Basics of Autoencoders

medium.com/@birla.deepak26/autoencoders-76bb49ae6a8f

Deepak Birla March 12, 2019

Deepak Birla

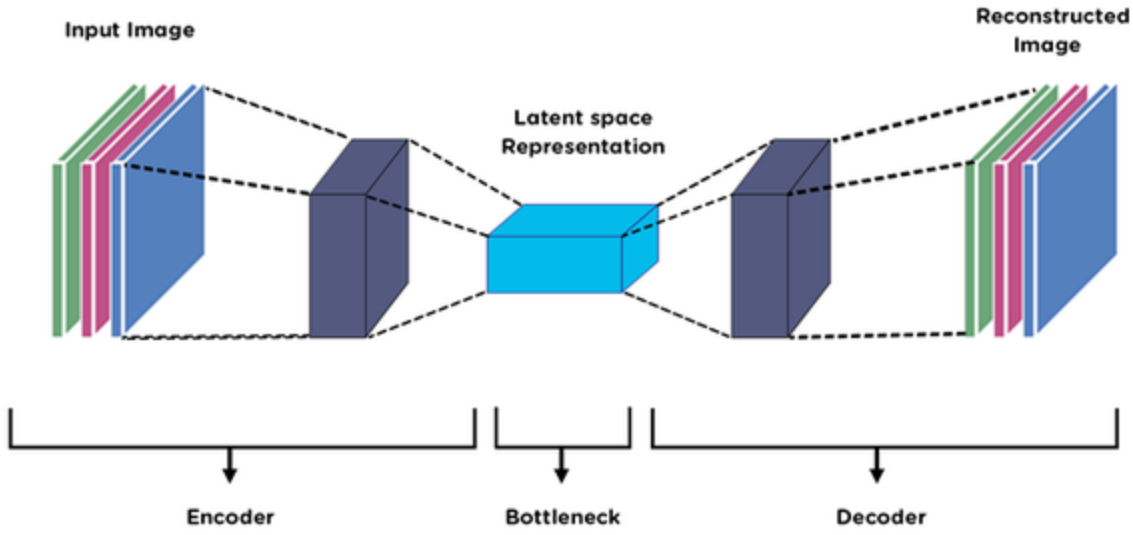
Looking to hide highlights? You can now hide them from the "•••" menu.

Autoencoders (AE) are type of artificial neural network that aims to copy their inputs to their outputs . They work by compressing the input into a **latent-space representation** also known as **bottleneck**, and then reconstructing the output from this representation. Autoencoder is an unsupervised machine learning algorithm. We can define autoencoder as **feature extraction algorithm**.

The input data may be in the form of speech, text, image, or video. An Autoencoder finds a representation or code in order to perform useful transformations on the input data.

Properties of Autoencoders

- Autoencoders are data-specific, which means that they will only be able to compress data similar to what they have been trained on. Example, an autoencoder trained on pictures of faces would do a rather poor job of compressing pictures of trees, because the features it would learn would be face-specific.
- Autoencoders are lossy, which means that the decompressed outputs will be degraded compared to the original inputs.
- Autoencoders are learned automatically from data examples, which is a useful property: it means that it is easy to train specialized instances of the algorithm that will perform well on a specific type of input. It doesn't require any new engineering, just appropriate training data.



Autoencoder Architecture

Autoencoder network is composed of two parts **Encoder** and **Decoder**.

Encoder : This part of the network encodes or compresses the input data into a latent-space representation. The compressed data typically looks garbled, nothing like the original data.

Decoder : This part of network decodes or reconstructs the encoded data(latent space representation) back to original dimension. The decoded data is a lossy reconstruction of the original data.

Idea behind Autoencoders:

Data compression is a big topic that's used in computer vision, computer networks, computer architecture, and many other fields. The point of data compression is to convert our input into a smaller(Latent Space) representation that we recreate, to a degree of quality. This smaller representation is what would be passed around, and, when anyone needed the original, they would reconstruct it from the smaller representation. And autoencoders are the networks which can be used for such tasks.

Why copy Input to Output?

Purpose of autoencoders in not to copy inputs to outputs, but to train autoencoders to copy inputs to outputs in such a way that **bottleneck** will learn useful information or properties.

We can make out latent space representation learn useful features by giving it smaller dimensions then input data. In this case autoencoder is **undercomplete.** By training an undercomplete representation, we force the autoencoder to learn the most salient features of

the training data. If we give autoencoder much capacity(like if we have almost same dimensions for input data and latent space), then it will just learn copying task without extracting useful features or information from data. If dimensions of latent space is equal to or greater then to input data, in such case autoencoder is **overcomplete**. In such case even linear encoder and linear decoder can learn to copy the input to the output without learning anything useful about the data distribution.

Why Autoencoders?

Data denoising and **Dimensionality reduction for data visualization** are considered as two main interesting practical applications of autoencoders. With appropriate dimensionality and sparsity constraints, autoencoders can learn data projections that are more interesting than PCA or other basic techniques.

Autoencoders also can be used for Image Reconstruction, Basic Image colorization, data compression, gray-scale images to colored images, generating higher resolution images etc. But you can only use them on data that is similar to what they were trained on, and making them more general thus requires *lots* of training data.

I pulse the readers interest through claps on the article.

This article is part of Series Autoencoders. For more about Autoencoders and there implementation you can go through series page(Link given below).

Autoencoders Series

Autoencoders (AE) are type of artificial neural network that aims to copy their inputs to their outputs . They work by…

medium.com