Introduction to Deep Learning

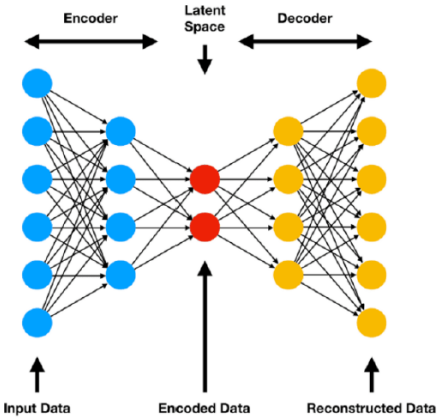
Tutorial 9

Gabriel Deza
Department of Industrial Engineering
Tel Aviv University

December 26, 2025

Autoencoders

- An autoencoder is a feed-forward neural net $\mathbf{f}$ whose job is to take an input $\mathbf{x}$ and predict $\mathbf{x}$.
- The goal is to reconstruct the input as accurately as possible, i.e., $\mathbf{f}(\mathbf{x}) \approx \mathbf{x}$.
- To prevent simply learning the identity map $\mathbf{f}(\mathbf{x}) = \mathbf{x}$, we add a bottleneck layer whose dimension is much smaller than the input.
- Key components:
  - Encoder: compresses the input into a low-dimensional representation (latent space)
  - Latent vector: the bottleneck representation of the input
  - Decoder: reconstructs the input from the latent vector

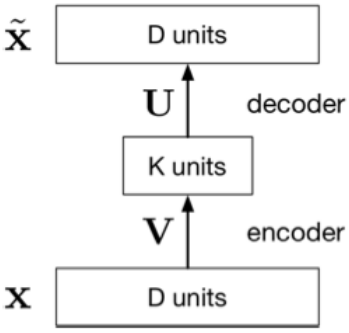


Principal Component Analysis (PCA) - Intro to ML

- The simplest kind of autoencoder has one hidden layer, linear activations, and squared error loss:

$$\mathcal{L}(\mathbf{x}, \tilde{\mathbf{x}}) = \|\mathbf{x} - \tilde{\mathbf{x}}\|^2.$$

- This network computes $\tilde{\mathbf{x}} = \mathbf{UVx}$, which is a linear function.
- When $K < D$:
  - $\mathbf{V}$ maps $\mathbf{x}$ to a K-dimensional space (dimensionality reduction).
  - The output must lie in a K-dimensional subspace, namely the column space of $\mathbf{U}$.
- Deep nonlinear autoencoders learn to project not onto a subspace, but on a nonlinear manifold



Autoencoders

Why autoencoders are useful?

- Map high-dimensional data to lower dimension (helpful for visualization)
- Compression (i.e. reducing file size)
- Learning abstract features in an unsupervised way so you can apply them to supervised task
- Learning a semantically meaningful representation (e.g, interpolating between different images)

Colab Notebook

- Google Colab link