Autoencoders

Pratik Prajapati (Pratikkumar.Prajapati@sjsu.edu)

Feb/25/2020

Autoencoders (AEs) Basics

- An artificial neural network that is trained to attempt to copy its input to its output with constrains.
- Unsupervised learning models.
- Types of AEs
  - Undercomplete AEs
  - Regularized AEs
    - Denoising AEs (DAEs)
    - Variational AEs (VAEs)
    - And many more…

Undercomplete AEs

- **Encoder**: encodes the input data by learning features of data
- **Latent Space**: hidden layer which holds the features of the input data in lower dimension.
- **Decoder**: decodes latent space to reconstruct the input image.

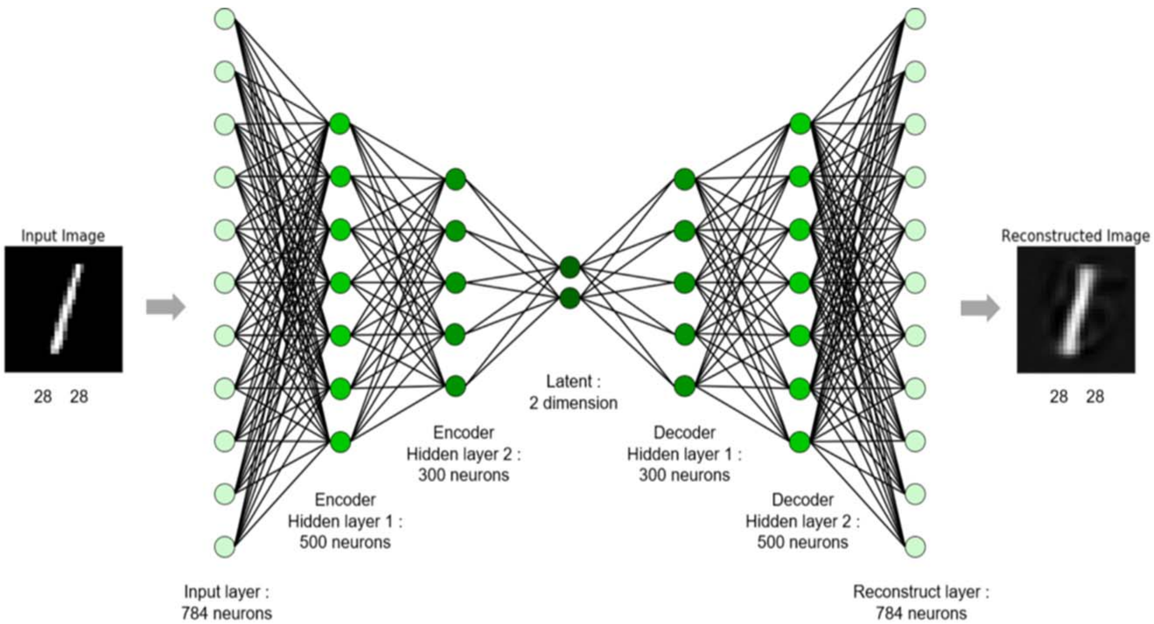

Image source: https://mc.ai/auto-encoder-in-biology/

Undercomplete AEs (cont.)

- **Loss function**: Typically MSE to make sure reconstructed image is same as input, to make sure encoders learn 'useful' latent space. (a.k.a. reconstruction loss)
- **Learning**: L(x, g(f(x)))
  where, x = input
  hidden code h = f(x) is learning objective of encoder
  x = g(h) is learning objective of decoder
- Undercomplete
  - Code dimension is less than the input dimension
- Overcomplete
  - Code dimension is more than the input dimension
  - Needs regularization to avoid learning trivial identity function
- AEs can be implemented with plain neurons and also using CNNs (for images)

The problem of AE as Generative model.

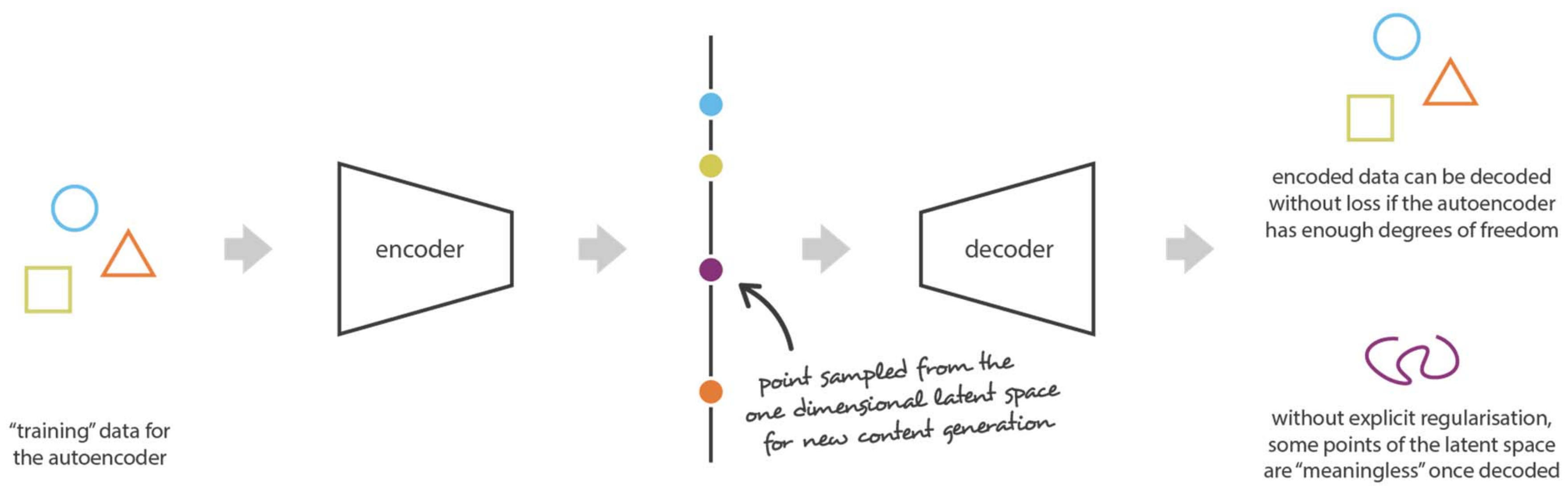

Image source: [4]

Possible solution space

- **instead of encoding an input as a single point or vector, we encode it as a distribution over the latent space**.

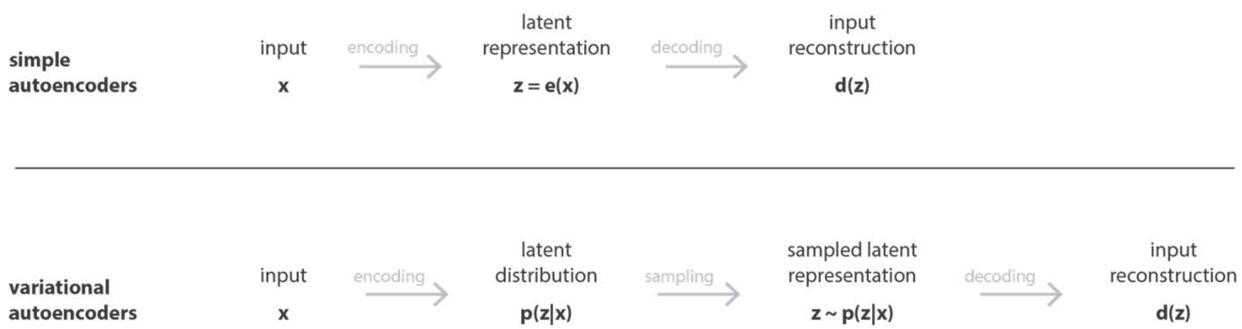

Image source: [4]

Variational Autoencoders (VAEs)

- VAE’s aim is to create a new image that is a member of the same class of images but recognizably new.
- Since we want to create new image, its not sufficient to minimize just reconstruction loss.
- Total loss = reconstruction loss + variational loss
- Variational loss (Kulback-Leibler divergence):

$$L_v(\mu, \sigma) = -\sum_i \frac{1}{2}(1 + 2\sigma[i] - \mu[i]^2 - e^{2\sigma[i]})$$

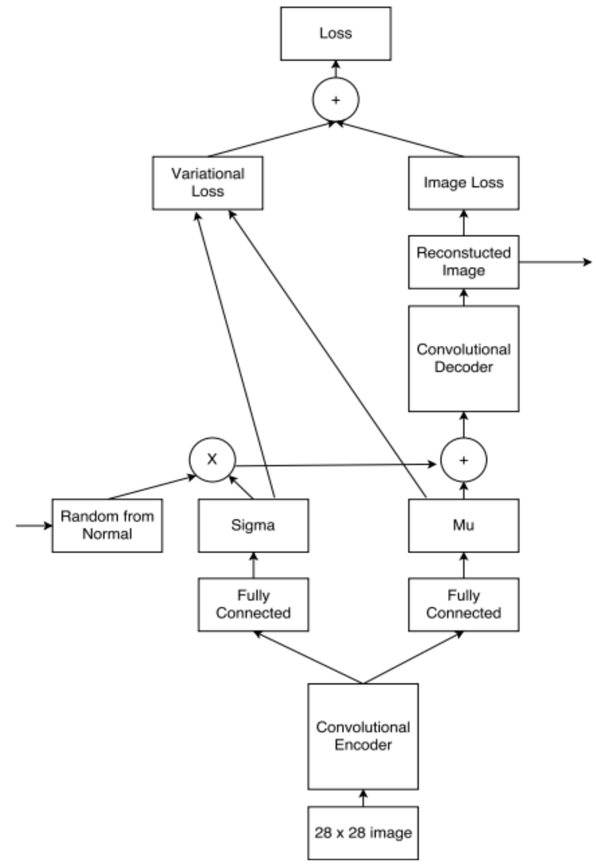

Image source: [3]

Regularized latent space

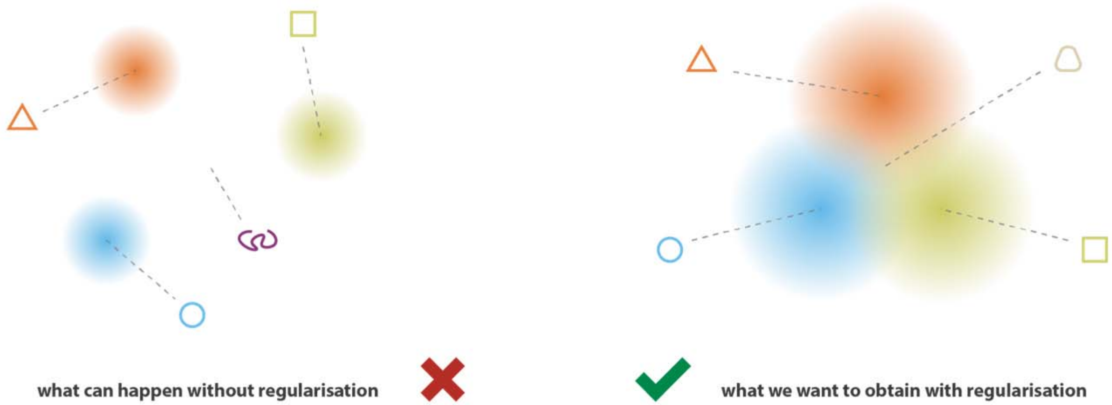

Image source: [4]

Applications of AEs

- Dimensionality reduction (Lower dimension representation can improve performance of tasks.)
- Information retrieval. E.g. semantic hashing
- Anomaly Detection. E.g. Reconstruction loss gets high for outliers
- Image Processing. E.g. lossy compression, denoising images
- VAEs and its variants are capable to generate new data.
  - new images/videos (DeepFakes)
  - Drug discovery
- Popularity prediction
- Machine Translation

References.

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. The MIT Press

[2] https://en.wikipedia.org/wiki/Autoencoder

[3] Eugene Charniak, Introduction to Deep Learning. 2019. The MIT Press

[4] https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73