Natural Language Generation with Neural Variational Models

Hareesh Bahuleyan

University of Waterloo

July 26, 2018

Overview

Plan

Natural Language Processing

- Enabling computers to effectively interact with humans using natural language
- Manipulation, understanding, interpretation and generation of textual and speech data
- Examples of NLP tasks - question answering, sentiment analysis, named entity recognition and machine translation
- This work: Natural Language Generation
  - Syntax
  - Semantics

Source: http://www.contrib.andrew.cmu.edu/~dyafei/NLP.html

Motivation

- Consider two dialog systems (conversational agent responding to user utterances)

Input: What are you doing?	
I don't know.	*Get out of here.*
I don't know!	*I'm going home.*
Nothing.	*Oh, my god!*
Get out of the way.	*I'm talking to you.*
Input: What is your name?	
I don't know.	*My name is Robert.*
I don't know!	*My name is John.*
I don't know, sir.	*My name's John.*
Oh, my god!	*My name is Alice.*
Input: How old are you?	
I don't know.	*Twenty-five.*
I'm fine.	*Five.*
I'm all right.	*Eight.*
I'm not sure.	*Ten years old.*

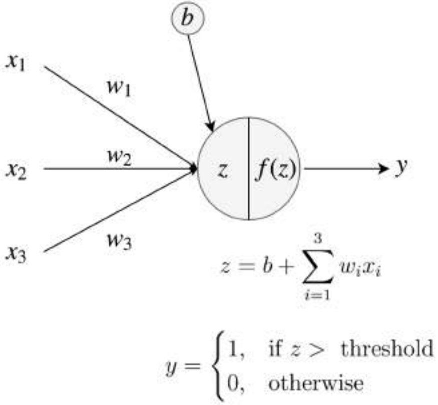
Table: Diversity of responses [Li et al., 2015]

- Objective is to generate a diverse set of responses ($\boldsymbol{y}$) for a given input line ($\boldsymbol{x}$)
- Approach - Neural variational models

Plan

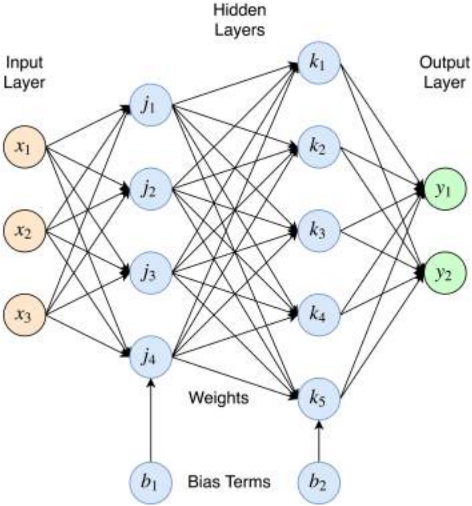
Deep Learning

- Subfield of machine learning
- Use of artificial neural networks
    - Inspired from neurons in the brain
    - Deep architectures
    - Outperform humans in a number of cognitive tasks
    - Massive amounts of data, powerful hardware
- Perceptron [Rosenblatt, 1958]



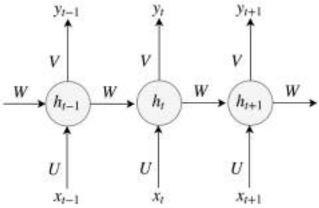
Feedforward Neural Networks

- Multiple layers
- Non-linear Activation functions
- Forward propagation
- Compute loss
- Weight update by Error Backpropagation
- Stochastic Gradient Descent (SGD) / ADAM



Recurrent Neural Networks

- Text data - expressed as a sequence
- RNNs
  - Feed inputs in a sequential manner
  - The hidden state contains info until t
  - $h_t = f(Ux_t + Wh_{t-1})$; $y_t = Vh_t$
  - Weight sharing

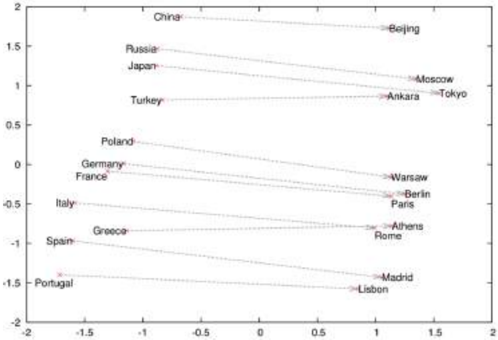


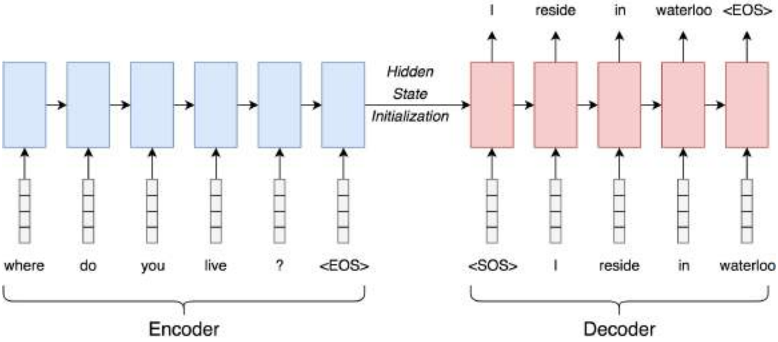
- Vanilla RNNs in practice
  - unable to remember the dependencies between inputs which are far apart in the sequence
- **Solution**: LSTM-RNNs [Hochreiter and Schmidhuber, 1997]
  - Better at capturing long term dependencies
  - An entire module (known as a *cell*) with a set of gates to replace f
  - Compute a hidden state h_t and a cell state c_t at each timestep

Word Embeddings

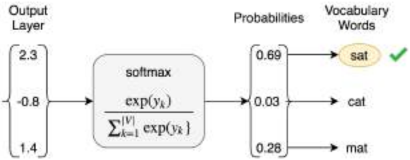
- Cannot directly input raw text into an ML algorithm
- Need to map the textual data into corresponding numeric representations
- **Solution**: `word2vec` - fixed vector representations for each word [Mikolov et al., 2013]

- Based on distributional similarity - "words that occur in similar contexts would have similar meaning" Eg. *sports* and *game*
- W: *words* $\rightarrow \mathbb{R}^n$, where n is the dimension of each word vector





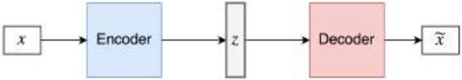
- Encoder and Decoder are RNNs with LSTM units
- Hidden state initialization
- Teacher Forcing
- Output Softmax layer

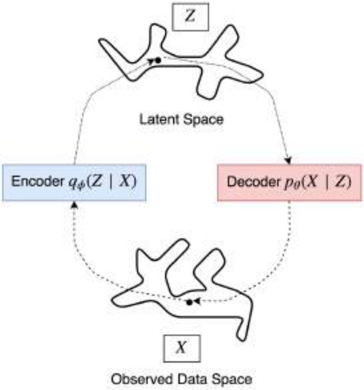


Plan

Autoencoding (Deterministic)

- Obtain a compressed representation of the data x from which it is possible to re-construct it
- Encoder $\boldsymbol{q_\phi}(\boldsymbol{z}|\boldsymbol{x})$ and Decoder $\boldsymbol{p_\theta}(\boldsymbol{x}|\boldsymbol{z})$ are jointly trained to maximize the conditional log-likelihood
- The latent representation $\boldsymbol{z}$ has an arbitrary distribution



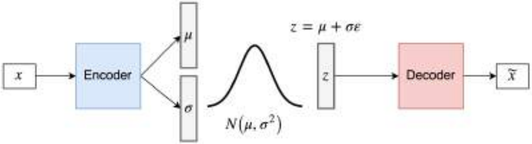


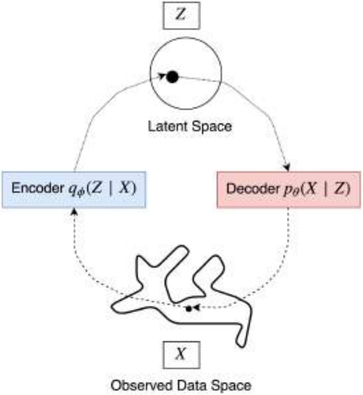
Minimize Reconstruction Loss

$$J = -\sum_{n=1}^{N} \sum_{t=1}^{|x^{(n)}|} \log p(x_t^{(n)} | z^{(n)}, x_{<t}^{(n)})$$

Variational Autoencoder [Kingma and Welling, 2013]

- Enforce a distribution on the latent space
- Minimize the Kullback-Leibler (KL) divergence between the learnt posterior and a pre-specified prior: $\mathrm{KL}(\mathcal{N}(\mu, \sigma)||\mathcal{N}(0, I))$
- Balance between reconstruction and KL penalty term
  - High λ - Ignores reconstruction
  - Low λ - Deterministic behaviour





Minimize Reconstruction Loss + KL Divergence

$$J = \sum_{n=1}^{N} \left[- \mathop{\mathbb{E}}_{z^{(n)} \sim q} \sum_{t=1}^{|x^{(n)}|} \log p(x_t^{(n)} | z^{(n)}, x_{<t}^{(n)}) + \lambda \cdot \mathrm{KL}(q(z^{(n)} | x^{(n)}) \| p(z)) \right]$$

Training Heuristics

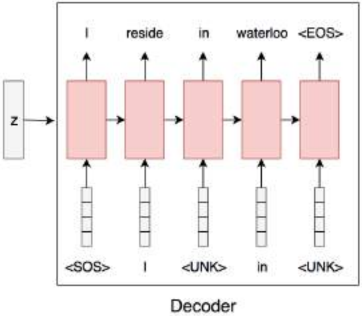
- Training VAEs for text generation is notoriously difficult
- Adopt two training strategies [Bowman et al., 2015]

KL Weight Annealing

- Gradually increase λ from zero to a threshold value
- Deterministic autoencoder $\rightarrow$ Variational autoencoder
- Experiment with different annealing schedules

Word Dropout

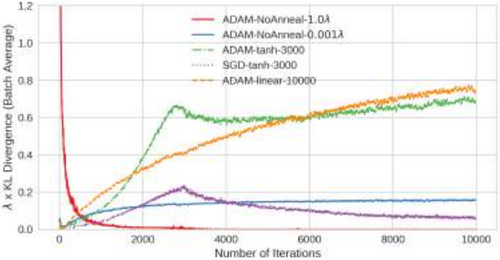
- Replace decoder inputs with $<$UNK$>$ with probability p
- Weakens the decoder and encourages the model to encode more information into $\boldsymbol{z}$



VAE Variants

- Trained on 80k sentences of the SNLI dataset
- Evaluating reconstruction performance with BLEU scores
- BLEU-$j = \min\left(1, \frac{\text{generated-length}}{\text{reference-length}}\right) * (\text{precision}_j)$

Model	BLEU-4
Deterministic AE	73.73
ADAM-NoAnneal-1.0	2.05
ADAM-NoAnneal-0.001	72.05
ADAM-tanh-3000	36.50
SGD-tanh-3000	2.70
ADAM-linear-10000	35.29

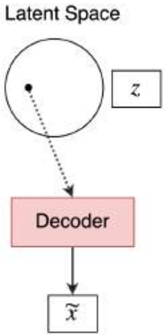


- Non-linear annealing $\lambda_i = \frac{\tanh\left(\frac{i-4500}{1000}\right)+1}{2}$
- Linear annealing $\lambda_i = \frac{i}{200000}$

Random Sampling

- VAEs exhibit interesting properties due to their learnt latent space
- Continuous latent space $\implies$ meaningful sentences
- Discard encoder; Sample from prior $\mathcal{N}(0, I)$ and generate
- New and interesting sentences unseen in the training data

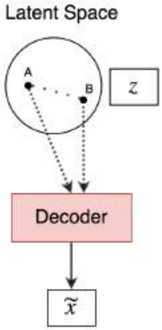
Deterministic AE	**ADAM-NoAnneal-1.0**
a men wears an umbrella waits to *a couple cows a monument* *there is sleeping and two rug .* *a man in a pick photos* *a boy are people at a lake escape .*	*a man is sitting on a bench .* *a man is sitting on a bench .* *a man is sitting on a bench .* *a man is sitting on a bench .* *a man is sitting on a bench .*
ADAM-NoAnneal-0.001	**ADAM-tanh-3000**
i woman who is on watch a factory *they are excited formation to ride a castle of a* *their janitor is leaving the dirt wearing his suits .* *two children in it exits a* *six people sitting are sorting at single radio in .*	*the dog is sleeping in the grass .* *the girls are being detained .* *the group of people are going to begin .* *a girl with blond-hair on a bike with a stick* *a woman and a man are walking on a street*



Linear Interpolation

- To test the continuity of the latent space
- $\boldsymbol{z}_{\alpha_i} = \alpha_i \cdot \boldsymbol{z}_A + (1 - \alpha_i) \cdot \boldsymbol{z}_B$ where $\alpha_i \in \left[0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\right]$
- VAE - Smooth transition maintaining syntax and semantics
- DAE - Transition is irregular and non-continuous

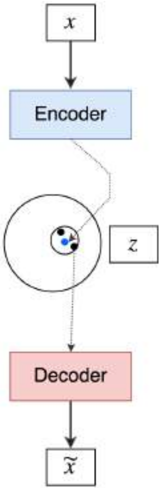
Deterministic AE	**Variational AE**
Sentence A: there is a couple eating cake .	
there is a couple eating cake . *there is a couple eating cake .* *there is a couple eating cake .* *there is a group of people eating a party .* *a group of men are watching a party .* *a group of men are watching a dance party .* *a group of men are watching a dance party .* *a group of men are watching a dance party .*	*there is a couple eating cake .* *there is a couple eating .* *there is a couple eating dinner .* *there is a couple of people eating dinner .* *a group of people are having a conversation .* *a group of men are having a discussion .* *a group of men are watching a movie .* *a group of men are watching a movie theater .*
Sentence B: a group of men are watching a dance party .	



Sampling from Neighborhood

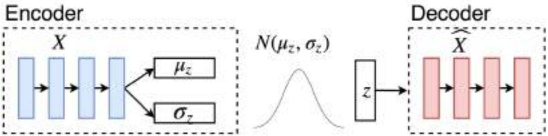
- For a given input $\boldsymbol{x}$, sample the latent vector as $\boldsymbol{z} = \boldsymbol{\mu} + 3\boldsymbol{\sigma} \otimes \epsilon$
- VAE - generates diverse sentences, however topically similar to the input.
- DAE - latent space has empty regions

Deterministic AE	Variational AE
Input Sentence: a dog with its mouth open is running .	
a dog with its mouth is open running . *a dog with its mouth is open running .* *a dog with its mouth is open running .*	*a dog with long hair is eating .* *a guy and the dogs are holding hands* *a dog with a toy at a rodeo .*
Input Sentence: there are people sitting on the side of the road	
there are people sitting on the side of the road *there are people sitting on the side of the road* *there are people sitting on the side of the road*	*the boy is walking down the street .* *there are people standing on the street outside* *the police are on the street corner .*

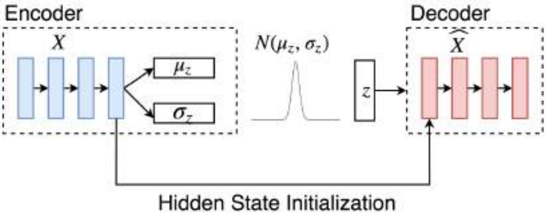


VAE Bypassing Phenomenon

- Design considerations
- z is sampled and fed to the decoder
- Encode useful information in the latent space



- With **bypass connection**, the decoder has direct deterministic access to the source info
- Latent space ignored, KL divergence doesn't act as a regularizer



Diversity Evaluation Metrics

For a given input $\boldsymbol{x}$, generate multiple outputs $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_k$

Entropy

- Compute unigram probability $p(w)$ of each word in the generated set
- $H = -\sum_w p(w) \log p(w)$
- More entropy $\implies$ more randomness $\implies$ more diversity

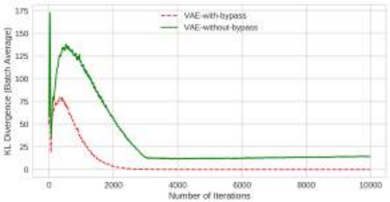
Distinct Scores

- Distinct-1 $= \frac{\text{Count of distinct unigrams}}{\text{Total unigram count}}$
- Distinct-2 $= \frac{\text{Count of distinct bigrams}}{\text{Total bigram count}}$

Effect on Latent Space

- VAE without hidden state initialization generates diverse outputs
- Bypass connection degrades the model to a deterministic AE

	VAE with Bypass	VAE without Bypass
Entropy	2.004	2.686
Distinct-1	0.099	0.302
Distinct-2	0.118	0.502

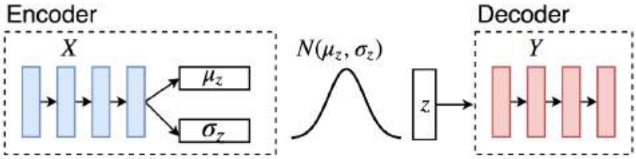


VAE with Bypass	VAE without Bypass
Input Sentence: the men are playing musical instruments	
the men are playing musical instruments *the man is playing musical instruments* *the men are playing musical instruments*	*the men are playing video games* *the men are playing musical instruments* *the musicians are playing musical instruments*
Input Sentence: a child holds a shovel on the beach .	
a child holds a shovel on the beach . *a child holds a shovel on the beach .* *a child holds a shovel on the beach .*	*a child playing with the ball on the beach .* *a child holding a toy on the water .* *a child holding a toy on the beach .*

Plan

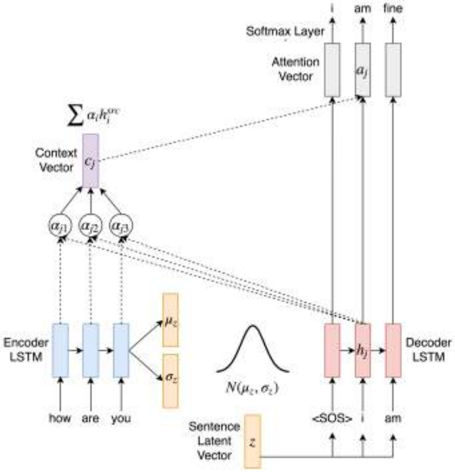
VED Introduction

- Transform an input sequence (X) into a different output sequence (Y)
- E.g., machine translation, text summarization, dialog generation



Deterministic Attention [Bahdanau et al., 2014]

- Performance improvements to existing Seq2Seq models
- Align source information on the encoder side to target information on the decoder side
- During each timestep j, the decoder weights the source tokens
- Pre-normalized score: $\widetilde{\alpha}_{ji} = \boldsymbol{h}_j^{(\text{tar})} W^T \boldsymbol{h}_i^{(\text{src})}$
- Attention weights: $\alpha_{ji} = \frac{\exp\{\widetilde{\alpha}_{ji}\}}{\sum_{i'=1}^{|\boldsymbol{x}|} \exp\{\widetilde{\alpha}_{ji'}\}}$
- Unfortunately, deterministic attention serves as a **bypass** connection



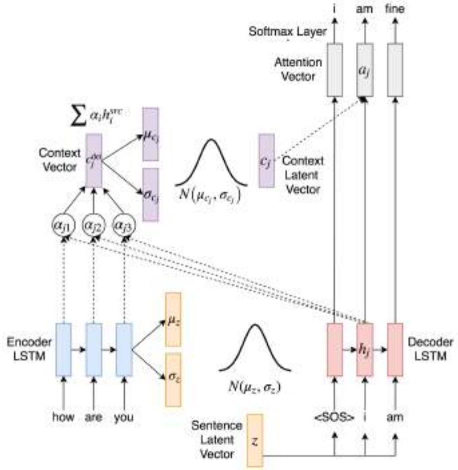
Variational Attention

- Treat the **context vector** as a random variable with a pre-defined prior distribution
- With 2 latent spaces:

Loss Function

$$J_{\text{rec}}(\boldsymbol{\theta}, \phi, y^{(n)}) +$$
$$\lambda \Big[\text{KL}\left(q_\phi^{(z)}(z|x^{(n)}) \| p(z) \right) +$$
$$\gamma_a \sum_{j=1}^{|y|} \text{KL}\left(q_\phi^{(c_j)}(c_j|x^{(n)}) \| p(c_j) \right) \Big]$$

- Two proposed priors $p(c_j)$:
  1. $\mathcal{N}(0, I)$
  2. $\mathcal{N}(\bar{h}^{(\text{src})}, I)$, where $\bar{h}^{(\text{src})} = \frac{1}{|x|} \sum_{i=1}^{|x|} h_i^{(\text{src})}$

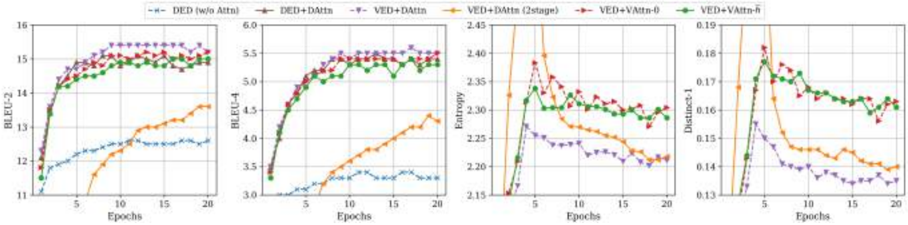


Experiment 1 - Question Generation

- Applications in ecommerce (generating FAQs), educational purposes
- Dataset: Stanford Question Answering Dataset (SQuAD)
- 100k question-answer pairs
- S: `zinc is a chemical element with symbol zn and atomic number 30`
  Q: `what is the symbol for zinc ?`

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

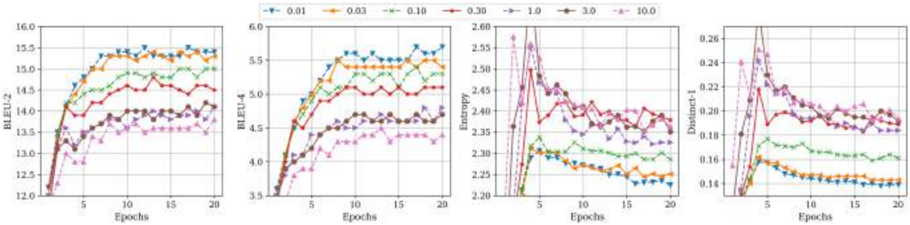


- Proposed models always have a higher diversity throughout training, while maintaining the BLEU scores

Strength of Attention KL Term

Loss Function

$$J_{\text{rec}}(\boldsymbol{\theta}, \phi, y^{(n)}) + \lambda\Big[\mathrm{KL}\left(q_\phi^{(z)}(z|x^{(n)}) \| p(z)\right) +$$
$$\gamma_a \sum_{j=1}^{|y|} \mathrm{KL}\left(q_\phi^{(c_j)}(c_j|x^{(n)}) \| p(c_j)\right) \Big]$$



- Low γ_a - model behaves *deterministically*
- High γ_a - achieves a higher diversity at the cost of output reconstruction performance

Experiment 2 - Dialog Systems

- Generative conversational agent
- Dataset: Cornell Movie-Dialogs Corpus
- 200k conversational exchanges from 617 movies
- M: `so what should i do with the pudding?`
  R: `lets just leave it there for now.`

Model	**Inference**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**Entropy**	**Distinct-1**	**Distinct-2**
DED+DAttn	MAP	5.75	1.84	0.99	0.64	-	-	-
VED+DAttn	MAP	5.33	1.68	0.88	0.57	-	-	-
	Sampling	5.34	1.68	0.89	0.57	2.113	0.311	0.450
VED+VAttn-$\bar{h}$	MAP	5.48	1.78	0.97	0.64	-	-	-
	Sampling	5.55	1.79	0.97	0.64	**2.167**	**0.324**	**0.467**

Qualitative Results

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 ?*

Human Evaluation Study for Comparing Language Fluency

- Each model - 100 generated questions
- 5 - Flawless, 4 - Good, 3 - Adequate, 2 - Poor, 1 - Incomprehensible
- VED+DAttn $\rightarrow$ 3.99 ; VED+VAttn-$\bar{h}$ $\rightarrow$ 4.01
- VAttn does not negatively affect the fluency of sentences

Plan

Summary and Conclusions

- VAE for text generation was first designed, trained successfully by adopting - (1) KL weight annealing, (2) Word dropout; Demonstrated the effectiveness of the latent space
- Negative impact of bypassing connections
- Traditional attention mechanism serves as bypassing. To circumvent this issue, variational attention is proposed
- Two possible priors to model the attention context vector
- Experiments on two tasks show that the proposed model yields higher diversity while retaining high quality of generated sentences.

References I

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

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

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

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

Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*, 2017.