**Report**

Assignment 2, Deep Learning

by Carmel Shablin 305812661

and Gregory Koushnir 321889479

## Task

Build Siamese convolutional neural network for one-shot image recognition, given two facial images of previously unseen persons, the model has to successfully determine whether they are the same person.

## Dataset

The Labeled Faces in the Wild-a (LWF) dataset was used for the task (can be found [here](https://talhassner.github.io/home/projects/lfwa/index.html)). This dataset contains gray-scaled 250x250 images of human faces which were alignmed using a commercial face alignment software. The data set was splitted by given distribution for [Train](http://vis-www.cs.umass.edu/lfw/pairsDevTrain.txt) \ [Test](http://vis-www.cs.umass.edu/lfw/pairsDevTest.txt) sets. This division is set up so that no subject from test set is included in the train set. When the images are loaded, input features (pixels) values are divided by 255 to normalize in range [0,1].

All samples are taken from labels that have more than 1 image – the reason is because if we use an image from label that has only one image it will not help the model to learn from this specific image in case of similarity task.

Test Data set:

Contains 500 sample of ‘same’ pairs

Contains 500 sample of ‘different’ pairs

Train Data set:

Contains 1100 sample of ‘same’ pairs

Contains 1100 sample of ‘different’ pairs

The new training set and the validation set are generated by the split of the original training set to 80% and 20% respectively.

## Environment

Since none of the students has or has an access to GPU physically, the assignment was done in the Google Colab environment in order to speed up the training process. All of the shown results are taken from the Jupyter notebook.

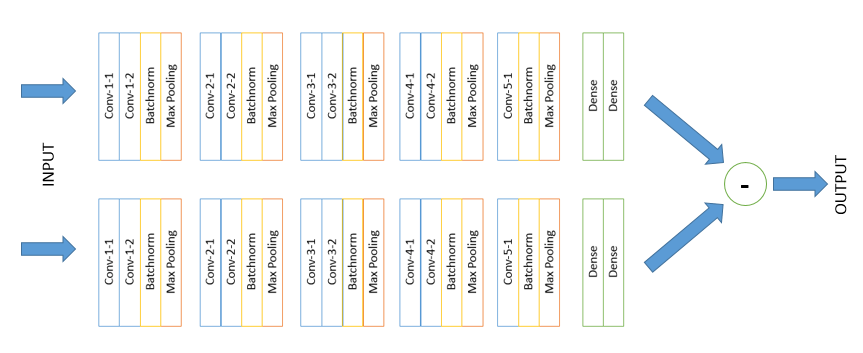
## Method

The general method is to build a Siamese Network to decide whether two given images are of the same or different persons.

The data is loaded with *load\_data\_2* function and *then prepare\_x\_y\_according\_to\_description* function is used to load, arrange and shuffle the data in a list of pairs of images attaching the label (1 for same person and 0 for different persons).

After implementing the architecture similar to the [given paper](https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf), the accuracy results were not stunning, reaching about 0.6. It was decided to change the CNN architecture, inspired by VGG knowing its ability in image classification task. Also the VGG is known for its ability to decrease the dimensionality which leads to a weights size reduction, and thus a quicker training.

The following architecture of the network is used for the classification task:



Every convolution layer is of size (3, 3), ReLU activation and stride of 1. Following additional parameters were used for convolution layers:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Number of filters | Padding | Initializer |
| Conv-1-1 | 32 | Same | Xavier normal |
| Conv-1-2 | 32 | Same | Xavier normal |
| Conv-2-1 | 64 | Valid | Xavier normal |
| Conv-2-2 | 64 | Valid | Xavier normal |
| Conv-3-1 | 128 | Valid | Xavier normal |
| Conv-3-2 | 128 | Valid | Xavier normal |
| Conv-4-1 | 256 | Valid | Normal distribution () |
| Conv-4-2 | 256 | Valid | Normal distribution () |
| Conv-5-1 | 512 | Valid | Normal distribution () |

Every max-pool layer has shape of (2, 2) and stride of 2.

Two dense layers have 1024 units each and their weights are initialized with the Normal distribution (). The first Dense layer is activated with ReLU and the last is activated with Sigmoid.

It might be seen that multiple changes were made to the oridinal VGG architecture, including the addition of batch normalization, changes in padding and initializers, addition of another convolution layer, change of last dense activation function. Also, many techniques were tried as described in the given paper, including but not limited by dropout, L2 layer-wise regularization, …. That did not brought a better result.

After the completion of the pass in twin-network by both images simultaneously, the resultant tensors are subtracted (neglecting the sign) and the result goes through the final sigmoid neuron, which classifies the similarity of the images.

In total, there are 16,512,673 trainable weights.

The fit was done with the following parameters: batch size = 8, validation split = 0.2, number of epochs = 20. Three callback functions were used:

* Early stopping for validation loss to interrupt training if there are no improvement for 3 epochs
* Learning rate scheduler, which decreases the learning rate each epoch by 5%
* Model checkpoint, to save the weights of model with highest validation accuracy

After finding the model with the best accuracy, it is loaded and retrained on the full dataset (including the validation set) for one epoch with the batch size of 8.

The loss that we are interested in is binary-crossentropy. The optimizer that was used is Adam with learning rate of 1e-4. The metric we are interested in is accuracy.

After the training and finding the best model parameters, the model is evaluated on a test set.

* 1. Explain the reasoning behind the choices made in answer to the previous section. If your choices were the result of trial and error, please state the fact and describe the changes made throughout your experiments. Choosing certain parameter combination because they appeared in a previously published paper is a perfectly valid reason.

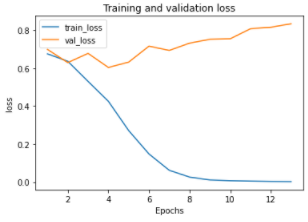
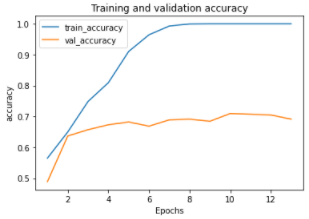
1. In addition to the details requested above, your report needs to include an analysis of your architecture’s performance. Please include the following information:
   1. Convergence times, final loss and accuracy on the test set and holdout set
   2. Graphs describing the loss on the training set throughout the training process
   3. Performance when experimenting with the various parameters
   4. Please include examples of accurate and misclassifications and try to determine why your model was not successful.

Please note the that report needs to reflect your decision-making process throughout the assignment. Please include all relevant information.

## Results

With the described architecture and method, the best epoch accuracy on a validation set is 0.71

Graphs that describe the accuracy and loss through the training process.

The training takes ~24 seconds per epoch. In total, we can see that the early stop worked after 13 epochs (313 seconds). Retraining for 3 more epochs takes another 87 seconds. After that the test set accuracy is also **0.71**.

Example of correctly classified “same” person:



Example of incorrectly classified “same” person:



The haircut, face exposure and face details quality are very different.

Example of correctly classified “different” person:



Example of incorrectly classified “different” person:

