Analysis of Neural Language Models for Artificial Data Generation

Thien Nguyen

May 7, 2019

Overview

Objective

- Survey the current developments in variational language models.
- Implement the VAD.
- Compare performances against their earlier counterparts.

Language Models

[Dyer, 2017] describes an unconditional language model as assigning a probability to a sequence of words, $w = (w_1, w_2, ..., w_{i-1})$. This probability can be decomposed using the chain rule:

$$p(w) = p(w_1) \times p(w_2|w_1) \times p(w_3|w_1, w_2) \times ... \times p(w_i|w_1, ..., w_{i-1}) \tag{1}$$

$$p(w) = \prod_{t=1}^{|w|} p(w_t|w_1, ..., w_{t-1}) \tag{2}$$

$$p(w|x) = \prod_{t=1}^{|w|} p(w_t|x, w_1, ..., w_{t-1}) \tag{3}$$

Recurrent Neural Networks

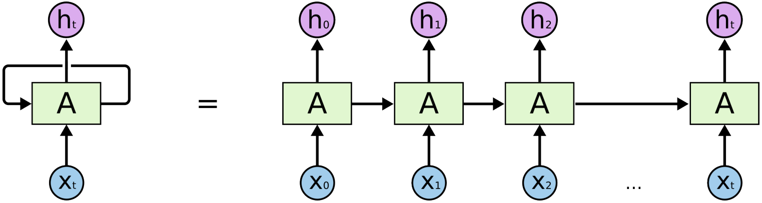

Figure: RNN and its unrolled form.

$$
\begin{aligned}
h_t &= tanh(W_{hh}h_{t-1} + W_{xh}x_t) \\
y_t &= W_{hy}h_t
\end{aligned}
\tag{4}
$$

Figure: Equations for the RNN cell.

Recurrent Language Model - Seq2Seq

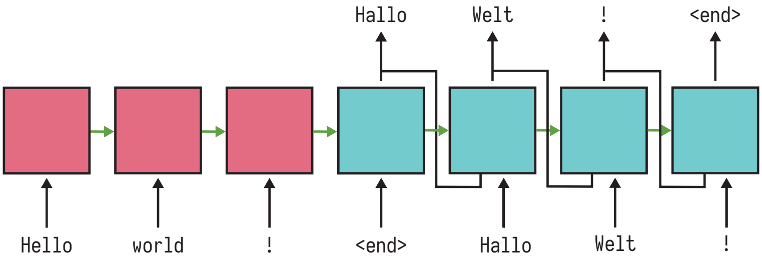

Figure: An abstracted model of the Seq2Seq architecture, where the encoder (pink) takes in the input sequence, and the decoder (blue) shows the output sequence. The encoder outputs are effectively ignored.

Variational Autoencoders

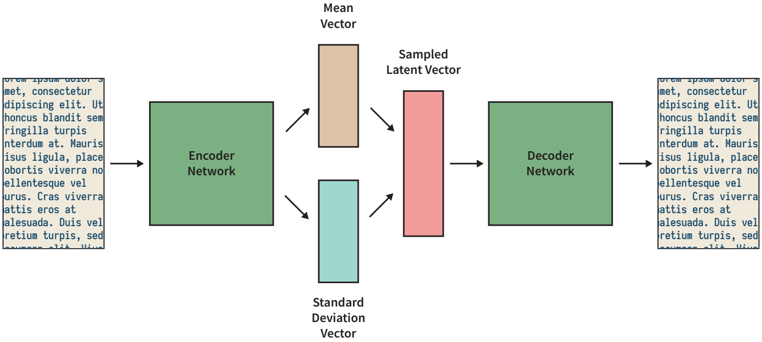

Figure: An abstracted model architecture for a variational autoencoder, which takes as input some text, and it's predicted output being the same text as the input.

Evidence Lower Bound (ELBO)

$$\mathcal{L}(\theta, \phi, x, z) = \mathbb{E}_{q\phi(z|x)}[log\ p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\,||\,p(z)) \leq log(p(x)) \tag{5}$$

Two Parts:

- Reconstruction Loss (Smaller is better.)
- KL divergence (Not actually a loss function - Larger is better!)

Variational Recurrent Language Model

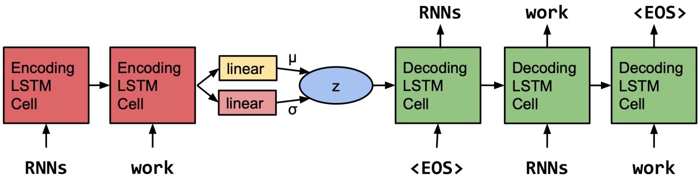

Figure: The core structure of the VAE language model - words are represented as embedded vectors (Diagram from [Bowman et al., 2015]).

- Does not feed the output as the next input.
- Not very useful for conditional response generation.
- Good for understanding how to maintain a non-zero KL divergence.

Conditional Variational Recurrent Language Model

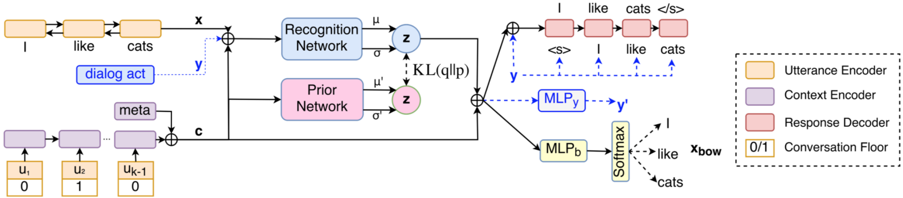

Figure: The structure of the CVAE language model. (Diagram from [Zhao et al., 2017]).

Conditional Variational Recurrent Language Model

Recognition and Prior Networks

Used to encode posterior and priors.

Bag of Words Loss

Mechanism to improve the KL divergence.

Still samples the latent variable once.

Considered to reduce expressivity of the responses.

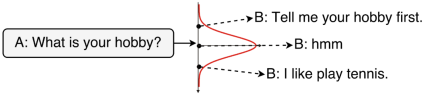

Figure: latent distributions of the CVAE.

The Variational Autoregressive Decoder[1]

- Extension of the CVAE Seq2Seq model.
- Utilises **multi-modal latent sampling**.

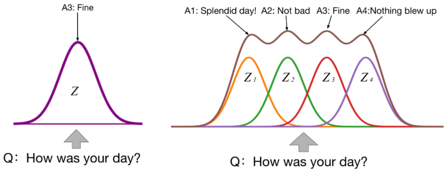

Figure: Unimodal (left) and multimodal latent distributions (right).

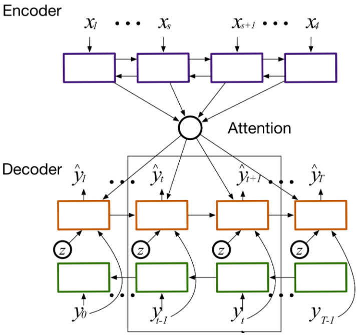

Figure: A High level diagram of the VAD.

[1] Diagrams from [Du et al., 2018]

Decoder

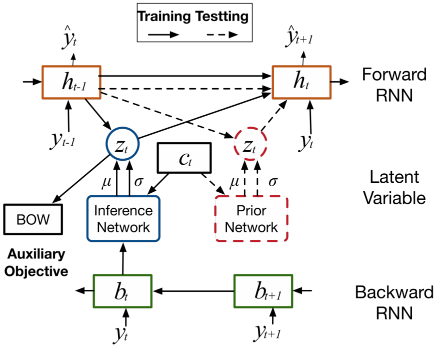

Figure: A High level diagram of the decoding component of the VAD. (Diagram from [Du et al., 2018])

Decoder (Latent Variable)

$$
\begin{aligned}
[\mu^i, \sigma^i] &= f_{infer}([\overrightarrow{h^d_{t-1}}, c_t, \overleftarrow{h^d_t}]) \\
q_\theta(z_t|\boldsymbol{y}, \boldsymbol{x}) &= \mathcal{N}(\mu^i, \sigma^i) \\
[\mu^p, \sigma^p] &= f_{prior}([\overrightarrow{h^d_{t-1}}, c_t]) \\
p_\phi(z_t|\boldsymbol{y}_{<t}, \boldsymbol{x}) &= \mathcal{N}(\mu^p, \sigma^p)
\end{aligned} \tag{6}
$$

Figure: Equations for Inference (left) and Prior (right) models.

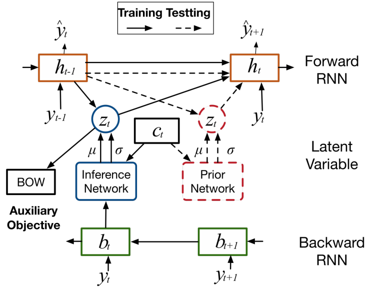

Figure: A High level diagram of the decoding component of the VAD. (Diagram from [Du et al., 2018])

Decoder (Forward RNN)

$$\overrightarrow{h_t^d} = \overrightarrow{GRU}([y_{t-1}, c_t, z_t], \overrightarrow{h_{t-1}^d})$$

$$p_\phi(y|\boldsymbol{y}_{<t}, \boldsymbol{z}_t, \boldsymbol{x}) = f_{output}([\overrightarrow{h_t^d}, c_t]) \quad (7)$$

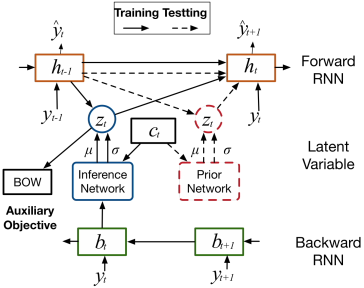

Figure: A High level diagram of the decoding component of the VAD. (Diagram from [Du et al., 2018])

Loss Function

$$\mathcal{L} = \sum_t [\mathcal{L}_{ELBO}(t) + \alpha \mathcal{L}_{AUX}(t)]$$
$$\mathcal{L} = \sum_t [(\mathcal{L}_{LL}(t) - \mathcal{L}_{KL}(t)) + \alpha \mathcal{L}_{AUX}(t)] \tag{8}$$

$$\mathcal{L}_{ELBO}(t) = \mathbb{E}_{q\phi(z|x)}[log\ p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\,||\,p_\theta(z|x)) \leq log(p(x)) \tag{9}$$

Datasets

Penn TreeBank

Model validation dataset. Models would recreate the input sequence. Example:

big investment banks refused to step up to the plate → big investment banks refused to step up to the plate

Open Subtitles

Conditioned sequences. Models would use one subtitle sentence to predict the next sentence. Example:

your paycheck ? → back off tucker , you don ' t sketch regulations .

Amazon Reviews

Conditioned and contextual sequences. Models would use one sentence to predict the next sentence. Example:

context the kindle is velcroed in so it ' s nice and secure . → very glad i brought this !

Optimisation Challenges

KL Collapse

AKA Vanishing KL, Posterior Collapse

$D_{KL}(q_\phi(z|x) \,||\, p_\theta(z|x)) = 0$

Methods

- KL Annealing ([Bowman et al., 2015])
- ~~Word Dropouts ([Bowman et al., 2015])~~
- Bag of Words Loss ([Zhao et al., 2017], [Du et al., 2018])

ELBO: Reconstruction Loss

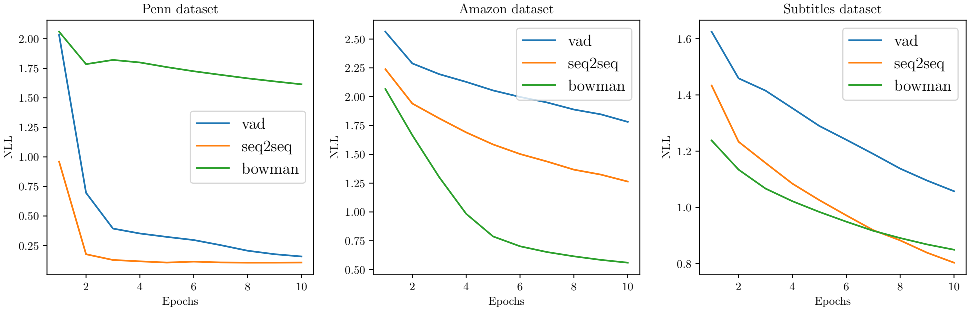

Figure: Reconstruction losses of the three models across the three datasets; lower loss is better.

ELBO: KL Ratio

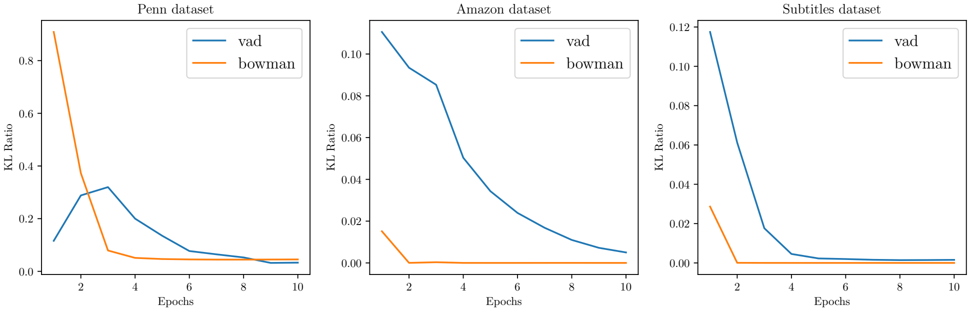

Figure: KL ratios of the three models across the three datasets; higher is better.

BLEU

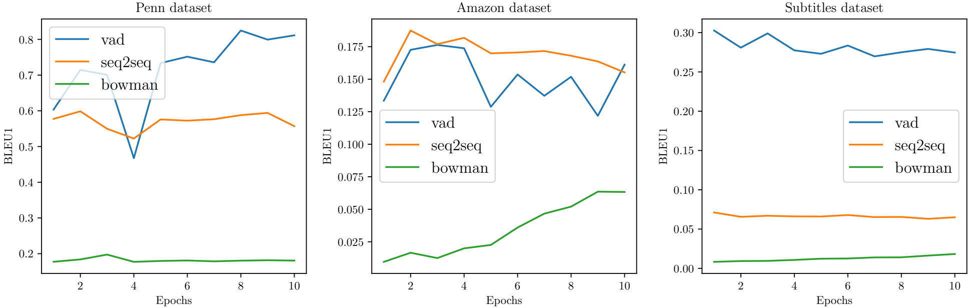

Figure: $BLEU_1$ scores (Modified Precision) of the three models across the three datasets; higher is better.

ROUGE

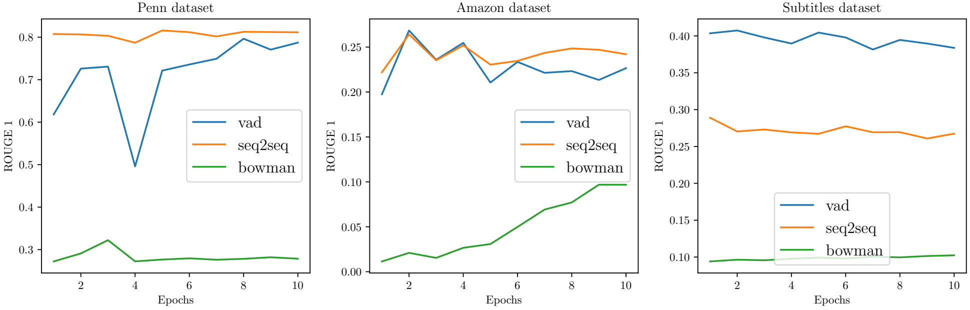

Figure: $ROUGE_1$ (Modified Recall) scores of the three models across the three datasets; higher is better.

F1

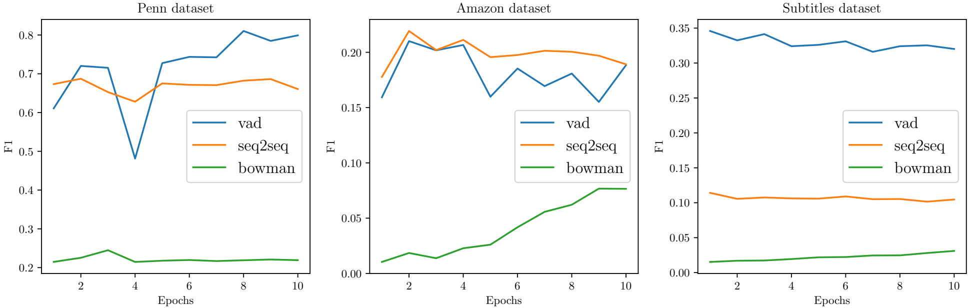

Figure: $F1_{\text{1-gram}}$ scores across the three models; higher is better.

Semantic Similarity

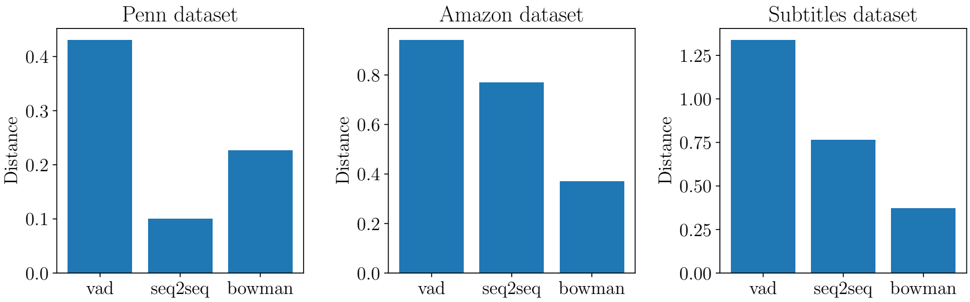

Figure: Semantic similarity scores across the datasets and models, higher is more varied.

Generated Examples - Penn TreeBank

these stocks eventually reopened

Bowman

the <unk> $ cents <unk> the <unk> open

Seq2Seq

these stocks eventually reopened

VAD

these stocks eventually reopened

as a result the link between the futures and stock markets , apart

---

Bowman

they is n't elaborate whether the <unk> is a we is s going to be a a years

Seq2Seq

as a result the link between the futures and stock markets , apart

VAD

as a result the link between the futures between and futures markets apart

Generated Examples - Open Subtitles

i give you ride.

Bowman

N/A

Seq2Seq

you ’ re going to take a delivery hour

VAD

i ’ m gon na take a walk .

not exactly.

Bowman

?

Seq2Seq

he ’ s hideous <unk> , he ’ s a musician

VAD

i ’ m gon na be able to finish the door .

Generated Examples - Amazon Reviews

b 0 0 d q d c 1 y 6 rating_4.0 polarity_0.8 i also have a background in it .

Bowman

i the : the the the , you 's the i , is the ,

Seq2Seq

i am going to get it to work well .

VAD

the email has a nice connection .

b 0 0 9 l l 9 v d g rating_5.0 polarity_-0.8 you will need the desktop to run larger things off of , and to use a printer remotely .

Bowman

N/A

Seq2Seq

i have a fit of the canon .

VAD

i a little to in distracting for the price .

Demo

Source code: https://github.com/thien/iso

Evaluation

Amazon performance is below expectations

Could be attributed to a variety of factors, ranging from data sanitisation, to contextual conditioning, to inappropiate word embeddings.

Dataset sizes were handicapped

Caused by computational constraints.

Spurious SBOW α weights

Attempts to contact the original authors of the VAD regarding missing information has been unsuccessful.

Conclusion

It is possible!

The VAD has shown to express variety in responses, greater so than the VAD and the Seq2Seq models.

Future Work

Experimentation on the SBOW, Adaptation for an Adversarial approach.

End of Presentation

Any Questions?

Please ask!

Source Code

https://github.com/thien/iso

References I

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2015).
Generating Sentences from a Continuous Space.
arXiv:1511.06349 [cs].
arXiv: 1511.06349.

Du, J., Li, W., He, Y., Xu, R., Bing, L., and Wang, X. (2018).
Variational Autoregressive Decoder for Neural Response Generation.
In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3154–3163, Brussels, Belgium. Association for Computational Linguistics.

Dyer, C. (2017).
Conditional Language Modelling.
original-date: 2017-02-06T11:32:46Z.

References II

Zhao, T., Zhao, R., and Eskenazi, M. (2017).

Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders.

arXiv:1703.10960 [cs].

arXiv: 1703.10960.