

Variational Autoencoders Write Poetry

(Generating Sentences from a
Continuous Space)

Elsbeth Turcan and Fei-Tzin Lee

Paper by Sam Bowman, Luke Vilnis et al.

2016



Motivation

- Generative models for natural language sentences
 - Machine translation
 - Image captioning
 - Dataset summarization
 - Chatbots
 - Etc.
- Want to capture high-level features of text such as topic and style and keep them consistent when generating text



Related work - RNNLM

- In the words of Bowman et al., “A standard RNN language model predicts each word of a sentence conditioned on the previous word and an evolving hidden state.”
- In other words, it only looks at the relationships between consecutive words, and so does not contain or observe any global features
- But what if we want global information?

A decorative graphic of a feather, rendered in a light beige or tan color, is positioned on the left side of the slide. It has a central rachis with numerous barbs extending outwards, creating a fan-like shape. The feather is oriented vertically, with the base at the bottom and the tip at the top.

Other related work

- Skip-thought
 - Generate sentence codes in the style of word embeddings to predict context sentences
- Paragraph vector
 - A vector representing the paragraph is incorporated into single-word embeddings



Autoencoders

- Typically composed of two RNNs
- The first RNN encodes a sentence into an intermediate vector
- The second RNN decodes the intermediate representation back into a sentence, ideally the same as the input



Variational Autoencoders (VAEs)

- Regular autoencoders learn only discrete mappings from point to point
- However, if we want to learn holistic information about the structure of sentences, we need to be able to fill sentence space better
- In a VAE, we replace the hidden vector \mathbf{z} with a posterior probability distribution $q(\mathbf{z}|x)$ conditioned on the input, and sample our latent \mathbf{z} from that distribution at each step
- We ensure that this distribution has a tractable form by enforcing its similarity to a defined prior distribution, typically some form of Gaussian



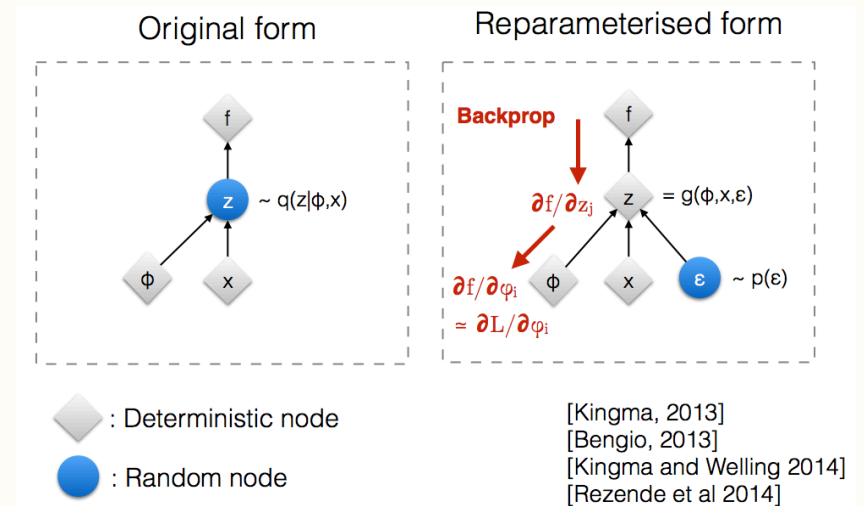
Modified loss function

- The regular autoencoder's loss function would encourage the VAE to learn posteriors as close to discrete as possible – in other words, Gaussians that are clustered extremely tightly around their means
- In order to enforce our posterior's similarity to a well-formed Gaussian, we introduce a KL divergence term into our loss, as below:

$$\begin{aligned}\mathcal{L}(\theta; x) &= -\text{KL}(q_\theta(\vec{z}|x) || p(\vec{z})) \\ &\quad + \mathbb{E}_{q_\theta(\vec{z}|x)}[\log p_\theta(x|\vec{z})] \\ &\leq \log p(x) .\end{aligned}$$

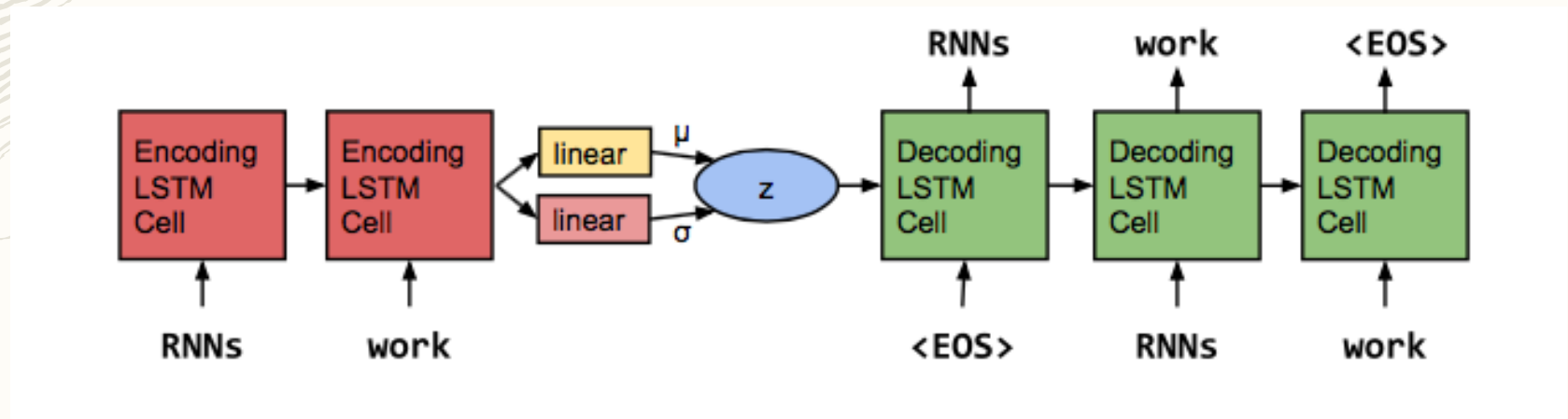
Reparameterization trick

- In the original formulation, the encoder net encodes the sentence into a probability distribution (usually Gaussian); practically speaking, it encodes the sentence into the parameters of the distribution (i.e. μ and σ)
- However, this poses challenges for us while backpropagating: we can't backpropagate over the jump from μ and σ to z , since it's random
- Solution: extract the randomness from the Gaussian by reformulating it as a function of μ , σ , and another separate random variable



Specific architecture

- Single-layer LSTM for encoder and decoder



Issues and fixes

- Decoder too strong, without any limitations just doesn't use z at all
- Fix: KL annealing
- Fix: word dropout

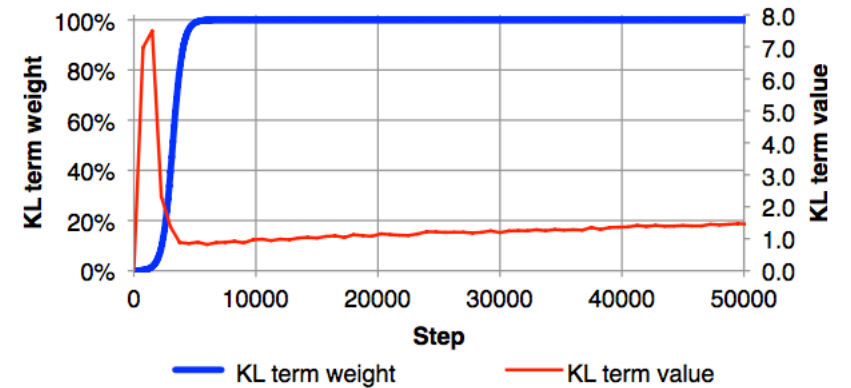



Figure 2: The weight of the KL divergence term of variational lower bound according to a typical sigmoid annealing schedule plotted alongside the (unweighted) value of the KL divergence term for our VAE on the Penn Treebank.



Experiments – Language modeling

- Used VAE to create language models on the Penn Treebank dataset, with RNNLM as baseline
 - Task: train an LM on the training set and have it designate the test set as highly probable
- RNNLM outperformed the VAE in the traditional setting
- However, when handicaps were imposed on both models (inputless decoder), the VAE was significantly better able to overcome them

Model	Standard				Inputless Decoder			
	Train NLL	Train PPL	Test NLL	Test PPL	Train NLL	Train PPL	Test NLL	Test PPL
RNNLM	100 –	95	100 –	116	135 –	600	135 –	> 600
VAE	98 (2)	100	101 (2)	119	120 (15)	300	125 (15)	380

Table 2: Penn Treebank language modeling results, reported as negative log likelihoods and as perplexities. Lower is better for both metrics. For the VAE, the KL term of the likelihood is shown in parentheses alongside the total likelihood.



Experiments – Imputing missing words

- Task: infer missing words in a sentence given some known words (imputation)
- Place the unknown words at the end of the sentence for the RNNLM
- RNNLM and VAE performed beam search (VAE decoding broken into three steps) to produce the most likely words to complete a sentence
- Precise evaluation of these results is computationally difficult

<i>but now , as they parked out front and owen stepped out of the car , he could see _ _ _ _ _</i>		
True: <i>that the transition was complete .</i>	RNNLM: <i>it , " i said .</i>	VAE: <i>through the driver 's door .</i>
<i>you kill him and his _ _</i>		
True: <i>men .</i>	RNNLM: <i>. "</i>	VAE: <i>brother .</i>
<i>not surprising , the mothers dont exactly see eye to eye with me _ _ _ _</i>		
True: <i>on this matter .</i>	RNNLM: <i>, i said .</i>	VAE: <i>, right now .</i>

Table 3: Examples of using beam search to impute missing words within sentences. Since we decode from right to left, note the stereotypical completions given by the RNNLM, compared to the VAE completions that often use topic data and more varied vocabulary.

Adversarial evaluation

- Instead, create an adversarial classifier, trained to distinguish real sentences from generated sentences, and score the model on how well it fools the adversary
- Adversarial error is defined as the gap between chance accuracy (50%) and the real accuracy of adversary – ideally this error will be minimized

Model	Adv. Err. (%)		NLL RNNLM
	Unigram	LSTM	
RNNLM (15 bm.)	28.32	38.92	46.01
VAE (3x5 bm.)	22.39	35.59	46.14

Table 4: Results for adversarial evaluation of imputations. Unigram and LSTM numbers are the *adversarial error* (see text) and RNNLM numbers are the negative log-likelihood given to entire generated sentence by the RNNLM, a measure of sentence typicality. Lower is better on both metrics. The VAE is able to generate imputations that are significantly more difficult to distinguish from the true sentences.

Experiments - Other

- Several other experiments in the appendix showed the VAE to be applicable to a variety of tasks
 - Text classification
 - Paraphrase detection
 - Question classification



Analysis

- Word dropout
 - Keep rate too low: sentence structure suffers
 - Keep rate too high: no creativity, stifles the variation
- Effects on cost function components:

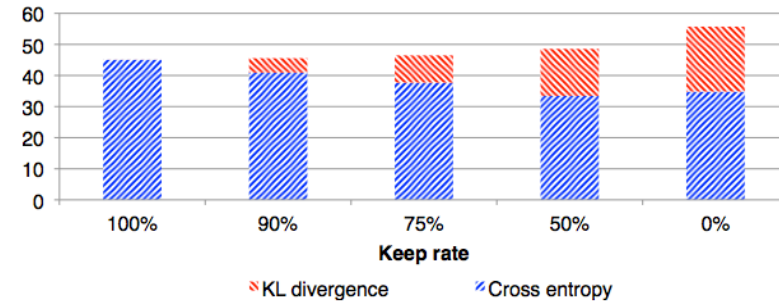


Figure 3: The values of the two terms of the cost function as word dropout increases.

Extras: sampling from the posterior and homotopies

- Sampling from the posterior: examples of sentences adjacent in sentence space

INPUT	we looked out at the setting sun .	i went to the kitchen .	how are you doing ?
MEAN	<i>they were laughing at the same time .</i>	<i>i went to the kitchen .</i>	<i>what are you doing ?</i>
SAMP. 1	<i>ill see you in the early morning .</i>	<i>i went to my apartment .</i>	<i>“ are you sure ?</i>
SAMP. 2	<i>i looked up at the blue sky .</i>	<i>i looked around the room .</i>	<i>what are you doing ?</i>
SAMP. 3	<i>it was down on the dance floor .</i>	<i>i turned back to the table .</i>	<i>what are you doing ?</i>

Table 7: Three sentences which were used as inputs to the VAE, presented with greedy decodes from the mean of the posterior distribution, and from three samples from that distribution.

- Homotopies: linear interpolations in sentence space between the codes for two sentences

i went to the store to buy some groceries . <i>i store to buy some groceries .</i> <i>i were to buy any groceries .</i> <i>horses are to buy any groceries .</i> <i>horses are to buy any animal .</i> <i>horses the favorite any animal .</i> <i>horses the favorite favorite animal .</i> horses are my favorite animal .
--

Table 1: Sentences produced by greedily decoding from points between two sentence encodings with a conventional autoencoder. The intermediate sentences are not plausible English.

“ i want to talk to you . ” <i>“i want to be with you . ”</i> <i>“i do n’t want to be with you . ”</i> <i>i do n’t want to be with you .</i> she did n’t want to be with him .
he was silent for a long moment . <i>he was silent for a moment .</i> <i>it was quiet for a moment .</i> <i>it was dark and cold .</i> <i>there was a pause .</i> it was my turn .

Table 8: Paths between pairs of random points in VAE space: Note that intermediate sentences are grammatical, and that topic and syntactic structure are usually locally consistent.

Even more homotopies

amazing , is n't it ?
so , what is it ?
it hurts , isnt it ?
why would you do that ?
" you can do it .
" i can do it .
i ca n't do it .
" i can do it .
" do n't do it .
" i can do it .
i could n't do it .

no .
he said .
" no , " he said .
" no , " i said .
" i know , " she said .
" thank you , " she said .
" come with me , " she said .
" talk to me , " she said .
" do n't worry about it , " she said .

i dont like it , he said .
i waited for what had happened .
it was almost thirty years ago .
it was over thirty years ago .
that was six years ago .
he had died two years ago .
ten , thirty years ago .
" it 's all right here .
" everything is all right here .
" it 's all right here .
it 's all right here .
we are all right here .
come here in five minutes .

this was the only way .
it was the only way .
it was her turn to blink .
it was hard to tell .
it was time to move on .
he had to do it again .
they all looked at each other .
they all turned to look back .
they both turned to face him .
they both turned and walked away .

there is no one else in the world .
there is no one else in sight .
they were the only ones who mattered .
they were the only ones left .
he had to be with me .
she had to be with him .
i had to do this .
i wanted to kill him .
i started to cry .
i turned to him .

im fine .
youre right .
" all right .
you 're right .
okay , fine .
" okay , fine .
yes , right here .
no , not right now .
" no , not right now .
" talk to me right now .
please talk to me right now .
i 'll talk to you right now .
" i 'll talk to you right now .
" you need to talk to me now .
" but you need to talk to me now .

Table 12: Selected homotopies between pairs of random points in the latent VAE space.

Thanks for listening!

– Any questions?

