0.1 Fundamentals

There are several types of Neural Networks architectures, but Convolutional Neural Networks (CNNs or Convnets) are probably the most widely implemented model overall [11, 6] with successful applications in the domains of Computer Vision [4, 7, 8, 13] or Natural Language Processing[1, 9, 10]. In the CNN category itself there are different variants, but they all abide the fundamental structure of a feedforward hierarchical multi-layer network. Feedforward because the information only flows in a singular direction without cycling [12], hierarchical because the higher complexity internal representations are learned from lower ones [5, 14] and multi-layer because it is composed of a series of stages, blocks or layers: the raw data is fed to an input layer, forwarded to a sequence of intercalating convolutional and pooling layers, proceeded to a stage of one or more fully-connected layers [5, 11, 3, 2].

Using CNNs for Computer Vision tasks is not an arbitrary choice, but due to the fact that the network design can benefit from the intrinsic characteristics of the input data [5] they perform really well in image related applications. In the first place, images have an array-like structure with numerous elements, namely, each pixel has an assigned value organized in a grid like manner [11], matching the type of input for these networks [5]. In the second place, there's an inherent correlation between local groups of values, which creates distinguishable motifs. Finally, the local values of images are invariant to location, that is, a certain composition should have the same value independently of the spatial location in the picture.

This model architecture is based on the visual cortex ventral pathway, therefore, it is capable of automatically extracting spatial feature hierarchies, from a lower to higher complexity (in contrast to conventional machine learning) [5, 11, 3, 2].

Key features such as local connections/receptive fields, shared weights, subsampling and the use of many layers allows this network to be invariant to shift, scale and distortions.