Usable ML Writeup

Convolutional Layer:

As the key components in a CNN, the convolutional layers perform two jobs: extract key features from the input while reducing its dimensionality, i.e. its size. To do this, a matrix smaller than the input is randomly initialized and iterates over the pixels while performing matrix multiplication with the given pixels. Depending on the matrix this might result in the detection of edges or calculation of the mean value for a pixel and its neighbors or other features. The output is a so-called feature map with highlighted features. There are multiple of these matrices in each convolutional layer whose values are adjusted during training to produce the best performance.

Maxpool layer:
Maxpooling is another operation that reduces dimensionality, thereby decreasing the number of calculations necessary. It is best explained by an example: Consider a 4x4 matrix. This matrix can be split into four quadrants of size 2x2. Maxpooling chooses the largest value in each quadrant (hence max pooling) and returns a 2x2 matrix with these values.

Linear layer:
Linear layers make up the final layers in neural networks for classification. As input, they take a vector which in a CNN is a flattened matrix from the previous layer. While data passes through the network, the linear layers perform a linear transformation by multiplying each input neuron by a weight and adding a bias. Several of these layers can be connected sequentially to reduce the number of inputs. Finally, the output consists of a neuron for each class and a corresponding probability determining how likely the input belongs to each class.