR generation, the problem definition, the evaluation protocol, and the comparison with standard autoregressive approaches. We use machine translation as the example generation task for the in-depth discussion as the first of its kind in NLP (Gu et al., 2018), and many follow-ups focus on this direction. Notably, we will show the underlying reasons (i.e., multi-modality problem) why NAR models generally perform worse and give some high-level instructions on improving NAR systems (Gu et al., 2018; Ren et al., 2020; Gu and Kong, 2021).

Methods Based on the high-level instructions, we will then dive into the detailed improvements from five aspects: *model architecture*, *objective function*, *training data*, *learning paradigm*, and *additional inference tricks*, respectively.

For *model architecture*, we divide existing approaches into four major categories according to the inference process: (1) **fully NAR models** that outputs the whole sequence in a single forward pass (Gu et al., 2018; Kaiser et al., 2018; Guo et al., 2019; Gu and Kong, 2021); (2) **iteration-based NAR models** which iteratively refine the parallel decoding results (Lee et al., 2018; Ghazvininejad et al., 2019, 2020b; Gu et al., 2019; Kasai et al., 2020); (3) **partially NAR models** where a sequence is still predicted autoregressively while each step multiple tokens are generated in parallel (Wang et al., 2018; Stern et al., 2018, 2019; Deng and Rush, 2020); (4) **locally AR models** which are, on the other hand, overall NAR while predict "phrases" autoregressively (Huang et al., 2017; Kong et al., 2020b). Aside from these major types, explicitly modeling NAR with **latent variables** is another useful approach that can boost the overall capability of all above NAR models. We will highlight several implementations including latent fertilities (Gu et al., 2018) and alignments (Saharia et al., 2020), VAEs with continuous (Shu et al., 2020; Lee et al., 2020; Gu and Kong, 2021) or discrete (Kaiser et al., 2018; Roy et al., 2018) latent variables, flow-based models (Ma et al., 2019b)

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 21 - 27

and stochastic diffusion models.

Next, we will discuss in-depth the *objective function* of NAR models starting from the standard cross-entropy (CE) loss which, however, leads to duplicated tokens in NAR outputs. To overcome this, we will introduce two types of advanced objective functions in this tutorial: (1) **loss function with latent information** which can be effectively marginalized/approximated through dynamic programming. For instance, we will cover latent alignments (CTC, AXE) (Graves et al., 2006; Libovický and Helcl, 2018; Saharia et al., 2020; Ghazvininejad et al., 2020a) and latent orders (OAXE) (Du et al., 2021); (2) the other type of objective function focuses on **loss beyond token-level**, which considers n-gram (Shao et al., 2020; Liu et al., 2021) or sequence-level (Sun et al., 2019; Shao et al., 2019; Tu et al., 2020) energy to optimize NAR models.

From the perspective of *training data*, we will first describe the sequence-level knowledge distillation (KD, Kim and Rush, 2016), and then explain its effectiveness of using KD on NAR generation (Zhou et al., 2020; Xu et al., 2021). In addition, we will also include the discussion about the drawbacks of over-relying on distillation for training NAR models (Ding et al., 2020) and propose potential alternatives.

For the fourth part, we will deepen the discussion on how to train NAR models more effectively. Due to the lack of modeling power, it may be crucial for NAR models to be trained with a more suitable *learning paradigm* to help match the performance of AR systems. In this tutorial, we will introduce the previous efforts from three primary directions: (1) **curriculum learning** where we train NAR models with tasks from easy to difficult progressively (Guo et al., 2020a; Liu et al., 2020; Qian et al., 2020); (2) **adversarial training** where a discriminator is jointly learned and the NAR model is forced to fool the discriminator. In this way, NAR models will not be directly exposed to the real training data, which is "too difficult" to fit. Adversarial training itself is not so popular in text generation in general. However, it is widely applied in other modalities such as NAR speech synthesis (Kong et al., 2020a). (3) **pre-training** where we will also show that combining with recent advances in self-supervised pre-training (e.g., BERT), we can naturally leverage the monolingual data to improve the learning of NAR models (Guo et al., 2020b; Qi et al., 2021; Jiang et al., 2021).

At the end of this part, we will also include additional discussions on valuable methods and tricks which help NAR models at inference time. For example, searching with length beams, reranking the AR model, incorporating the n-gram language model, etc.

Applications In the third section, we review some typical tasks that adopt non-autoregressive sequence generation, including *text generation* and *speech generation*. For *text generation*, we cover several tasks: (1) **neural machine translation** (Gu et al., 2018; Lee et al., 2018; Wang et al., 2018; Kong et al., 2020b; Gu and Kong, 2021); (2) **text summarization** (Gu et al., 2019; Qi et al., 2021; Jiang et al., 2021); (3) **text error correction** (Awasthi et al., 2019; Mallinson et al., 2020; Leng et al., 2021a,b); (4) **automatic speech recognition** (Chen et al., 2019; Higuchi et al., 2020; Chan et al., 2020). For *speech generation*, we cover: (1) **text to speech** (Ren et al., 2019; Peng et al., 2020; Oord et al., 2018; Kim et al., 2020, 2021); (2) **voice conversion** (Hayashi et al., 2021; Kameoka et al., 2021).

Beyond the introduction of task-level characteristics for non-autoregressive sequence generation, we also introduce some *advanced topics in applications*, including: (1) some advanced length prediction methods for text summarization (Qi et al., 2021) and speech recognition (Chen et al., 2019); (2) alignment modeling between source and target sequence in text to speech, e.g., duration prediction (Ren et al., 2019) or source-target attention (Peng et al., 2020); (3) analysis on the dependency among target tokens that can influence the modeling difficulty of non-autoregressive generation models (Ren et al., 2020); (4) the relationship between non-autoregressive sequence generation and streaming sequence generation (Ma et al., 2019a), considering they are both for inference speedup.

Conclusion At the end of the tutorial, we will describe several research challenges and list the comparison with other speed-up approaches for AR models (e.g., quantization, pruning, distillation). Finally, we will also discuss the potential future research directions to close this tutorial.

2 Type of the Tutorial

Cutting-edge.

3 Target Audience

This tutorial targets those audiences who work on 1) neural sequence generation (e.g., neural machine translation, etc.); 2) natural language and speech processing; 3) deep learning and artificial intelligence in general. Some prerequisites for the attendees are:

- Math: calculus, linear algebra, and probability theory.
- Machine learning: basic machine learning paradigms and basic deep learning models such as MLP, RNN, CNN, and Transformer.
- Neural sequence generation: Be familiar with at least one sequence generation task, such as neural machine translation, text summarization, automatic speech recognition, text to speech, etc.

4 Tutorial Outline

PART I Introduction ($\sim$ 20 minutes)

1.1 Problem definition

1.2 Evaluation protocol

1.3 Multi-modality problem

PART II Methods ($\sim$ 90 minutes)

2.1 Model architectures

2.1.1 Fully NAR models

2.1.2 Iteration-based NAR models

2.1.3 Partially NAR models

2.1.4 Locally AR models

2.1.5 NAR models with latent variables

2.2 Objective functions

2.2.1 Loss with latent variables

2.2.2 Loss beyond token-level

2.3 Training data

2.4 Learning paradigms

2.4.1 Curriculum learning

2.4.2 Adversarial training

2.4.3 Self-supervised pre-training

2.5 Inference methods and tricks

PART III Applications ($\sim$ 50 minutes)

3.1 Text generation

3.1.1 Neural machine translation

3.1.2 Text summarization

3.1.3 Text error correction

3.1.4 Automatic speech recognition

3.2 Speech generation

3.2.1 Text to speech

3.2.2 Voice conversion

3.3 Advanced topics in applications

3.3.1 Advanced length prediction

3.3.2 Alignment (duration vs attention)

3.3.3 Target token dependency

3.3.4 Relationship with streaming

PART IV Open problems, future directions, Q&A ($\sim$20 minutes)

5 How the tutorial includes other people's work

We organize our tutorial content from a broad view of non-autoregressive sequence generation, spanning from basic methods to applications, which cover diverse work in this area, most of which are other people's work.

6 Diversity Considerations

Methods We introduce the methods of non-autoregressive sequence generation in a comprehensive and diverse view, covering model architectures, objective functions, training data, learning paradigms, and additional tricks. These methods are general and not limited to specific languages or domains.

Applications We introduce a variety of non-autoregressive sequence generation tasks, spanning from the text (e.g., neural machine translation, text error correction) to speech (e.g., text to speech, voice conversion).

Instructors We are from different institutions (Facebook and Microsoft) and work on diverse topics in machine learning, NLP, and non-autoregressive sequence generation.

Audiences Due to the diversity in the methods and applications of our tutorial and the tutorial instructors, we can attract audiences interested in diverse sequence generation tasks and modalities (text and speech) and from both academia and industry.

7 Reading List

Please see the citations in Section 1. For participants interested in reading important studies before this tutorial, we recommend the following basic papers: (1) the typical AR model (Transformer) (Vaswani et al., 2017); (2) the vanilla NAR model (Gu et al., 2018); (3) the typical iteration-based NAR model (Ghazvininejad et al., 2019); (4) a study on NAR models for both text and speech tasks (Ren et al., 2020).

8 Bio of Speakers

8.1 Jiatao Gu

Dr. Jiatao Gu is a Research Scientist at Facebook AI Research (FAIR). Jiatao received his Ph.D. degree in 2018 from the University of Hong Kong and B.Eng from Tsinghua University in 2014. His research interests cover representation learning and generative models and their applications on NLP, speech, computer vision, and multi-modal learning. Particularly, his research focuses on developing efficient learning and inference algorithms and applying them successfully to neural machine translation and 3D-aware image synthesis. He has over 40 papers published at top-tier conferences and journals, including ACL, EMNLP, NeurIPS, ICLR, and TACL. Jiao has also served as an area chair for several top conferences. Jiatao has rich research experience on the topic of non-autoregressive sequence generation. He published the first of its kind paper for non-autoregressive neural machine translation in 2018 and has led the following exploration and extensions. Website: https://jiataogu.me/.

8.2 Xu Tan

Xu Tan is a Senior Researcher at Microsoft Research Asia (MSRA). His research interests cover deep learning and its applications in language/speech/music, including neural machine translation, text to speech, automatic speech recognition, pre-training, music generation, etc. The machine translation systems have achieved human parity on Chinese-English news translation in 2018 and won several champions on WMT machine translation competition in 2019. He has designed several popular language/speech/music models, and systems (e.g., MASS, FastSpeech, and Muzic) and has transferred many research works to the products in Microsoft (e.g., Azure, Bing). He has rich research experiences on non-autoregressive sequence generation and has designed several models such as FastCorrect 1/2, FastSpeech 1/2. He has given several tutorials on language/speech/music at international conferences: 1) A tutorial on text to speech at IJCAI 2021; 2) A tutorial on AI music composition at ACM Multimedia 2021. Website: https://www.microsoft.com/en-us/research/people/xuta/.

9 Ethics Statement

Non-autoregressive sequence generation can improve the inference speed of various sequence generation tasks in text and speech. Unfortunately, this technology may be misused to generate deepfake content (Thies et al., 2016) such as mimicking one's writing style or speaking style. However, great attempts have been made to detect the deepfake content (Kaggle, 2019), which can minimize or avoid its potential negative impact.

References

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel iterative edit models for local sequence transduction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4259–4269, Hong Kong, China. Association for Computational Linguistics.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, pages 1171–1179.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

William Chan, Chitwan Saharia, Geoffrey Hinton, Mohammad Norouzi, and Navdeep Jaitly. 2020. Imputer: Sequence modelling via imputation and dynamic programming. In *International Conference on Machine Learning*, pages 1403–1413. PMLR.

Nanxin Chen, Shinji Watanabe, Jesús Villalba, and Najim Dehak. 2019. Listen and fill in the missing letters: Non-autoregressive transformer for speech recognition. *arXiv preprint arXiv:1911.04908*.

Yuntian Deng and Alexander Rush. 2020. Cascaded text generation with markov transformers. *Advances in Neural Information Processing Systems*, 33.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2020. Understanding and improving lexical choice in non-autoregressive translation. *arXiv preprint arXiv:2012.14583*.

Cunxiao Du, Zhaopeng Tu, and Jing Jiang. 2021. Order-agnostic cross entropy for non-autoregressive machine translation. *arXiv preprint arXiv:2106.05093*.

Marjan Ghazvininejad, Vladimir Karpukhin, Luke Zettlemoyer, and Omer Levy. 2020a. Aligned cross entropy for non-autoregressive machine translation. In *International Conference on Machine Learning*, pages 3515–3523. PMLR.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6114–6123.

Marjan Ghazvininejad, Omer Levy, and Luke Zettlemoyer. 2020b. Semi-autoregressive training improves mask-predict decoding. *arXiv preprint arXiv:2001.08785*.

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*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Jiatao Gu and Xiang Kong. 2021. Fully non-autoregressive neural machine translation: Tricks of the trade. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 120–133, Online. Association for Computational Linguistics.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems*, volume 32, pages 11181–11191. Curran Associates, Inc.

Junliang Guo, Xu Tan, Di He, Tao Qin, Linli Xu, and Tie-Yan Liu. 2019. Non-autoregressive neural machine translation with enhanced decoder input. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3723–3730.

Junliang Guo, Xu Tan, Linli Xu, Tao Qin, Enhong Chen, and Tie-Yan Liu. 2020a. Fine-tuning by curriculum learning for non-autoregressive neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7839–7846.

Junliang Guo, Linli Xu, and Enhong Chen. 2020b. Jointly masked sequence-to-sequence model for non-autoregressive neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 376–385.

Tomoki Hayashi, Wen-Chin Huang, Kazuhiro Kobayashi, and Tomoki Toda. 2021. Non-autoregressive sequence-to-sequence voice conversion. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7068–7072. IEEE.

Yosuke Higuchi, Shinji Watanabe, Nanxin Chen, Tetsuji Ogawa, and Tetsunori Kobayashi. 2020. Mask ctc: Non-autoregressive end-to-end asr with ctc and mask predict. *arXiv preprint arXiv:2005.08700*.

Po-Sen Huang, Chong Wang, Sitao Huang, Dengyong Zhou, and Li Deng. 2017. Towards neural phrase-based machine translation. *CoRR*, abs/1706.05565.

Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2021. Improving non-autoregressive generation with mixup training. *arXiv preprint arXiv:2110.11115*.

Kaggle. 2019. Deepfake detection challenge | kaggle. https://www.kaggle.com/c/deepfake-detection-challenge.

Lukasz Kaiser, Samy Bengio, Aurko Roy, Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, and Noam Shazeer. 2018. Fast decoding in sequence models using discrete latent variables. In *International Conference on Machine Learning*, pages 2395–2404.

Hirokazu Kameoka, Kou Tanaka, and Takuhiro Kaneko. 2021. Fasts2s-vc: Streaming non-autoregressive sequence-to-sequence voice conversion. *arXiv preprint arXiv:2104.06900*.

Jungo Kasai, James Cross, Marjan Ghazvininejad, and Jiatao Gu. 2020. Non-autoregressive machine translation with disentangled context transformer. In *International Conference on Machine Learning*, pages 5144–5155. PMLR.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *arXiv preprint arXiv:2106.06103*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020a. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33.

Xiang Kong, Zhisong Zhang, and Eduard Hovy. 2020b. Incorporating a local translation mechanism into non-autoregressive translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1067–1073, Online. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 1173–1182.

Jason Lee, Raphael Shu, and Kyunghyun Cho. 2020. Iterative refinement in the continuous space for non-autoregressive neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1006–1015.

Yichong Leng, Xu Tan, Rui Wang, Linchen Zhu, Jin Xu, Linquan Liu, Tao Qin, Xiang-Yang Li, Edward Lin, and Tie-Yan Liu. 2021a. Fastcorrect 2: Fast error correction on multiple candidates for automatic speech recognition. *arXiv preprint arXiv:2109.14420*.

Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiang-Yang Li, Ed Lin, and Tie-Yan Liu. 2021b. Fastcorrect: Fast error correction with edit alignment for automatic speech recognition. In *NeurIPS*.

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.

Guangyi Liu, Zichao Yang, Tianhua Tao, Xiaodan Liang, Zhen Li, Bowen Zhou, Shuguang Cui, and Zhiting Hu. 2021. Don't take it literally: An edit-invariant sequence loss for text generation. *arXiv preprint arXiv:2106.15078*.

Jinglin Liu, Yi Ren, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Task-level curriculum learning for non-autoregressive neural machine translation. *arXiv preprint arXiv:2007.08772*.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, et al. 2019a. Stacl: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036.

Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. 2019b. Flowseq: Non-autoregressive conditional sequence generation with generative flow. *arXiv preprint arXiv:1909.02480*.

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. Felix: Flexible text editing through tagging and insertion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1244–1255.

Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR.

Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao. 2020. Non-autoregressive neural text-to-speech. In *International conference on machine learning*, pages 7586–7598. PMLR.

Weizhen Qi, Yeyun Gong, Jian Jiao, Yu Yan, Weizhu Chen, Dayiheng Liu, Kewen Tang, Houqiang Li, Jiusheng Chen, Ruofei Zhang, et al. 2021. Bang: Bridging autoregressive and non-autoregressive generation with large scale pretraining. In *International Conference on Machine Learning*, pages 8630–8639. PMLR.

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. 2020. Glancing transformer for non-autoregressive neural machine translation. *arXiv preprint arXiv:2008.07905*.

Yi Ren, Jinglin Liu, Xu Tan, Zhou Zhao, Sheng Zhao, and Tie-Yan Liu. 2020. A study of non-autoregressive model for sequence generation. *arXiv preprint arXiv:2004.10454*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: fast, robust and controllable text to speech. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3171–3180.

Aurko Roy, Ashish Vaswani, Arvind Neelakantan, and Niki Parmar. 2018. Theory and experiments on vector quantized autoencoders. *arXiv preprint arXiv:1805.11063*.

Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. 2020. Non-autoregressive machine translation with latent alignments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1098–1108, Online. Association for Computational Linguistics.

Chenze Shao, Yang Feng, Jinchao Zhang, Fandong Meng, Xilin Chen, and Jie Zhou. 2019. Retrieving sequential information for non-autoregressive neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3013–3024.

Chenze Shao, Jinchao Zhang, Yang Feng, Fandong Meng, and Jie Zhou. 2020. Minimizing the bag-of-ngrams difference for non-autoregressive neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 198–205.

Raphael Shu, Jason Lee, Hideki Nakayama, and Kyunghyun Cho. 2020. Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. 2019. Insertion transformer: Flexible sequence generation via insertion operations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5976–5985, Long Beach, California, USA. PMLR.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. 2018. Blockwise parallel decoding for deep autoregressive models. In *Advances in Neural Information Processing Systems*, pages 10107–10116.

Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, and Zhihong Deng. 2019. Fast structured decoding for sequence models. In *Advances in Neural Information Processing Systems*, pages 3016–3026.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395.

Lifu Tu, Richard Yuanzhe Pang, Sam Wiseman, and Kevin Gimpel. 2020. ENGINE: Energy-based inference networks for non-autoregressive machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2826, Online. Association for Computational Linguistics.

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 *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Chunqi Wang, Ji Zhang, and Haiqing Chen. 2018. Semi-autoregressive neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 479–488.

Weijia Xu, Shuming Ma, Dongdong Zhang, and Marine Carpuat. 2021. How does distilled data complexity impact the quality and confidence of non-autoregressive machine translation? *arXiv preprint arXiv:2105.12900*.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, Conference Track Proceedings*.