Analysis of Neural Language Models for Artificial Data Generation

Thien Nguyen

May 7, 2019

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

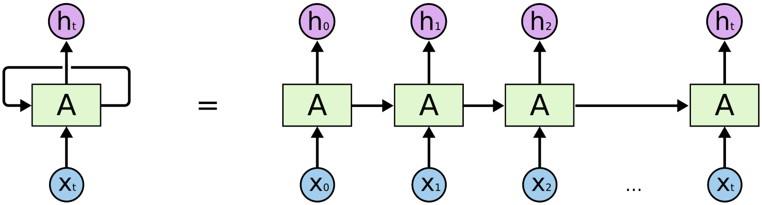


Figure: RNN and its unrolled form.

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

Figure: Equations for the RNN cell.

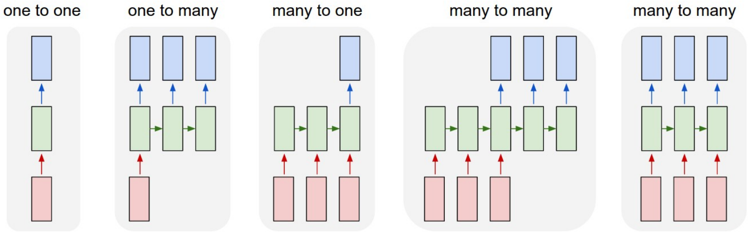
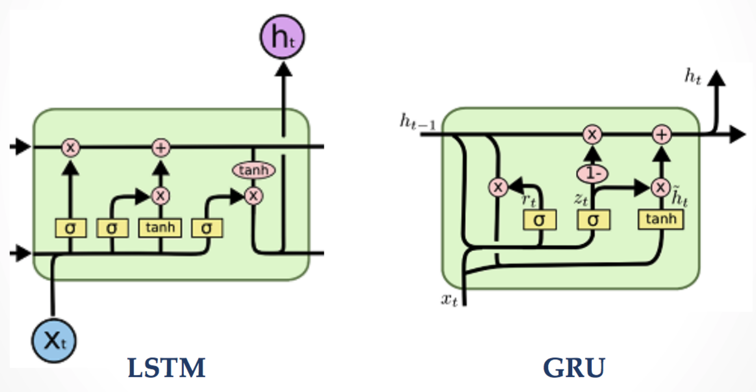


Figure: From left to right: (1) an MLP. (2,3,4,5) examples of different styles of recurrent neural networks, describing the different types of input and output combinations. (Diagram from [Karpathy, 2015]).

LSTMs and GRUs



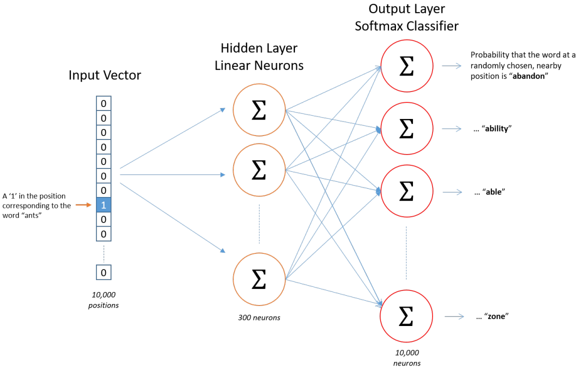
Word Embeddings

How do you represent words?

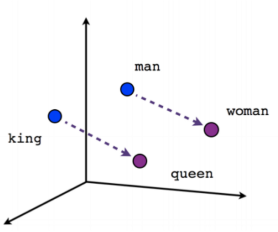
- You have tens of thousands of words.
- How do you mark the relationships between them?
- Feeding them into neural networks is not necessarily feasible.

Word Embeddings

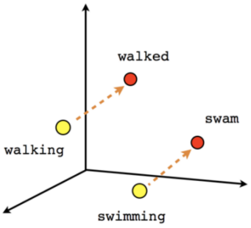
- Converts words to vectors.
- Models relationships of words based their co-occurence.
- Trained in an unsupervised skip-gram neural network.



Word Embeddings



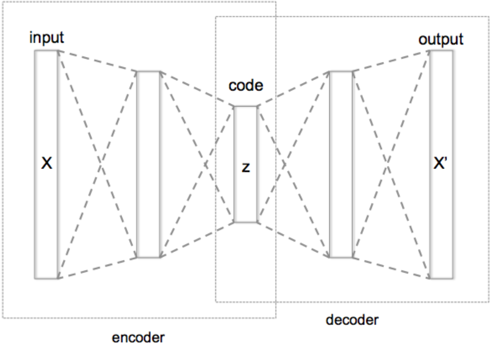
Male-Female



Verb tense

Autoencoders

- Attempts to faithfully recreate the inputs at the output.
- Learns the properties of the input data.
- Typically has a layer where its dimension is smaller than the input space - called the latent layer.



Recurrent Language Model - Seq2Seq

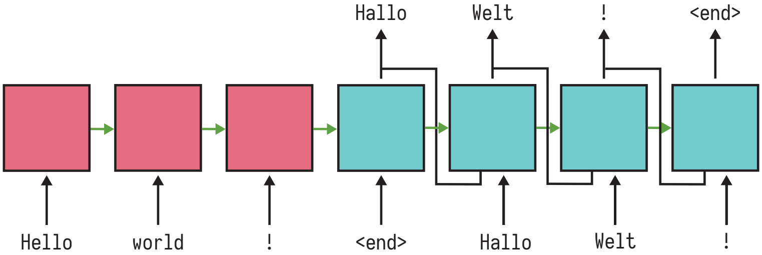


Figure: An abstracted model of the Seq2Seq architecture, where the encoder (pink) takes in the input sequence, and the decoder (blue) shows the output sequence. The encoder outputs are effectively ignored.

Variational Autoencoders

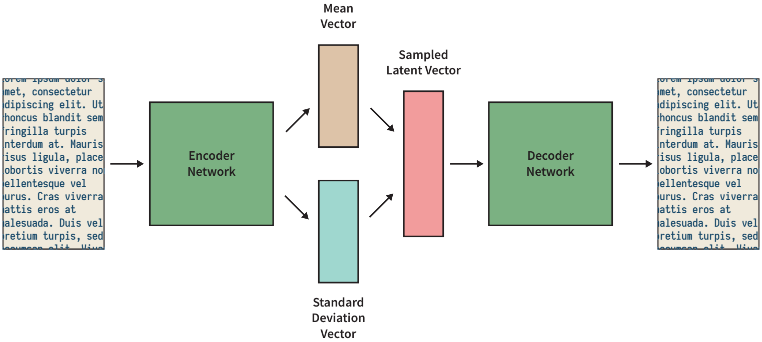


Figure: An abstracted model architecture for a variational autoencoder, which takes as input some text, and it's predicted output being the same text as the input.

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.)
- Distribution divergence (Not actually a loss function - Larger is better!)

Variational Recurrent Language Model

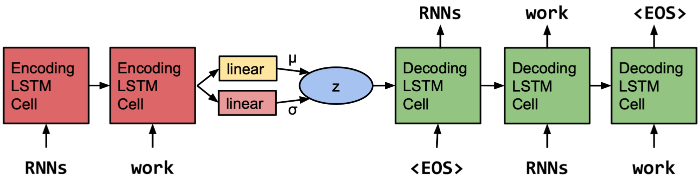


Figure: The core structure of the VAE language model - words are represented as embedded vectors (Diagram from [Bowman et al., 2015]).

- 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

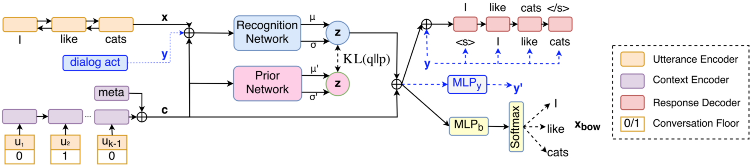


Figure: The structure of the CVAE language model. (Diagram from [Zhao et al., 2017]).

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.

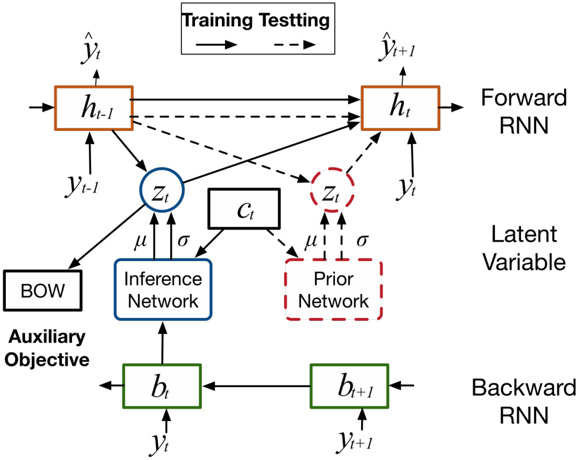
Still samples the latent variable once.

Considered to reduce expressivity of the responses.

The Variational Autoregressive Decoder

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

Decoder


Training Testting
$\hat{y}_t$
$\hat{y}_{t+1}$
h_{t-1}
h_t
Forward RNN
y_{t-1}
z_t
c_t
z_t
y_t
μ
σ
μ
σ
Latent Variable
BOW
Inference Network
Prior Network
Auxiliary Objective
b_t
b_{t+1}
Backward RNN
y_t
y_{t+1}

Decoder (Latent Variable)

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

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

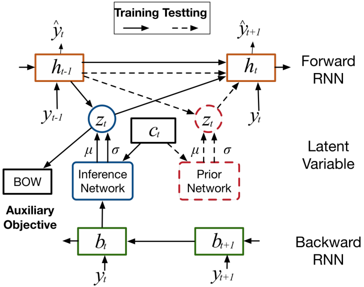


Figure: A High level diagram of the decoding component of the VAD. (Diagram from [Du et al., 2018])

Decoder (Forward RNN)

$$\overrightarrow{h^d_t} = \overrightarrow{GRU}([y_{t-1}, c_t, z_t], \overrightarrow{h^d_{t-1}})$$

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

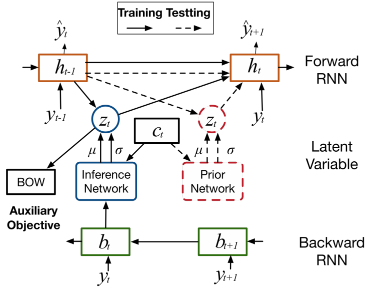


Figure: A High level diagram of the decoding component of the VAD. (Diagram from [Du et al., 2018])

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.

Open Subtitles

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

Amazon Reviews

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

Optimisation Challenges

KL Collapse

AKA Vanishing KL, Posterior Collapse

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

Methods

- Bag of Words Loss
- KL Annealing
- Word Dropouts

Quantitative Measurements

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

$$BLEU_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{clip}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)} \qquad ROUGE_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

$$f1_n = 2 \cdot \frac{BLEU_n \cdot ROUGE_n}{BLEU_n + ROUGE_n}$$

Semantic Variance

procedure SEMANTIC VARIANCE
  $query \leftarrow$ input $sequence$
  $resp \leftarrow []$
  for $i = 1$ to n **do**
    $r_i \leftarrow \text{model}(query)$
    $r_i \leftarrow [\text{embedding}(token) \text{ for } token\ in\ r_i]$
    $resp[i] \leftarrow mean(r_i)$
  $m \leftarrow mean(resp)$
  return $max(euclidean(m, r_{i=1 \text{ to } n}))$

ELBO: Reconstruction Loss

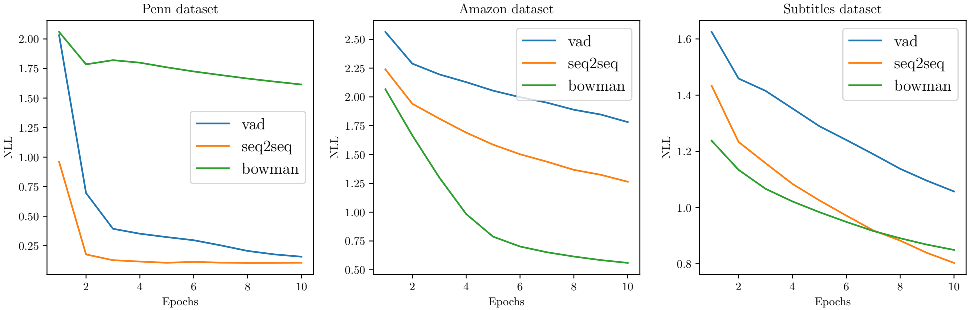


Figure: Reconstruction losses of the three models across the three datasets; lower loss is better.

ELBO: KL Ratio

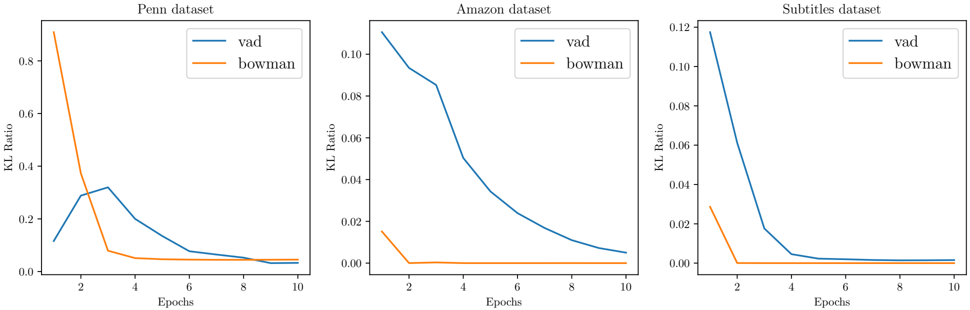


Figure: KL ratios of the three models across the three datasets; higher is better.

BLEU

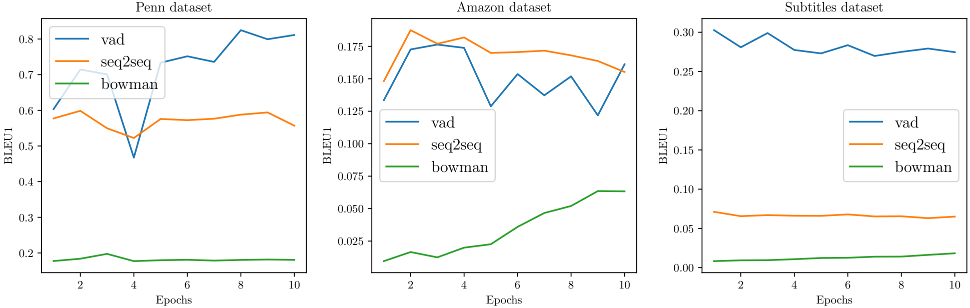


Figure: $BLEU_1$ scores of the three models across the three datasets; higher is better.

ROUGE

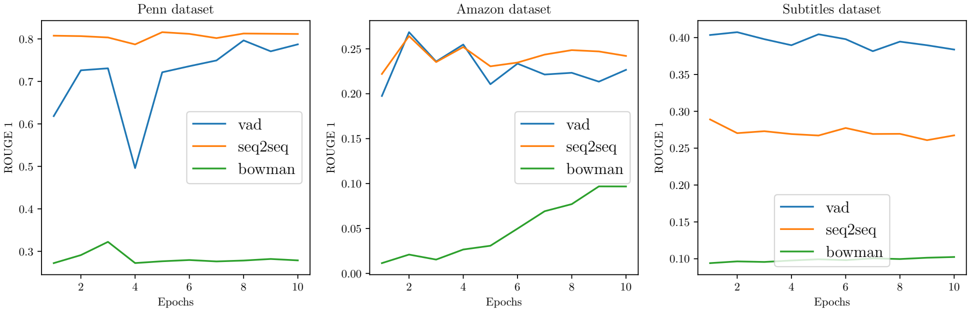


Figure: ROUGE_1 scores of the three models across the three datasets; higher is better.

F1

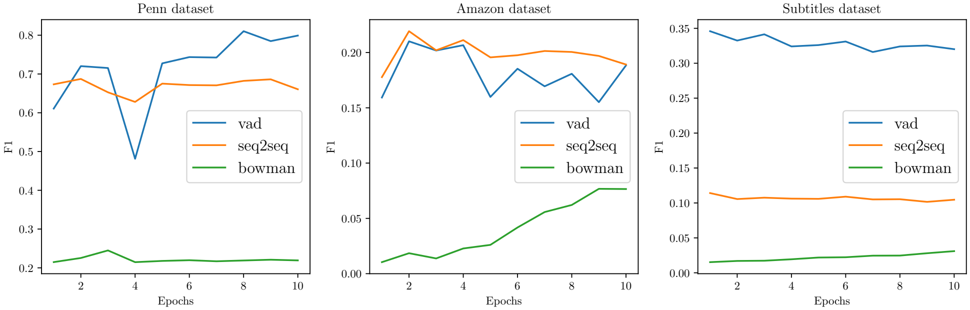


Figure: $F1_{\text{1-gram}}$ scores across the three models; higher is better.

Semantic Similarity

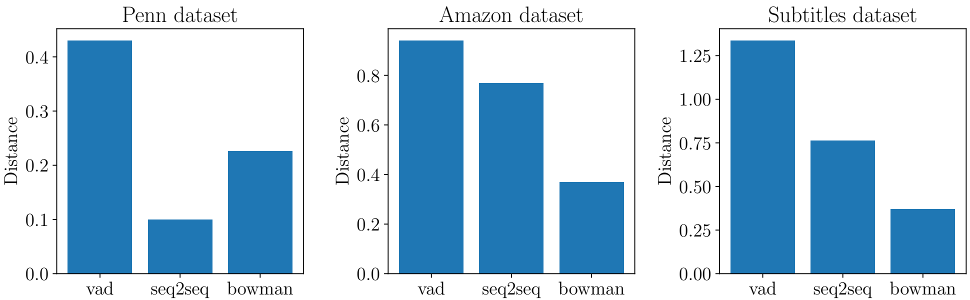


Figure: Semantic similarity scores across the datasets and models, higher is more varied.

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 inappropriate 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.

Performance could be improved

The amazon dataset could be better augmented to improve VAD performance.

Future Work

Experimentation on the SBOW, Augmentation for an Adversarial approach.

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

Karpathy, A. (2015).
The Unreasonable Effectiveness of Recurrent Neural Networks.

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.