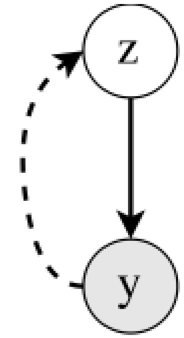
Variational Attention for Sequence-to-Sequence Models

Hareesh Bahuleyan, Lili Mou, Olga Vechtomova, Pascal Poupart
University of Waterloo, ON, Canada

Variational Autoencoder

❑ Variational Autoencoder (VAE)

- Encode data Y as hidden random variables Z, then reconstruct Y.
- VAE models both $q_\phi(z|y)$ and $p_\theta(y|z)$ with neural networks.
- Training objective:

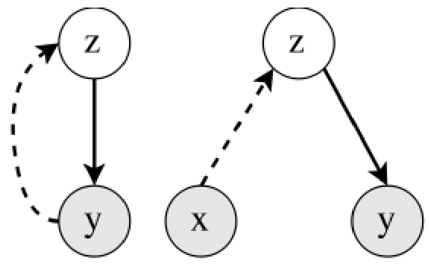


$$\underbrace{J_{\text{rec}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{y}^{(n)})}_{\text{reconstruction loss}} + \underbrace{\mathrm{KL}\Big(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)}) \| p(\boldsymbol{z})\Big)}_{\text{regularization}}$$

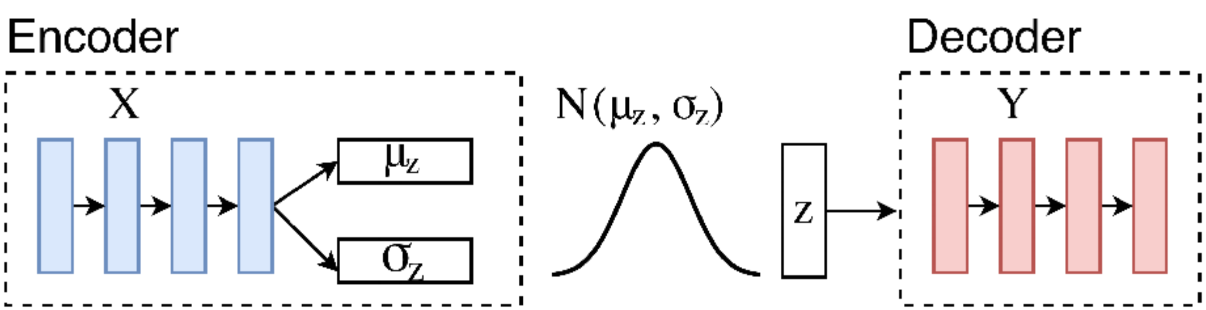
Variational Encoder-Decoder

❑ Variational Encoder-Decoder (VED)

- Extend VAE to VED.



- Model:

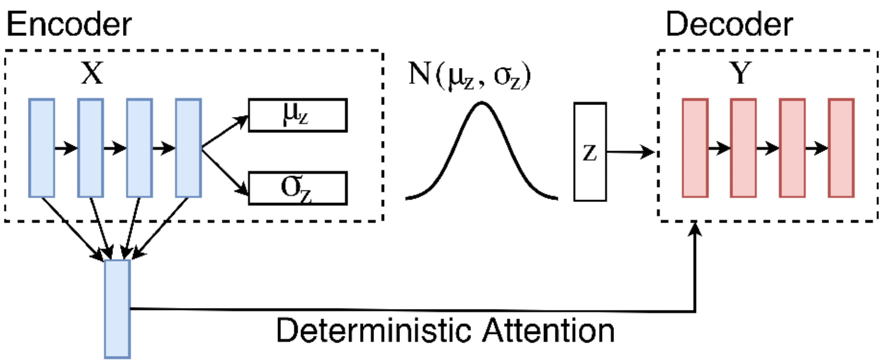


Attention Mechanism

- Attention significantly improves Seq2Seq performance in translation, summarization, etc.

$$\alpha_{ji} = \frac{\exp\{\widetilde{\alpha}_{ji}\}}{\sum_{i'=1}^{|\boldsymbol{x}|} \exp\{\widetilde{\alpha}_{ji'}\}} \qquad \boldsymbol{a}_j = \sum_{i=1}^{|\boldsymbol{x}|} \alpha_{ji} \boldsymbol{h}_i^{(\mathrm{src})}$$

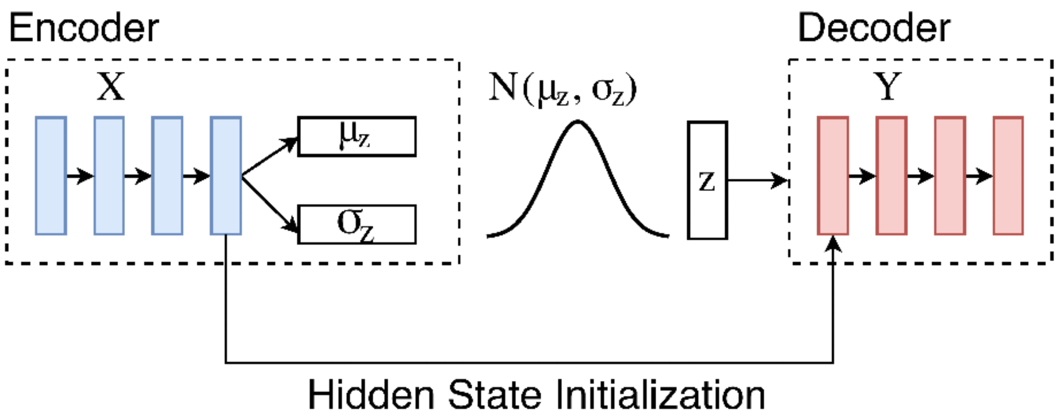
- Add attention to variational Seq2Seq model?



"Bypassing" Phenomenon

- Observation: if the decoder has a direct, deterministic access to the encoder, the latent variables Z might not capture much information, so that VED does not learn much and play a role in the process.

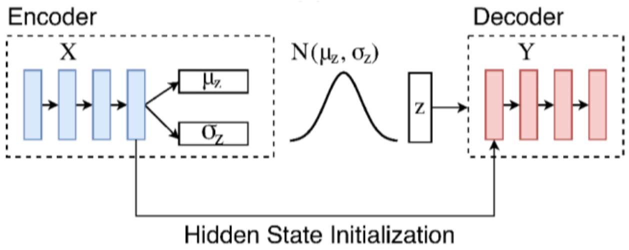
- Example: variational Seq2Seq with hidden state initialization



A Pilot Study

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*

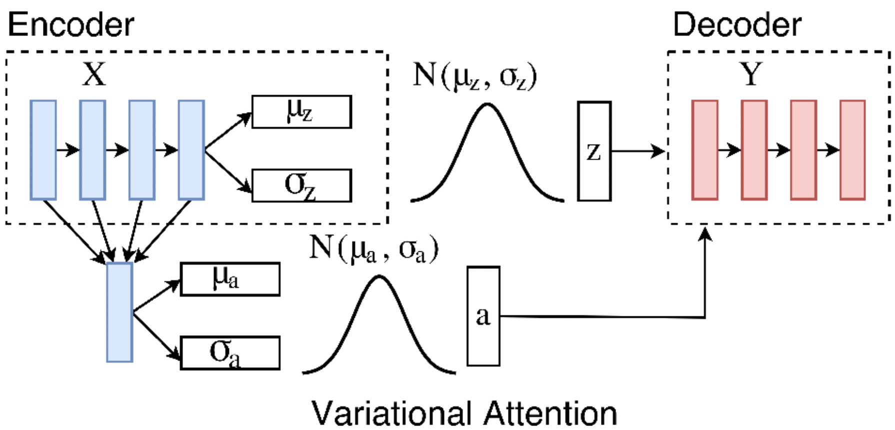
Table 1: Sentences obtained by sampling from the VAE latent space. (a) VAE without hidden state initialization. (b) VAE with hidden state initialization.

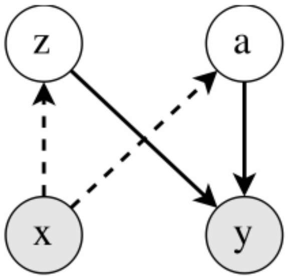


Less diversified!

Variational Attention

- Treat latent space Z and attention vector a_j as random variables.





- Variational lower bound

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

z and a are conditional independent given x, also marginally independent.

Variational Attention

- Prior $p(a_j)$
  - The same as z, set $p(a_j) = N(0, I)$.
  - $p(a_j) = N(\bar{h}^{(src)}, I)$, where $\bar{h}^{(src)} = \frac{1}{|x|}\sum_{i=1}^{|x|} h_i^{(src)}$.
- Training Objective

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

Experiments

- ❑ Task: generate questions based on a sentence in a paragraph.
- ❑ Generated questions do need some variety.
- ❑ Dataset: follow Du et al. (2017) and use SQuAD.
- ❑ Metrics:
  - Measure accuracy: BLEU-1 to BLEU-4.
  - Measure diversity: entropy and distinct metrics.

The percentage of distinct unigrams or bigrams.

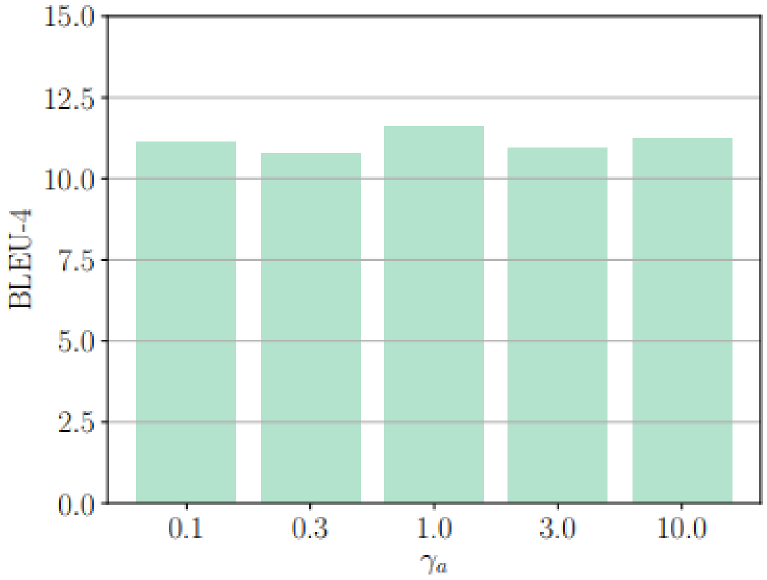
Results

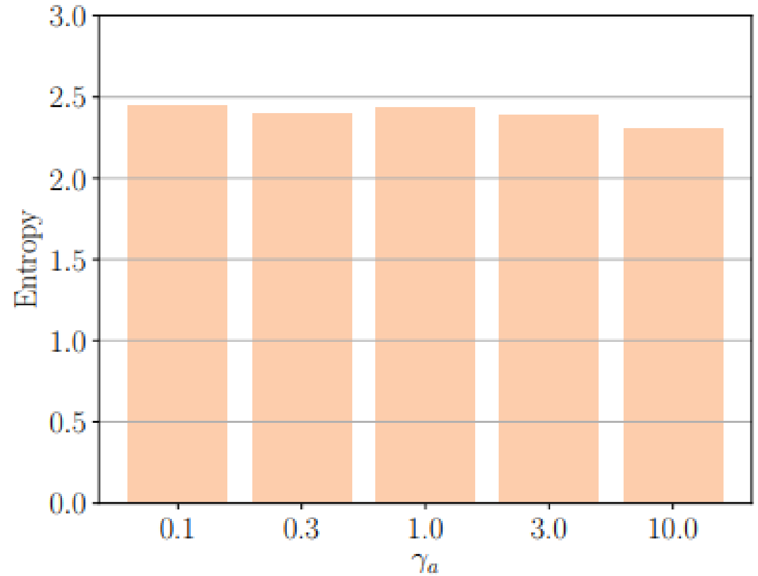
Model	Inference	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Entropy	Dist-1	Dist-2
Previous work (Du et al., 2017)	MAP	43.09	25.96	17.50	12.28	-	-	-
DED (w/o Attn)	MAP	39.46	28.49	20.74	8.10	-	-	-
DED+DAttn	MAP	42.34	30.86	22.74	11.60	-	-	-
VED+DAttn	MAP	**42.50**	**31.13**	**23.09**	**12.38**	-	-	-
	Sampling	42.48	31.10	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.

Sampling: draw 10 samples each time.
2-stage training: train VED without attention for first 20 epochs.

Strength of Attention's KL Loss





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

$\boxed{\gamma_a \text{ has little effects.}}$