Variational Attention for Sequence-to-Sequence Models

Hareesh Bahuleyan,[1] **Lili Mou**,[1] Olga Vechtomova, Pascal Poupart

University of Waterloo

March, 2018

Outline

Outline

Variational Autoencoder (VAE)

Kingma and Welling (2013)
A combination of (neural) autoencoders and variational inference

- Compared with traditional **variational inference**

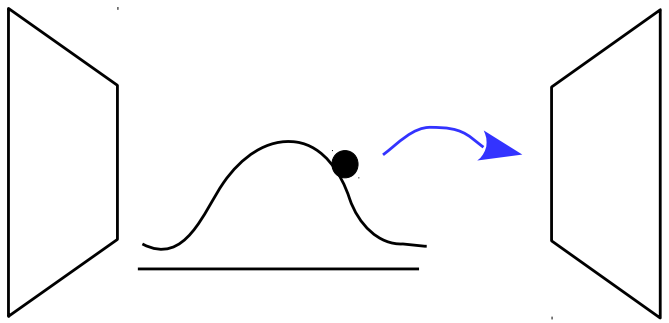
  Takes use of neural networks as a powerful density estimator
- Compared with traditional **autoencoders**

  Imposes a probabilistic distribution on latent representations

Why VAE?

- Learns the distribution of data
  - Instead of the MAP sample
  - Implicitly capture the diversity of data
- Generating samples from scratch
  - Image generation, sentence generation, etc.
  - Controlling the generated samples

  $\Rightarrow$ Not quite feasible in deterministic AE
- Regularization



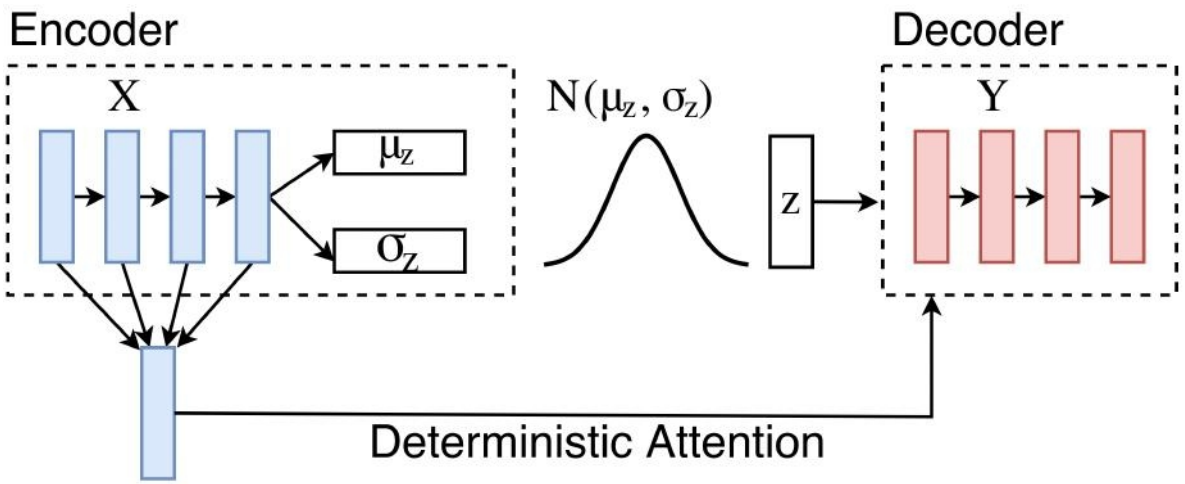
Variational Encoder-Decoder (VED)

- Encoder-decoder frameworks
  - Machine translation
  - Dialog systems
  - Summarization
- VAE $\Rightarrow$ VED (Variational encoder-decoder)

VAE/VED in NLP

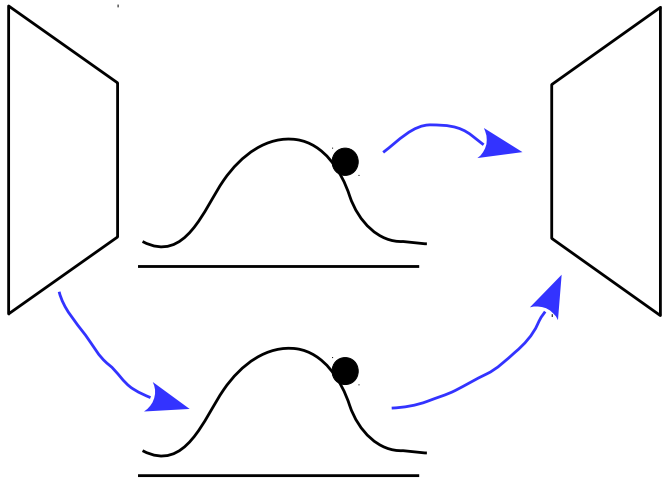
- Seq2Seq models as encoders & decoders
- Attention mechanism serving as dynamic alignment

However, we observe the bypassing phenomenon



Our Proposal

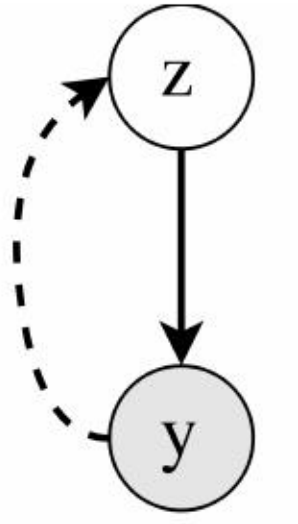
- Variational attention mechanism
  - Modeling attention probability/vector as random variables
  - Imposing some prior and posterior distributions over attention



- Experimental results show that the variational space is more effective when combined with variational attention.

Outline

Variational Inference



- $\boldsymbol{Z}$ Hidden variables
- $\boldsymbol{Y}$ Observable variables

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)}) = \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)})}\left[\log\left\{\frac{p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)})}\right\}\right] + \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)})\|p(\boldsymbol{z}|\boldsymbol{y}))$$

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

$$\text{loss} = \mathbb{E}[\text{reconstruction loss}] + \text{KL divergence}$$

Variational Autoencoder

loss = $\mathbb{E}$[reconstruction loss] + KL divergence

- Prior

$$p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

- Posterior

$$q_\phi(\boldsymbol{z}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}\{\boldsymbol{\sigma}\})$$

where $\boldsymbol{\mu}, \boldsymbol{\sigma} = \mathrm{NN}(\boldsymbol{y})$

Variational Encoder-Decoder

- Encoder-Decoder: Transforming $\boldsymbol{X}$ to $\boldsymbol{Y}$
- **Attempt#1**: Condition any distribution on $\boldsymbol{X}$
  (Zhang et al., 2016; Cao and Clark, 2017; Serban et al., 2017)

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)}|\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)},\boldsymbol{x})}\left[\log\left\{\frac{p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)},\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)},\boldsymbol{x})}\right\}\right]$$

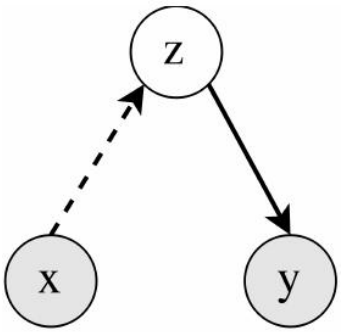
$$= \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)},\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(n)}|\boldsymbol{z},\boldsymbol{x})\right]$$

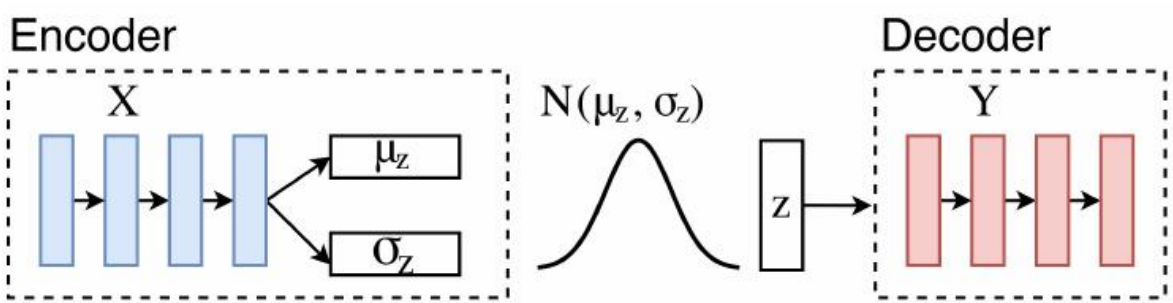
$$- \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}^{(n)},\boldsymbol{x})\|p(\boldsymbol{z}|\boldsymbol{x})\right) \triangleq \mathcal{L}^{(n)}(\boldsymbol{\theta},\boldsymbol{\phi})$$

- Doubts:
  - The posterior contains $\boldsymbol{Y}$
  - Cannot have fine-grained (word/character-level) variational modeling

Variational Encoder-Decoder

- Encoder-Decoder: Transforming $\boldsymbol{X}$ to $\boldsymbol{Y}$
- **Attempt#2** (Cao and Clark, 2017):
  Assuming $\boldsymbol{Y}$ is some function of $\boldsymbol{X}$ i.e., $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{X})$, then
  $q(\boldsymbol{z}|\boldsymbol{y}) = q(\boldsymbol{z}|\boldsymbol{Y}(\boldsymbol{x})) = \widetilde{q}(\boldsymbol{z}|\boldsymbol{x})$





Attention Mechanism

At each step of decoding

- Attention probability

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

- Attention vector

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

Attention Mechanism

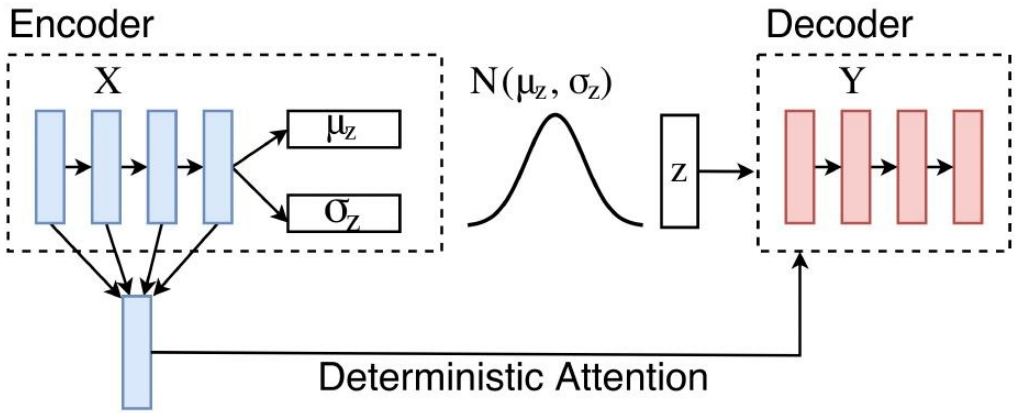
At each step of decoding

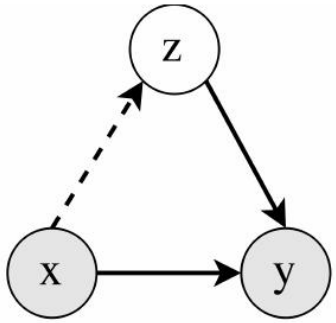
- Attention probability

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

- Attention vector

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



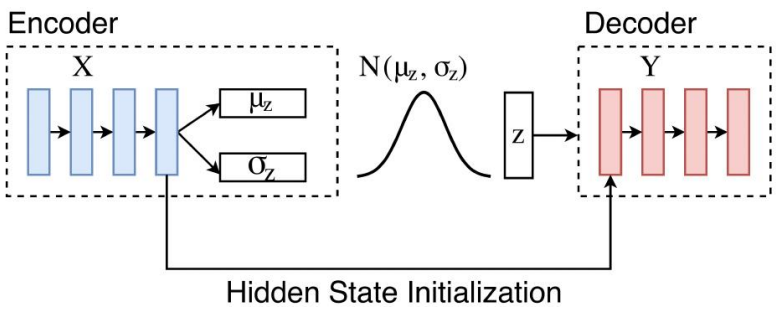


(Bypassing phenomenon)

Bypassing Phenomenon

A deterministic connection bypasses variational space

Hidden state initialization:

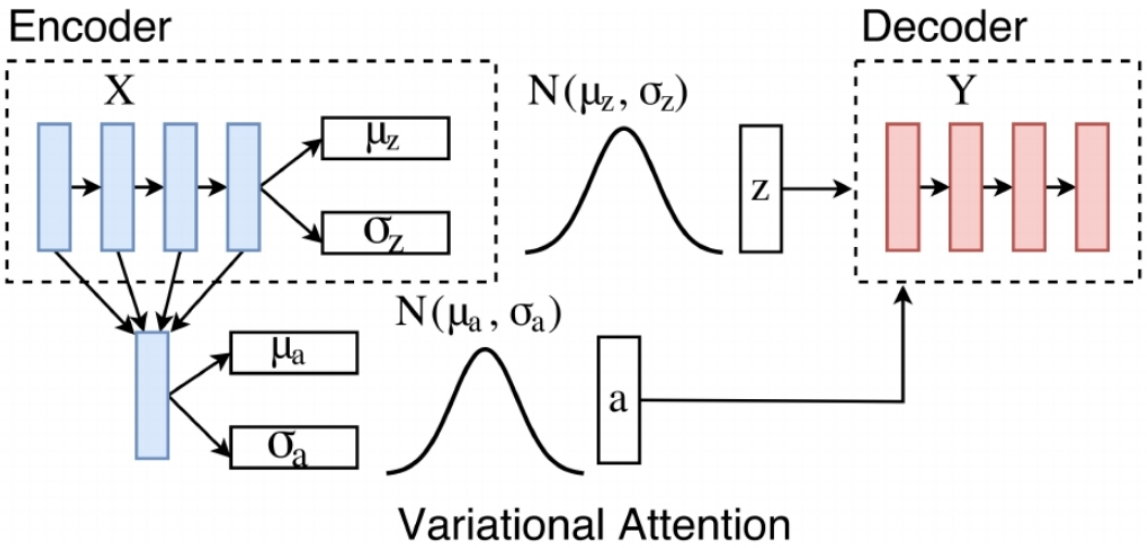


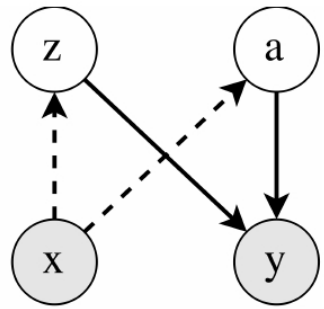
Input: *the men are playing musical instruments*
(a) VAE w/o hidden state init. (Avg entropy: 2.52)
the men are playing musical instruments
the men are playing video games
the musicians are playing musical instruments
the women are playing musical instruments
(b) VAE w/ hidden state init. (Avg entropy: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

Outline

Intuition

General idea: Model attention as random variables



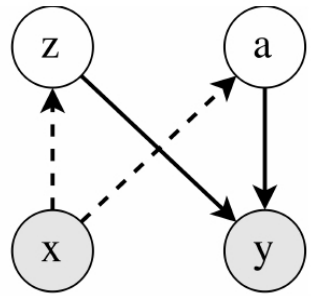


Lower Bound

$$\mathcal{L}_j^{(n)}(\boldsymbol{\theta}, \boldsymbol{\phi})$$

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

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



Prior

- Standard norm: $p(\boldsymbol{a}_j) = \mathcal{N}(\mathbf{0}, \mathbf{I})$
- Norm centered at $\bar{\boldsymbol{h}}^{(\text{src})}$:

$$p(\boldsymbol{a}_j) = \mathcal{N}(\bar{\boldsymbol{h}}^{(\text{src})}, \mathbf{I})$$

where $\bar{\boldsymbol{h}}^{(\text{src})} = \frac{1}{|\boldsymbol{x}|} \sum_{i=1}^{|\boldsymbol{x}|} \boldsymbol{h}_i^{(\text{src})}$

Posterior

- Attention probability

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

- Attention vector

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

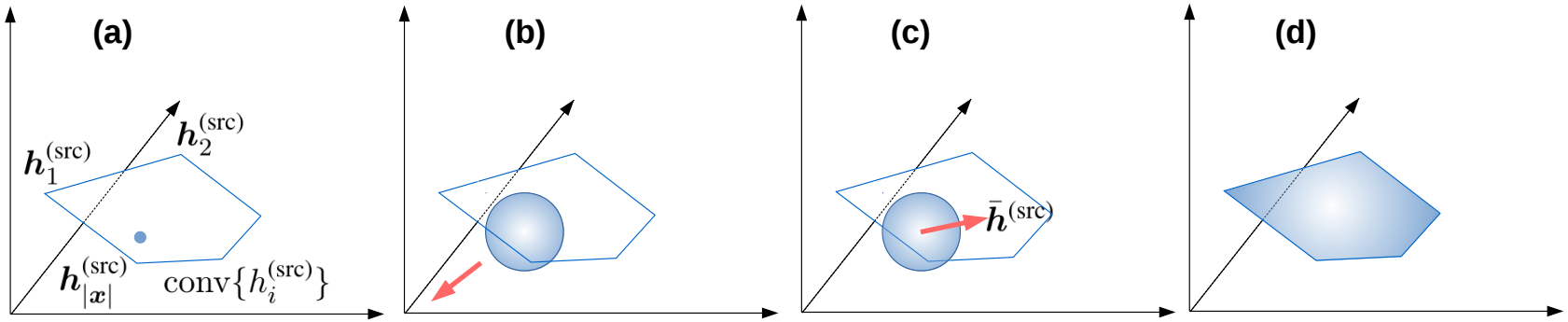
- Posterior: $\mathcal{N}(\boldsymbol{\mu}_{a_j}, \boldsymbol{\sigma}_{a_j})$, where

$$\boldsymbol{\mu}_{a_j} \equiv \boldsymbol{a}_j^{\text{det}}, \qquad \boldsymbol{\sigma}_{a_j} = \text{NN}(\boldsymbol{a}_j^{\text{det}})$$

Training Objective

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

Geometric Interpretation



Outline

Task

Question generation (Du et al., 2017)

- Given some information (a sentence), to generate a related question
- Dataset from the Stanford Question Answering Dataset (Rajpurkar et al., 2016, SQuAD)

Settings

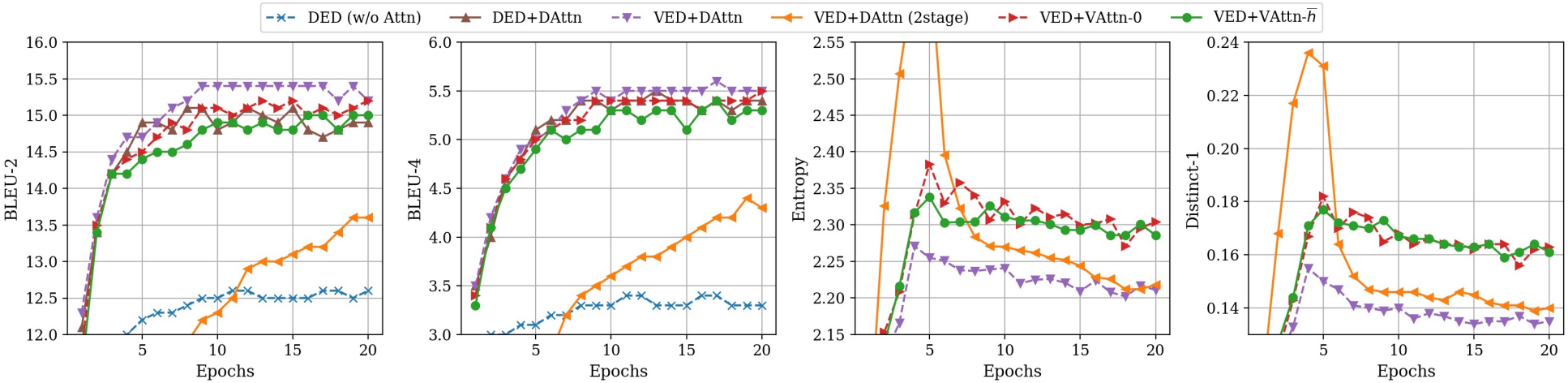
- KL-annealing: logistic schema
- Word dropout: 25%

Hyperparaemters tuned on VAE and adopted to all VED models

Overall Performance

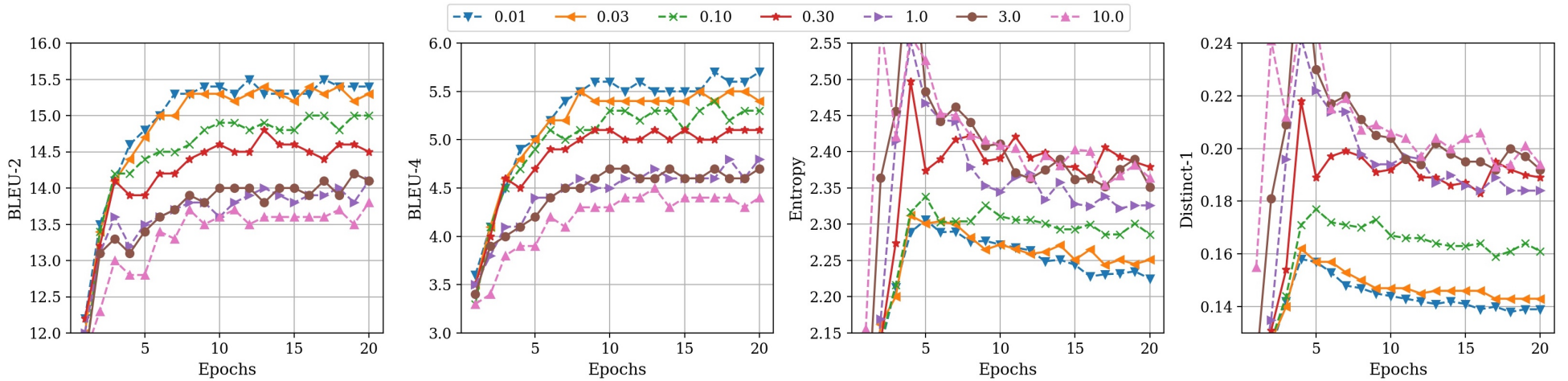
Model	Inference	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Entropy	Dist-1	Dist-2
DED (w/o Attn) Du et al. (2017)	MAP	31.34	13.79	7.36	4.26	-	-	-
DED (w/o Attn)	MAP	29.31	12.42	6.55	3.61	-	-	-
DED+DAttn	MAP	30.24	14.33	8.26	4.96	-	-	-
VED+DAttn	MAP	**31.02**	14.57	8.49	5.02	-	-	-
	Sampling	30.87	**14.71**	**8.61**	**5.08**	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

Learning Curves



Effect of γ_a

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



Case Study

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 ?

Outline

Conclusion

Our work

- Address the bypassing phenomenon
- Propose a variational attention

Future work

- Probabilistic modeling of attention probability

Lesson learned

- Design philosophy of VAE/VED

References

Kris Cao and Stephen Clark. 2017. Latent variable dialogue models and their diversity. In *EACL*. pages 182–187.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *ACL*. pages 1342–1352.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*. pages 2383–2392.

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 *AAAI*. pages 3295–3301.

Biao Zhang, Deyi Xiong, jinsong su, Qun Liu, Rongrong Ji, Hong Duan, and Min Zhang. 2016. Variational neural discourse relation recognizer. In *EMNLP*. pages 382–391.

Thanks for listening

Question?