Anomaly Detection with Conditional VAEs

Borealis AI

Hareesh Bahuleyan, Garrin McGoldrick

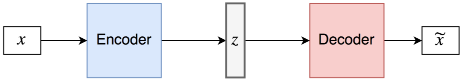
November 8, 2018

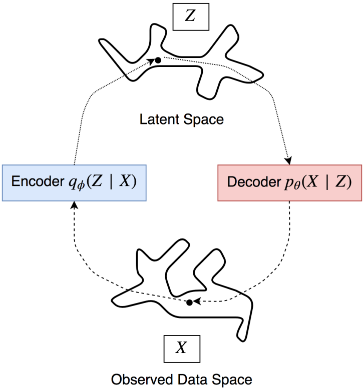
Overview

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



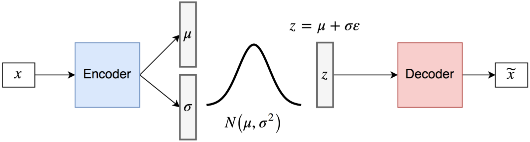


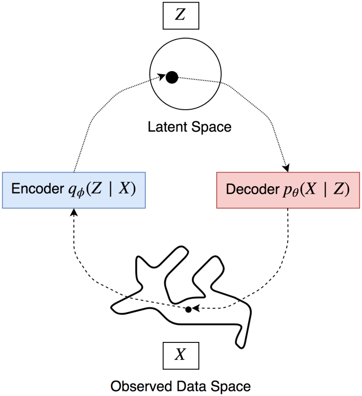
Minimize Reconstruction Loss

$J = -\sum_{n=1}^{N} \log p(x^{(n)}|z^{(n)})$

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

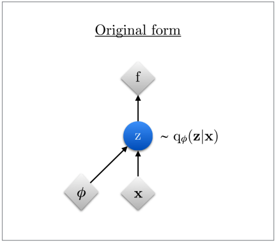
$$J = \sum_{n=1}^{N} \left[- \mathop{\mathbb{E}}_{z^{(n)} \sim q} \left[\log p(x^{(n)} | z^{(n)}) \right] + \lambda \cdot \mathrm{KL}(q(z^{(n)} | x^{(n)}) \| p(z)) \right]$$

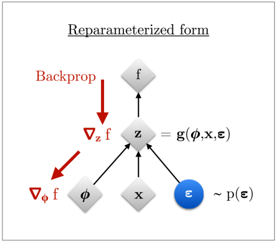
Reparameterization Trick

KL Divergence between posterior and standard normal prior

$$\mathrm{KL}(\mathcal{N}(\mu,\sigma)||\mathcal{N}(0,I)) = \frac{1}{2}(1+\log((\boldsymbol{\sigma}^{(n)})^2) - (\boldsymbol{\mu}^{(n)})^2 - (\boldsymbol{\sigma}^{(n)})^2)$$

- Model training via SGD and error backpropagation
- Cannot sample directly from the approximate posterior distribution $\mathcal{N}(\mu,\sigma)$
- Stochastic Node - disconnect in the graph
- **Solution**: Sample from fixed distribution $\mathcal{N}(0,I)$ and reparameterize
- $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \otimes \epsilon$ where $\epsilon \sim \mathcal{N}(0,I)$





MNIST Experiments

- Toy Example - Compress image to 2d latent space and reconstruct

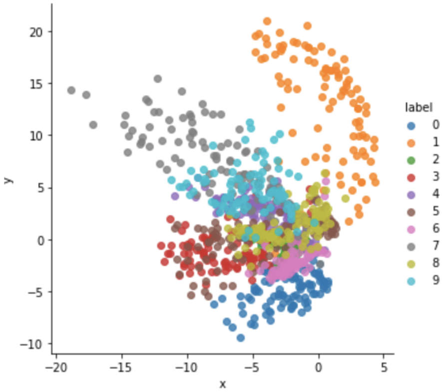


Figure: Deterministic AE

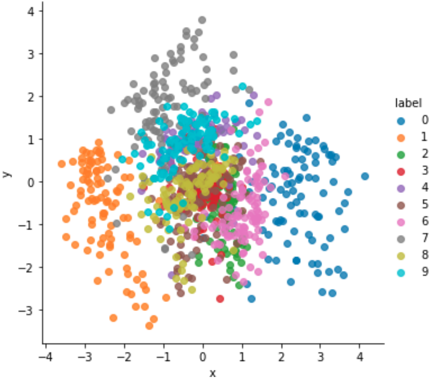


Figure: Variational AE

MNIST Experiments

- Toy Example - Compress image to 2d latent space and reconstruct

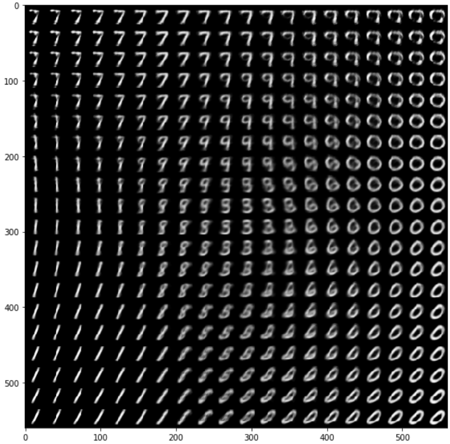


Figure: VAE Reconstructions from different parts of the latent space

Text VAEs

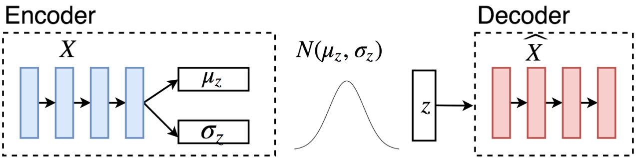


Figure: Model Architecture

- Trained on a subset of SNLI Dataset [Bowman et al., 2015a]

Training Heuristics

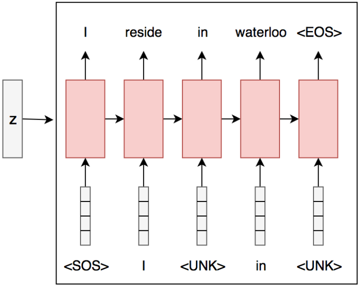
- Training VAEs for text generation is notoriously difficult
- Adopt two training strategies [Bowman et al., 2015b]

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 $\boldsymbol{z}$

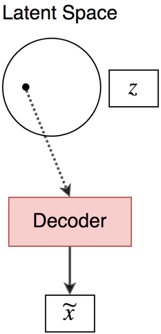


Decoder

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

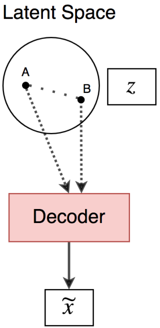
Deterministic AE	**Variational AE**
a men wears an umbrella waits to	*the dog is sleeping in the grass .*
a couple cows a monument	*the girls are being detained .*
there is sleeping and two rug .	*the group of people are going to begin .*
a man in a pick photos	*a girl with blond-hair on a bike with a stick*
a boy are people at a lake escape .	*a woman and a man are walking on a street*



Linear Interpolation

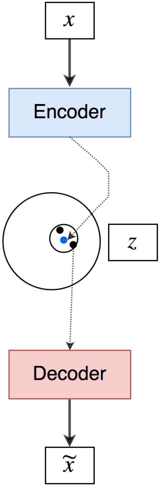
- To test the continuity of the latent space
- $\boldsymbol{z}_{\alpha_i} = \alpha_i \cdot \boldsymbol{z}_A + (1 - \alpha_i) \cdot \boldsymbol{z}_B$ where $\alpha_i \in \left[0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\right]$
- VAE - Smooth transition maintaining syntax and semantics
- 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 couple eating .*
there is a couple eating cake .	*there is a couple eating dinner .*
there is a group of people eating a party .	*there is a couple of people eating dinner .*
a group of men are watching a party .	*a group of people are having a conversation .*
a group of men are watching a dance party .	*a group of men are having a discussion .*
a group of men are watching a dance party .	*a group of men are watching a movie .*
a group of men are watching a dance party .	*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 .*

Plan

CVAEs [Sohn et al., 2015]

- Regular VAE - no control over the *class* of data being generated
- CVAEs - flexibility to synthesize data from the desired *class*

Minimize Reconstruction Loss + KL Divergence

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

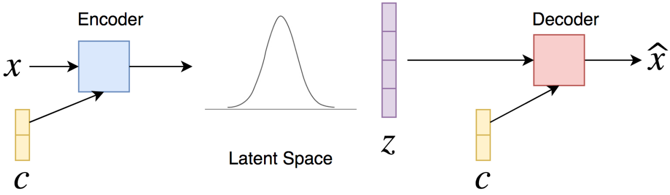


Figure: CVAE Model Architecture

Hypothesis for Outlier Detection

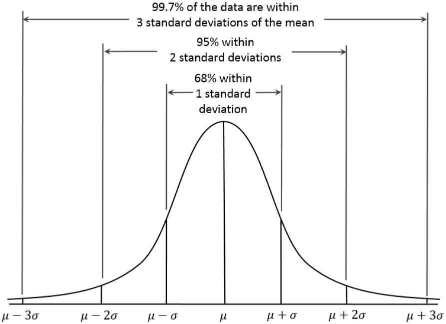


Figure: Univariate Normal distribution

- Data points further away from the mean are less probable
- More likely to be outliers
- For Apollo: Novel news detection
- Why CVAE: News articles conditioned on specific companies or sectors or even news history

Preliminary Experiments with MNIST

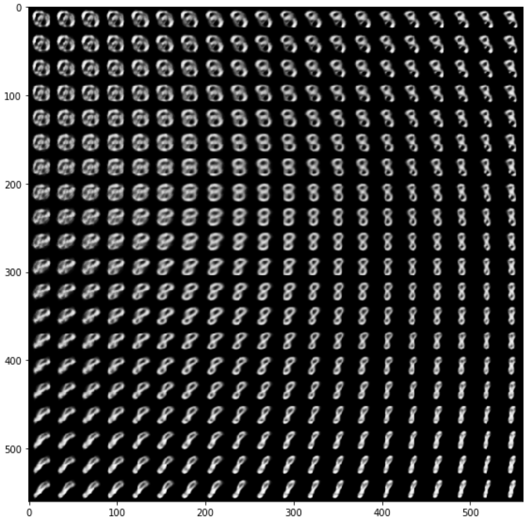


Figure: Samples from conditional prior of digit '8' - Blurry images when sampled away from the mean (centre)

Preliminary Experiments with MNIST

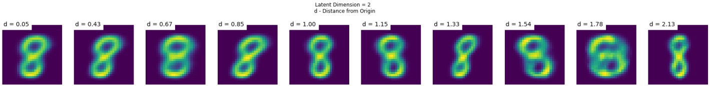


Figure: Sampling at different distances from the mean (origin)

Text Data

Yahoo Questions Dataset

- Conditioning Variable - Topic Label Embedding
- Subset of 100k questions

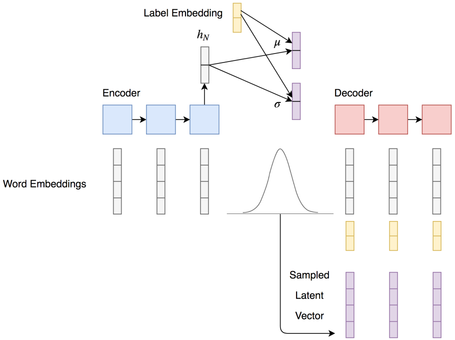


Figure: CVAE Model Architecture

Conditionally Generated Text

Samples drawn from the conditional posterior distribution of two topics:

Health	**Sports**
how would i find my molar reaction ?	*can you find a alternative mountain bike ?*
what type of oily skin ?	*why is the superbowl so amazing by brand ?*
how does spinach go for the fat and vegetable ?	*whats your favorite swim team on each ?*
how to control the swelling for this burning ?	*why is the boxing championships ?*
what is that mental disorder that i have one ?	*what is a club to be playing from the computer ?*

Sentences and Distances

- Created *fake* questions
- Topic: Health

Table: Success Cases

how do you get rid of herpes ?	28.8
do lawyers cause herpes ?	**42.1**
how soon can you know if you are pregnant ?	43.2
how soon can you touch fresh paint ?	**53.4**
how can i grow my hair back ?	42.0
are bald people good at doing business ?	**50.7**

- Lower distances for shorter sentences
- Use of rare words results in higher distances

Table: Failure Cases

how old were you before you were able to grow a good looking beard ?	**63.5**
do you need a hammer to construct a good looking beard ?	56.5
how to relieve severe itchy skin ?	**52.4**
how do police relieve severe criminals ?	47.9
should human genetic engineering be allowed ?	**47.1**
should human build artificial intelligence ?	38.4

Plan

1. Introduction
2. Variational Autoencoders
3. Conditional VAEs
4. Conclusions and Future Work

Conclusions and Future Work

- **Summary**
  - VAEs are generative models from which it is possible to synthesize new data
  - The usage of CVAEs for novelty/anomaly detection based on euclidean distance
- **Issues**
  - For VAEs with textual data, the basis for clustering probably has to do more with syntax rather than semantics
  - Gaussian latent space and euclidean distance may not be appropriate in high dimensions
- **Next Steps**
  - Spherical VAEs based on von Mises Fisher Distribution - data is distributed on a unit hypersphere - cosine similarity as distance metric Xu and Durrett [2018]

References I

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

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015a.

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*, 2015b.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders. *arXiv preprint arXiv:1808.10805*, 2018.