Experimentation with the latent space of a variational autoencoder

Priyank Pathak [1], **Vihang Agarwal**[2] , **Anshumali Shrivastava** [3]

[1] ,[2]IIT Kanpur , [3]Rice University,

{ppriyank,vihang}@iitk.ac.in , anshumali@rice.edu

Abstract. *The current architecture of VAE suffers from latent space saturation (with inefficient packing) and mode collapse. In this project we believe if the dataset is distributed among different architectures, mode collapse can be easily dealt with, yet retaining the properties of a normal VAE and getting a reasonable reconstruction. Furthermore, we postulate if the architecture shares the latent space over the modes the network would result in more efficient packing. Hencee, we aim to introduce more than one encoder, with the latent spaces mapped to a single decoder. In order to get the model to work, we are currently trying to incorporate different techniques to get a perfect reconstruction.*

Keywords*: VAE, Encoders, Latent Space .*

Introduction

As shown in the figure, a VAE is a beural net which encodes a feature vector to a low dimensional vector, also known as *z-latent vectors*, which can be further decoded to reconstruct the output. It is a probabilistic version of an autoencoder, that is, we generate an output, using a suitable distribution (Gaussian) and use it as a prior to generate the output. In the process, the VAE learns the modal cluster representation of the data in terms of the low dimensional representation. However, only a fixed amount of data can be encoded before z-latent space saturates, partially due to inefficient clustering, i.e. the current representation is insufficent to give discernable clusters of data as we wish for.
Besides, VAEs and GANs often suffer from **mode-collapse**, a situation in which the generator tends to learn data of a single dominant modal representation, while ignoring the others. This usually occurs due to a disproportionate amount of data corresponding to a distribution.

Previous Works

1 **Paper** - BEGAN: Boundary Equilibrium Generative Adversarial Networks [2]
Author - David Berthelot, Thomas Schumm, Luke Metz (2017)
Idea - Proposed a new equilibrium enforcing method paired with a loss derived from the Wasserstein distance for training auto-encoder based Generative Adversarial Networks. This method balances the generator and discriminator during training.

2 **Paper** - Multi-Agent Diverse Generative Adversarial Networks [3]
Auther - Arnab Ghosh, Viveka Kulharia, Vinay Namboodiri (2017)

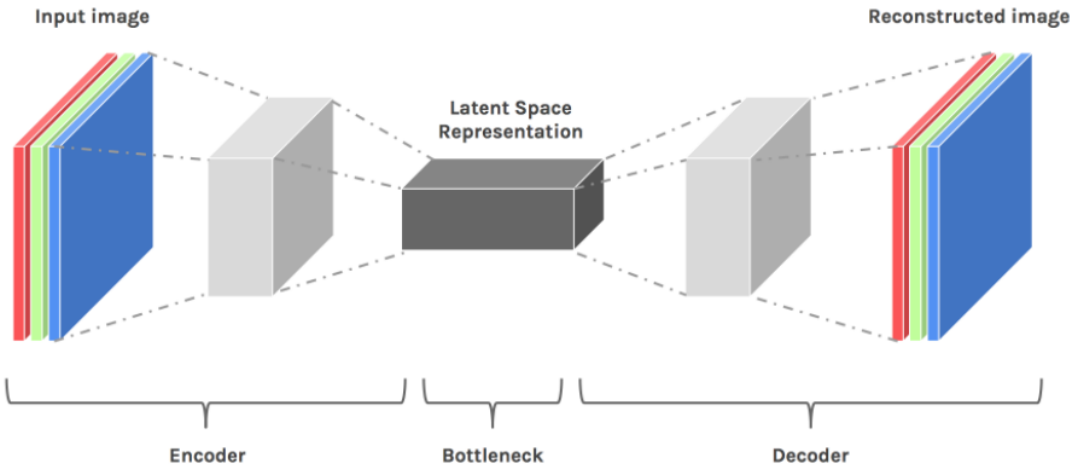

Figure 1. Typical VAE taken from Siraj Raval' blog [1]

Idea - Described a generalization to the Generative Adversarial Networks (GANs) to generate samples while capturing diverse modes of the true data distribution.

3 **Paper** -Disentangling Variational Autoencoders for Image Classification [4]
Author - Chris Varano (2017)
Idea - investigated the use of a disentangled VAE for downstream image classification tasks. They learn the VAE encoder that maps images into a disentangled latent space. The weights of the encoder are frozen and it is then used as a feature extractor for the downstream supervised task.

Background Idea

The Architecture

Given below is the proposed variation of our VAE. We attempt to train such a model that encompasses a class dataset as by one encoder, which is a subset of the overall data, the problem of mode collapse on the latent variables may mitigate. If the decoder successfully decodes to give the proper image then it is proved that the encoders share a common latent space, because decoder is unaware of which encoder has provided the z vector.

The Toy Problem

We have modeled the above challenge by a toy example: We considered the paintings of Monet and five of his associates who were all well known painters. All the Paintings have been taken from wiki art [5]

Eugene Boudin (560 paintings) taught Monet (1349) oil painting in his early years. Monet's friend is Manet (232 paintings). Monnet main teacher is Charles Gleyre (32). Most influential friends Pierre-Auguste Renoir (1405) and Alfred Sisley (473). Pierre and Monet timeline period is well known as Impressionism. Fig 3. The Dataset is partitioned into dated and undated images. Source is wiki-art.

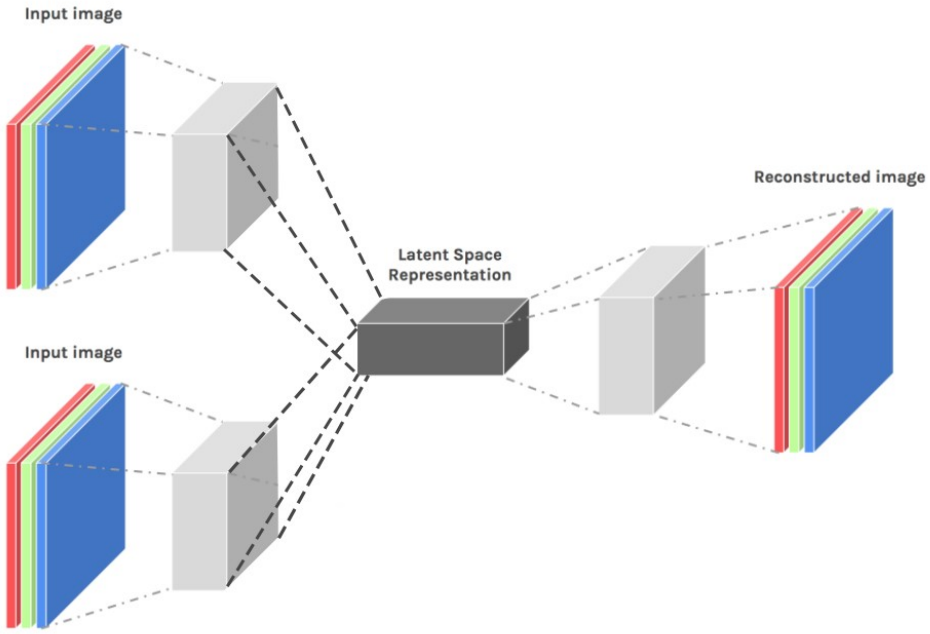

Figure 2. Proposed Variant of VAE

As all the above painters have been professionally associated with Monet, we expect a general trend between the artistic style of these paintings. Our model, if trained correctly will capture this very trend.

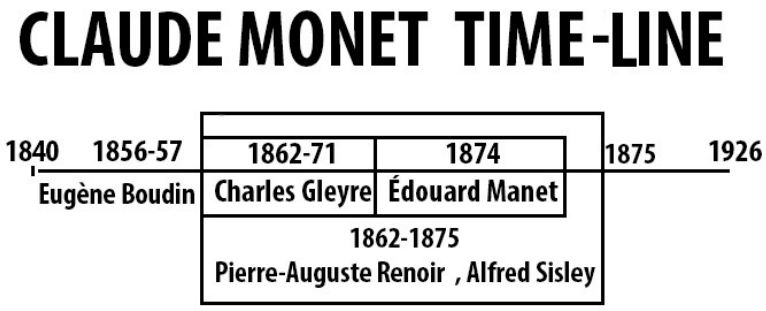

Figure 3. Claude Monet Time-line, information source Wikipedia [6]

The Posit

On successful training of the model, we will have six different encoders for each of the painters. Our current discussion is on two encoders sharing a decoder.

- The latent space exists independently (our aim)
- The latent space of encoder overlaps

1) If the latent space exists independently, possibly ensuring denser encoding, and hence efficient clustering. This ensures increased capacity for the bounds of the dimensions to represent the images, which was not adequately represented previously.

In other words, the model which has never seen Class A painting will be able to retain only those features that were common among the paintings it was trained on, i.e. it will throw away features it has never seen, applying them to the painting style.
Also, as the model shows distinct clusters for different artists, a given painting can be passed through the VAE, giving the z-vector. Using this, we can find the nearest cluster to get a notion of similarity. Further, we can trace out a time-line of this 'sample' painting and deduce the painter using our architecture.

2) The second possibility is when encoders overlaps the latent space, which is equivalent to one encoder-decoder model. This will prevent mode collapse, and will enable us to visualize how Monet's painting changed in his career. An important issue of the memory overhead will be dealt only when comparing the training time of both the model.

Irrespective of the path model chooses we can further postulate as the encoder discarding the qualities of the data which has no mode attached to it. A model which has seen all the data does not leave any scope of experiments in this direction, whereas the proposed model involves encoders which have not seen each other's data, hence we need to conduct few experiments to cross Monet's paintings in all encoders, since encoders have seen paintings based on similar style.

Execution

Since the painting reflect the style of the painter and hence we limited the size of images to be 256 x 256 as VAEs are known to explode beyond this size.

Our implementations are adapted from:

- Cycle GAN available on github [7] *(tensorflow)*
  Unresolved error, coudn't make it work
- VAE version of Pix2Pix architecture [8] *(tensorflow)*
- VAE style transfer on high resolution [9] *(tensorflow)*
- Non-convolutional fully connected VAE [10] *(Keras)*
  On 150 x 150 input, first model to work, advised by Vishak (PG graduate)
- Non-convolutional fully connected AE on 150 x 150 input *(Keras)* [10]
- Convolutional Auto Encoder on a 150 x 150 input *(Keras)* [10]
- Convolutional VAE on a 256 x 256 input, z-space of 100 dim, *(Keras)* [10]
- Convolutional VAE on a 256 x 256 input, z-space of 100 dim *(Keras)*[11], based on architecture suggested in Deep Feature Consistent Variational Autoencoder, coded from skratch (current implementation)

We have used Adam optimizer with L_1 loss multiplied by a factor of 1000 times for the reconstruction along with the KL divergence Loss and Dropout on initial layers, along with, FC while converting from z to convolution layers. The training of the encoders is done alternatively, in a batch of 50 images or alternatively batch of 10 both getting trained 5 times each. In the current architecture we are training the model in a ratio of 2:5 images, since the set A has only 20 images and set B has around 430 images, which has shown the best output, preventing the overpowering of all units.

Results

The initial results from the existing repository have failed miserably. The cycle GAN couldn't be implemented, so the output is not present. The VAE version of pix2pix architecture and Style Transfer architecture has also failed

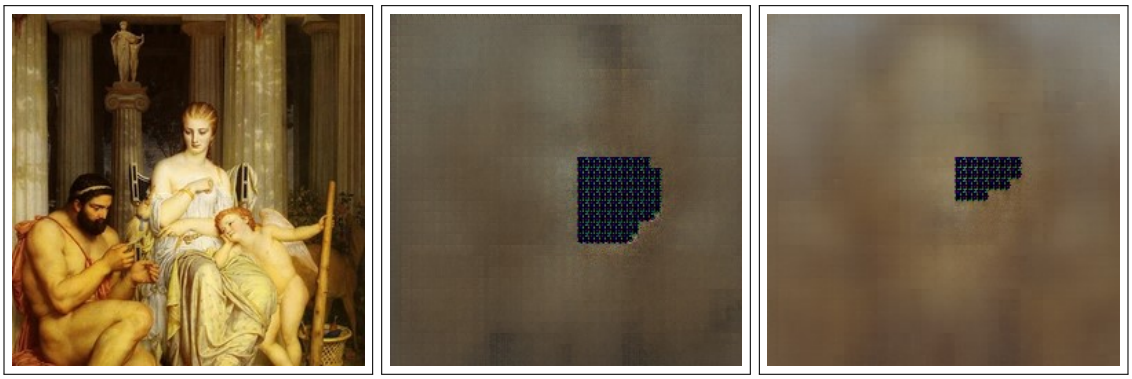

Figure 4. Failed Result obtained from the Pix2Pix architecture and style transfer git repository. Extreme left is the input image and corresponding output from pix2pix in the model and style transfer VAE in the extreme right

Figure 5. 1st image is the input image, 2nd image is the non-convolutional fully-connected Variantional Autoencoder and 3rd is the non-convolutional fully-connected Autoencoder, 4th image is the convolutional autoencoder (entropy loss) 5th image is the Convolutional Autoencoder (L1 loss) and last image is the Convolutional Variational Autoencoder (though not of the same input)

The result shown above shows that we can not use the previous architectures since style of paining is lost hence the model is not able to learn the style of paining, the premise of our posit. Interestingly all these models were first implemented in Tensorflow and results were so bad that the model were discarded. After getting first successful result on Keras we re implemented the **Deep Feature Consistent Variational Autoencoder** architecture, a model which has proved to work in the past and the result are provided below.

Training time remained roughly the same. Note that once Dropout was implemented the result have since improved, though by not much, and clarity has increased. This means that the decoder is getting overpowered, since its being trained per iteration where as the the training of encoder takes place alternatively, where it's possible that the decoder have moved to such a high state that for the encoder its equivalent to a random initialized weight. Besides Dropout we have are now training the encoders in 2:5 since encoder 1 is seeing only 20 images and encoder 2 is seeing around 430 images, in addition, thus training in batches leaves decoder learn some meaningful reconstruction aiding each encoder.

Figure 6. 1st image is the input image, 2nd image is perfect reconstruction from VAE on 20 images and 3rd output on 600 images (encoder 1) and last image is the single encoder-decoder on 600 images, is 150 x 150 therefore cropped a little more and appears zoomed

The 2 encoder and decoder model does not suffer from mode collapse because it has seen only 20 images where as the single encoder-decoder has seen all 450 images, thus a uniform poor result on all paintings.

Comparing the results of the second encoder and single encoder when both the models are suffering from mode collapse, we have :

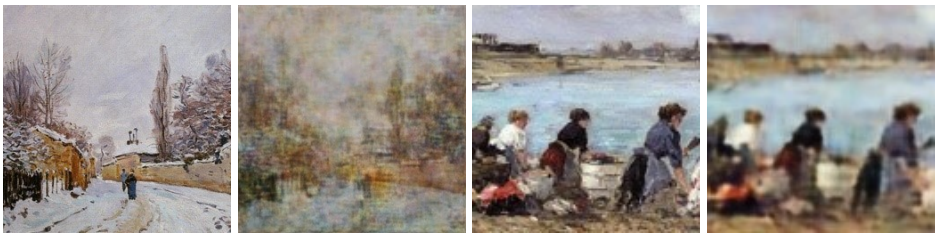

Figure 7. 1st image is the input image, 2nd image is output from 2nd encoder trained on 450 images and 3rd is the input image to single encoder decoder model and 4th is the corresponding output from the single encoder-decoder model

On the advice of Prof. Namboodiri, we tried training the encoder and decoder

with two different loss functions. Encoder with KL divergence and L1 loss multiplied by a factor of 100 and decoder with a L1 loss with a factor of 1000. When we trained the decoder with the cross entropy loss output got corrupted. This time we trained with 800 images.

Figure 8. 1st image is the input image, 2nd image is output single encoder-decoder model and 3rd is the input image and 4th is the corresponding output based on the difference loss function on decoder and encoder using L1 loss on decoder, and last pair is the output when decoder is trained using cross entropy loss, all trained on 800 images

Conclusion and Future Proposals

- We believe there is a imbalance in training two encoder and a decoder since decoder gets trained each time while encoders lag behind, trying to cope up. Need to average the weights of decoder after each epoch. **(This is evident from the fact that the dropout (0.8) on decoder performed better than normal architecture)**
- In order to force model to learn the ”style” of painting, we need to add what is similar to a Discriminator (VAE-GAN). The only difference is that it shall output the period in which that painting was painted. Thus this should help the model to learn the similar style, the semantic details rather learning the object of paintings like ”girl”, ”sky”....
- Work on Purushottam Kar's advice on technique paper on Bridge Correlational Technique to separate mode clusters
- Try to incorporate the MAD-GAN technique which separates modal clusrers [3]

References

[1] VAE Blog
https://github.com/llSourcell/how_to_generate_images_with_tensorflow_LIVE/blob/master/demo.ipynb

[2] David Berthelot, Thomas Schumm, Luke Metz
BEGAN: Boundary Equilibrium Generative Adversarial Networks
https://arxiv.org/pdf/1703.10717.pdf

[3] Arnab Ghosh , Viveka Kulharia, Vinay Namboodiri Philip H. S. Torr, Puneet K. Dokania,
MAD-GAN: Multi-Agent Diverse Generative Adversarial Networks
https://arxiv.org/pdf/1704.02906.pdf

[4] Chris Varano: Disentangling Variational Autoencoders for Image Classification
http://cs231n.stanford.edu/reports/2017/pdfs/3.pdf

[5] Wiki Arts.
https://www.wikiart.org

[6] Information about Claude Monet
https://en.wikipedia.org/wiki/Claude_Monet

[7] Cycle-GAN github architecture
https://github.com/xhujoy/CycleGAN-tensorflow

[8] Pix2Pix github architecture
https://github.com/yenchenlin/pix2pix-tensorflow

[9] Style Transfer Architecture
https://github.com/sunsided/vae-style-transfer

[10] Keras Blog for AE and VAE
https://blog.keras.io/building-autoencoders-in-keras.html

[11] Deep Feature Consistent Variational Autoencoder
https://arxiv.org/pdf/1610.00291.pdf

Figure 9. 1st image is the input image, 2nd image is almost perfect reconstruction form 1st encoder on 600 images and last image is the single encoder-decoder reconstruction on 600 images, trained on 150 x 150 size input

TO BE CONTINUED....