Advanced Institute for Artificial Intelligence

Autoencoders

Advanced Institute for Artificial Intelligence – AI2

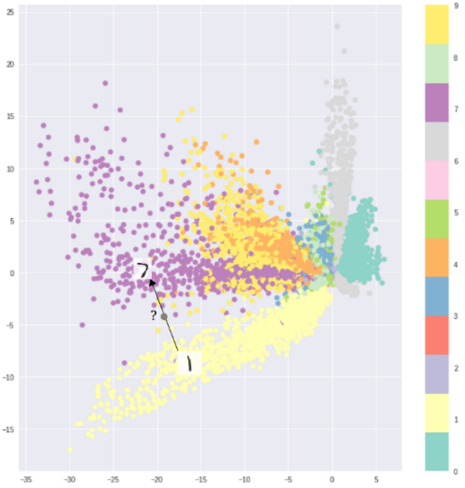
https://advancedinstitute.ai

Variational Autoencoders

Variational Autoencoders

Recap: The problem with standard autoencoders

- Asides from a few applications like denoising autoencoders, they are fairly limited;
- The latent space they convert their inputs to and where their encoded vectors lie, may not be continuous, or allow easy interpolation.
- For example, training an autoencoder on the MNIST dataset, and visualizing the encodings from a 2D latent space reveals the formation of distinct clusters:



Recap: The problem with standard autoencoders

- When building a generative model, you don't want to prepare to replicate the same image you put in:
    - Randomly sample from the latent space, or
    - Generate variations on an input image, from a continuous latent space;
- If the space has discontinuities and you sample/generate a variation from there, the decoder will simply generate an unrealistic output;
    - The decoder has no idea how to deal with that region of the latent space;
    - During training, it never saw encoded vectors coming from that region of latent space;

Definitions

- Variational Autoencoders (VAEs) have one fundamentally unique property that separates them from regular autoencoders:
  - Their latent spaces are, by design, continuous;
  - The continuity of the latent space allows for easy random sampling and interpolation.
- Its encoder not output an encoding vector of size n;
- Instead, it outputs two vectors of size n:
  - a vector of means, μ, and
  - another vector of standard deviations, σ.
  - The mean and standard deviation of the i-th random variable, X_i from which we sample, to obtain the sampled encoding which we pass onward to the decoder;

Variational Autoencoders

Definitions

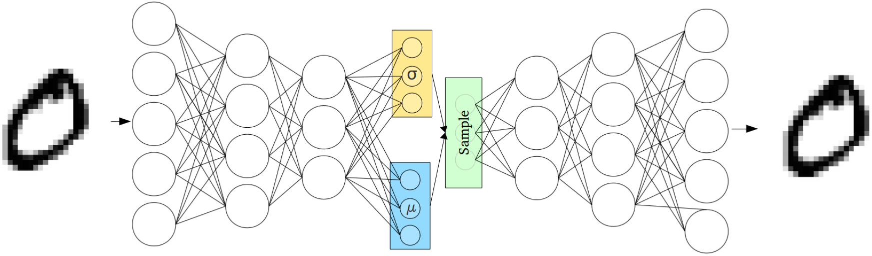


Figure: Variational Autoencoder with the μ and σ vectors

Image from Variational Autoencoder architecture by I

Example

In the scenario where we have an input signal with 500 features and we aim to reduce this signal to just 30, we could think of building a VAE just like this:

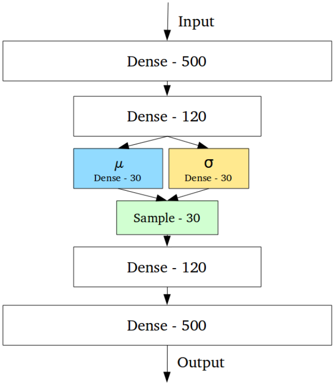


Figure: VAE that reduces a 500 dimensional input to a 30 dimensions latent space

Variational Autoencoders

Definitions

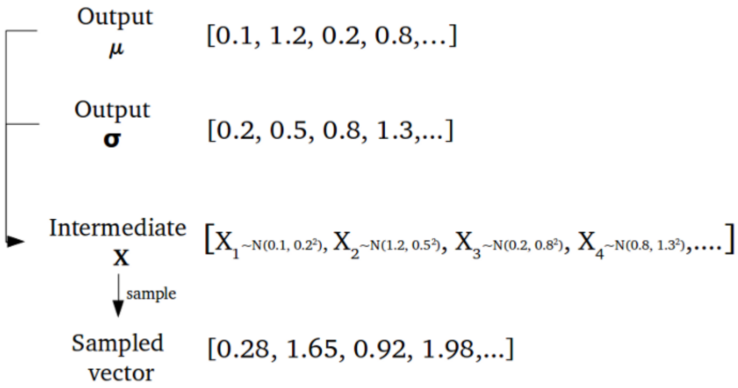


Figure: How the forward pass works

Definitions

- Stochastic generation of encoding vectors
  - For the same input, keeping the mean and standard deviation the same the actual encoding will vary on every single pass due to sampling.
- The mean vector controls where the encoding of an input should be centered;
- the standard deviation controls how much from the mean the encoding can vary (the area)

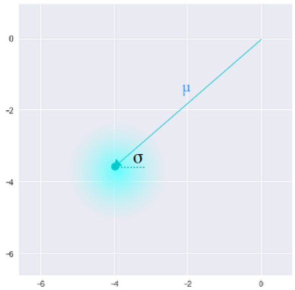


Figure: μ and σ to control the sampling

Definitions

- Not only is a single point in latent space refers to a sample of that class.
- All nearby points refer to the same as well in a $sigma$-radius;
- The goal here is to try to create more homogeneous latent space, getting rid of the discontinuity;
  - The model is now exposed to a certain degree of local variation by varying the encoding of one sample;
  - We want overlap between samples that are not very similar too;
    - Interpolation between classes;

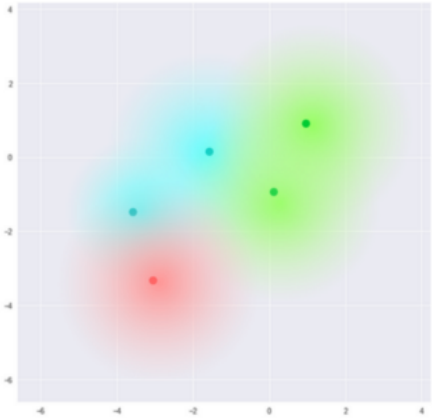
Definitions

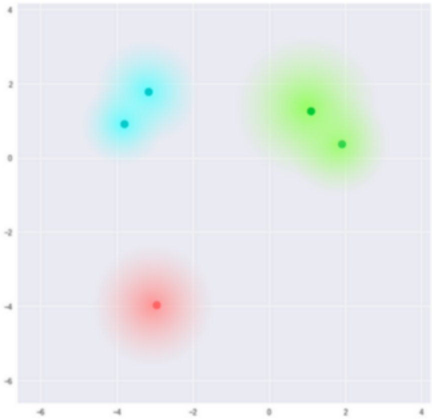
- There are no limits on what values vectors μ and σ can take on:
  - Encoder can learn to generate very different μ for different classes, clustering them apart, and minimizing σ
  - Can come to a point that it looks like a single dot.
- Desirable: Encodings that are as close as possible while still being distinct, allowing smooth interpolation, and enabling the construction of new samples

Variational Autoencoders

Definitions

What we want, and what we may get:





Definitions - The KL Divergence

- Kullback-Leible divergence
- Measures how much they diverge from each other;
- For VAEs, the KL loss is equivalent to the sum of all the KL divergences between the component $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and the standard normal.
  - This measure is minimized when $\mu_i = 0$ and $\sigma_i = 0$
- When the divergence is calculated between univariate distributions it can be simplified to [1]:

$$\sum_{i=1}^{n} \sigma_i^2 + \mu_i^2 - log(\sigma_i^2) - 1$$

[1] Deriving the KL divergence loss for VAEs

Definitions - The KL Divergence

- This loss forces the encoder to distribute all encodings evenly around the center of the latent space;
- Using purely KL loss results in a latent space results in encodings densely placed randomly, near the center of the latent space
- The decoder finds it impossible to decode anything meaningful from this space;

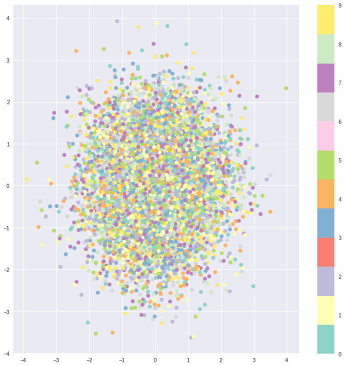


Figure: Latent space produced by VAE trained only with KL Loss

Putting it all together

- Use the KL divergence as a penalization mechanism;
- Optimizing the two together (reconstruction - e.g., crossentropy) and the KL divergence;
  - Generation of a latent space which maintains the similarity of nearby encodings;
  - Globally, is very densely packed near the latent space origin
  - Equilibrium reached by the cluster-forming nature of the reconstruction loss and the dense packing nature of the KL loss;

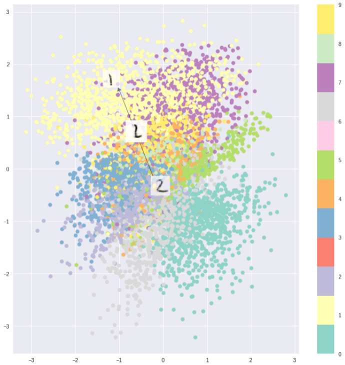


Figure: Using the composed loss