Fast Decoding in Sequence Models Using Discrete Latent Variables

Łukasz Kaiser 1  Aurko Roy 1  Ashish Vaswani 1  Niki Parmar 1  Samy Bengio 1  Jakob Uszkoreit 1  Noam Shazeer 1

Abstract

Autoregressive sequence models based on deep neural networks, such as RNNs, Wavenet and the Transformer attain state-of-the-art results on many tasks. However, they are difficult to parallelize and are thus slow at processing long sequences. RNNs lack parallelism both during training and decoding, while architectures like WaveNet and Transformer are much more parallelizable during training, yet still operate sequentially during decoding. We present a method to extend sequence models using discrete latent variables that makes decoding much more parallelizable. We first autoencode the target sequence into a shorter sequence of discrete latent variables, which at inference time is generated autoregressively, and finally decode the output sequence from this shorter latent sequence in parallel. To this end, we introduce a novel method for constructing a sequence of discrete latent variables and compare it with previously introduced methods. Finally, we evaluate our model end-to-end on the task of neural machine translation, where it is an order of magnitude faster at decoding than comparable autoregressive models. While lower in BLEU than purely autoregressive models, our model achieves higher scores than previously proposed non-autoregressive translation models.

1. Introduction

Neural networks have been applied successfully to a variety of tasks involving natural language. In particular, recurrent neural networks (RNNs) with long short-term memory (LSTM) cells (Hochreiter & Schmidhuber, 1997) in a sequence-to-sequence configuration (Sutskever et al., 2014) have proven successful at tasks including machine translation (Sutskever et al., 2014; Bahdanau et al., 2014; Cho et al., 2014), parsing (Vinyals et al., 2015), and many others. RNNs are inherently sequential, however, and thus tend to be slow to execute on modern hardware optimized for parallel execution. Recently, a number of more parallelizable sequence models were proposed and architectures such as WaveNet (van den Oord et al., 2016), ByteNet (Kalchbrenner et al., 2016) and the Transformer (Vaswani et al., 2017) can indeed be trained faster due to improved parallelism.

When actually generating sequential output, however, their autoregressive nature still fundamentally prevents these models from taking full advantage of parallel computation. When generating a sequence $y_1 \ldots y_n$ in a canonical order, say from left to right, predicting the symbol $y_t$ first requires generating all symbols $y_1 \ldots y_{t-1}$ as the model predicts

$$P(y_t | y_{t-1} y_{t-2} \ldots y_1).$$

During training, the ground truth is known so the conditioning on previous symbols can be parallelized. But during decoding, this is a fundamental limitation as at least $n$ sequential steps need to be made to generate $y_1 \ldots y_n$.

To overcome this limitation, we propose to introduce a sequence of discrete latent variables $l_1 \ldots l_m$, with $m < n$, that summarizes the relevant information from the sequence $y_1 \ldots y_n$. We will still generate $l_1 \ldots l_m$ autoregressively, but it will be much faster as $m < n$ (in our experiments we mostly use $m = \frac{n}{8}$). Then, we reconstruct each position in the sequence $y_1 \ldots y_n$ from $l_1 \ldots l_m$ in parallel.

For the above strategy to work, we need to autoencode the target sequence $y_1 \ldots y_n$ into a shorter sequence $l_1 \ldots l_m$. Autoencoders have a long history in deep learning (Hinton & Salakhutdinov, 2006; Salakhutdinov & Hinton, 2009a; Vincent et al., 2010; Kingma & Welling, 2013). Autoencoders mostly operate on continuous representations, either by imposing a bottleneck (Hinton & Salakhutdinov, 2006), requiring them to remove added noise (Vincent et al., 2010), or adding a variational component (Kingma & Welling, 2013). In our case though, we prefer the sequence $l_1 \ldots l_m$ to be discrete, as we use standard autoregressive models to predict it. Despite some success (Bowman et al., 2016; Yang et al., 2017), predicting continuous latent representations does not work as well as the discrete case in our setting.

---

1Google Brain, Mountain View, California, USA. Correspondence to: Łukasz Kaiser <lukaszkaiser@google.com>, Aurko Roy <aurkor@google.com>.

Fast Decoding in Sequence Models Using Discrete Latent Variables

However, using discrete latent variables can be challenging when training models end-to-end. Three techniques recently have shown how to successfully use discrete variables in deep models: the Gumbel-Softmax (Jang et al., 2016; Maddison et al., 2016), VQ-VAE (van den Oord et al., 2017) and improved semantic hashing (Kaiser & Bengio, 2018). We compare all these techniques in our setting and introduce another one: decomposed vector quantization (DVQ) which performs better than VQ-VAE for large latent alphabet sizes.

Using either DVQ or improved semantic hashing, we are able to create a neural machine translation model that achieves good BLEU scores on the standard benchmarks while being an order of magnitude faster at decoding time than autoregressive models. A recent paper (Gu et al., 2017) reported similar gain for neural machine translation. But when training models end-to-end, so it can be applied to any sequence problem. Despite being more generic, our model outperforms the hand-tuned technique from (Gu et al., 2017) yielding better BLEU. To summarize, our main contributions are:

1. A method for fast decoding for autoregressive models.
2. An improved discretization technique: the DVQ.
3. The resulting Latent Transformer model, achieving good results on translation while decoding much faster.

2. Discretization Techniques

In this section we introduce various discretization bottlenecks used to train discrete autoencoders for the target sequence. We will use the notation from (van den Oord et al., 2017) where the target sequence \( y \) is passed through an encoder, \( \text{enc} \), to produce a continuous latent representation \( \text{enc}(y) \in R^D \), where \( D \) is the dimension of the latent space. Let \( K \) be the size of the discrete latent space and let \( [K] \) denote the set \( \{1, 2, \ldots, K\} \). The continuous latent \( \text{enc}(y) \) is subsequently passed through a discretization bottleneck to produce a discrete latent representation \( z_d(y) \in [K] \), and an input \( z_q(y) \) to be passed to the decoder \( \text{dec} \). For integers \( i, m \) we will use \( \tau_m(i) \) to denote the binary representation of \( i \) using \( m \) bits, with the inverse operation, i.e. conversion from binary to decimal denoted by \( \tau_m^{-1} \).

2.1. Gumbel-Softmax

A discretization technique that has recently received a lot of interest is the Gumbel-Softmax trick proposed by (Jang et al., 2016; Maddison et al., 2016). In this case one simply projects the encoder output \( \text{enc}(y) \) using a learnable projection \( W \in R^K \times D \) to get the logits \( l = W \text{enc}(y) \) with the discrete code \( z_d(y) \) being defined as

\[
z_d(y) = \arg \max_{i \in [K]} l_i. \tag{1}
\]

The decoder input \( z_q(y) \in R^D \) during evaluation and inference is computed using an embedding \( e \in R^K \times D \) where \( z_q(y) = e_j \), where \( j = z_d(y) \). For training, the Gumbel-Softmax trick is used by generating samples \( g_1, g_2, \ldots, g_K \) i.i.d samples from the Gumbel distribution: \( g_i \sim -\log(-\log u) \), where \( u \sim U(0, 1) \) are uniform samples. Then as in (Jang et al., 2016; Maddison et al., 2016), one computes the log-softmax of \( l \) to get \( w \in R^K \):

\[
w_i = \frac{\exp((l_i + g_i)/\tau)}{\sum_j \exp((l_j + g_j)/\tau)}, \tag{2}
\]

with the input to the decoder \( z_q(y) \) being simply the matrix-vector product \( w e \). Note that the Gumbel-Softmax trick makes the model differentiable and thus it can be trained using backpropagation.

For low temperature \( \tau \) the vector \( w \) is close to the 1-hot vector representing the maximum index of \( l \), which is what is used during evaluation and testing. But at higher temperatures, it is an approximation (see Figure 1 in Jang et al. (2016)).

2.2. Improved Semantic Hashing

Another discretization technique proposed by (Kaiser & Bengio, 2018) that has been recently explored stems from semantic hashing (Salakhutdinov & Hinton, 2009b). The main idea behind this technique is to use a simple rounding bottleneck after squashing the encoder state \( z_q(y) \) using a saturating sigmoid. Recall the saturating sigmoid function from (Kaiser & Sutskever, 2016; Kaiser & Bengio, 2016):

\[
\sigma'(x) = \max(0, \min(1, 1.2\sigma(x) - 0.1)). \tag{3}
\]

During training, a Gaussian noise \( \eta \sim \mathcal{N}(0, 1)^D \) is added to \( z_q(y) \) which is then passed through a saturating sigmoid to get the vector \( f_e(y) \):

\[
f_e(y) = \sigma'(z_e(y) + \eta). \tag{4}
\]

To compute the discrete latent representation, the binary vector \( g_e(y) \) is constructed via rounding, i.e.:

\[
g_e(y)_i = \begin{cases} 1 & \text{if } f_e(y)_i > 0.5 \\ 0 & \text{otherwise}, \end{cases} \tag{5}
\]

with the discrete latent code \( z_d(y) \) corresponding to \( \tau_{\log R^K}(g(y)) \). The input to the decoder \( z_q(y) \in R^D \) is computed using two embedding spaces \( e^1, e^2 \in R^K \times D \), with \( z_q(y) = e^1_{h_e} + e^2_{1-h_e}(y) \), where the function \( h_e \) is randomly chosen to be \( f_e \) or \( g_e \) half of the time during training, while \( h_e \) is set equal to \( g_e \) during inference.
2.3. Vector Quantization

The Vector Quantized - Variational Autoencoder (VQ-VAE) discretization bottleneck method was proposed in (van den Oord et al., 2017). Note that vector quantization based methods have a long history of being used successfully in various Hidden Markov Model (HMM) based machine learning models (see e.g., (Huang & Jack, 1989; Lee et al., 1989)). In VQ-VAE, the encoder output $\text{enc}(y) \in \mathbb{R}^D$ is passed through a discretization bottleneck using a nearest-neighbor lookup on embedding vectors $e \in \mathbb{R}^{K \times D}$.

More specifically, the decoder input $z_q(y)$ is defined as

$$z_q(y) = e_k, \quad k = \arg \min_{j \in [K]} \| \text{enc}(y) - e_j \|_2,$$  \hspace{1cm} (6)

The corresponding discrete latent $z_d(y)$ is then the index $k$ of the embedding vector closest to $\text{enc}(y)$ in $\ell_2$ distance. Let $l_r$ be the reconstruction loss of the decoder $\text{dec}$ given $z_q(y)$, (e.g., the cross entropy loss); then the model is trained to minimize

$$L = l_r + \beta \| \text{enc}(y) - \text{sg}(z_q(y)) \|_2,$$  \hspace{1cm} (7)

where $\text{sg}(\cdot)$ is the stop gradient operator defined as follows:

$$\text{sg}(x) = \begin{cases} x & \text{forward pass} \\ 0 & \text{backward pass} \end{cases}$$  \hspace{1cm} (8)

We maintain an exponential moving average (EMA) over the following two quantities: 1) the embeddings $e_j$ for every $j \in [K]$ and, 2) the count $c_j$ measuring the number of encoder hidden states that have $e_j$ as it’s nearest neighbor. The counts are updated over a mini-batch of targets $\{y_1, \ldots, y_l, \ldots\}$ as:

$$c_j \leftarrow \lambda c_j + (1 - \lambda) \sum_l 1[z_q(y_l) = e_j],$$  \hspace{1cm} (9)

with the embedding $e_j$ being subsequently updated as:

$$e_j \leftarrow \lambda e_j + (1 - \lambda) \sum_l 1[z_q(y_l) = e_j] \frac{\text{enc}(y_l)}{c_j},$$  \hspace{1cm} (10)

where $1[\cdot]$ is the indicator function and $\lambda$ is a decay parameter which we set to 0.999 in our experiments.

2.4. Decomposed Vector Quantization

When the size of the discrete latent space $K$ is large, then an issue with the approach of Section 2.3 is index collapse, where only a few of the embedding vectors get trained due to a rich getting richer phenomena. In particular, if an embedding vector $e_j$ is close to a lot of encoder outputs $\text{enc}(y_1), \ldots, \text{enc}(y_l)$, then it receives the strongest signal to get even closer via the EMA update of Equations (9) and (10). Thus only a few of the embedding vectors will end up actually being used. To circumvent this issue, we propose two variants of decomposing VQ-VAE that make more efficient use of the embedding vectors for large values of $K$.

2.4.1. Sliced Vector Quantization

The main idea behind this approach is to break up the encoder output $\text{enc}(y)$ into $n_d$ smaller slices

$$\text{enc}^1(y) \circ \text{enc}^2(y) \cdots \circ \text{enc}^{n_d}(y),$$  \hspace{1cm} (11)

where each $\text{enc}^i(y)$ is a $D/n_d$ dimensional vector and $\circ$ denotes the concatenation operation. Corresponding to each $\text{enc}^i(y)$ we have an embedding space $e_i \in \mathbb{R}^{K_i \times D/n_d}$, where $K_i = 2^{(\log_2 K)/n_d}$. Note that the reason for the particular choice of $K_i$ is information theoretic: using an embedding space of size $K$ from Section 2.3 allows us to express discrete codes of size $\log_2 K$. In the case when we have $n_d$ different slices, we want the total expressible size of the discrete code to be still $\log_2 K$ and so $K_i$ is set to $2^{(\log_2 K)/n_d}$. We now compute nearest neighbors for each subspace as:

$$z_q^i(y) = e_k^i, \quad k_i = \arg \min_{j \in [K_i]} \| \text{enc}^i(y) - e_j^i \|_2,$$  \hspace{1cm} (12)

with the decoder input $z_q(y) = z_q^1(y) \circ \cdots \circ z_q^{n_d}(y)$.

The training objective $L$ is the same as in Section 2.3, with each embedding space $e_i$ trained individually via EMA updates from $\text{enc}^i(y)$ over a mini-batch of targets $\{y_1, \ldots, y_l, \ldots\}$:

$$e_j^i \leftarrow \lambda e_j^i + (1 - \lambda) \sum_l 1[z_q^i(y_l) = e_j^i] \quad \text{enc}^i(y_l),$$  \hspace{1cm} (13)

$$e_j^i \leftarrow \lambda e_j^i + (1 - \lambda) \sum_l 1[z_q^i(y_l) = e_j^i] \frac{\text{enc}^i(y_l)}{c_j^i},$$  \hspace{1cm} (14)

where $1[\cdot]$ is the indicator function as before, and $\lambda$ is the decay parameter.

Then the discrete latent code $z_d(y)$ is now defined as

$$z_d(y) = \tau_{\log_2 K}^{-1}(\tau_{\log_2 K_1}^{-1}(\tau_{\log_2 K_2}^{-1}(\cdots \tau_{\log_2 K_{n_d}}^{-1}(y)))).$$

(15)

Observe that when $n_d = 1$, the sliced Vector Quantization reduces to the VQ-VAE of (van den Oord et al., 2017). On the other hand, when $n_d = \log_2 K$, sliced DVQ is equivalent to improved semantic hashing of Section 2.2 loosely speaking: the individual table size $K_i$ for each slice $\text{enc}^i(y)$ is 2, and it gets rounded to 0 or 1 depending on which embedding is closer. However, the rounding bottleneck in semantic hashing of Section 2.2 proceeds via a saturating sigmoid and thus strictly speaking, the two techniques are different.
Note that similar decomposition approaches to vector quantization in the context of HMMs have been studied in the past under the name multiple code-books, see for instance (Huang et al., 1989; Rogina & Waibel, 1994; Peinado et al., 1996). The approach of sliced Vector Quantization has also been studied more recently in the context of clustering, under the name of Product or Cartesian Quantization in (Jegou et al., 2011; Norouzi & Fleet, 2013). A more recent work (Shu & Nakayama, 2018) explores a similar quantization approach coupled with the Gumbel-Softmax trick to learn compressed word embeddings (see also (Lam, 2018)).

2.4.2. Projected Vector Quantization

Another natural way to decompose Vector Quantization is to use a set of fixed randomly initialized projections

$$\pi_i \in \mathbb{R}^{D \times D/n_d} \mid i \in [n_d]$$

(16)

to project the encoder output $\text{enc}(y)$ into a $\mathbb{R}^{D/n_d}$-dimensional subspace. For $\text{enc}(y) = \pi(y) \in \mathbb{R}^{D/n_d}$ we have an embedding space $e^{(i)} \in \mathbb{R}^{K_i \times D/n_d}$, where $K_i = 2^{(\log_2 K)/n_d}$ as before. The training objective, embeddings update, the input $z_d(y)$ to the decoder, and the discrete latent representation $z_d(y)$ is computed exactly as in Section 2.4.1. Note that when $n_d = 1$, projected Vector Quantization reduces to the VQ-VAE of (van den Oord et al., 2017) with an extra encoder layer corresponding to the projections $\pi_i$. Similarly, when $n_d = \log_2 K$, projected DVQ is equivalent to improved semantic hashing of Section 2.2 with the same analogy as in Section 2.4.1, except the encoder now has an extra layer. The VQ-VAE paper (van den Oord et al., 2017) also use multiple latents in the experiments reported on CIFAR-10 and in Figure 5, using an approach similar to what we call projected DVQ.

3. Latent Transformer

Using the discretization techniques from Section 2 we can now introduce the Latent Transformer (LT) model. Given an input-output pair $(x, y) = (x_1, \ldots, x_k, y_1, \ldots, y_n)$ the LT will make use of the following components.

- The function $a(e(y, x))$ will autoencode $y$ into a shorter sequence $l = l_1, \ldots, l_m$ of discrete latent variables using the discretization bottleneck from Section 2.
- The latent prediction model $l_{lp}(x)$ (a Transformer) will autoregressively predict $l$ based on $x$.
- The decoder $ad(l, x)$ is a parallel model that will decode $y$ from $l$ and the input sequence $x$.

The functions $a(e(y, x))$ and $ad(l, x)$ together form an autoencoder of the targets $y$ that has additional access to the input sequence $x$. For the autoregressive latent prediction we use a Transformer (Vaswani et al., 2017), a model based on multiple self-attention layers that was originally introduced in the context of neural machine translation. In this work we focused on the autoencoding functions and did not tune the Transformer: we used all the defaults from the baseline provided by the Transformer authors (6 layers, hidden size of 512 and filter size of 4096) and only varied parameters relevant to $ae$ and $ad$, which we describe below. The three components above give rise to two losses:

- The autoencoder reconstruction loss $l_r$ coming from comparing $ad(a(e(y, x), x))$ to $y$.
- The latent prediction loss $l_{lp}$, that comes from comparing $l = ae(y, x)$ to the generated $lp(x)$.

We train the LT model by minimizing $l_r + l_{lp}$. Note that the final outputs $y$ are generated only depending on the latents $l$ but not on each other, as depicted in Figure 1. In an autoregressive model, each $y_i$ would have a dependence on all previous $y_j, j < i$, as is the case for $l$s in Figure 1.

The function $ae(y, x)$. The autoencoding function $ae(y, x)$ we use is a stack of residual convolutions followed by an attention layer attending to $x$ and a stack of strided convolutions. We first apply to $y$ a 3-layer block of 1-dimensional convolutions with kernel size 3 and padding with 0s on both sides (SAME-padding). We use ReLU non-linearities between the layers and layer-normalization (Ba et al., 2016). Then, we add the input to the result, forming a residual block. Next we have an encoder-decoder attention layer with dot-product attention, same as in (Vaswani et al., 2017), with a residual connection. Finally, we process the result with a convolution with kernel size 2 and stride 2, effectively halving the size of $s$. We do this strided processing $c$ times so as to decrease the length $C = 2^c$ times
(later $C = \frac{n}{m}$). The result is put through the discretization bottleneck of Section 2. The final compression function is given by $\text{ae}(y,x) = z_q(y)$ and the architecture described above is depicted in Figure 2.

![Figure 2. Architecture of ae(y, x).](image)

The function $\text{ad}(l, x)$. To decode from the latent sequence $l = \text{ae}(y, x)$, we use the function $\text{ad}(l, x)$. It consists of $c$ steps that include up-convolutions that double the length, so effectively it increases the length $2^c = C = \frac{n}{m}$ times. Each step starts with the residual block, followed by an encoder-decoder attention to $x$ (both as in the ae function above). Then it applies an up-convolution, which in our case is a feed-forward layer (equivalently a kernel-1 convolution) that doubles the internal dimension, followed by a reshape to twice the length. The result after the $c$ steps is then passed to a self-attention decoder, same as in the Transformer model (Vaswani et al., 2017).

Note that at the beginning of training (for the first 10K steps), we give the true targets $y$ to the transformer-decoder here, instead of the decompressed latents $l$. This pre-training ensures that the self-attention part has reasonable gradients that are then back-propagated to the convolution stack and then back to the ae function and the discretization bottleneck of Section 2.

4. Related Work

**Neural Machine Translation.** Machine translation using deep neural networks achieved great success with sequence-to-sequence models (Sutskever et al., 2014; Bahdanau et al., 2014; Cho et al., 2014) that used recurrent neural networks (RNNs) with LSTM cells (Hochreiter & Schmidhuber, 1997). The basic sequence-to-sequence architecture is composed of an RNN encoder which reads the source sentence one token at a time and transforms it into a fixed-sized state vector. This is followed by an RNN decoder, which generates the target sentence, one token at a time, from the state vector. While a pure sequence-to-sequence recurrent neural network can already obtain good translation results (Sutskever et al., 2014; Cho et al., 2014), it suffers from the fact that the whole input sentence needs to be encoded into a single fixed-size vector. This clearly manifests itself in the degradation of translation quality on longer sentences and was overcome in (Bahdanau et al., 2014) by using a neural model of attention. Convolutional architectures have been used to obtain good results in word-level neural machine translation starting from (Kalchbrenner & Blunsom, 2013) and later in (Meng et al., 2015). These early models used a standard RNN on top of the convolution to generate the output, which creates a bottleneck and hurts performance. Fully convolutional neural machine translation without this bottleneck was first achieved in (Kaiser & Bengio, 2016) and (Kalchbrenner et al., 2016). The model in (Kaiser & Bengio, 2016) (Extended Neural GPU) used a recurrent stack of gated convolutional layers, while the model in (Kalchbrenner et al., 2016) (ByteNet) did away with recursion and used left-padded convolutions in the decoder. This idea, introduced in WaveNet (van den Oord et al., 2016), significantly improves efficiency of the model. The same technique was improved in a number of neural translation models recently, including (Gehring et al., 2017) and (Kaiser et al., 2017). Instead of convolutions, one can use stacked self-attention layers. This was introduced in the Transformer model (Vaswani et al., 2017) and has significantly improved state-of-the-art in machine translation while also improving the speed of training. Thus, we use the Transformer model as a baseline in this work.

**Autoencoders and discretization bottlenecks.** Variational autoencoders were first introduced in (Kingma & Welling, 2013; Rezende et al., 2014), however training them for discrete latent variable models has been challenging. The NVIL estimator of (Mnih & Gregor, 2014) proposes using a single sample objective to optimize the variational lower bound, while VIMCO (Mnih & Rezende, 2016) proposes using a multi-sample objective of (Burda et al., 2015) which...
Fast Decoding in Sequence Models Using Discrete Latent Variables

further speeds up convergence by using multiple samples from the inference network. There have also been several discretization bottlenecks proposed recently that have been used successfully in various learning tasks, see Section 2 for a more detailed description of the techniques directly relevant to this work. Other recent works with similar approach to autoencoding include (Subakan et al., 2018).

**Non-autoregressive Neural Machine Translation.** Much of the recent state of the art models in Neural Machine Translation are auto-regressive, meaning that the model consumes previously generated tokens to predict the next one. A recent work that attempts to speed up decoding by training a non-autoregressive model is (Gu et al., 2017). The approach of (Gu et al., 2017) is to use the self-attention Transformer model of (Vaswani et al., 2017), together with the REINFORCE algorithm (Williams, 1992) to model the *fertilities* of words to tackle the multi-modality problem in translation. However, the main drawback of this work is the need for extensive fine-tuning to make policy gradients work for REINFORCE, as well as the issue that this approach only works for machine translation and is not generic, so it cannot be directly applied to other sequence learning tasks.

**Graphical models.** The core of our approach to fast decoding consists of finding a sequence $l$ of latent variables such that we can predict the output sequence $y$ in parallel from $l$ and the input $x$. In other words, we assume that each token $y_i$ is conditionally independent of all other tokens $y_j$ ($j \neq i$) given $l$ and $x$: $y_i \perp y_j \mid l, x$. Our autoencoder is thus learning to create a one-layer graphical model with $m$ variables ($l_1 \ldots l_m$) that can then be used to predict $y_1 \ldots y_n$ independently of each other.

**5. Experiments**

We train the Latent Transformer with the base configuration to make it comparable to both the autoregressive baseline (Vaswani et al., 2017) and to the recent non-autoregressive NMT results (Gu et al., 2017). We used around 33K subword units as vocabulary and implemented our model in TensorFlow (Abadi et al., 2015). Our implementation, together with hyper-parameters and everything needed to reproduce our results is available as open-source$^1$.

For non-autoregressive models, it is beneficial to generate a number of possible translations and re-score them with an autoregressive model. This can be done in parallel, so it is still fast, and it improves performance. This is called *noisy parallel decoding* in (Gu et al., 2017) and we include results both with and without it. The best BLEU scores obtained by different methods are summarized in Table 1. As you can see, our method with re-scoring almost matches the baseline autoregressive model without beam search.

To get a better understanding of the non-autoregressive models, we focus on performance without rescoring and investigate different variants of the Latent Transformer. We include different discretization bottlenecks, and report the final BLEU scores together with decoding speeds in Table 2. The LT is slower in non-batch mode than the simple NAT baseline of (Gu et al., 2017), which might be caused by system differences (our code is in TensorFlow and has not been optimized, while their implementation is in Torch). Latency at higher batch-size is much smaller, showing that the speed of the LT can still be significantly improved with batching. The choice of the discretization bottleneck seems to have a small impact on speed and both DVQ and improved semantic hashing yield good BLEU scores, while VQ-VAE fails in this context (see below for a discussion).

<table>
<thead>
<tr>
<th><strong>Model</strong></th>
<th><strong>BLEU</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Transformer [1]</td>
<td>27.3</td>
</tr>
<tr>
<td>Baseline Transformer [2]</td>
<td>23.5</td>
</tr>
<tr>
<td>Baseline Transformer [2] (no beam-search)</td>
<td>22.7</td>
</tr>
<tr>
<td>DT (no NPD) [2]</td>
<td>17.7</td>
</tr>
<tr>
<td>DT without rescoring ($n_d = 8$)</td>
<td>19.8</td>
</tr>
<tr>
<td>DT with rescoring top-10 ($n_d = 8$)</td>
<td>18.7</td>
</tr>
<tr>
<td>DT with rescoring top-100 ($n_d = 8$)</td>
<td>21.0</td>
</tr>
</tbody>
</table>

Table 1. BLEU scores (the higher the better) on the WMT English-German translation task on the newstest2014 test set. The acronym NAT corresponds to the Non-Autoregressive Transformer, while FT denotes an additional Fertility Training and NPD denotes Nosy Parallel Decoding, all of them from (Gu et al., 2017). The acronym LT denotes the Latent Transformer from Section 3. Results reported for LT are from this work, the others are from (Gu et al., 2017) [2] except for the first baseline Transformer result which is from (Vaswani et al., 2017) [1].

**6. Discussion**

Since the discretization bottleneck is critical to obtaining good results for fast decoding of sequence models, we focused on looking into it, especially in conjunction with the size $K$ of the latent vocabulary, the dimension $D$ of the latent space, and the number of decompositions $n_d$ for DVQ.

An issue with the VQ-VAE of (van den Oord et al., 2017) that motivated the introduction of DVQ in Section 2.3 is *index collapse*, where only a few embeddings are used and subsequently trained. This can be visualized in the histogram of Figure 4, where the $x$-axis corresponds to the possible values of the discrete latents (in this case $\{1, \ldots, K\}$), and the $y$-axis corresponds to the training progression of

$^1$The code is available under redacted.
Fast Decoding in Sequence Models Using Discrete Latent Variables

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (no beam-search)</td>
<td>22.7</td>
<td>408 ms</td>
</tr>
<tr>
<td>NAT</td>
<td>17.7</td>
<td>39 ms</td>
</tr>
<tr>
<td>NAT+NPD=10</td>
<td>18.7</td>
<td>79 ms</td>
</tr>
<tr>
<td>NAT+NPD=100</td>
<td>19.2</td>
<td>257 ms</td>
</tr>
<tr>
<td>LT, Improved Semhash</td>
<td>19.8</td>
<td>105 ms</td>
</tr>
<tr>
<td>LT, VQ-VAE</td>
<td>2.78</td>
<td>148 ms</td>
</tr>
<tr>
<td>LT, s-DVQ</td>
<td>19.7</td>
<td>177 ms</td>
</tr>
<tr>
<td>LT, p-DVQ</td>
<td>19.8</td>
<td>182 ms</td>
</tr>
</tbody>
</table>

Table 2. BLEU scores and decode times on the WMT English-German translation task on the newstest2014 test set for different variants of the LT with \( \frac{d}{m} = 8 \) and \( D = 512 \) and \( n_d = 2 \) with s-DVQ and p-DVQ representing sliced and projected DVQ respectively. The LT model using improved semantic hashing from Section 2.2 uses \( \log_2 K = 14 \), while the one using VQ-VAE and DVQ from Sections 2.3 and 2.4 uses \( \log_2 K = 16 \). For comparison, we include the baselines from (Gu et al., 2017). We report the time to decode per sentence averaged over the whole test set as in (Gu et al., 2017); decoding is implemented in Tensorflow on a Nvidia GeForce GTX 1080. The batch size used during decoding is denoted by \( b \) and we report both \( b = 1 \) and \( b = 64 \).

The other variables for DVQ are the choice of the decomposition, and the number \( n_d \) of decompositions. For the projected DVQ, we use fixed projections \( \pi_i \)'s initialized using the Glorot initializer (Glorot & Bengio, 2010). We also found that the optimal number of decompositions for our choice of latent vocabulary size \( \log_2 K = 14 \) and 16 was \( n_d = 2 \), with \( n_d = 1 \) (i.e., regular VQ-VAE) performing noticeably worse (see Table 2 and Figure 4). Setting higher values of \( n_d \) led to a decline in performance, possibly because the expressive power \( \left( \log_2 K \right) \) was reduced for each decomposition, and the model also ended up using fewer latents (see Table 3).

Another important point about LT is that it allows making different trade-offs by tuning the \( \frac{d}{m} \) fraction of the length of the original output sequence to the length of the latent sequence. As \( \frac{d}{m} \) increases, so does the parallelism and decoding speed, but the latents need to encode more and more information to be able to decode the outputs in parallel. To study this tradeoff, we measure the reconstruction loss (the perplexity of the reconstructed \( y \) vs the original) for different \( \frac{d}{m} \) and varying the number of bits in the latent variables. The results, presented in Table 4, show clearly that reconstruction get better, as expected, if the latent state has more bits or is used to compress less subword units.

7. Conclusions

Autoregressive sequence models based on deep neural networks were made successful due to their applications in machine translation (Sutskever et al., 2014) and have since yielded state-of-the-art results on a number of tasks. With models like WaveNet and Transformer, it is possible to train them fast in a parallel way, which opened the way to applications to longer sequences, such as WaveNet for sound generation (van den Oord et al., 2016) or Transformer for long text summarization (Liu et al., 2018) and image generation (Vaswani et al., 2018). The key problem appearing in these new applications is the slowness of decoding: it is not practical to wait for minutes to generate a single example. In this work, while still focusing on the original problem of machine translation, we lay the groundwork for fast decoding for sequence models in general. While the latent transformer does not yet recover the full performance of the autoregressive model, it is already an order of magni-
Table 3. Percentage of latent codewords used by the Decomposed Vector Quantization (DVQ) of Section 2.4.1 for \( \log_2 K = 16 \) and \( D = 512 \) after 500,000 steps. Note that when \( n_d = 1 \), i.e. for vanilla VQ-VAE, only 5% of the available \( 2^{16} \) discrete latents (roughly 3000) are used. The latent usage is maximized for \( n_d = 2 \).

<table>
<thead>
<tr>
<th>( \frac{n}{m} )</th>
<th>( \log_2 K = 8 )</th>
<th>( \log_2 K = 14 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.33</td>
<td>0.64</td>
</tr>
<tr>
<td>4</td>
<td>2.04</td>
<td>1.26</td>
</tr>
<tr>
<td>8</td>
<td>2.44</td>
<td>1.77</td>
</tr>
</tbody>
</table>

Table 4. Log-perplexities of autoencoder reconstructions on the development set (newstest2013) for different values of \( \frac{n}{m} \) and numbers of bits in latent variables (LT trained for 250K steps).

... faster and performs better than a heavily hand-tuned, task-specific non-autoregressive model.

In the future, we plan to improve both the speed and the accuracy of the latent transformer. A simple way to improve speed that we did not yet try is to use the methods from this work in a hierarchical way. As illustrated in Figure 1, the latents are still generated autoregressively which takes most of the time for longer sentences. In the future, we will apply the LT model to generate the latents in a hierarchical manner, which should result in further speedup. To improve the BLEU scores, on the other hand, we intend to investigate methods related to Gibbs sampling or even make the model partially autoregressive. For example, one could generate only the odd-indexed outputs, \( y_1, y_3, y_5, \ldots \), based on the latent symbols \( l \), and then generate the even-indexed ones based on both the latents and the odd-indexed outputs. We believe that including such techniques has the potential to remove the gap between fast-decoding models and purely autoregressive ones and will lead to many new applications.

References


Fast Decoding in Sequence Models Using Discrete Latent Variables


