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 generated sentences less diversified. In our paper, we propose a variational attention mechanism for VED, where the attention vector is modeled as normally distributed random variables. Experiments show that variational attention increases diversity while retaining high quality. We also show that the model is not sensitive to hyperparameters.

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. 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 the variational Seq2Seq, the attention mechanism unfortunately may serve 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 to follow a Gaussian distribution. However, our prior has a mean being the average of source infor-

* The first two authors contributed equally.

mation, as opposed to a vector of all zeros. This is more suited in our scenario because attention is a weighted sum of source information. Experiments show that the proposed models with variational attention have a higher diversity than variational Seq2Seq models with deterministic attention, while retaining the 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

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

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

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

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 likelihood $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$, which is equivalent to minimizing the objective:

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

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 $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

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 proposed 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.

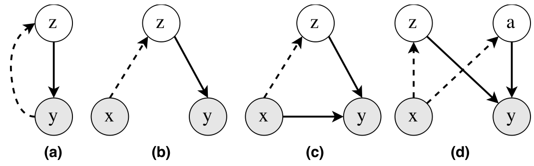

Figure 1: Graphical model representations. **(a)** Variational autoencoder (VAE). **(b)** Variational encoder-decoder (VED). **(c)** VED with deterministic attention (VED+DAttn). **(d)** VED with variational attention (VED+VAttn). **Dashed lines**: Recognition phase. **Solid lines**: Reconstruction phase.

Another approach is to build a recognition model on $\boldsymbol{X}$ (Zhou and Neubig, 2017). Taking the assumption $\boldsymbol{Y}$ 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., sentences). 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 step j in the decoder, the attention mechanism computes a probabilistic distribution by

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

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.

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

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

which is fed to the decoder RNN at the jth step. Figure 2c 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 $q_{\boldsymbol{\phi}}(\boldsymbol{Y}|\boldsymbol{Z})$ is aware of $\boldsymbol{X}$ by itself, then $q_{\boldsymbol{\phi}}(\boldsymbol{Y}|\boldsymbol{Z})$ might be learned as $q_{\boldsymbol{\phi}}(\boldsymbol{Y}|\boldsymbol{X})$ 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 with a subset ($\sim$80k samples) of the massive dataset provided by Bowman et al. (2015). We show generated sentences and entropy in Table 1. 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 needs to fit to its prior in order 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 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, a_j is computed by Eq. (5) in a

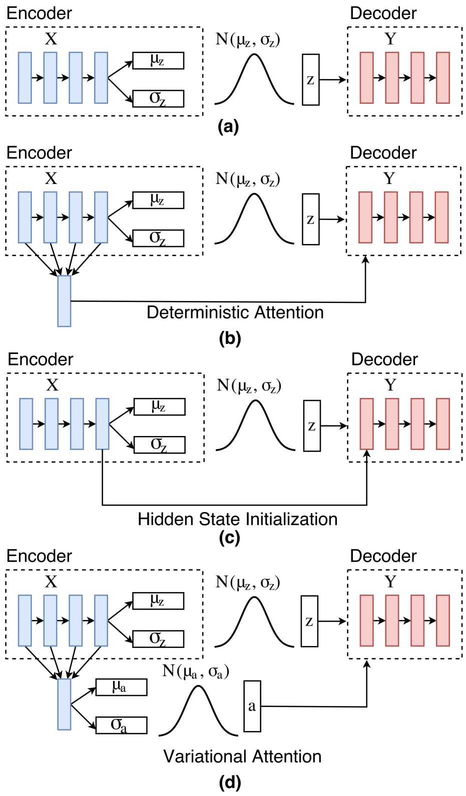

Figure 2: (a) Variational Seq2Seq model. (b) Variational Seq2Seq with deterministic attention. (c) Variational Seq2Seq with hidden state initialization. (d) Variational Seq2Seq with variational attention.

deterministic fashion in traditional attention.

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

$$\begin{aligned}
&\mathcal{L}_j^{(n)}(\boldsymbol{\theta}, \boldsymbol{\phi}) \\
=&\mathbb{E}_{\boldsymbol{z},\boldsymbol{a}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{a}|\boldsymbol{x}^{(n)})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)}|\boldsymbol{z},\boldsymbol{a})\right] \\
&\quad - \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{a}|\boldsymbol{y}^{(n)})\|p(\boldsymbol{z},\boldsymbol{a})\right) \qquad (6) \\
=&\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}^{(z)}(\boldsymbol{z}|\boldsymbol{x}^{(n)}),\boldsymbol{a}\sim q_{\boldsymbol{\phi}}^{(a)}(\boldsymbol{a}|\boldsymbol{x}^{(n)})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)}|\boldsymbol{z},\boldsymbol{a})\right] \\
&\quad - \mathrm{KL}\left(q_{\boldsymbol{\phi}}^{(z)}(\boldsymbol{z}|\boldsymbol{y}^{(n)})\|p(\boldsymbol{z})\right) \\
&\quad - \mathrm{KL}\left(q_{\boldsymbol{\phi}}^{(a)}(\boldsymbol{a}|\boldsymbol{y}^{(n)})\|p(\boldsymbol{a})\right) \qquad (7)
\end{aligned}$$

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 have 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., $\boldsymbol{a}_j \in \mathrm{conv}\{\boldsymbol{h}_i^{(\mathrm{src})}\}$. We impose a normal prior whose mean is the average of $\boldsymbol{h}_i^{(\mathrm{src})}$, i.e, $p(\boldsymbol{a}_j) = \mathcal{N}(\bar{\boldsymbol{h}}^{(\mathrm{src})}, \mathbf{I})$, where $\bar{\boldsymbol{h}}^{(\mathrm{src})} = \frac{1}{|\boldsymbol{x}|}\sum_{i=1}^{|\boldsymbol{x}|}\boldsymbol{h}_i^{(\mathrm{src})}$.

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) and then transform it by another layer, shown in Figure 2d. As $\boldsymbol{\mu}_{a_j}$ are real-values, its transformation is done with a feed-forward neural network with $\tanh$ activation, followed by a linear layer. Likewise, a similar transformation is carried out to obtain $\boldsymbol{\sigma}_{a_j}$, followed by an additional $\exp$ activation function to ensure that the values are positive.

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

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

Here, we have a hyperparameter λ_{KL} to balance the reconstruction loss and KL losses. γ_a

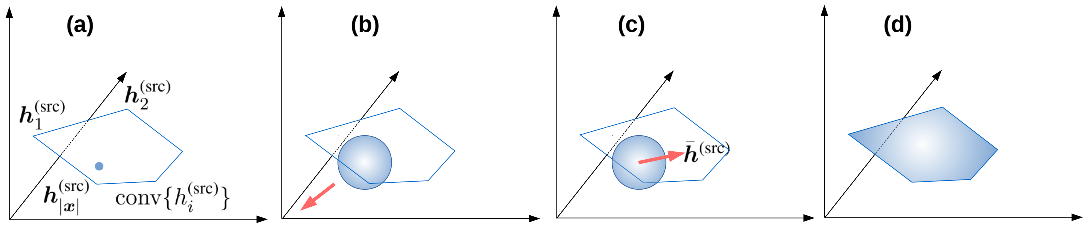

Figure 3: Geometric interpretation of attention mechanisms.

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 or 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 the 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 null $\text{conv}\{\boldsymbol{h}_i^{(\text{src})}\}$, shown in Figure 3a.

For variational attention in Figures 3b and 3c, the mean 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 aims to generate questions based on a sentence in a paragraph. We followed Du et al. (2017) and used the Stanford Question Answering Dataset (SQuAD) dataset (Rajpurkar et al., 2016), except that we had a different split of 1k and 1k samples for validation and testing, respectively, from 86k pairs of sentence-question pairs in total. As reported by Du et al. (2017), 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 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

Model	Inference	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Entropy	Dist-1	Dist-2
Previous work (Du et al., 2017)	MAP	43.09	25.96	17.50	12.28	-	-	-
DED (w/o Attn)	MAP	39.46	28.49	20.74	8.10	-	-	-
DED+DAttn	MAP	42.34	30.86	22.74	11.60	-	-	-
VED+DAttn	MAP	**42.50**	**31.13**	**23.09**	**12.38**	-	-	-
	Sampling	42.48	31.10	23.08	12.30	2.37	0.18	0.26
VED+DAttn (2-stage training)	MAP	42.17	30.96	22.95	11.98	-	-	-
	Sampling	41.98	30.82	22.81	11.78	2.41	0.19	0.27
VED+VAttn-0	MAP	41.77	30.54	22.53	11.37	-	-	-
	Sampling	41.73	30.51	22.49	11.27	**2.44**	**0.20**	0.28
VED+VAttn-$\bar{h}$	MAP	42.10	30.71	22.70	11.55	-	-	-
	Sampling	42.03	30.62	22.66	11.50	**2.44**	**0.20**	**0.29**

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.

300d pretrained word embeddings from Mikolov et al. (2013). For both the source and target side, the vocabulary was limited to the most frequent 40k tokens. We use the Adam optimizer (Kingma and Ba, 2014) to train all models, with an initial learning rate of 0.005, decay of 0.95, 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.

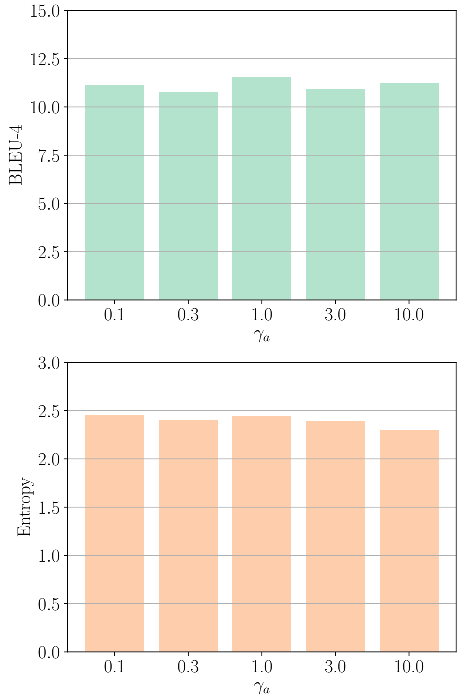

Figure 4: The effect of the strength of attention's KL loss.

4.3 Performance

Table 2 represents the performance of various models. We first implemented a traditional Seq2Seq with attention (DED+DAttn) and generally replicated the results in Du et al. (2017), showing that our implementation is fair. We also tried a Seq2Seq model (DED) without attention, and we see the performance is degraded by 3.5 BLEU points, which is large, showing that the task is suited for testing the attention mechanism.

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 $\boldsymbol{a}$) from the posterior given $\boldsymbol{x}$ for each data point, and report average BLEU scores. We see that VED with deterministic attention (VED+DAttn) yields the best performance in terms of BLEU scores. However, it is not satisfactory if we would like to take diversity into account. Although it is better than variational Seq2Seq without attention (shown in Section 2.4), it is less diversified than our proposed variational attention with a fair margin, since entropy is a logarithmic measure. Variational attention models also generate more distinct unigrams and bigrams, as indicated by *distinct* metrics.

We also tried a heuristic of 2-stage training to correct the lack of diversity. In this baseline, we first trained VED without attention for 20 epochs, since the variational latent space is more difficult to train. Then we added the attention mechanism to the model. This yields an entropy value in between the variational attention and the deterministic attention applied at the beginning.

The proposed variational attention achieves the best diversity while maintaining high quality. The prior of $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{N}(\bar{\boldsymbol{h}}^{(\mathrm{src})}, \mathbf{I})$ yield similar diversity. Their resulting models are denoted as VED+VAttn-0 and VED+VAttn-$\bar{h}$. Regarding quality, $\mathcal{N}(\bar{\boldsymbol{h}}^{(\mathrm{src})}, \mathbf{I})$ is better than $\mathcal{N}(\mathbf{0}, \mathbf{I})$ in terms of all BLEU scores, being a more reasonable prior. Its corresponding model VED+VAttn-$\bar{h}$ has less than 1 BLEU score degradation compared with VED+DAttn, but is comparable to the deterministic Seq2Seq with attention (DED+DAttn), showing that variational attention with a proper prior does not hurt performance much, despite its diversity.

Strength of Attention's KL Loss. We tuned the KL loss's strength of variational attention, i.e., γ_a in Eq. (8), and plot the BLEU-4 and entropy metrics in Figure 4. In this experiment, we used the VED+DAttn-$\bar{h}$ variant. As shown, the strength of attention's KL loss does not have a large effect on both BLEU and entropy. The result is expected because in the limit of λ_a to infinity, the model is mathematically equivalent (regardless of computational issues) to mean pooling of the source's hidden states with a noise drawn from standard normal $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The experiment shows that the proposed variational attention is not sensitive to this hyperparameter.

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, among which, choosing a normal distribution centered at the mean of source representations is appropriate in our scenario. The proposed model was evaluated on a question generation task, showing that variational attention yields more diversified samples while retaining high quality. We also show that the model is not sensitive to the strength of attention's KL term.

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.