What is Autoencoder?

- An autoencoder is a neural network that is trained to copy its input to its output, with the typical purpose of dimension reduction - the process of reducing the number of random variables under consideration.
- It features an encoder function to create a hidden layer (or multiple layers) which contains a code to describe the input.
- There is then a decoder which creates a reconstruction of the input from the hidden layer.

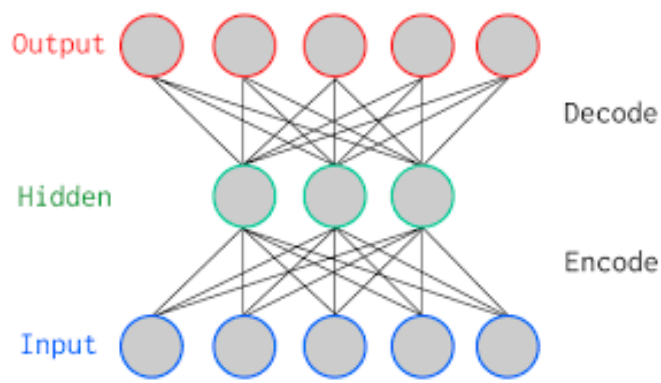

Figure 1

The most intuitive application of autoencoders is data compression. Given a 256 x 256 pixel image for example, a representation of a 28 x 28 pixel may be learned, which is easier to handle.

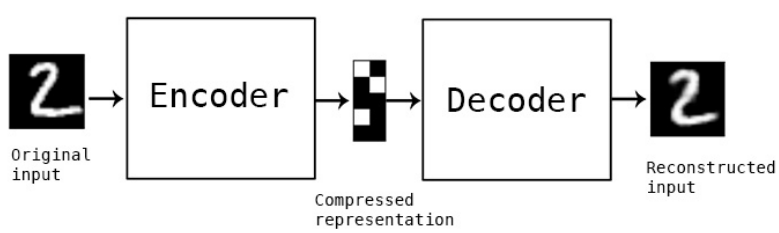

Figure 2

An autoencoder can then become useful by having a hidden layer smaller than the input layer, forcing it to create a compressed representation of the data in the hidden layer by learning correlations in the data.

This facilitates the classification, visualisation, communication and storage of data. Autoencoders are a form of unsupervised learning, meaning that an autoencoder only needs unlabelled data - a set of input data rather than input-output pairs.

There are variety of autoencoders, such as the convolutional autoencoder, denoising autoencoder, deep autoencoder, variational autoencoder and sparse autoencoder.

References:

For more details on AutoEncoders refer below:
`https://en.wikipedia.org/wiki/Autoencoder`
`http://www.math.snu.ac.kr/~hichoi/machinelearning/lecturenotes/Autoencoder.pdf`