

Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement

Jason Lee*

New York University
jason@cs.nyu.edu

Elman Mansimov*

New York University
mansimov@cs.nyu.edu

Kyunghyun Cho

New York University
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Abstract

We propose a conditional non-autoregressive neural sequence model based on iterative refinement. The proposed model is designed based on the principles of latent variable models and denoising autoencoders, and is generally applicable to any sequence generation task. We extensively evaluate the proposed model on machine translation (En \leftrightarrow De and En \leftrightarrow Ro) and image caption generation, and observe that it significantly speeds up decoding while maintaining the generation quality comparable to the autoregressive counterpart.

1 Introduction

Conditional neural sequence modeling has become a *de facto* standard in a variety of tasks (see, e.g., Cho et al., 2015, and references therein). Much of this recent success is built on top of autoregressive sequence models in which the probability of a target sequence is factorized as a product of conditional probabilities of next symbols given all the preceding ones. Despite its success, neural autoregressive modeling has its weakness in decoding, i.e., finding the most likely sequence. Because of intractability, we must resort to sub-optimal approximate decoding, and due to its sequential nature, decoding cannot be easily parallelized and results in a large latency (see, e.g., Cho, 2016). This has motivated the recent investigation into non-autoregressive neural sequence modeling by Gu et al. (2017) in the context of machine translation and Oord et al. (2017) in the context of speech synthesis.

In this paper, we propose a non-autoregressive neural sequence model based on iterative refinement, which is generally applicable to any sequence generation task beyond machine translation. The proposed model can be viewed as both

a latent variable model and a conditional denoising autoencoder. We thus propose a learning algorithm that is hybrid of lowerbound maximization and reconstruction error minimization. We further design an iterative inference strategy with an adaptive number of steps to minimize the generation latency without sacrificing the generation quality.

We extensively evaluate the proposed conditional non-autoregressive sequence model and compare it against the autoregressive counterpart, using the state-of-the-art Transformer (Vaswani et al., 2017), on machine translation and image caption generation. In the case of machine translation, the proposed deterministic non-autoregressive models are able to decode approximately 2 – 3 \times faster than beam search from the autoregressive counterparts on both GPU and CPU, while maintaining 90-95% of translation quality on IWSLT’16 En \leftrightarrow De, WMT’16 En \leftrightarrow Ro and WMT’14 En \leftrightarrow De. On image caption generation, we observe approximately 3 \times and 5 \times faster decoding on GPU and CPU, respectively, while maintaining 85% of caption quality.¹

2 Non-Autoregressive Sequence Models

Sequence modeling in deep learning has largely focused on autoregressive modeling. That is, given a sequence $Y = (y_1, \dots, y_T)$, we use some form of a neural network to parametrize the conditional distribution over each variable y_t given all the preceding variables, i.e.,

$$\log p(y_t|y_{<t}) = f_\theta(y_{<t}),$$

where f_θ is for instance a recurrent neural network. This approach has become a *de facto* standard in language modeling (Mikolov et al.,

¹ We release the implementation, preprocessed datasets as well as trained models online at <https://github.com/nyu-dl/dl4mt-nonauto>.

* Equal Contribution

2010). When this is conditioned on an extra variable X , it becomes a conditional sequence model $\log p(Y|X)$ which serves as a basis on which many recent advances in, e.g., machine translation (Bahdanau et al., 2014; Sutskever et al., 2014; Kalchbrenner and Blunsom, 2013) and speech recognition (Chorowski et al., 2015; Chiu et al., 2017) have been made.

Despite the recent success, autoregressive sequence modeling has a weakness due to its nature of sequential processing. This weakness shows itself especially when we try to decode the most likely sequence from a trained model, i.e.,

$$\hat{Y} = \arg \max_Y \log p(Y|X).$$

There is no known polynomial algorithm for solving it exactly, and practitioners have relied on approximate decoding algorithms (see, e.g., Cho, 2016; Hoang et al., 2017). Among these, beam search has become the method of choice, due to its superior performance over greedy decoding, which however comes with a substantial computational overhead (Cho, 2016).

As a solution to this issue of slow decoding, two recent works have attempted non-autoregressive sequence modeling. Gu et al. (2017) have modified the Transformer (Vaswani et al., 2017) for non-autoregressive machine translation, and Oord et al. (2017) a convolutional network (Oord et al., 2016) for non-autoregressive modeling of waveform. Non-autoregressive modeling factorizes the distribution over a target sequence given a source into a product of conditionally independent per-step distributions:

$$p(Y|X) = \prod_{t=1}^T p(y_t|X),$$

breaking the dependency among the target variables across time. This allows us to trivially find the most likely target sequence by taking $\arg \max_{y_t} p(y_t|X)$ for each t , effectively bypassing the computational overhead and suboptimality of decoding from an autoregressive sequence model.

This desirable property of exact and parallel decoding however comes at the expense of potential performance degradation (Kaiser and Bengio, 2016). The potential modeling gap, which is the gap between the underlying, true model and the neural sequence model, could be larger with the

non-autoregressive model compared to the autoregressive one due to challenge of modeling the factorized conditional distribution above.

3 Iterative Refinement for Deterministic Non-Autoregressive Sequence Models

3.1 Latent variable model

Similarly to two recent works (Oord et al., 2017; Gu et al., 2017), we introduce latent variables to implicitly capture the dependencies among target variables. We however remove any stochastic behavior by interpreting this latent variable model, introduced immediately below, as a process of iterative refinement.

Our goal is to capture the dependencies among target symbols given a source sentence without auto-regression by introducing L intermediate random variables and marginalizing them out:

$$p(Y|X) = \sum_{Y^0, \dots, Y^L} \left(\prod_{t=1}^T p(y_t|Y^L, X) \right) \left(\prod_{t=1}^T p(y_t^L|Y^{L-1}, X) \right) \dots \left(\prod_{t=1}^T p(y_t^0|X) \right). \quad (1)$$

Each product term inside the summation is modelled by a deep neural network that takes as input a source sentence and outputs the conditional distribution over the target vocabulary V for each t .

Deterministic Approximation The marginalization in Eq. (1) is intractable. In order to avoid this issue, we consider two approximation strategies; deterministic and stochastic approximation. Without loss of generality, let us consider the case of single intermediate latent variable, that is $L = 1$. In the *deterministic* case, we set \hat{y}_t^0 to the most likely value according to its distribution $p(y_t^0|X)$, that is $\hat{y}_t^0 = \arg \max_{y_t^0} p(y_t^0|X)$. The entire lower bound can then be written as:

$$\log p(Y|X) \geq \left(\sum_{t=1}^T \log p(y_t|\hat{Y}^L, X) \right) + \dots + \left(\sum_{t=1}^T \log p(y_t^1|\hat{Y}^0, X) \right) + \left(\sum_{t=1}^T \log p(\hat{y}_t^0|X) \right).$$

Stochastic Approximation In the case of *stochastic* approximation, we instead sample \hat{y}_t^0 from the distribution $p(y_t^0|X)$. This results in the unbiased estimate of the marginal log-probability $\log p(Y|X)$. Other than the difference in whether

most likely values or samples are used, the remaining steps are identical.

Latent Variables Although the intermediate random variables could be anonymous, we constrain them to be of the same type as the output Y is, in order to share an underlying neural network. This constraint allows us to view each conditional $p(Y^l|\hat{Y}^{l-1}, X)$ as a single-step of refinement of a rough target sequence \hat{Y}^{l-1} . The entire chain of L conditionals is then the L -step iterative refinement. Furthermore, sharing the parameters across these refinement steps enables us to dynamically adapt the number of iterations per input X . This is important as it substantially reduces the amount of time required for decoding, as we see later in the experiments.

Training For each training pair (X, Y^*) , we first approximate the marginal log-probability. We then minimize

$$J_{\text{LVM}}(\theta) = - \sum_{l=0}^{L+1} \left(\sum_{t=1}^T \log p_{\theta}(y_t^* | \hat{Y}^{l-1}, X) \right), \quad (2)$$

where $\hat{Y}^{l-1} = (\hat{y}_1^{l-1}, \dots, \hat{y}_T^{l-1})$, and θ is a set of parameters. We initialize \hat{y}_t^0 (t -th target word in the first iteration) as $x_{t'}$, where $t' = (T'/T) \cdot t$. T' and T are the lengths of the source X and target Y^* , respectively.

3.2 Denoising Autoencoder

The proposed approach could instead be viewed as learning a conditional denoising autoencoder which is known to capture the gradient of the log-density. That is, we implicitly learn to find a direction Δ_Y in the output space that maximizes the underlying true, data-generating distribution $\log P(Y|X)$. Because the output space is discrete, much of the theoretical analysis by [Alain and Bengio \(2014\)](#) are not strictly applicable. We however find this view attractive as it serves as an alternative foundation for designing a learning algorithm.

Training We start with a corruption process $C(Y|Y^*)$, which introduces noise to the correct output Y^* . Given the reference translation Y^* , we sample $\tilde{Y} \sim C(Y|Y^*)$ which becomes as an input to each conditional in Eq. (1). Then, the goal of learning is to maximize the log-probability of the original reference Y^* given the corrupted version. That is, to minimize

$$J_{\text{DAE}}(\theta) = - \sum_{t=1}^T \log p_{\theta}(y_t^* | \tilde{Y}, X). \quad (3)$$

Once this cost J_{DAE} is minimized, we can recursively perform the maximum-a-posterior inference, i.e., $\hat{Y} = \arg \max_Y \log p_{\theta}(Y|X)$, to find \hat{Y} that (approximately) maximizes $\log p(Y|X)$.

Corruption Process C There is little consensus on the best corruption process for a sequence, especially of discrete tokens. In this work, we use a corruption process proposed by [Hill et al. \(2016\)](#), which has recently become more widely adopted (see, e.g., [Artetxe et al., 2017](#); [Lample et al., 2017](#)). Each y_t^* in a reference target $Y^* = (y_1^*, \dots, y_T^*)$ is corrupted with a probability $\beta \in [0, 1]$. If decided to corrupt, we either (1) replace y_{t+1}^* with this token y_t^* , (2) replace y_t^* with a token uniformly selected from a vocabulary of all unique tokens at random, or (3) swap y_t^* and y_{t+1}^* . This is done sequentially from y_1^* until y_T^* .

3.3 Learning

Cost function Although it is possible to train the proposed non-autoregressive sequence model using either of the cost functions above (J_{LVM} or J_{DAE}) we propose to stochastically mix these two cost functions. We do so by randomly replacing each term \hat{Y}^{l-1} in Eq. (2) with \tilde{Y} in Eq. (3):

$$J(\theta) = - \sum_{l=0}^{L+1} \left(\alpha_l \sum_{t=1}^T \log p_{\theta}(y_t^* | \hat{Y}^{l-1}, X) \right. \\ \left. + (1 - \alpha_l) \sum_{t=1}^T \log p_{\theta}(y_t^* | \tilde{Y}, X) \right), \quad (4)$$

where $\tilde{Y} \sim C(Y|Y^*)$, and α_l is a sample from a Bernoulli distribution with the probability p_{DAE} . p_{DAE} is a hyperparameter. As the first conditional $p(Y^0|X)$ in Eq. (1) does not take as input any target Y , we set $\alpha_0 = 1$ always.

Distillation [Gu et al. \(2017\)](#), in the context of machine translation, and [Oord et al. \(2017\)](#), in the context of speech generation, have recently discovered that it is important to use knowledge distillation ([Hinton et al., 2015](#); [Kim and Rush, 2016](#)) to successfully train a non-autoregressive sequence model. Following [Gu et al. \(2017\)](#), we also use knowledge distillation by replacing the reference target Y^* of each training example

(X, Y^*) with a target Y^{AR} generated from a well-trained autoregressive counterpart. Other than this replacement, the cost function in Eq (4) and the model architecture remain unchanged.

Target Length Prediction One difference between the autoregressive and non-autoregressive models is that the former naturally models the length of a target sequence without any arbitrary upper-bound, while the latter does not. It is hence necessary to separately model $p(T|X)$, where T is the length of a target sequence, although during training, we simply use the length of each reference target sequence.

3.4 Inference: Decoding

Inference in the proposed approach is entirely deterministic. We start from the input X and first predict the length of the target sequence $\hat{T} = \arg \max_T \log p(T|X)$. Then, given X and \hat{T} we generate the initial target sequence by $\hat{y}_t^0 = \arg \max_{y_t} \log p(y_t^0|X)$, for $t = 1, \dots, \hat{T}$. We continue refining the target sequence by $\hat{y}_t^l = \arg \max_{y_t} \log p(y_t^l|\hat{Y}^{l-1}, X)$, for $t = 1, \dots, \hat{T}$.

Because these conditionals, except for the initial one, are modeled by a single, shared neural network, this refinement can be performed as many iterations as necessary until a pre-defined stopping criterion is met. A criterion can be based either on the amount of change in a target sequence after each iteration (i.e., $D(\hat{Y}^{l-1}, \hat{Y}^l) \leq \epsilon$), or on the amount of change in the conditional log-probabilities (i.e., $|\log p(\hat{Y}^{l-1}|X) - \log p(\hat{Y}^l|X)| \leq \epsilon$) or on the computational budget. In our experiments, we use the first criterion and use Jaccard distance as our distance function D .

4 Related Work

Non-Autoregressive Neural Machine Translation Schwenk (2012) proposed a continuous-space translation model to estimate the conditional distribution over a target phrase given a source phrase, while dropping the conditional dependencies among target tokens. The evaluation was however limited to reranking and to short phrase pairs (up to 7 words on each side) only. Kaiser and Bengio (2016) investigated neural GPU (Kaiser and Sutskever, 2015), for machine translation. They evaluated both non-autoregressive and autoregressive approaches, and found that the non-

autoregressive approach significantly lags behind the autoregressive variants. It however differs from our approach that each iteration does not output a refined version from the previous iteration. The recent paper by Gu et al. (2017) is most relevant to the proposed work. They similarly introduced a sequence of discrete latent variables. They however use supervised learning for inference, using the word alignment tool (Dyer et al., 2013). To achieve the best result, Gu et al. (2017) stochastically sample the latent variables and rerank the corresponding target sequences with an external, autoregressive model. This is in contrast to the proposed approach which is fully deterministic during decoding and does not rely on any extra reranking mechanism.

Parallel WaveNet Simultaneously with Gu et al. (2017), Oord et al. (2017) presented a non-autoregressive sequence model for speech generation. They use inverse autoregressive flow (IAF, Kingma et al., 2016) to map a sequence of independent random variables to a target sequence. They apply the IAF multiple times, similarly to our iterative refinement strategy. Their approach is however restricted to continuous target variables, while the proposed approach in principle could be applied to both discrete and continuous variables.

Post-Editing for Machine Translation Novak et al. (2016) proposed a convolutional neural network that iteratively predicts and applies token substitutions given a translation from a phrase-based translation system. Unlike their system, our approach can edit an intermediate translation with a higher degree of freedom. QuickEdit (Grangier and Auli, 2017) and deliberation network (Xia et al., 2017) incorporate the idea of refinement into neural machine translation. Both systems consist of two autoregressive decoders. The second decoder takes into account the translation generated by the first decoder. We extend these earlier efforts by incorporating more than one refinement steps without necessitating extra annotations.

Infusion Training Bordes et al. (2017) proposed an unconditional generative model for images based on iterative refinement. At each step l of iterative refinement, the model is trained to maximize the log-likelihood of target Y given the weighted mixture of generated samples from the previous iteration \hat{Y}^{l-1} and a corrupted target \tilde{Y} . That is, the corrupted version of target is “infused”

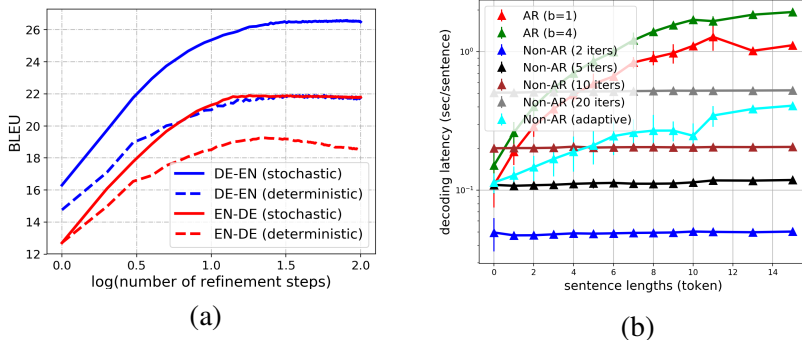


Figure 1: (a) BLEU scores on WMT’14 En-De w.r.t. the number of refinement steps (up to 10²). The x-axis is in the logarithmic scale. (b) the decoding latencies (sec/sentence) of different approaches on IWSLT’16 En→De. The y-axis is in the logarithmic scale.

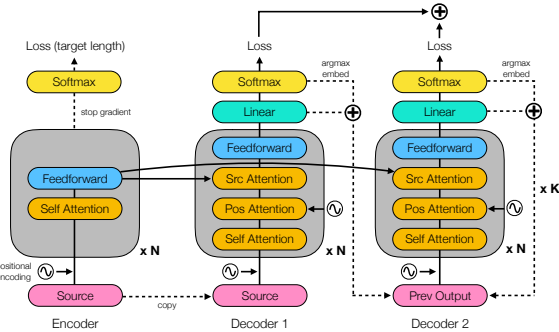


Figure 2: We compose three transformer blocks (“Encoder”, “Decoder 1” and “Decoder 2”) to implement the proposed non-autoregressive sequence model.

into generated samples during training. In the domain of text, however, computing a weighted mixture of two sequences of discrete tokens is not well defined, and we propose to stochastically mix denoising and lowerbound maximization objectives.

5 Network Architecture

We use three transformer-based network blocks to implement our model. The first block (“Encoder”) encodes the input X , the second block (“Decoder 1”) models the first conditional $\log p(Y^0|X)$, and the final block (“Decoder 2”) is shared across iterative refinement steps, modeling $\log p(Y^l|\hat{Y}^{l-1}, X)$. These blocks are depicted side-by-side in Fig. 2. The encoder is identical to that from the original Transformer (Vaswani et al., 2017). We however use the decoders from Gu et al. (2017) with additional positional attention and use the highway layer (Srivastava et al., 2015) instead of the residual layer (He et al., 2016).

The original input X is padded or shortened to fit the length of the reference target sequence before being fed to Decoder 1. At each refinement step l , Decoder 2 takes as input the predicted target sequence \hat{Y}^{l-1} and the sequence of final activation vectors from the previous step.

(b)

6 Experimental Setting

We evaluate the proposed approach on two sequence modeling tasks: machine translation and image caption generation. We compare the proposed non-autoregressive model against the autoregressive counterpart both in terms of generation quality, measured in terms of BLEU (Papineni et al., 2002), and generation efficiency, measured in terms of (source) tokens and images per second for translation and image captioning, respectively.

Machine Translation We choose three tasks of different sizes: IWSLT’16 En↔De (196k pairs), WMT’16 En↔Ro (610k pairs) and WMT’14 En↔De (4.5M pairs). We tokenize each sentence using a script from Moses (Koehn et al., 2007) and segment each word into subword units using BPE (Sennrich et al., 2016). We use 40k tokens from both source and target for all the tasks. For WMT’14 En-De, we use newstest-2013 and newstest-2014 as development and test sets. For WMT’16 En-Ro, we use newsdev-2016 and newstest-2016 as development and test sets. For IWSLT’16 En-De, we use test2013 for validation.

We closely follow the setting by Gu et al. (2017). In the case of IWSLT’16 En-De, we use the small model ($d_{\text{model}} = 278, d_{\text{hidden}} = 507, p_{\text{dropout}} = 0.1, n_{\text{layer}} = 5$ and $n_{\text{head}} = 2$).² For WMT’14 En-De and WMT’16 En-Ro, we use the base transformer by Vaswani et al. (2017) ($d_{\text{model}} = 512, d_{\text{hidden}} = 512, p_{\text{dropout}} = 0.1, n_{\text{layer}} = 6$ and $n_{\text{head}} = 8$). We use the warm-up learning rate scheduling (Vaswani et al., 2017) for the WMT tasks, while using linear annealing (from 3×10^{-4} to 10^{-5}) for the IWSLT task. We do not use label smoothing nor average multiple check-pointed models. These decisions were made based on the preliminary experiments. We train each model either on a single

² Due to the space constraint, we refer readers to (Vaswani et al., 2017; Gu et al., 2017) for more details.

| | IWSLT'16 En-De | | | | WMT'16 En-Ro | | | | WMT'14 En-De | | | | MS COCO | | | |
|-----------|-----------------------|-------|-------|-------|--------------|-------|-------|-------|--------------|-------|-------|-------|---------|-------|------|-----|
| | En→ | De→ | GPU | CPU | En→ | Ro→ | GPU | CPU | En→ | De→ | GPU | CPU | BLEU | GPU | CPU | |
| AR | $b = 1$ | 28.64 | 34.11 | 70.3 | 32.2 | 31.93 | 31.55 | 55.6 | 15.7 | 23.77 | 28.15 | 54.0 | 15.8 | 23.47 | 4.3 | 2.1 |
| | $b = 4$ | 28.98 | 34.81 | 63.8 | 14.6 | 32.40 | 32.06 | 43.3 | 7.3 | 24.57 | 28.47 | 44.9 | 7.0 | 24.78 | 3.6 | 1.0 |
| NAT | FT | 26.52 | – | – | – | 27.29 | 29.06 | – | – | 17.69 | 21.47 | – | – | – | – | – |
| | FT+NPD | 28.16 | – | – | – | 29.79 | 31.44 | – | – | 19.17 | 23.30 | – | – | – | – | – |
| Our Model | $i_{\text{dec}} = 1$ | 22.20 | 27.68 | 573.0 | 213.2 | 24.45 | 25.73 | 694.2 | 98.6 | 13.91 | 16.77 | 511.4 | 83.3 | 20.12 | 17.1 | 8.9 |
| | $i_{\text{dec}} = 2$ | 24.82 | 30.23 | 423.8 | 110.9 | 27.10 | 28.15 | 332.7 | 62.8 | 16.95 | 20.39 | 393.6 | 49.6 | 20.88 | 12.0 | 5.7 |
| | $i_{\text{dec}} = 5$ | 26.58 | 31.85 | 189.7 | 52.8 | 28.86 | 29.72 | 194.4 | 29.0 | 20.26 | 23.86 | 139.7 | 23.1 | 21.12 | 6.2 | 2.8 |
| | $i_{\text{dec}} = 10$ | 27.11 | 32.31 | 98.8 | 24.1 | 29.32 | 30.19 | 93.1 | 14.8 | 21.61 | 25.48 | 90.4 | 12.3 | 21.24 | 2.0 | 1.2 |
| | Adaptive | 27.01 | 32.43 | 125.9 | 29.3 | 29.66 | 30.30 | 118.3 | 16.5 | 21.54 | 25.43 | 107.2 | 20.3 | 21.12 | 10.8 | 4.8 |

Table 1: Generation quality (BLEU \uparrow) and decoding efficiency (tokens/sec \uparrow for translation, images/sec \uparrow for image captioning). Decoding efficiency is measured sentence-by-sentence. AR: autoregressive models. b : beam width. i_{dec} : the number of refinement steps taken during decoding. Adaptive: the adaptive number of refinement steps. NAT: non-autoregressive transformer models (Gu et al., 2017). FT: fertility. NPD reranking using 100 samples.

P40 (WMT'14 En-De and WMT'16 En-Ro) or on a single P100 (IWSLT'16 En-De) with each minibatch consisting of approximately 2k tokens. We use four P100's to train non-autoregressive models on WMT'14 En-De.

Image Caption Generation: MS COCO We use MS COCO (Lin et al., 2014). We use the publicly available splits (Karpathy and Li, 2015), consisting of 113,287 training images, 5k validation images and 5k test images. We extract 49 512-dimensional feature vectors for each image, using a ResNet-18 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009). The average of these vectors is copied as many times to match the length of the target sentence (reference during training and predicted during evaluation) to form the initial input to Decoder 1. We use the base transformer (Vaswani et al., 2017) except that n_{layer} is set to 4. We train each model on a single 1080ti with each minibatch consisting of approximately 1,024 tokens.

Target Length Prediction We formulate the target length prediction as classification, predicting the difference between the target and source lengths for translation and the target length for image captioning. All the hidden vectors from the n_{layer} layers of the encoder are summed and fed to a softmax classifier after affine transformation. We however do not tune the encoder's parameters for target length prediction. We use this length predictor only during test time. We find it important to accurately predict the target length for good overall performance. See Appendix A for an analysis on our length prediction model.

Training and Inference We use Adam (Kingma and Ba, 2014) and use $L = 3$ in Eq. (1) during training ($i_{\text{train}} = 4$ from hereon.) We use $p_{\text{DAE}} = 0.5$. We use the deterministic strategy for IWSLT'16 En-De, WMT'16 En-Ro and MS COCO, while the stochastic strategy is used for WMT'14 En-De. These decisions were made based on the validation set performance. After both the non-autoregressive sequence model and target length predictor are trained, we decode by first predicting the target length and then running iterative refinement steps until the outputs of consecutive iterations are the same (or Jaccard distance between consecutive decoded sequences is 1). To assess the effectiveness of this adaptive scheme, we also test a fixed number of steps (i_{dec}). In machine translation, we remove any repetition by collapsing multiple consecutive occurrences of a token.

7 Results and Analysis

We make some important observations in Table 1. First, the generation quality improves across all the tasks as we run more refinement steps i_{dec} even beyond that used in training ($i_{\text{train}} = 4$), which supports our interpretation as a conditional denoising autoencoder in Sec. 3.2. To further verify this, we run decoding on WMT'14 (both directions) up to 100 iterations. As shown in Fig. 1 (a), the quality improves well beyond the number of refinement steps used during training.

Second, the generation efficiency decreases as more refinements are made. We plot the average seconds per sentence in Fig. 1 (b), measured on GPU while sequentially decoding one sentence at a time. As expected, decoding from the autoregressive model linearly slows down as the sen-

| | | | En→De | | De→En | |
|--------------------|------------------|---------|-------|--------|-------|--------|
| i_{train} | p_{DAE} | distill | rep | no rep | rep | no rep |
| AR | $b = 1$ | | 28.64 | | 34.11 | |
| | $b = 4$ | | 28.98 | | 34.81 | |
| Our Models | 1 | 0 | 14.62 | 18.03 | 16.70 | 21.18 |
| | 2 | 0 | 17.42 | 21.08 | 19.84 | 24.25 |
| | 4 | 0 | 19.22 | 22.65 | 22.15 | 25.24 |
| | 4 | 1 | 19.83 | 22.29 | 24.00 | 26.57 |
| | 4 | 0.5 | 20.91 | 23.65 | 24.05 | 28.18 |
| | 4 | 0.5 | ✓ | 26.17 | 27.11 | 31.92 |

Table 2: Ablation study on the dev set of IWSLT’16.

tence length grows, while decoding from the non-autoregressive model with a fixed number of iterations has the constant complexity. However, the generation efficiency of non-autoregressive model decreases as more refinements are made. To make a smooth trade-off between the quality and speed, the adaptive decoding scheme allows us to achieve near-best generation quality with a significantly lower computational overhead. Moreover, the adaptive decoding scheme automatically increases the number of refinement steps as the sentence length increases, suggesting that this scheme captures the amount of information in the input well. The increase in latency is however less severe than that of the autoregressive model.

We also observe that the speedup in decoding is much clearer on GPU than on CPU. This is a consequence of highly parallel computation of the proposed non-autoregressive model, which is better suited to GPUs, showcasing the potential of using the non-autoregressive model with a specialized hardware for parallel computation, such as Google’s TPUs (Jouppi et al., 2017). The results of our model decoded with adaptive decoding scheme are comparable to the results from (Gu et al., 2017), without relying on any external tool. On WMT’14 En-De, the proposed model outperforms the best model from (Gu et al., 2017) by two points.

Lastly, it is encouraging to observe that the proposed non-autoregressive model works well on image caption generation. This result confirms the generality of our approach beyond machine translation, unlike that by Gu et al. (2017) which was for machine translation or by Oord et al. (2017) which was for speech synthesis.

Ablation Study We use IWSLT’16 En-De to investigate the impact of different number of refinement steps during training (denoted as i_{train}) as well as probability of using denoising autoencoder objective during training (denoted as p_{DAE}). The

| | | IWSLT’16 (En→) | | WMT’14 (En→) | |
|------------|---------|----------------|--------|--------------|--------|
| stochastic | distill | rep | no rep | rep | no rep |
| ✓ | | 23.65 | | 7.56 | |
| | | 22.80 | | 16.56 | |
| ✓ | ✓ | 27.11 | | 18.91 | |
| | ✓ | 25.39 | | 21.22 | |

Table 3: Deterministic and stochastic approximation

results are presented in Table 2.

First, we observe that it is beneficial to use multiple iterations of refinement during training. By using four iterations (one step of decoder 1, followed by three steps of decoder 2), the BLEU score improved by approximately 1.5 points in both directions. We also notice that it is necessary to use the proposed hybrid learning strategy to maximize the improvement from more iterations during training ($i_{\text{train}} = 4$ vs. $i_{\text{train}} = 4, p_{\text{DAE}} = 1.0$ vs. $i_{\text{train}} = 4, p_{\text{DAE}} = 0.5$.) Knowledge distillation was crucial to close the gap between the proposed deterministic non-autoregressive sequence model and its autoregressive counterpart, echoing the observations by Gu et al. (2017) and Oord et al. (2017). Finally, we see that removing repeating consecutive symbols improves the quality of the best trained models ($i_{\text{train}} = 4, p_{\text{DAE}} = 0.5$) by approximately +1 BLEU. This suggests that the proposed iterative refinement is not enough to remove repetitions on its own. Further investigation is necessary to properly tackle this issue, which we leave as a future work.

We then compare the deterministic and stochastic approximation strategies on IWSLT’16 En→De and WMT’14 En→De. According to the results in Table 3, the stochastic strategy is crucial with a large corpus (WMT’14), while the deterministic strategy works as well or better with a small corpus (IWSLT’16). Both of the strategies benefit from knowledge distillation, but the gap between the two strategies when the dataset is large is much more apparent without knowledge distillation.

7.1 Qualitative Analysis

Machine Translation In Table 4, we present three sample translations and their iterative refinement steps from the development set of IWSLT’16 (De→En). As expected, the sequence generated from the first iteration is a rough version of translation and is iteratively refined over multiple steps. By inspecting the underlined sub-sequences, we see that each iteration does not monotonically improve the translation, but overall modifies the

| | | |
|------------|---|---|
| Src | seitdem habe ich sieben Häuser in der Nachbarschaft mit den Lichtern versorgt und sie funktionieren wirklich gut . | |
| Iter 1 | and I 've <u>been seven homes since in</u> neighborhood with the lights and they 're really functional . | |
| Iter 2 | and I 've <u>been seven homes in the</u> neighborhood with the lights , and they 're a really functional . | |
| Iter 4 | and I 've <u>been seven homes in</u> neighborhood with the lights , and they 're a really functional . | |
| Iter 8 | and I 've <u>been providing seven homes in the</u> neighborhood with the lights and they 're a really functional . | |
| Iter 20 | and I 've <u>been providing seven homes in the</u> neighborhood with the lights , and they 're a very good functional . | |
| Ref | since now , I 've set up seven homes around my community , and they 're really working . | |
| Src | er sah sehr glücklich aus , was damals ziemlich ungewöhnlich war , da ihn die Nachrichten meistens deprimierten . | |
| Iter 1 | he looked very happy , which was pretty <u>unusual the</u> , because the news was were usually depressing . | |
| Iter 2 | he looked very happy , which was pretty <u>unusual at the</u> , because <u>the news was s</u> depressing . | |
| Iter 4 | he looked very happy , which was pretty <u>unusual at the</u> , because <u>news was mostly</u> depressing . | |
| Iter 8 | he looked very happy , which was pretty <u>unusual at the time</u> because <u>the news was mostly</u> depressing . | |
| Iter 20 | he looked very happy , which was pretty <u>unusual at the time</u> , because <u>the news was mostly</u> depressing . | |
| Ref | there was a big smile on his face which was unusual then , because the news mostly depressed him . | |
| Src | furchtlos zu sein heißt für mich , heute ehrlich zu sein . | Table 4: Three sample De→En translations from the non-autoregressive sequence model. Source sentences are from the dev set of IWSLT'16. The first iteration corresponds to Decoder 1, and from thereon, Decoder 2 is repeatedly applied. Sub-sequences with changes across the refinement steps are underlined. |
| Iter 1 | to be , <u>for me</u> , to be honest today . | |
| Iter 2 | to be <u>fearless , me , is</u> to be honest today . | |
| Iter 4 | to be <u>fearless for me</u> , is to be honest today . | |
| Iter 8 | to be <u>fearless for me , me</u> to be honest today . | |
| Iter 20 | to be <u>fearless for me , is</u> to be honest today . | |
| Ref | so today , for me , being fearless means being honest . | |

translation towards the reference sentence. Missing words are added, while unnecessary words are dropped. For instance, see the second example. The second iteration removes the unnecessary “were”, and the fourth iteration inserts a new word “mostly”. The phrase “at the time” is gradually added one word at a time.

Image Caption Generation Table 5 shows two examples of image caption generation. We observe that each iteration captures more and more details of the input image. In the first example (left), the bus was described only as a “yellow bus” in the first iteration, but the subsequent iterations refine it into “yellow and black bus”. Similarly, “road” is refined into “lot”. We notice this behavior in the second example (right) as well. The first iteration does not specify the place in which “a woman” is “standing on”, which is fixed immediately in the second iteration: “standing on a tennis court”. In the final and fourth iteration, the proposed model captures the fact that the “woman” is “holding” a racquet.

8 Conclusion

Following on the exciting, recent success of non-autoregressive neural sequence modeling by Gu et al. (2017) and Oord et al. (2017), we proposed a deterministic non-autoregressive neural sequence model based on the idea of iterative refinement. We designed a learning algorithm specialized to the proposed approach by interpreting the entire

model as a latent variable model and each refinement step as denoising.

We implemented our approach using the Transformer and evaluated it on two tasks: machine translation and image caption generation. On both tasks, we were able to show that the proposed non-autoregressive model performs closely to the autoregressive counterpart with significant speedup in decoding. Qualitative analysis revealed that the iterative refinement indeed refines a target sequence gradually over multiple steps.

Despite these promising results, we observed that proposed non-autoregressive neural sequence model is outperformed by its autoregressive counterpart in terms of the generation quality. The following directions should be pursued in the future to narrow this gap. First, we should investigate better approximation to the marginal log-probability. Second, the impact of the corruption process on the generation quality must be studied. Lastly, further work on sequence-to-sequence model architectures could yield better results in non-autoregressive sequence modeling.

Acknowledgement

We thank support by AdeptMind, eBay, TenCent and NVIDIA. This work was partly supported by Samsung Advanced Institute of Technology (Next Generation Deep Learning: from pattern recognition to AI) and Samsung Electronics (Improving Deep Learning using Latent Structure). We also thank Jiatao Gu for valuable feedback.



Generated Caption

Iter 1 a yellow bus parked on parked in of parking road .
 Iter 2 a yellow and black on parked in a parking lot .
 Iter 3 a yellow and black bus parked in a parking lot .
 Iter 4 a yellow and black bus parked in a parking lot .

Reference Captions

a tour bus is parked on the curb waiting
 city bus parked on side of hotel in the rain .
 bus parked under an awning next to brick sidewalk
 a bus is parked on the curb in front of a building .
 a double decked bus sits parked under an awning



Generated Caption

Iter 1 a woman standing on playing tennis on a tennis racquet .
 Iter 2 a woman standing on a tennis court a tennis racquet .
 Iter 3 a woman standing on a tennis court a racquet .
 Iter 4 a woman standing on a tennis court holding a racquet .

Reference Captions

a female tennis player in a black top playing tennis
 a woman standing on a tennis court holding a racquet .
 a female tennis player preparing to serve the ball .
 a woman is holding a tennis racket on a court
 a woman getting ready to reach for a tennis ball on the ground

Table 5: Two sample image captions from the proposed non-autoregressive sequence model. The images are from the development set of MS COCO. The first iteration is from decoder 1, while the subsequent ones are from decoder 2. Subsequences with changes across the refinement steps are underlined.

References

- Guillaume Alain and Yoshua Bengio. 2014. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1).
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Florian Bordes, Sina Honari, and Pascal Vincent. 2017. Learning to generate samples from noise through infusion training. In *ICLR*.
- Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjali Kannan, Ron J Weiss, Kanishka Rao, Katya Gonina, et al. 2017. State-of-the-art speech recognition with sequence-to-sequence models. *arXiv preprint arXiv:1712.01769*.
- Kyunghyun Cho. 2016. Noisy parallel approximate decoding for conditional recurrent language model. *arXiv preprint arXiv:1605.03835*.
- Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. 2015. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11).
- Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *NIPS*.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.
- Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *ACL*.
- David Grangier and Michael Auli. 2017. Quikedit: Editing text & translations via simple delete actions. *arXiv preprint arXiv:1711.04805*.
- Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
- Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *NAACL*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. 2017. Decoding as continuous optimization in neural machine translation. *arXiv preprint arXiv:1701.02854*.
- Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. 2017. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*.
- Łukasz Kaiser and Samy Bengio. 2016. Can active memory replace attention? In *NIPS*.
- Łukasz Kaiser and Ilya Sutskever. 2015. Neural GPUs learn algorithms. *arXiv preprint arXiv:1511.08228*.

- Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*.
- Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.
- Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved variational inference with inverse autoregressive flow. In *NIPS*.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*.
- Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.
- T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*.
- Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.
- Roman Novak, Michael Auli, and David Grangier. 2016. Iterative refinement for machine translation. *arXiv preprint arXiv:1610.06602*.
- Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.
- Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C Cobo, Florian Stimberg, et al. 2017. Parallel wavenet: Fast high-fidelity speech synthesis. *arXiv preprint arXiv:1711.10433*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.
- Holger Schwenk. 2012. Continuous space translation models for phrase-based statistical machine translation. *Proceedings of COLING 2012: Posters*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.
- Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
- Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *NIPS*.