# Notes on Autoencoders

compiled by D. Gueorguiev 3/25/24

An autoencoder is a specific type of a neural network, which is mainly designed to encode the input into a compressed and meaningful representation, and then decode it back such that the reconstructed input is as similar as possible to the original one. The purpose of autoencoders is learning in an unsupervised manner an “informative” representation of the data that can be used for various tasks such as clustering.

The problem is defined as to learn the functions (encoder) and (decoder) that satisfy:

(1)

where is the expectation over the distribution of , and is the reconstruction loss function, which measures the distance between the output of the decoder and the input. The latter is usually set to be the -norm. The Figure provides an example of an Autoencoder application:

A diagram of a coder

Description automatically generatedFigure: an Autoencoder example with an image

Historically, and are neural networks. In the special case that A and B are linear operations, we get a linear autoencoder (see [3]).

## Appendix

### Linear Autoencoders and Principal Component Analysis

//TODO: this is mainly review and discussion of [3]

Literature

[1] [Autoencoders, Dor Bank et al, 2021](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/Autoencoders.pdf)

[2] [Autoencoders, Unsupervised Learning, and Deep Architectures, Pierre Baldi, 2012](https://github.com/dimitarpg13/deep_learning_and_neural_networks/blob/main/literature/articles/Autoencoders_Unsupervised_Learning_and_Deep_Architectures_Baldi_2012a.pdf)

[2] [Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima, Pierre Baldi, Kurt Hornik, 1988](https://github.com/dimitarpg13/deep_learning_and_neural_networks/blob/main/literature/articles/Neural_Networks_and_Principal_Component_Analysis-Learning_from_Examples_Without_Local_Minima_Baldi_Hornik-89.pdf)