Generating Sentences from a Continuous Space

Samuel R. Bowman, Luke Vilnis, et al.

Presenting: Yevgeny Tkach

https://qdata.github.io/deep2Read/

Executive Summary

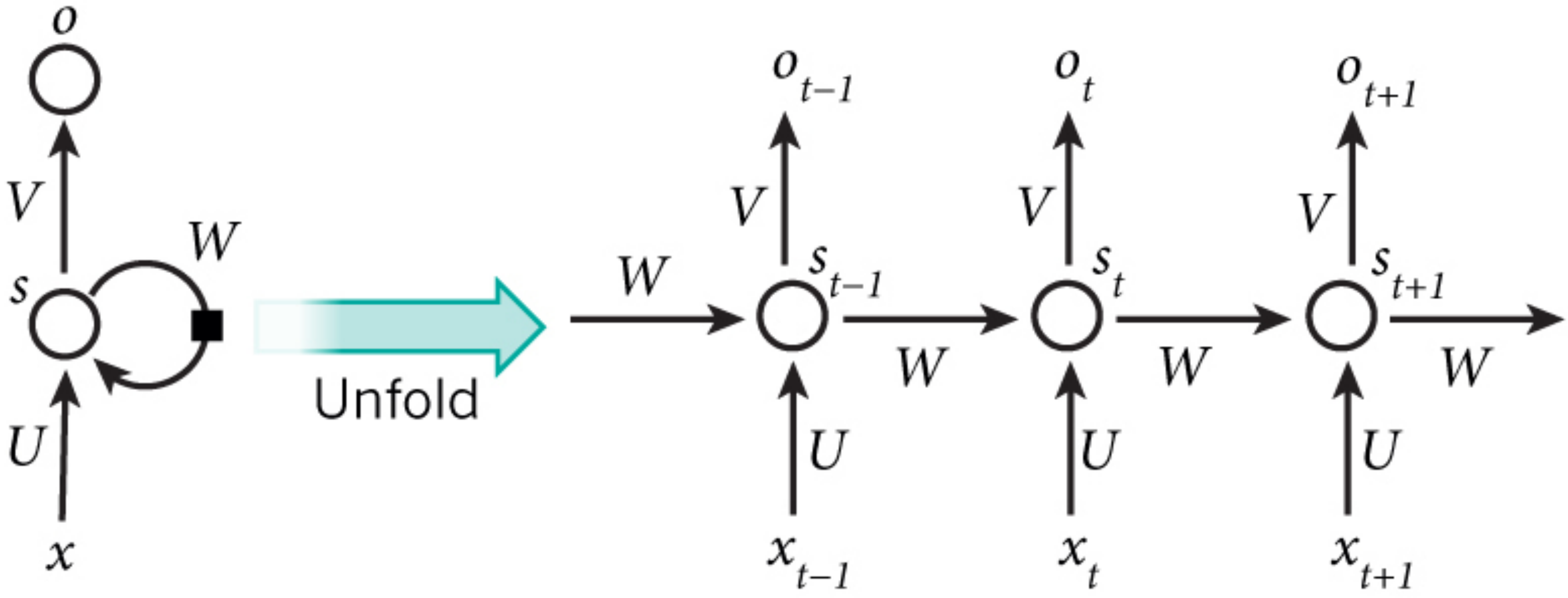
- The authors present an RNN-based variational autoencoder for language modeling.
- A linear layer that predicts the parameters of a Gaussian distribution.
- Straight forward implementation of the VAE fails to learn a latent representation of sentences. The solutions that were used by the authors are presented.
- The model is evaluated on Language Modeling and Word Imputation tasks.
- Qualitative analysis of the latent space is provided.

Outline

- RNNLM
- Autoencoder
- VAE – Variational Auto Encoders
- RNNLM + VAE
- Language Modeling
- Word Imputation
- More analysis

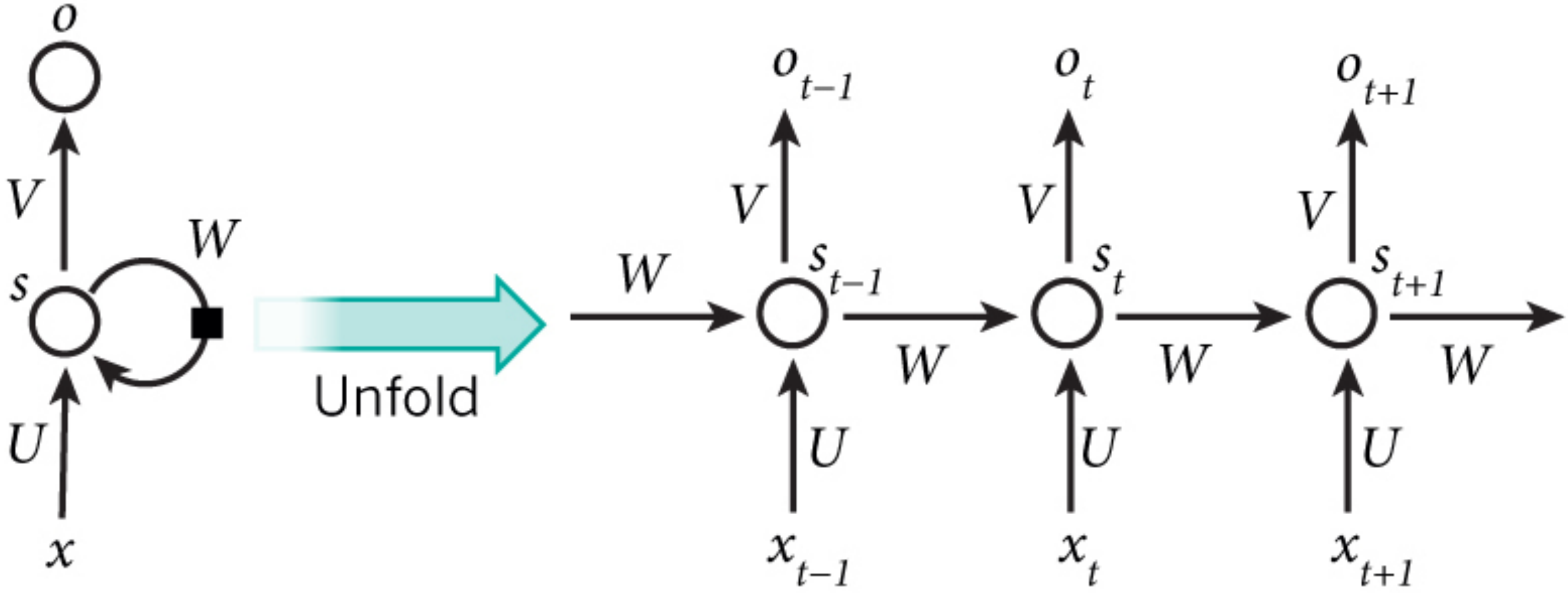
RNNLM

- A statistical **Language Model** is a probability distribution over sequences of words. Given such a sequence, say of length m, it assigns a probability $P(w_1, w_2, \ldots, w_m)$ to the hole sequence.
- Language models are successfully trained using RNNs and specifically LSTMs.



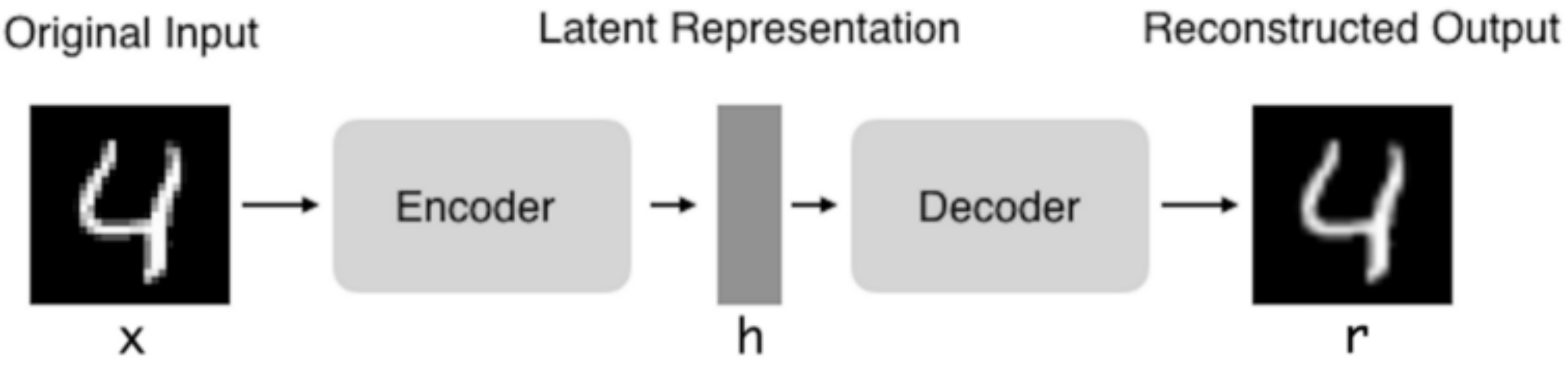
RNNLM

- RNNLMs are trained on some text corpus.
- The input and output are words in a "vocabulary", one-hot encoded.
- Cross entropy loss



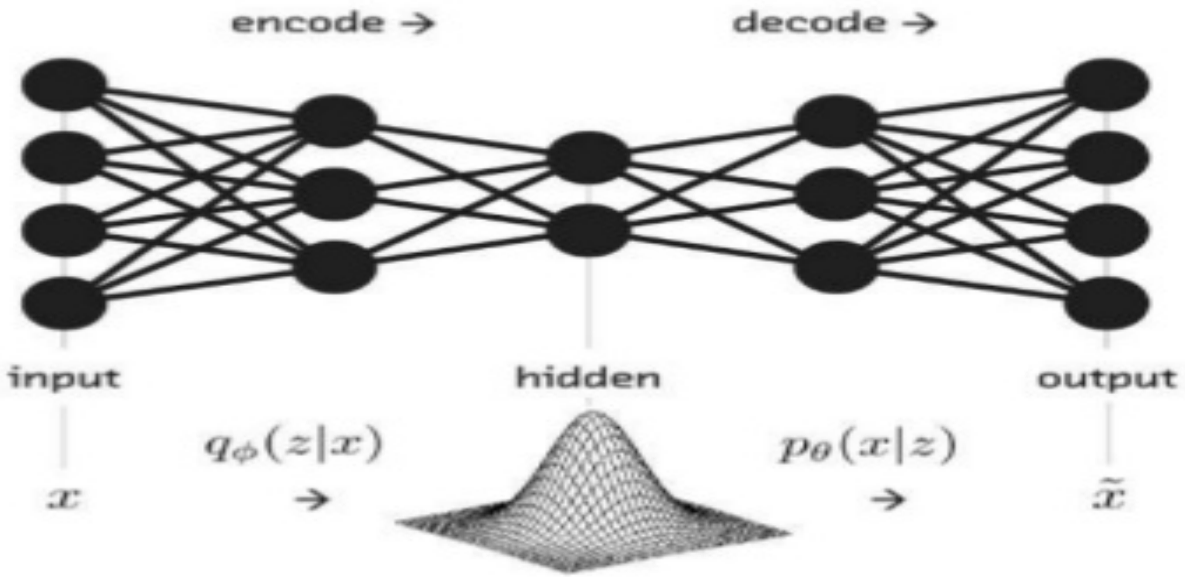
Autoencoder

- Unsupervised algorithm to create a representation of the original data (usually low dimensional)
- Encoder and Decoder are both neural network
- To train an auto encoder we minimize a distance measure between the input and the output



VAE – Variational Auto Encoders

- Basic idea: encoder will define a distribution over the latent representations of the input
- The Decoder samples from the latent distribution and can generate new previously unseen outputs.



VAE – Variational Auto Encoders

- The regular autoencoder's loss function would encourage the VAE to learn deterministic representations – 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:

KL loss

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

Reconstruction loss

ELBO

VAE – Variational Auto Encoders

- We want to minimize over λ:

$$KL(q_\lambda(z|x)||p(z|x)) =$$

$$\mathbf{E}_q[\log q_\lambda(z|x)] - \mathbf{E}_q[\log p(x,z)] + \log p(x)$$

- Which is equivalent to maximize over λ*:*

$$ELBO(\lambda) = \mathbf{E}_q[\log p(x,z)] - \mathbf{E}_q[\log q_\lambda(z|x)].$$

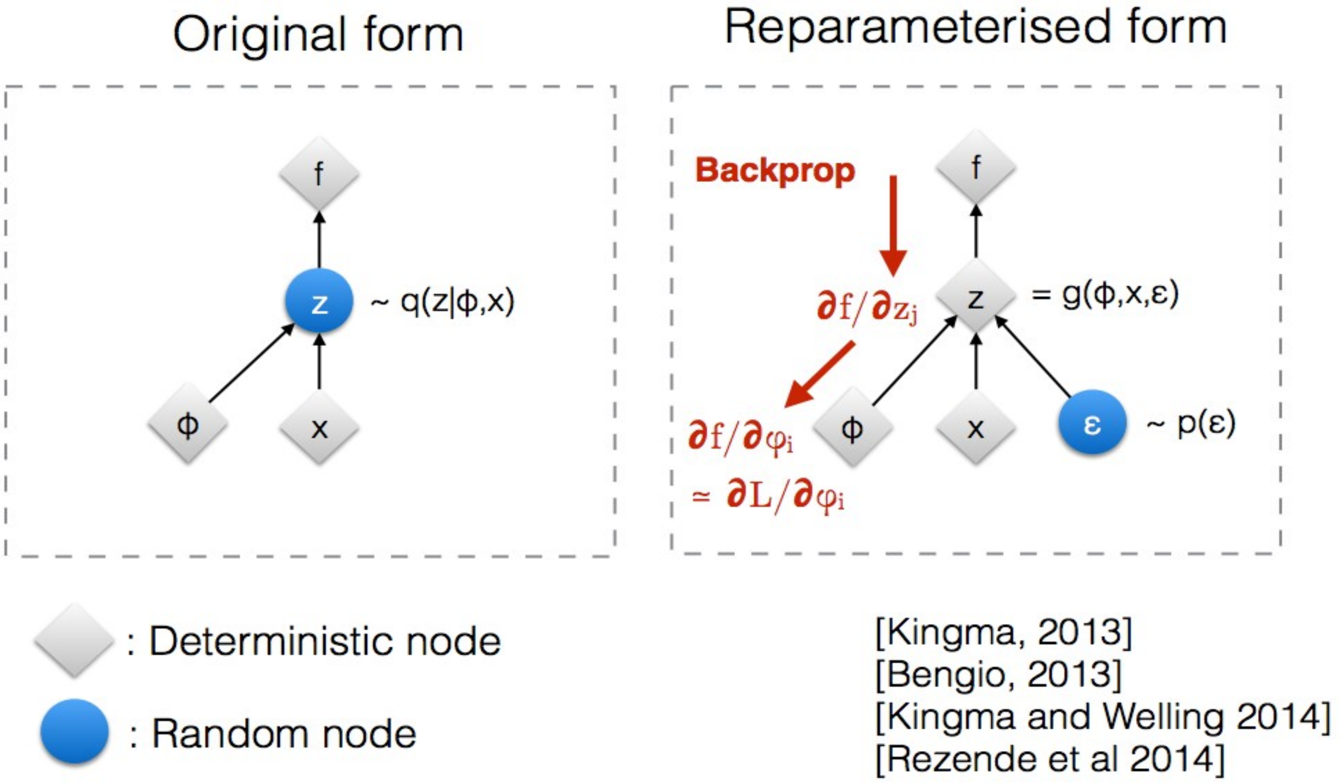
- That can be rewritten as:

$$ELBO_i(\lambda) = E_{q_\lambda(z|x_i)}[\log p(x_i|z)] - KL(q_\lambda(z|x_i)||p(z)).$$

- Maximizing the Evidence Lower BOund on the true log likelihood of the data.

VAE – Variational Auto Encoders

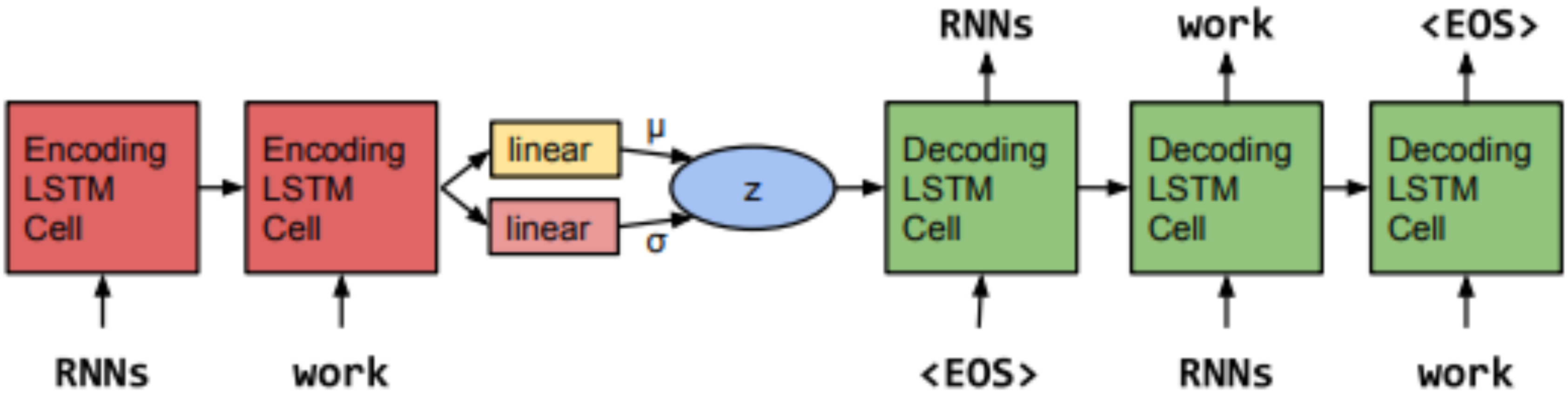
- To train VAE we can calculate the KL loss analytically for each data point
- In order to back propagate we use a reparameterization trick:



[Kingma, 2013]
[Bengio, 2013]
[Kingma and Welling 2014]
[Rezende et al 2014]

RNNLM + VAE

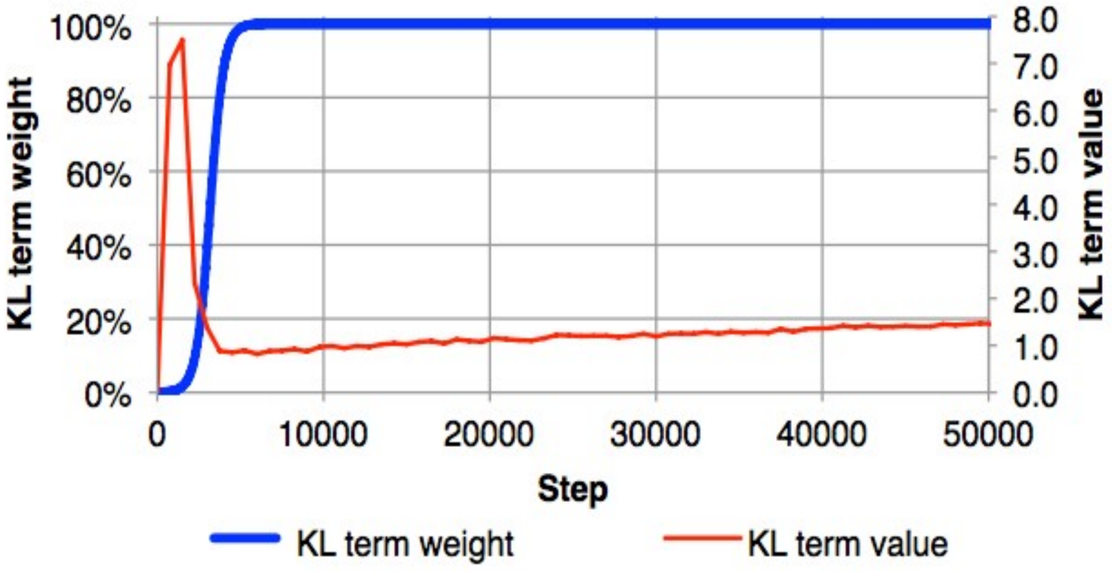
- Single-layer LSTM for encoder and decoder for language modeling with VAE.



- The use of linear model between the decoder and encoder is not mandatory.
- The authors reported to experiment with other architectures as well, however without gains over the simple linear encoding.

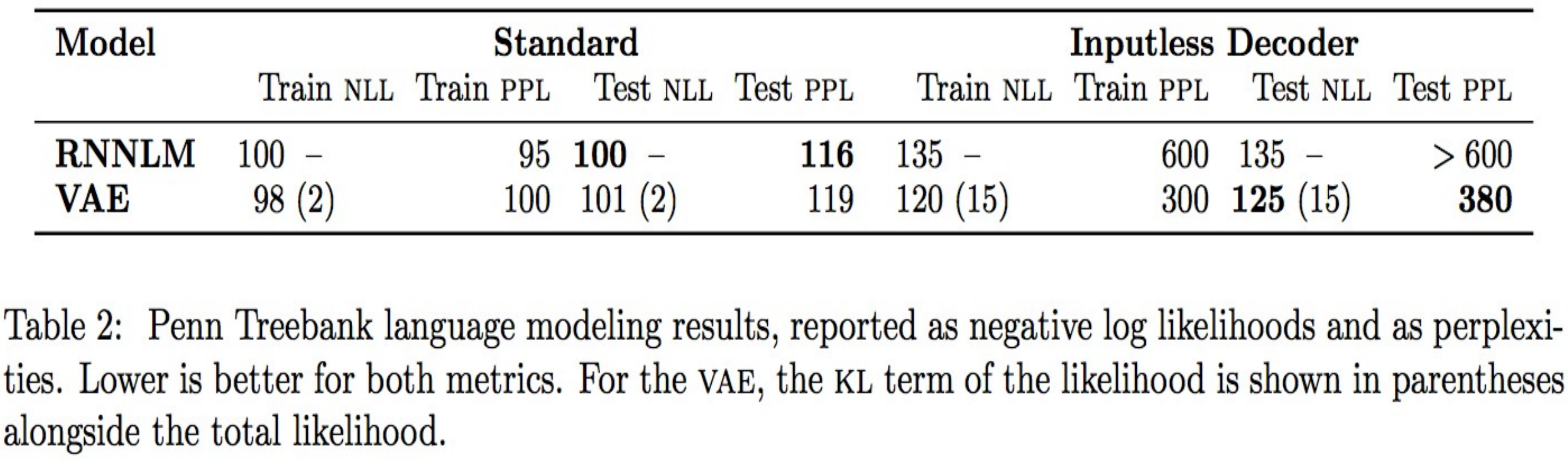
RNNLM + VAE

- VAE **fails** to learn a latent representation of the sentence content.
- $q(\vec{z}|x)$ is equal to the prior p$(\vec{z})$ bringing the KL divergence term to 0.
- Solution 1: KL cost annealing
- Solution 2: Word dropout



Language modeling

- Used VAE to create language models on the Penn Treebank dataset, with RNNLM as baseline
- RNNLM outperformed the VAE in the traditional setting
- However, when handicaps were imposed on both models (inputless decoder), the VAE performed better

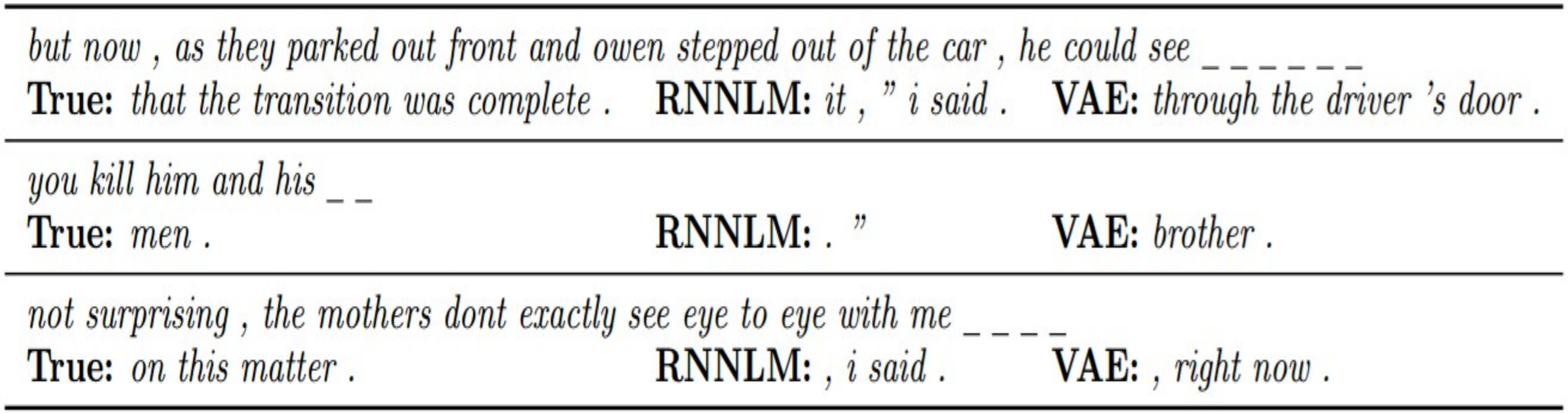


Model	Standard				Inputless Decoder			
	Train NLL	Train PPL	Test NLL	Test PPL	Train NLL	Train PPL	Test NLL	Test PPL
RNNLM	100 –	95	**100** –	**116**	135 –	600	135 –	> 600
VAE	98 (2)	100	101 (2)	119	120 (15)	300	**125** (15)	**380**

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.

Word Imputation

- Task: infer missing words in a sentence given some known words (imputation)
- RNNLM works well only when the unknoun words are at the end of the sentence
- RNNLM and VAE performed a beam search. VAE decoding broken into 3 samples with shorter beam



but now , as they parked out front and owen stepped out of the car , he could see _ _ _ _ _ _

True: *that the transition was complete .* **RNNLM:** *it , " i said .* **VAE:** *through the driver 's door .*

you kill him and his _ _

True: *men .* **RNNLM:** *. "* **VAE:** *brother .*

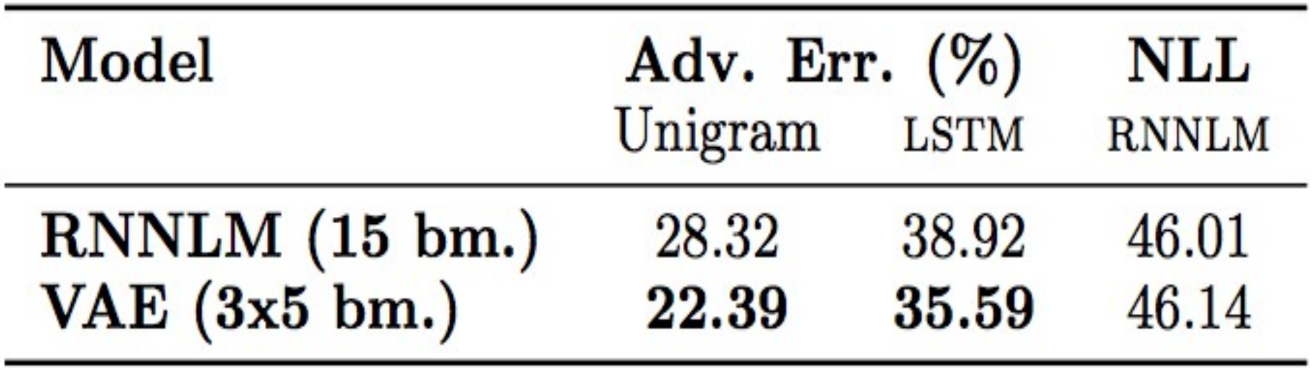
not surprising , the mothers dont exactly see eye to eye with me _ _ _ _

True: *on this matter .* **RNNLM:** *, i said .* **VAE:** *, right now .*

  - We can see that that VAE is more diverse and keeps the topic of the sentence

Word Imputation

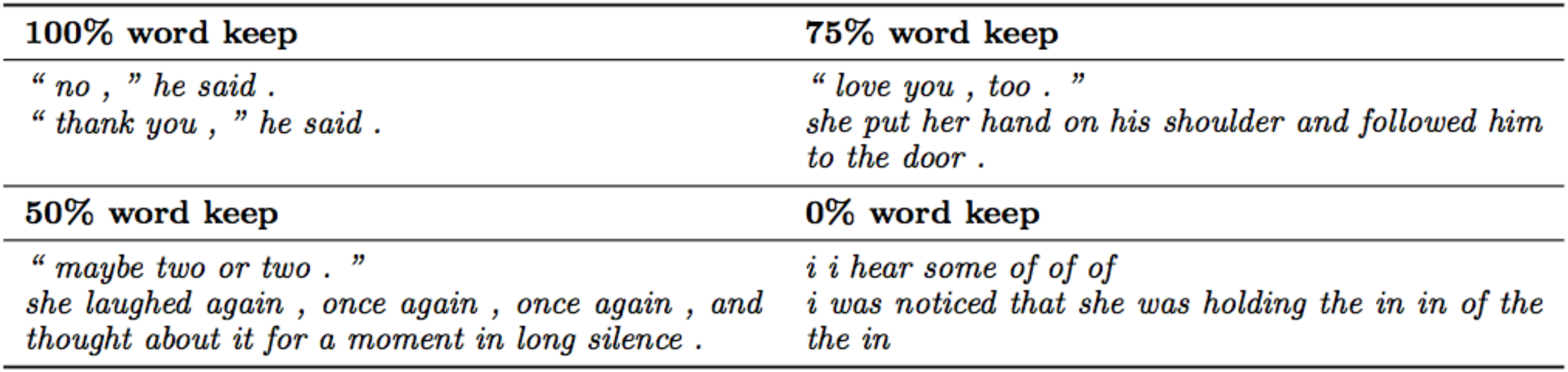
- Precise evaluation of these results is computationally difficult
- Adversarial classifier, is trained to distinguish real sentences from imputed sentences, and score the model on how well it fools the adversary
- Adversarial error is defined as the gap between chance accuracy (50%) and the real accuracy of adversary – ideally this error will be minimized



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

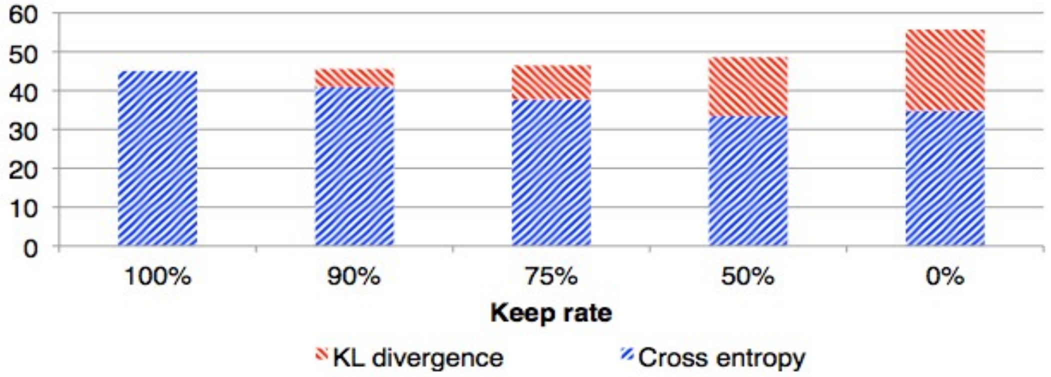
More Analysis

- Word dropout
  - Keep rate too low: sentence structure suffers
  - Keep rate too high: no creativity, stifles the variation



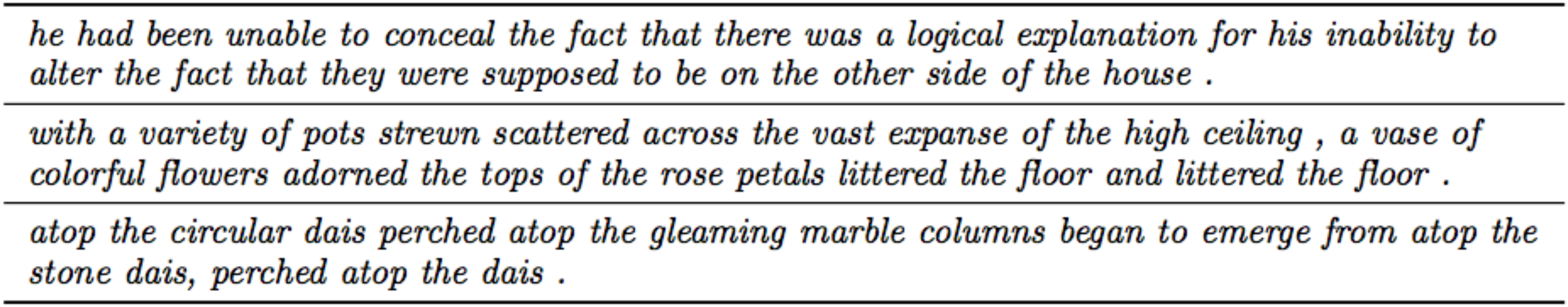
100% word keep	**75% word keep**
“ no , ” he said . *“ thank you , ” he said .*	*“ love you , too . ”* *she put her hand on his shoulder and followed him to the door .*
50% word keep	**0% word keep**
“ maybe two or two . ” *she laughed again , once again , once again , and thought about it for a moment in long silence .*	*i i hear some of of of* *i was noticed that she was holding the in in of the the in*

- Effects on cost function components:



More Analysis

- Sampling from the lower likelihood areas of the latent space, with 75% dropout



he had been unable to conceal the fact that there was a logical explanation for his inability to alter the fact that they were supposed to be on the other side of the house .

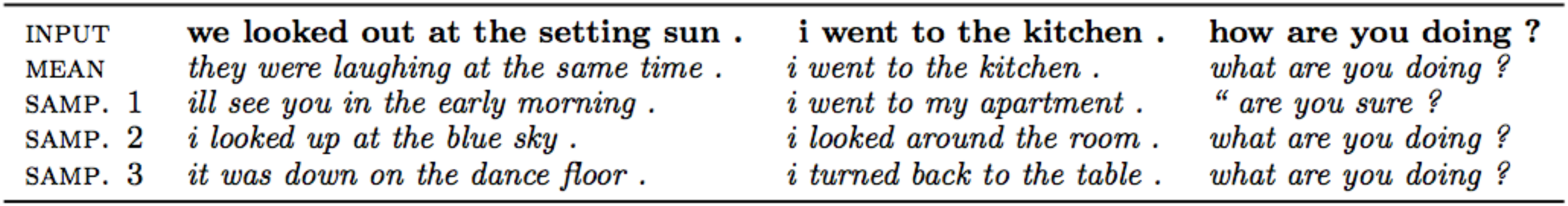
with a variety of pots strewn scattered across the vast expanse of the high ceiling , a vase of colorful flowers adorned the tops of the rose petals littered the floor and littered the floor .

atop the circular dais perched atop the gleaming marble columns began to emerge from atop the stone dais, perched atop the dais .

- Sentences are less typical but for the most part grammatical and maintain a clear topic

More Analysis

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

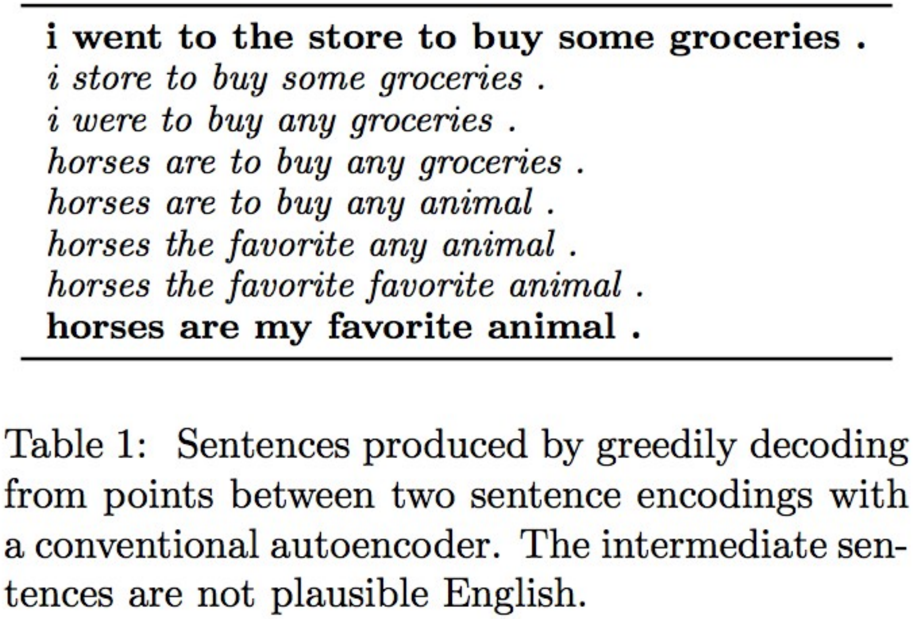


INPUT	**we looked out at the setting sun .**	**i went to the kitchen .**	**how are you doing ?**
MEAN	*they were laughing at the same time .*	*i went to the kitchen .*	*what are you doing ?*
SAMP. 1	*ill see you in the early morning .*	*i went to my apartment .*	*“ are you sure ?*
SAMP. 2	*i looked up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

- Codes appear to capture info about number of tokens, parts of speech for each token, and topic information.
- Longer sentences are less likely to be reproduced

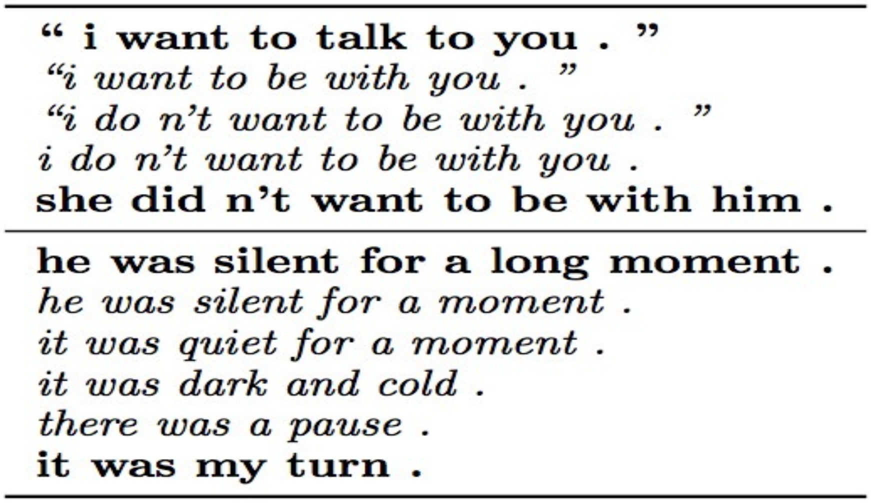
More Analysis

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



i went to the store to buy some groceries .
i store to buy some groceries .
i were to buy any groceries .
horses are to buy any groceries .
horses are to buy any animal .
horses the favorite any animal .
horses the favorite favorite animal .
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 want to be with you . ”
“i do n’t want to be with you . ”
i do n’t want to be with you .
she did n’t want to be with him .

he was silent for a long moment .
he was silent for a moment .
it was quiet for a moment .
it was dark and cold .
there was a pause .
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.

Discussion