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 [12, 7] with successful applications in the domains of Computer Vision [5, 8, 9, 14] or Natural Language Processing[1, 10, 11]. 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 [13], hierarchical because the higher complexity internal representations are learned from lower ones [6, 15] 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, transmitted to a stage of one or more fully-connected layers [6, 12, 4, 2]. The convolutional layer is designed to extract feature representations by being composed of kernels (or filter banks [6]) that compute feature maps through element-wise product, to which is applied a nonlinear activation function [4, 12]. Next is the pooling layer, that's responsible for reducing the spatial size of the input data [4] and joining identical features [6]. Finally, the fully connected layers and their core function is to perform high logic and generate semantic information [4]. Finally, the output layer

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, consequently performing really well in image related applications [6, 3]. 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 [12]. In the second place, there's an inherent correlation between local groups of values, which creates distinguishable motifs [6]. 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 [6]. Therefore, the following key, unique features potentiate the previously stated efficient performance [3]:

1. Designed to process multidimensional arrays [6];
2. Shared weights between the same features in different locations;
3. Automatically identifies the relevant features without any human supervision, hence, small amounts of preprocessing [2, 7];
4. Local connections/receptive fields/sparse connectivity [2];
5. Pooling layers that reduces the spatial size of the input data.

The ensemble of features 2, 4 and 5 enable an invariance to small shifts, distortions and rotations citations needed, while 2, 3, 4 and 5 helps to reduce the complexity of the model, as a result, it is easier to train the network.