Variational Attention for Sequence-to-Sequence Models

Hareesh Bahuleyan[*], Lili Mou[*], Olga Vechtomova, Pascal Poupart
University of Waterloo, ON, Canada
{hpallika, ovechtomova, ppoupart}@uwaterloo.ca
doublepower.mou@gmail.com

Abstract

The variational encoder-decoder (VED) encodes source information as a set of random variables using a neural network, which in turn is decoded into target data using another neural network. In natural language processing, sequence-to-sequence (Seq2Seq) models typically serve as encoder-decoder networks. When combined with a traditional (deterministic) attention mechanism, the variational latent space may be bypassed by the attention model, making the variational space ineffective. In our paper, we propose a variational attention mechanism for VED, where the attention vector is modeled as Gaussian distributed random variables. Experiments show that, without loss of quality, our proposed method alleviates the bypassing phenomenon as it increases diversity of generated sentences.

1 Introduction

The variational autoencoder (VAE), proposed by Kingma and Welling (2013), *encodes* data to latent (random) variables, and then *decodes* the latent variables to reconstruct data. Theoretically, it optimizes a variational lower bound of the log-likelihood of data. Compared with traditional variational methods such as mean-field approximation (Wainwright et al., 2008), VAE leverages modern neural networks and hence is a more powerful density estimator. Compared with traditional autoencoders (Hinton and Salakhutdinov, 2006), which are *deterministic*, VAE populates hidden representations to a region (instead of a single point), making it possible to generate diversified data from the vector space (Bowman et al., 2016) or even control the generated samples (Hu et al., 2017).

In natural language processing (NLP), recurrent neural networks (RNNs) are typically used as both encoder and decoder, known as sequence-to-sequence (Seq2Seq) models. Although variational Seq2Seq models are much trickier to train in comparison to the image domain, Bowman et al. (2016) succeed in training a sequence-to-sequence VAE and generating sentences from a continuous latent space. Such an architecture can further be extended to variational encoder-decoder (VED) to transform one sequence into another utilizing the "variational" property (Serban et al., 2017; Zhou and Neubig, 2017).

When applying attention mechanisms (Bahdanau et al., 2014) to variational Seq2Seq models, however, we find the generated sentences are of less variety, implying that the variational latent space is ineffective. The attention mechanism summarizes source information as an *attention vector* by weighted sum, where the weights are a learned probabilistic distribution; then the attention vector is fed to the decoder. Evidence shows that attention significantly improves Seq2Seq performance in translation (Bahdanau et al., 2014), summarization (Rush et al., 2015), etc. In variational Seq2Seq, however, the attention mechanism unfortunately serves as a "bypassing" mechanism. In other words, the variational latent space does not need to learn much, as long as the attention mechanism itself is powerful enough to capture source information.

In this paper, we propose a variational attention mechanism to address this problem. We model the attention vector as random variables by imposing a probabilistic distribution. We follow traditional VAE and model the prior of the attention vector by a Gaussian distribution. We further propose two priors of the attention vector: Gaussian distribu-

[*] The first two authors contributed equally.

tions with a mean being (1) all zeros, and (2) the average of source information.

We evaluate our approach on a question generation task. Our goal is to generate diversified questions, given some information (a sentence) in a paragraph, with the variational mechanism. Experiments show that the proposed variational attention yields a higher diversity than variational Seq2Seq with deterministic attention, while retaining high quality of generated sentences.

2 Background and Motivation

In this section, we introduce the variational autoencoders and attention mechanism. We also present a pilot experiment motivating our variational attention model.

2.1 Variational Autoencoder (VAE)

VAE encodes data $\boldsymbol{Y}$ (e.g., a sentence) as hidden random variables $\boldsymbol{Z}$, based on which VAE reconstructs data $\boldsymbol{Y}$. Consider a generative model, parameterized by $\boldsymbol{\theta}$, as

$$p_{\boldsymbol{\theta}}(\boldsymbol{Z},\boldsymbol{Y}) = p_{\boldsymbol{\theta}}(\boldsymbol{Z})p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{Z}) \qquad (1)$$

Given a dataset $\mathcal{D} = \{\boldsymbol{y}^{(n)}\}_{n=1}^{N}$, the likelihood of a data point is

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)}) \geq \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)})}\left[\log\left\{\frac{p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)})}\right\}\right]$$
$$= \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)}|\boldsymbol{z})\right]$$
$$- \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)})\|p(\boldsymbol{z})\right) \triangleq \mathcal{L}^{(n)}(\boldsymbol{\theta},\boldsymbol{\phi}) \qquad (2)$$

VAE models both $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z})$ with neural networks, parametrized by $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively. Figure 1a shows the graphical model of this process. The training objective is to maximize the lower bound of the likelihood $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi})$, which can be rewritten as minimizing

$$J_{\text{rec}}(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{y}^{(n)}) + \mathrm{KL}\Big(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)})\|p(\boldsymbol{z})\Big) \qquad (3)$$

The first term, called *reconstruction loss*, is the (expected) negative log-likelihood of data, similar to traditional deterministic autoencoders. The expectation is obtained by Monte Carlo sampling. The second term is the KL-divergence between $\boldsymbol{z}$'s posterior and prior distributions. Typically the prior is set to standard normal $\mathcal{N}(\mathbf{0},\mathbf{I})$.

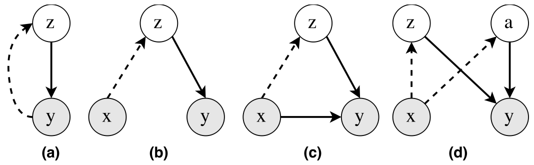

Figure 1: Graphical model representations. **(a)** Variational autoencoder (VAE). **(b)** Variational encoder-decoder (VED). **(c)** VED with deterministic attention (VED+DAttn). **(d)** VED with variational attention (VED+VAttn). **Dashed lines**: Encoding phase. **Solid lines**: Decoding phase.

2.2 Variational Encoder-Decoder (VED)

In some applications, we would like to transform source information to target information, e.g., machine translation, dialogue systems, and text summarization. In these tasks, "auto"-encoding is not sufficient, and an encoding-decoding framework is needed. Different efforts have been made to extend VAE to variational encoder-decoder (VED) frameworks, which transform an input $\boldsymbol{X}$ to output $\boldsymbol{Y}$. One possible extension is to condition all probabilistic distributions further on $\boldsymbol{X}$ (Zhang et al., 2016; Cao and Clark, 2017; Serban et al., 2017). This, however, introduces a discrepancy between training and prediction, since $\boldsymbol{Y}$ is not available during prediction.

Another approach is to build a recognition model on $\boldsymbol{X}$ (Zhou and Neubig, 2017). Taking the assumption that $\boldsymbol{Y}$ is a function of $\boldsymbol{X}$, i.e., $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{X})$, we have $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}) = q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{Y}(\boldsymbol{x})) \triangleq q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$. In this work, we follow Zhou and Neubig (2017) and adopt this extension. Figure 1b shows the graphical model of the VED used in our work.

2.3 Attention Mechanism

In NLP, sequence-to-sequence recurrent neural networks are typically used as the encoder and decoder, as they are suitable for modeling a sequence of words (i.e., sentence). Figure 2a shows a basic Seq2Seq model in the VAE/VED scenario (Bowman et al., 2016). The encoder has an input $\boldsymbol{x}$ and outputs $\boldsymbol{\mu}_z$ and $\boldsymbol{\sigma}_z$ as the parameters of $\boldsymbol{z}$'s posterior normal distribution. Then a decoder generates $\boldsymbol{y}$ based on a sample $\boldsymbol{z}$ drawn from its posterior distribution.

Attention mechanisms are proposed to dynamically align $\boldsymbol{y} = (y_1, \cdots, y_{|\boldsymbol{y}|})$ and $\boldsymbol{x} = (x_1, \cdots, x_{|\boldsymbol{x}|})$ during generation. At each time

Input: *the men are playing musical instruments*
(a) VAE w/o hidden state init. (Avg entropy: 2.52)
the men are playing musical instruments *the men are playing video games* *the musicians are playing musical instruments* *the women are playing musical instruments*
(b) VAE w/ hidden state init. (Avg entropy: 2.01)
the men are playing musical instruments *the men are playing musical instruments* *the men are playing musical instruments* *the man is playing musical instruments*

Table 1: Sentences obtained by sampling from the VAE latent space. (a) VAE without hidden state initialization. (b) VAE with hidden state initialization.

step j in the decoder, the attention mechanism computes a probabilistic distribution by

$$\alpha_{ji} = \frac{\exp\{\widetilde{\alpha}_{ji}\}}{\sum_{i'=1}^{|\boldsymbol{x}|} \exp\{\widetilde{\alpha}_{ji'}\}} \tag{4}$$

where $\widetilde{\alpha}_{ji}$ is a pre-normalized score, computed by $\widetilde{\alpha}_{ji} = \boldsymbol{h}_j^{(\text{tar})} W^T \boldsymbol{h}_i^{(\text{src})}$ in our model. Here, $\boldsymbol{h}_j^{(\text{tar})}$ and $\boldsymbol{h}_i^{(\text{src})}$ are the hidden representations of the jth step in target and ith in the source, and W is a learnable weight matrix.

Then the source information $\{\boldsymbol{h}_i^{(\text{src})}\}_{i=1}^{|\boldsymbol{x}|}$ is summed by weights α_{ji} to obtain the attention vector

$$\boldsymbol{a}_j = \sum_{i=1}^{|\boldsymbol{x}|} \alpha_{ji} \boldsymbol{h}_i^{(\text{src})} \tag{5}$$

which is fed to the decoder RNN at the jth step. Figure 2b shows the variational Seq2Seq model with such traditional attention.

2.4 "Bypassing" Phenomenon

In this part, we explain the "bypassing" phenomenon in VAE or VED, if the network is not designed properly; this motivates our variational attention described in Section 3.

We observe that, if the decoder has a direct, deterministic access to the encoder, the latent variables $\boldsymbol{Z}$ might not capture much information so that the VAE or VED does not play a role in the process. We call this a *bypassing phenomenon.*

Theoretically, if $q_\phi(\boldsymbol{Y}|\boldsymbol{Z})$ is aware of $\boldsymbol{X}$ by itself, then $q_\phi(\boldsymbol{Y}|\boldsymbol{Z})$ might be learned as $q_\phi(\boldsymbol{Y}|\boldsymbol{X})$ without hurting the reconstruction loss J_{rec}, but the KL term in Eq. (3) can be minimized. This degrades a variational Seq2Seq model to a deterministic one.

The phenomenon can be best shown with a bypassing connection between the encoder and decoder for hidden state initialization. Some previous studies set the decoder's initial state to be the encoder's final state (Cao and Clark, 2017), shown in Figure 2c. We conducted a pilot study with a Seq2Seq VAE using a subset (∼80k samples) of the massive dataset provided by Bowman et al. (2015). Table 1 shows examples of generated sentences and their entropy. We see that the variational Seq2Seq can only generate very similar sentences with such bypassing connections (Table 1b), as opposed to generating diversified samples from the latent space only (Table 1a). Quantitatively, the entropy decreases by 0.5 over 1k unseen samples on average, showing a significant difference since entropy is a logarithmic metric. This analysis provides design philosophy of neural architectures in VAE or VED.

Since attention largely improves model performance for deterministic Seq2Seq models, it is tempting to include attention in the variational Seq2Seq as well. However, our pilot experiment raises the doubt if a traditional attention mechanism, which is deterministic, may bypass the latent space in VED, as illustrated by a graphical model in Figure 1c. Also, evidence in Zheng et al. (2017) shows the attention mechanism is so powerful that removing other connections between the encoder and decoder has little effect on BLEU scores in machine translation. In other words, VED with deterministic attention might learn reconstruction mostly from attention, whereas the posterior of the latent space may fit to its prior so as to minimize the KL term.

To alleviate this problem, we propose a variational attention mechanism for variational Seq2Seq models, as is described in detail in the next section.

3 The Proposed Variational Attention

Let us consider the decoding process of an RNN. At a time step j, it adjusts its hidden state $\boldsymbol{h}_j^{(\text{tar})}$ with an input of a word embedding $\boldsymbol{y}_{j-1}$ (typically the groundtruth during training and the prediction from the previous step during testing). This is given by $\boldsymbol{h}_j^{(\text{tar})} = \text{RNN}_{\boldsymbol{\theta}}(\boldsymbol{h}_{j-1}^{(\text{tar})}, \boldsymbol{y}_{j-1})$. In our experiments, we use long short-term memory units (Hochreiter and

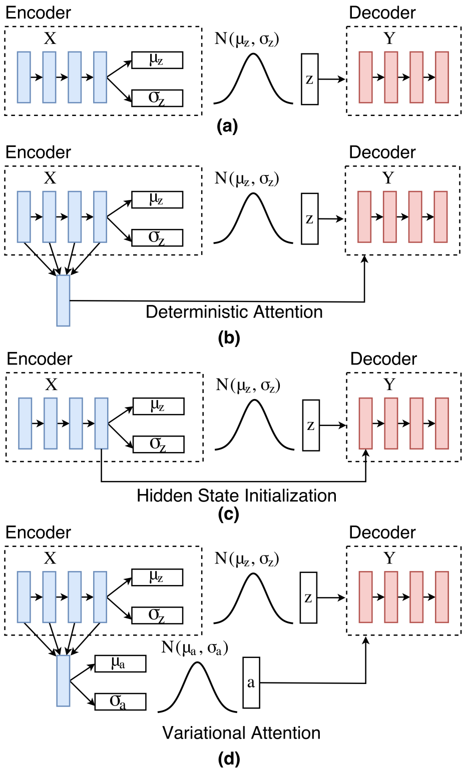

Figure 2: (a) Variational Seq2Seq model. (b) Variational Seq2Seq with deterministic attention. (c) Variational Seq2Seq with hidden state initialization. (d) Variational Seq2Seq with variational attention.

Schmidhuber, 1997) as RNN's transition. Enhanced with attention, the RNN is computed by $\boldsymbol{h}_j^{\text{(tar)}} = \text{RNN}_{\boldsymbol{\theta}}(\boldsymbol{h}_{j-1}^{\text{(tar)}}, \boldsymbol{y}_{j-1}, \boldsymbol{a}_j)$. The predicted word is given by a softmax layer $p(y_j) = \text{softmax}(W_{\text{out}}\boldsymbol{h}_j^{\text{(tar)}})$, where W_{out} is weight. As discussed earlier, traditional attention computes $\boldsymbol{a}_j$ in a deterministic fashion by Eq. (5).

To build a variational attention, we treat both traditional latent space $\boldsymbol{z}$ and the attention vector $\boldsymbol{a}_j$ as random variables. The recognition and reconstruction graphical models are shown in Figure 1d.

3.1 Lower Bound

Since the likelihood of the nth data point decomposes for different time steps, we consider the lower bound $\mathcal{L}_j^{(n)}(\boldsymbol{\theta}, \boldsymbol{\phi})$ at the jth step. The variational lower bound, i.e., Eq. (2), becomes

$$\begin{aligned}&\mathcal{L}_j^{(n)}(\boldsymbol{\theta}, \boldsymbol{\phi})\\ =&\mathbb{E}_{\boldsymbol{z},\boldsymbol{a}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{a}|\boldsymbol{x}^{(n)})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)}|\boldsymbol{z},\boldsymbol{a})\right]\\ &\quad - \text{KL}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{a}|\boldsymbol{y}^{(n)})\|p(\boldsymbol{z},\boldsymbol{a})\right) \qquad (6)\end{aligned}$$

$$\begin{aligned}=&\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}^{(z)}(\boldsymbol{z}|\boldsymbol{x}^{(n)}),\boldsymbol{a}\sim q_{\boldsymbol{\phi}}^{(a)}(\boldsymbol{a}|\boldsymbol{x}^{(n)})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)}|\boldsymbol{z},\boldsymbol{a})\right]\\ &- \text{KL}\left(q_{\boldsymbol{\phi}}^{(z)}(\boldsymbol{z}|\boldsymbol{y}^{(n)})\|p(\boldsymbol{z})\right)\\ &- \text{KL}\left(q_{\boldsymbol{\phi}}^{(a)}(\boldsymbol{a}|\boldsymbol{y}^{(n)})\|p(\boldsymbol{a})\right) \qquad (7)\end{aligned}$$

The second step is due to the independence in both recognition and reconstruction phrases. The posterior factorizes as $q_\phi(\boldsymbol{z}, \boldsymbol{a}|\cdot) = q_\phi^{(z)}(\boldsymbol{z}|\cdot)\, q_\phi^{(a)}(\boldsymbol{a}|\cdot)$ because $\boldsymbol{z}$ and $\boldsymbol{a}$ are conditional independent given $\boldsymbol{x}$ (dashed lines in Figure 1d), whereas the prior factorizes because $\boldsymbol{z}$ and $\boldsymbol{a}$ are marginally independent (solid lines in Figure 1d). In this way, the sampling procedure can be done separately and the KL loss can also be computed independently.

3.2 Prior

We propose two plausible prior distributions for $\boldsymbol{a}_j$.

- The simplest prior, perhaps, is the standard normal, i.e., $p(\boldsymbol{a}_j) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. This follows the prior of the latent space $\boldsymbol{z}$ (Kingma and Welling, 2013; Bowman et al., 2016).
- We observe that the attention vector has to be inside the convex hull of hidden representations of the source sequence, i.e., $\boldsymbol{a}_j \in \text{conv}\{\boldsymbol{h}_i^{\text{(src)}}\}$. We thus impose a Gaussian prior whose mean is the average of $\boldsymbol{h}_i^{\text{(src)}}$, i.e., $p(\boldsymbol{a}_j) = \mathcal{N}(\bar{\boldsymbol{h}}^{\text{(src)}}, \mathbf{I})$, where $\bar{\boldsymbol{h}}^{\text{(src)}} = \frac{1}{|\boldsymbol{x}|}\sum_{i=1}^{|\boldsymbol{x}|}\boldsymbol{h}_i^{\text{(src)}}$.

3.3 Posterior

We model the posterior of $q_{\boldsymbol{\phi}}^{(a)}(\boldsymbol{a}_j)$ as a normal distribution $\mathcal{N}(\boldsymbol{\mu}_{a_j}, \boldsymbol{\sigma}_{a_j})$, where the parameters $\boldsymbol{\mu}_{a_j}$ and $\boldsymbol{\sigma}_{a_j}$ are obtained by a recognition neural network. Following VAE, we compute parameters as if they are deterministic attention in Eq. (5) (denoted by $\boldsymbol{a}_j^{\text{det}}$ in this part) and then transform them by another layer, shown in Figure 2d.

For the mean $\boldsymbol{\mu}_{a_j}$, we apply an identity transformation, i.e., $\boldsymbol{\mu}_{a_j} \equiv \boldsymbol{a}_j^{\text{det}}$. The identify transformation makes much sense because we assume the variational attention still performs (probabilistic) alignment between source and target, similar to deterministic attention in functionality. To compute $\boldsymbol{\sigma}_{a_j}$, $\boldsymbol{a}_j^{\text{det}}$ is first transformed through a neural layer with $\tanh$ activation. The resulting vector then undergoes a linear transformation followed by an $\exp$ activation function to ensure that the values are positive.

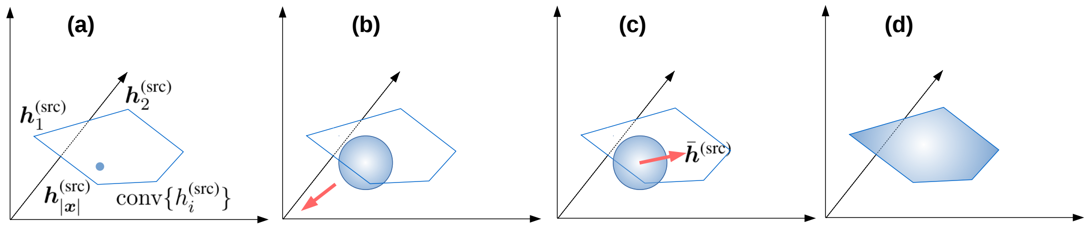
(a)
$h_1^{(\mathrm{src})}$
$h_2^{(\mathrm{src})}$
$h_{|x|}^{(\mathrm{src})}$
$\mathrm{conv}\{h_i^{(\mathrm{src})}\}$
(b)
(c)
$\bar{h}^{(\mathrm{src})}$
(d)

Figure 3: Geometric interpretation of attention mechanisms.

3.4 Training Objective

The overall training objective of Seq2Seq with both variational latent space $\boldsymbol{z}$ and variational attention $\boldsymbol{a}$ is to minimize

$$
\begin{aligned}
J^{(n)}(\boldsymbol{\theta}, \boldsymbol{\phi}) = & J_{\text{rec}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{y}^{(n)}) \\
& + \lambda_{\text{KL}} \Big[\mathrm{KL}\left(q_{\phi}^{(z)}(\boldsymbol{z}) \| p(\boldsymbol{z}) \right) \\
& + \gamma_a \sum_{j=1}^{|\boldsymbol{y}|} \mathrm{KL}\left(q_{\phi}^{(a)}(\boldsymbol{a}_j) \| p(\boldsymbol{a}_j) \right) \Big]
\end{aligned} \tag{8}
$$

Here, we have a hyperparameter λ_{KL} to balance the reconstruction loss and KL losses. γ_a further balances the attention's KL loss and $\boldsymbol{z}$'s KL loss. Since VAE and VED are tricky with Seq2Seq models (e.g., requiring KL annealing), we tie the change of both KL terms and only anneal λ_{KL}. (Training details will be presented in Section 4.2.)

Notice that if $\boldsymbol{a}_j$ has a prior of $\mathcal{N}(\bar{\boldsymbol{h}}^{(\text{src})}, \mathbf{I})$, the derivative of the KL term also goes to $\bar{\boldsymbol{h}}^{(\text{src})}$. This can be computed straightforwardly by auto-differentiation tools, e.g., TensorFlow.

3.5 Geometric Interpretation

We present a geometric interpretation of both deterministic and variational attention mechanisms in Figure 3.

Suppose the hidden representations $\boldsymbol{h}_i^{(\text{src})}$ is of k-dimensional space (represented as a 3-d space in Figure 3). In the deterministic mechanism, the attention model is a convex combination of $\{\boldsymbol{h}_i^{(\text{src})}\}_{i=1}^{|\boldsymbol{x}|}$, as the weights in Eq. (5) are a probabilistic distribution. The attention vector $\boldsymbol{a}_j$ is a point in the convex hull $\mathrm{conv}\{\boldsymbol{h}_i^{(\text{src})}\}$, shown in Figure 3a.

For variational attention in Figures 3b and 3c, the mean of posterior is still in the convex hull, but the sample drawn from the posterior is populated over the entire space (although mostly around the mean, shown as a ball). The difference between the two is that the standard normal prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ pulls the posterior to the origin, whereas the prior $\mathcal{N}(\bar{\boldsymbol{h}}^{(\text{src})}, \mathbf{I})$ pulls the posterior to the mean of $\boldsymbol{h}_1^{(\text{src})}, \boldsymbol{h}_2^{(\text{src})}, \cdots, \boldsymbol{h}_{|\boldsymbol{x}|}^{(\text{src})}$. They are shown as red arrows.

Finally we would like to present a (potential) alternative of modeling variational attention. Instead of treating $\boldsymbol{a}_j$ as random variables, we might also treat $\boldsymbol{\alpha}_j$ as random variables. Since $\boldsymbol{\alpha}_j$ is the parameter of a categorical distribution, its conjugate prior is a Dirichlet distribution. In this case, the resulting attention vector populates the entire convex hull (Figure 3d). However, it relies on a reparametrization trick to propagate reconstruction error's gradient back to the recognition neural network (Kingma and Welling, 2013). In other words, the sampling of latent variables should be drawn from a fixed distribution (without parameters) and then transformed to a desired sample with the distribution's parameters. This is nontrivial for Dirichlet distributions and further research is needed to address this problem.

4 Experiments

4.1 Task, Dataset, and Metrics

We evaluated our approach on a question generation task. It uses the Stanford Question Answering Dataset (Rajpurkar et al., 2016, SQuAD), and aims to generate questions based on a sentence in a paragraph. We used the same train-validation-test split as in Du et al. (2017). According to Du et al. (2017), the attention mechanism is especially critical in this task in order to generate relevant questions. Also, generated questions do need some variety (e.g., in the creation of reading comprehension datasets), as opposed to machine translation, which is typically deterministic.

We followed Du et al. (2017) and used BLEU-1 to BLEU-4 scores (Papineni et al., 2002) to evaluated the quality (in the sense of accuracy) of generated sentences. Besides, we adopted entropy and

Model	Inference	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Entropy	Dist-1	Dist-2
DED (w/o Attn) (Du et al., 2017)	MAP	31.34	13.79	7.36	4.26	-	-	-
DED (w/o Attn)	MAP	29.31	12.42	6.55	3.61	-	-	-
DED+DAttn	MAP	30.24	14.33	8.26	4.96	-	-	-
VED+DAttn	MAP	**31.02**	14.57	8.49	5.02	-	-	-
	Sampling	30.87	**14.71**	**8.61**	**5.08**	2.214	0.132	0.176
VED+DAttn (2-stage training)	MAP	28.88	13.02	7.33	4.16	-	-	-
	Sampling	29.25	13.21	7.45	4.25	2.241	0.140	0.188
VED+VAttn-0	MAP	29.70	14.17	8.21	4.92	-	-	-
	Sampling	30.22	14.22	8.28	4.87	**2.320**	**0.165**	**0.231**
VED+VAttn-$\bar{h}$	MAP	30.23	14.30	8.28	4.93	-	-	-
	Sampling	30.47	14.35	8.39	4.96	2.316	0.162	0.228

Table 2: BLEU, entropy, and distinct scores. We compare the deterministic encoder-decoder (DED) and variational encoder-decoders (VEDs). For VED, we have several variates: deterministic attention (DAttn) and the proposed variational attention (VAttn). We evaluate the sentences obtained by both max *a posteriori* (MAP) inference and sampling.

distinct metrics to measure the diversity. The entropy is computed as $-\sum_w p(w) \log p(w)$, where $p(\cdot)$ is the unigram probability in generated sentences. Distinct metrics—used in previous work to measure diversity (Li et al., 2016)—compute the percentage of distinct unigrams or bigrams (denoted as *Dist-1* and *Dist-2*, respectively).

4.2 Training Details

We used LSTM-RNNs with 100 hidden units for both the encoder and decoder; the dimension of the latent vector $\boldsymbol{z}$ was also 100d. We adopted 300d word embeddings (Mikolov et al., 2013), pretrained on the SQuAD dataset. The vocabulary was limited to the most frequent 40k tokens for the source side and 30k tokens for the target side. We use the Adam optimizer (Kingma and Ba, 2014) to train all models, with an initial learning rate of 0.005, decay of 0.75, and other default hyperparameters. The batch size was set to be 100.

As shown in Bowman et al. (2016), Seq2Seq VAE is hard to train because of the issues associated with the KL term vanishing to zero. Following Bowman et al. (2016), we adopted KL cost annealing and word dropout during training. The coefficient of the KL term λ_{KL} was gradually increased using a logistic annealing schedule, allowing the model to learn to reconstruct the input accurately during the early stages of training. A fixed word dropout rate of 25% was used.

All the hyperparameter tuning was based on validation performance on the motivating Seq2Seq VAE discussed in Section 2.4, and the same hyperparameters were used for all of the models described in Section 3.

4.3 Performance

Table 2 represents the performance of various models. We first implemented a traditional vanilla Seq2Seq model, which we call a deterministic encoder-decoder (DED), and generally replicated the results on the question generation task as reported in Du et al. (2017), showing that our implementation is fair.[1] Incorporating attention mechanism to this model (DED+DAttn) improves BLEU scores, as expected.

In the variational encoder-decoder (VED) framework, we report results obtained by both max *a posterior* (MAP) inference as well as sampling. In the sampling setting, we draw 10 samples ($\boldsymbol{z}$ and/or $\boldsymbol{a}$) from the posterior given $\boldsymbol{x}$ for each data point, and report average BLEU scores.

The proposed variational attention model (VED+VAttn) largely outperforms deterministic attention (VED+DAttn) in terms of all diversity metrics. It should be noted that entropy is a logarithmic measure, and hence a difference of 0.1 in Table 2 is significant; VED+VAttn also generates more distinct unigrams and bigrams than VED+DAttn.

Regarding the prior of variational attention, we propose two variants: $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{N}(\bar{\boldsymbol{h}}^{(\text{src})}, \mathbf{I})$, denoted as VED+VAttn-0 and VED+VAttn-$\bar{h}$, re-

[1] We fixed a bug when computing the metric in our preprint `https://arxiv.org/pdf/1712.08207.pdf`. Notice that we do not compare more complicated models in Du et al. (2017) as they used postprocessing to deal with <UNK>, since its effect in a variational setting is unknown.

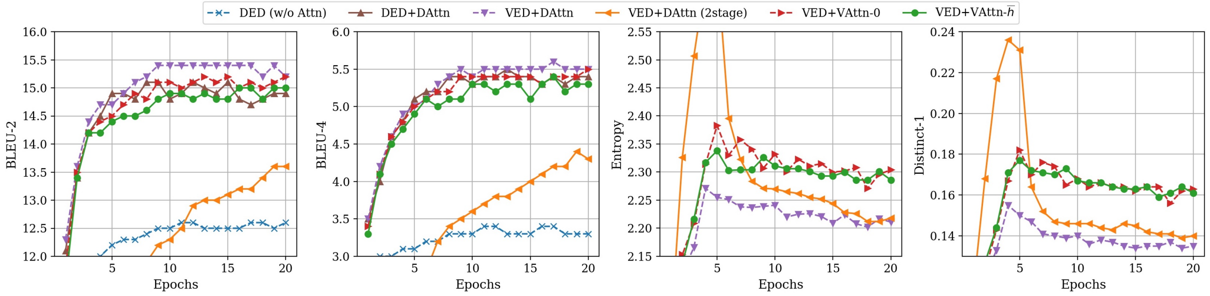

Figure 4: Learning curves of the BLEU-2, BLEU-4, Entropy, and *Dist-1* metrics.

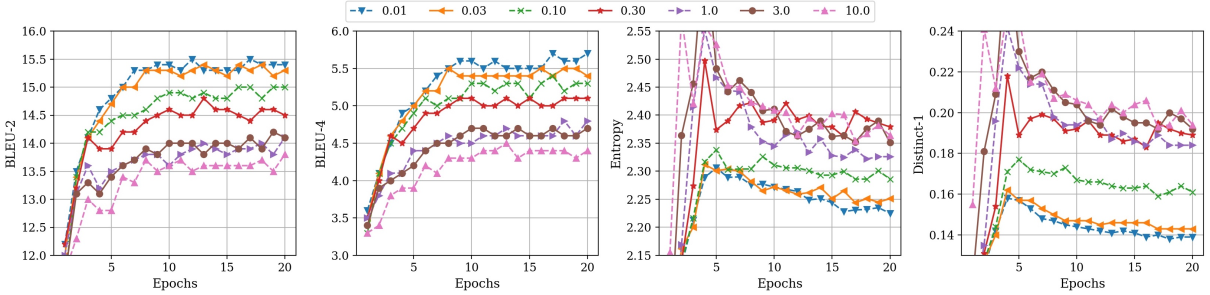

Figure 5: BLEU-2, BLEU-4, Entropy, and *Dist-1* with different γ_a values.

spectively. VED+VAttn-0 has slightly lower BLEU but higher diversity. The results are generally comparable, showing both priors are reasonable.

We also tried a heuristic of 2-stage training (VED+DAttn 2-stage), which first trained VED without attention for 6 epochs, and then adds the attention mechanism to the model. This heuristic is proposed in hopes of better training the variational latent space at the beginning stages. However, the experiments show that such simple heuristic does not help much, and is worse than the principled variational attention mechanism in all BLEU and diversity metrics.

Learning curves. Figure 4 shows the trends of sentence quality (BLEU-2 and BLEU-4) and diversity (entropy and *Dist-1*) of all models on the validation set, as training progresses.[2] We see that BLEU and diversity are conflicting objectives: a high BLEU score indicates resemblance to the groundtruth, resulting in low diversity. However, the variational attention mechanisms (red and green lines in Figure 4) remain high in both aspects, showing the effectiveness of our model.

For the 2-stage heuristic, although it has high diversity before attention is added, the quality of generated sentences is unreasonably low. After attention, BLEU scores are improving, but diversity decreases to the level of VED+DAttn. This shows that such simple heuristic does not alleviate the bypassing phenomenon along the training process, and is worse than our statistically motivated variational attention.

Strength of Attention's KL Loss. We tuned the KL term's strength variational attention, i.e., γ_a in Eq. (8), and plot the BLEU and diversity metrics in Figure 5. In this experiment, we used the VED+DAttn-$\bar{h}$ variant. As shown, a decrease in γ_a increases the quality of generated sentences at the cost of diversity. This is expected because a lower γ_a gives the model less incentive to optimize the KL term of attention, which then causes the model to behave more "deterministic." Based on this experiment, we chose a value of 0.1 for γ_a, as it yields a learning curve in the middle, among those of different hyperparameters, being a good balance between quality and diversity.

It should be further mentioned that, with a milder γ_a (e.g., 0.01), VED+VAttn outperforms VED+DAttn in terms of both quality and diversity (on the validation set). This is consistent with the evidence that variational latent space may serve as a way of regularization and improve quality (Zhang et al., 2016). However, a small γ_a only slightly improves diversity, and hence we did not choose this hyperparameter in Table 2.

Case study. We show in Table 3 two exam-

[2] Other metrics are omitted because the trend is the similar.

Source	*when the british forces evacuated at the close of the war in 1783 , they transported 3,000 freedmen for resettlement in nova scotia .*
Reference	*in what year did the american revolutionary war end ?*
VED+DAttn	*how many people evacuated in newfoundland ?*
	how many people evacuated in newfoundland ?
	what did the british forces seize in the war ?
VED+VAttn-$\bar{h}$	*how many people lived in nova scotia ?*
	where did the british forces retreat ?
	when did the british forces leave the war ?
Source	*downstream , more than 200,000 people were evacuated from mianyang by june 1 in anticipation of the dam bursting .*
Reference	*how many people were evacuated downstream ?*
VED+DAttn	*how many people evacuated from the mianyang basin ?*
	how many people evacuated from the mianyang basin ?
	how many people evacuated from the mianyang basin ?
VED+VAttn-$\bar{h}$	*how many people evacuated from the tunnel ?*
	how many people evacuated from the dam ?
	how many people were evacuated from fort in the dam ?

Table 3: Case study of question generation.

ples of generated sentences by VED+DAttn and VED+VAttn-$\bar{h}$, each containing three random sentences drawn from the variational latent space(s). In both examples, the variational attention generates more diversified sentences than deterministic attention. However, the quality of generated sentences is close.

5 Conclusion

In this paper, we proposed a variational attention mechanism for variational encoder-decoder (VED) frameworks. We observe that, in VED, if the decoder has direct access to the encoder, the connection may bypass the variational space. Traditional attention mechanisms might serve as such bypassing connection, making the output less diverse. Our variational attention imposes a probabilistic distribution on the attention vector. We also proposed different priors for the attention vector. The proposed model was evaluated on a question generation task, showing that variational attention yields more diversified samples while retaining high quality.

Acknowledgments

We thank Hao Zhou for helpful discussion. The Titan Xp GPU used for this research was donated by the NVIDIA Corporation to Olga Vechtomova.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 632–642. https://doi.org/10.18653/v1/D15-1075.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. pages 10–21. https://doi.org/10.18653/v1/K16-1002.

Kris Cao and Stephen Clark. 2017. Latent variable dialogue models and their diversity. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. pages 182–187. https://doi.org/10.18653/v1/E17-2029.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pages 1342–1352. https://doi.org/10.18653/v1/P17-1123.

Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*. volume 70, pages 1587–1596.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114* .

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 110–119. https://doi.org/10.18653/v1/N16-1014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. pages 3111–3119.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 311–318.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 2383–2392. https://doi.org/10.18653/v1/D16-1264.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 379–389. https://doi.org/10.18653/v1/D15-1044.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. pages 3295–3301.

Martin J Wainwright, Michael I Jordan, et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* pages 1–305.

Biao Zhang, Deyi Xiong, jinsong su, Qun Liu, Rongrong Ji, Hong Duan, and Min Zhang. 2016. Variational neural discourse relation recognizer. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 382–391. https://doi.org/10.18653/v1/D16-1037.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lili Mou, Xinyu Dai, Jiajun Chen, and Zhaopeng Tu. 2017. Modeling past and future for neural machine translation. *arXiv preprint arXiv:1711.09502 (to appear in TACL)* .

Chunting Zhou and Graham Neubig. 2017. Morphological inflection generation with multi-space variational encoder-decoders. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*. pages 58–65.