CHAPTER 37

Autoencoders

Autoencoder is an unsupervised learning algorithm that uses neural networks to reconstruct the features of a dataset. Just like the unsupervised algorithms that we earlier discussed in the chapter on machine learning, autoencoders can be used to reduce the dimensionality of a dataset and to extract relevant features. Moreso, peculiar to autoencoders is the ability to generate more examples of the dataset after learning an internal representation (also called coding) that reconstructs the features of the inputs to the neural network.

An autoencoder receives as input the features of the dataset. These features are passed through a set of encoders, which are the hidden layers of a neural network to create an internal representation called codings. The learned coding is then used to reconstruct the output through a set of decoders, which are also hidden neural network layers. The autoencoder cannot merely do a trivial memorization of the inputs, because a constraint is placed on the encoders by reducing the input dimension to force the network to learn an efficient set of representation from which the decoders use to reconstruct the inputs.

Autoencoders with restricted Encoders and Decoders are called **undercomplete**. A reconstruction error term is used to evaluate the performance of an autoencoder by testing how well the output corresponds with the input. Of course, just like other neural networks, the neurons of the Encoders and Decoders have non-linear activation functions for learning complex patterns. An example of a simple autoencoder network architecture is shown in Figure 37-1.

© Ekaba Bisong 2019
E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_37

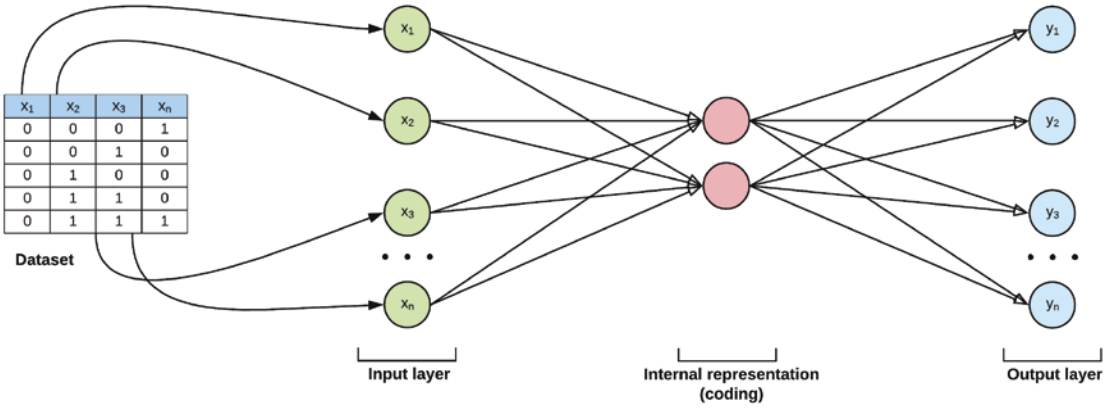

Figure 37-1. *A simple autoencoder architecture*

Stacked Autoencoders

Stacked autoencoder is when the simple autoencoder architecture as shown in Figure 37-1 is enhanced with multiple hidden layers. Just like other deep neural network architectures with hidden layers, the hidden layers of an autoencoder enable the network to learn more complex patterns of the input dataset.

The hidden layers of a stacked or deep autoencoder are added symmetrically at both the Encoder and Decoder part of the network as shown in Figure 22-2. The neurons of the hidden layers are restricted to be less than that of the input layer. This formulation places a restriction on the network, so it doesn't merely memorize the input. Moreso, care must be taken not to create too many deep layers, so the autoencoder does not overfit the input data and fail to generalize to out-of-sample examples. To optimize the training of a deep autoencoder, the weights of the symmetrical neural layers are shared in a technique called *tying*.