Advances on Variational Autoencoders

Danilo Comminiello

Generative Deep Learning 2021/2022

June 15, 2022

Sapienza Università di Roma

Lecture content highlights

- Variational autoencoders are a powerful tool in the generative modeling toolbox.
- Transforming an autoencoder into a variational autoencoder gives it the power to be a generative model.
- By performing vector arithmetic within the latent space, we can achieve some amazing effects, such as face morphing and feature manipulation.
- Some tips in implementing VAEs for face generation is provided.

Table of contents

1 FURTHER "DECODING" THE VAE

Understanding the Variational Latent Space

Vector Arithmetic

Summary of the variational autoencoder

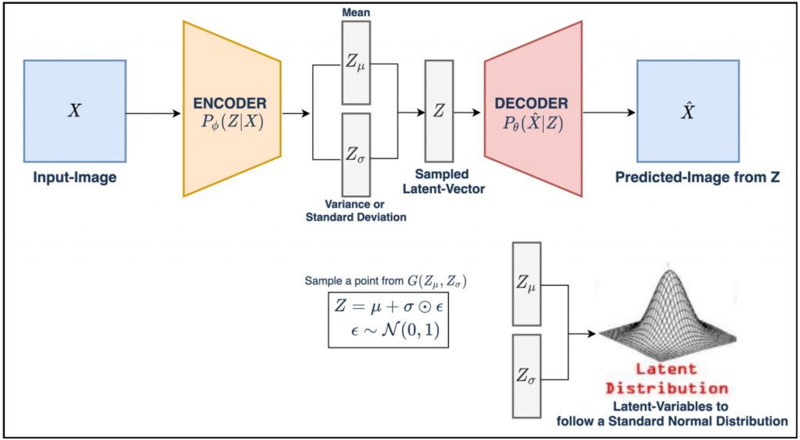


Figure 1: Diagram of a variational autoencoder (VAE). The VAE is a generative model that enforces a prior on the latent vector. Image source: learnopencv.com.

Improvement over standard autoencoders

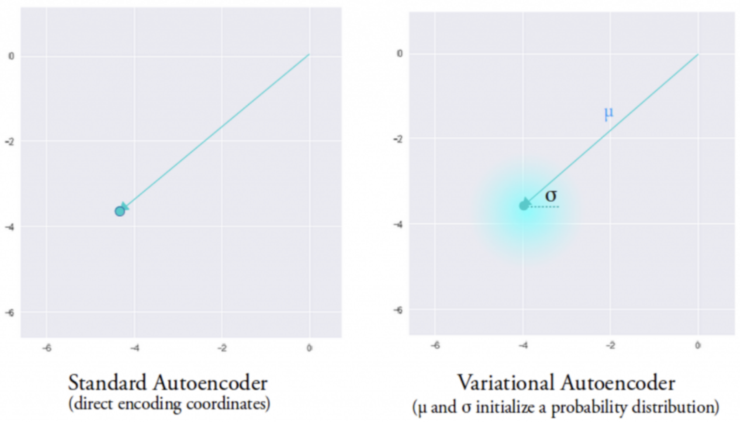
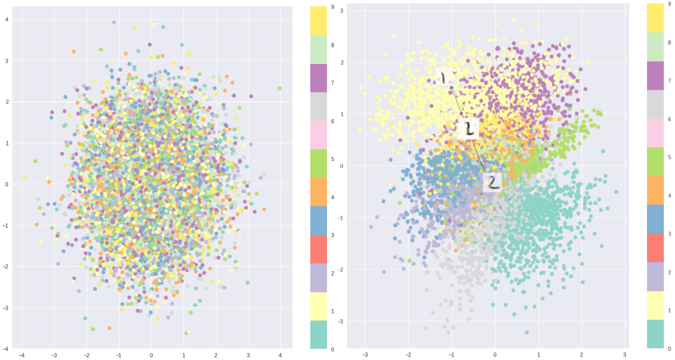


Figure 2: Instead of a single point in the latent space as in vanilla autoencoder, the VAE covers a certain “area” centered around the mean value with a size corresponding to the standard deviation. This gives the decoder a lot more to work with — a sample from anywhere in the area will be very similar to the original input. Image source: TowardsDataScience.

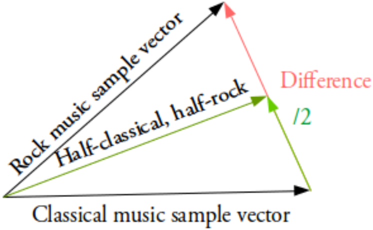
Optimizing the VAE loss



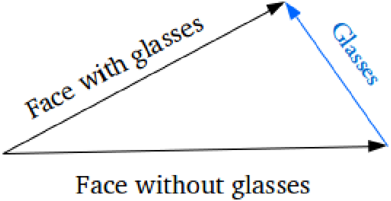
(a) (b)

Figure 3: The KL loss encourages the encoder to distribute all encodings (for all types of inputs, eg. all MNIST numbers), evenly around the center of the latent space. (a) Thus, using purely KL loss results in a latent space results in encodings densely placed randomly, near the center of the latent space. (b) Optimizing the KL together with the reconstruction loss, however, results in the generation of a latent space which maintains the similarity of nearby encodings on the local scale via clustering, yet globally, is very densely packed near the latent space origin (compare the axes with the original). Image source: TowardsDataScience.

If we interpolate between latent variables, there are no sudden gaps between clusters, but a smooth mix of features a decoder can understand.



(a)



(b)

Figure 4: (a) If we wish to generate a new sample halfway between two samples, just find the difference between their mean vectors $\boldsymbol{\mu}$, add half the difference to the original, and then simply decode it. (b) If we want to generate a specific feature (e.g., glasses on a face) we need to find two samples, one with glasses, one without, obtain their encoded vectors from the encoder, and save the difference. Add this new "glasses" vector to any other face image, and decode it. Image source: TowardsDataScience.

Morphing between faces

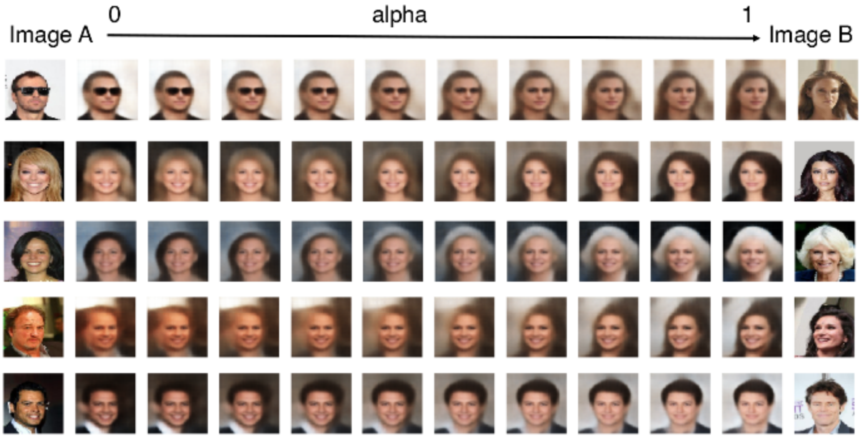


Figure 5: Let's see this in action. In a first example, we take two images, encode them into the latent space, and then decode points along the straight line between them at regular intervals [1].

Latent space arithmetic on faces

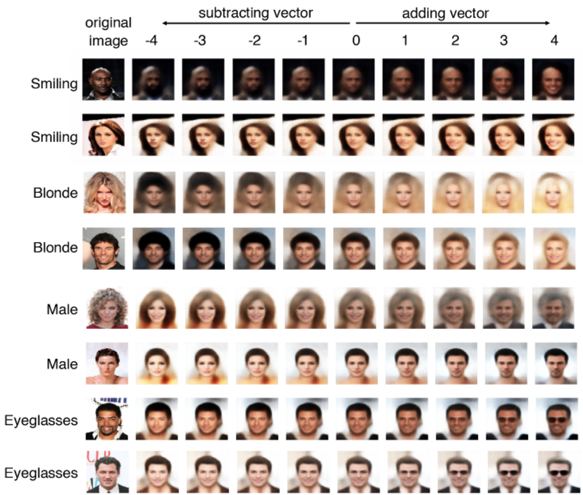


Figure 6: Here several images that have been encoded into the latent space. We then add or subtract multiples of a certain vector (e.g., smile, blonde, male, eyeglasses) to obtain different versions of the image, with only the relevant feature changed [1].

2 VAEs for Specific Applications

Using VAEs to Generate Faces

VAE for Music Generation

There are plenty of further improvements that can be made over the VAE.

Figure 7: Replacing a standard fully-connected dense encoder-decoder with a convolutional-deconvolutional encoder-decoder pair, yelds to produce great synthetic human face photos. Image source: TowardsDataScience.

Training the VAE for face generation

1. Color images have three input channels (RGB) instead of one (grayscale). This means we need to use 3 channels in the final convolutional transpose layer of the decoder.
2. Since faces are much more complex than digits, we increase the dimensionality of the latent space so that the network can encode a satisfactory amount of detail
3. Batch normalization layers are often used after each convolution layer to speed up training.
4. We increase the reconstruction loss factor to ten thousand.
5. We use a generator to feed images to the VAE from a folder, rather than loading all the images into memory up front.

Analysis of the VAE

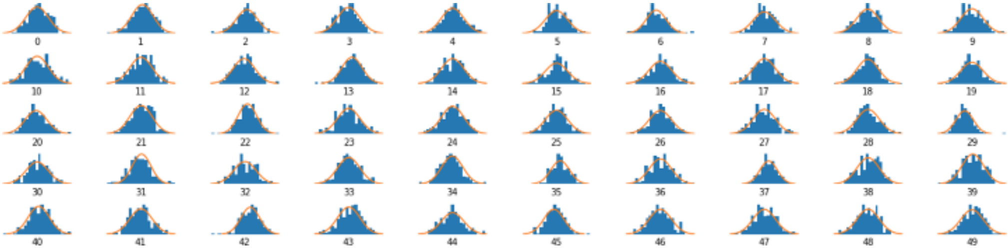


Figure 8: Distributions of points for the first 50 dimensions in the latent space. There aren't any distributions that stand out as being significantly different from the standard normal. This means that the model is ready to generate faces! [1]

VAEs can even work with sequential data, to produce synthetic text, or even interpolate between MIDI samples.

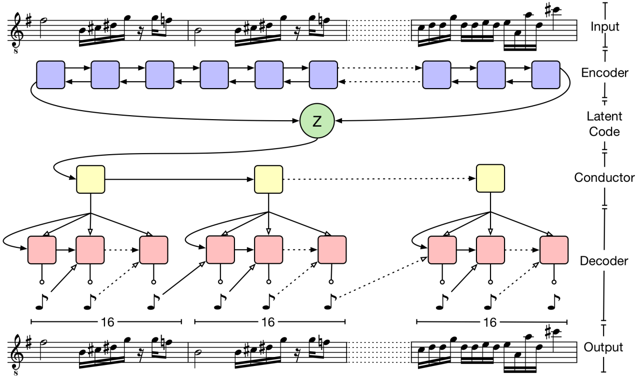


Figure 9: The MusicVAE is a machine learning model that lets us create palettes for blending and exploring musical scores. Image source: Magenta. Watch this video: Drum 2-bar Performance Interpolation.

References

[1] D. Foster, *Generative Deep Learning – Teaching Machines to Paint, Write, Compose and Play*. O'Reilly Media, Inc., Jun. 2019.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv Preprint: arXiv:1312.6114v10*, May 2014.

[3] ——, "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, Nov. 2019.

[4] D. J. Rezende, L. Metz, and S. Chintala, "Stochastic backpropagation in approximate inference in deep generative models," in *Int. Conf. on Machine Learning (ICML)*, Beijing, China, Jun. 2014, pp. 1278–1286.

Advances on Variational Autoencoders

Generative Deep Learning
2021/2022

Danilo Comminiello

PhD Course in Information and Communication Technology (ICT)

http://danilocomminiello.site.uniroma1.it

danilo.comminiello@uniroma1.it