Convolutional Neural Network for Fruits Classification

Ylenia Messina YLENIA.MESS@ICLOUD.COM

1. Model Description

The following model, **CNN3**, is a neural network consisting of 3 Convolutional Layers, followed by 1 Fully Connected Layer and a final Classifier. It was used for a multiclass classification problem: in particular, the task for the model involved recognizing the *fruit* type – among the 10 considered – and therefore assigning the correct label.

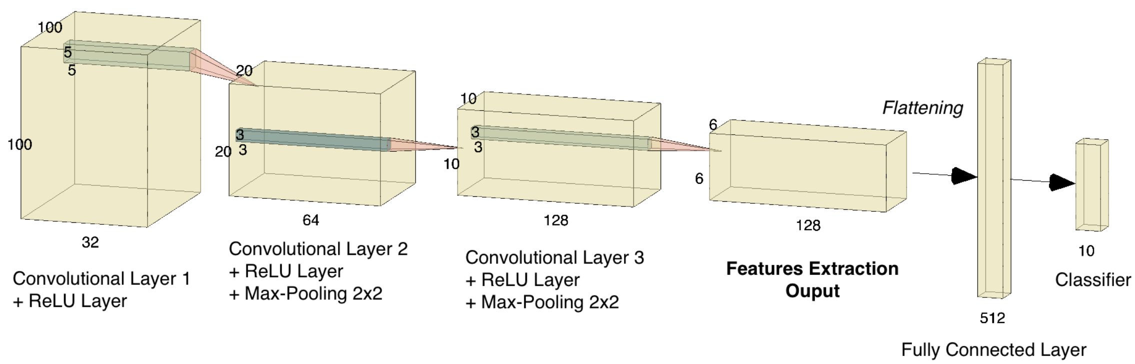

Figure 1: A simple representation of the proposed model

The features extraction is performed by the initial sequence of 3 Convolutional Layers. The first convolutional layer takes in the 3 channels of the input image – as the model is set to process RGB images – and produces 32 out channels, applying convolutional kernels of size 5x5 with a stride of 3. The remaining two convolutional layers produce respectively 64 and 128 out channels, and both apply kernels of size 3x3 with a stride of 1. Each convolutional layer is followed by a ReLU activation function and also by a Max Pooling Layer with kernel size 2x2 and stride of 2 (except for the first convolutional layer which has no Max Pooling Layer after).

The output of such convolutional layers is flattened into 4608 features which are taken as input from the first and only one fully connected layer, consisting of 512 nodes. The fully connected layer is followed by a ReLU activation function as well, and finally the Classifier layer which is responsible for ultimately classifying the input image sample with the correct class according to the highest value among the 10 outputs of the model – where 10 is the number of classes in the dataset used.

2. Dataset

The dataset is "**Fruits-360**", provided by Horea Muresan and Mihai Oltean from their paper "*Fruit recognition from images using deep learning*" (Acta Univ. Sapientiae, Informatica Vol. 10, Issue 1, pp. 26-42, 2018). The dataset is available on Kaggle but also on GitHub.

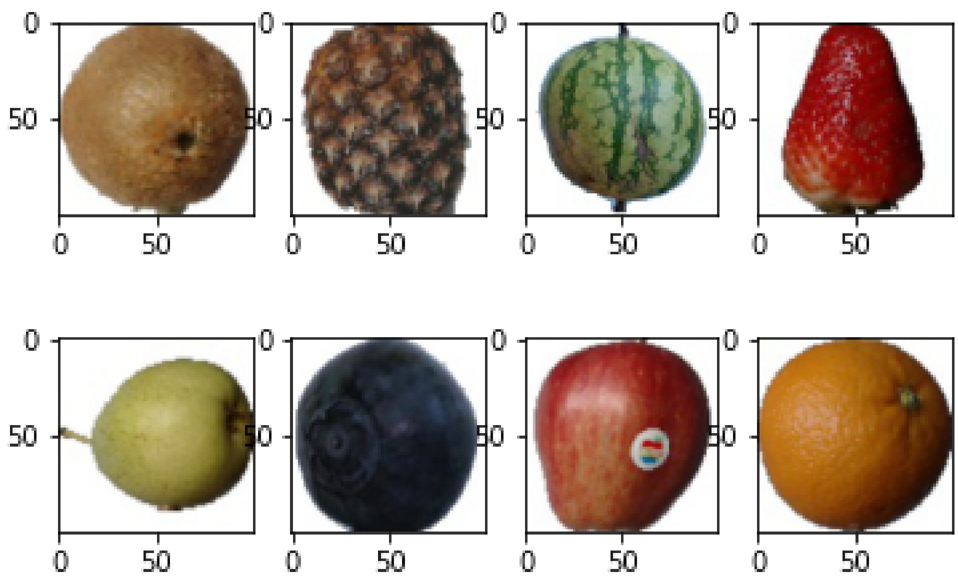

Figure 2: Some example images from the Fruits-360 dataset.

The original dataset is made up of 100x100 size images of fruits or vegetables: each image represents a fruit or vegetable among a total of 131 classes, on white background. In particular, it consists of 67692 images intended to be used as *training set* and 22688 images as *test set*. Both training set and test set are balanced in terms of classes: each class has more or less the same number of images.

However, a smaller subset of such dataset was used for the task: both training set and test set were reduced in order to contain only image samples of 10 fruits (*Apple Red, Banana, Blueberry, Kiwi, Orange, Peach, Pear, Pineapple, Strawberry, Watermelon*) – therefore only 10 classes considered. For model selection purposes, a validation set was obtained as a fraction of 20% from the training set.

Also, data augmentation was performed, applying some transformations to images such as horizontal and vertical flipping (both with probability 0.5).

3. Training procedure

A starting model (CNN4) – structured similarly to CNN3 but with 4 convolutional layers rather than 3 – was trained and validated, to then proceed with an ablation study, gradually removing layers to check whether results would improve or not.

As training loss, the *Cross Entropy Loss* – which measures the cross-entropy between the predicted class label and the actual class label – has been employed. This criterion combines the *Softmax* activation function principle and the negative log-likelihood loss.

All the models are trained from scratch for 30 epochs on Google Colab. Batch size is set to 64 for all the models. *Stochastic Gradient Descent* is chosen as optimizer algorithm, using a learning rate of 0.01.

4. Experimental Results

Models were trained as described in the previous section. Table (1) shows test results for the proposed architecture as well as of ablation studies (i.e., different variants of the final architecture when removing layers).

Model	**Validation Loss**	**Validation Accuracy**	**Test Accuracy**
CNN4	0.055	99.32%	95.88%
- 1 Conv Layer (CNN3)	0.017	100%	98.30%
- 2 Conv Layers (CNN2)	0.023	100%	96.39%
- 3 Conv Layers (CNN1)	0.027	100%	100%

Table 1: Performance of the models.

Even though in this case the ablation studies experiments seem to be a little pointless since the starting model CNN4 already provides an accuracy of 99% on validation set, removing layers further improves performance allowing to reach 100% validation accuracy.

Although it would be better to choose the Baseline model (CNN1, with only 1 convolutional layer) due to it its lower complexity, CNN3 is the one that the results with the least loss on validation set. Therefore, CNN3 is the final selected model.