HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

GRADUATION RESEARCH 1 REPORT

Supervisor: Do Phan Thuan
Academic year: 2020/2021
Student: Ta Dinh Son – 20176862

Date: July 2021

Contents

1. Introduction

During recent years, deep learning has become somewhat of a buzzword in the tech community. We always seem to hear about it in news regarding AI, and yet most people don't actually know what it is! In this article, I'll be demystifying the buzzword that is deep learning, and providing an intuition of how it works.

The most important field of Deep Learning, which is called as "Classification" with the role of classifying things. With classification, deep learning is able to establish correlations between, say, pixels in an image and the name of a person. You might call this a static prediction. By the same token, exposed to enough of the right data, deep learning is able to establish correlations between present events and future events

In this project, I used classification to do face recognize

2. Theory

2.1 Deep learning course in Coursera website

2.1.1 Neural Networks and Deep Learning

- Introduction to Deep Learning
- Neural Networks Basics
- Shallow Neural Networks
- Deep Neural Networks

2.1.2 Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization

- Practical Aspects of Deep Learning
- Optimization Algorithms
- Hyperparameter Tuning, Batch Normalization and programming Frameworks

2.1.3 Structuring Machine Learning Projects

Machine learning strategy:

2.1.4 Convolution Neural Networks

- Foundations of Convolution Neural Networks
- Deep Convolution Models: Case Studies
- Object Detection
- Special Applications: Face recognition and Neural Style Transfer

2.1.5 Sequence Model

- Recurrent Neural Networks
- Natural Language Processing and Word Embedding
- Sequence Model and Attention Mechanism
- Transformer Networks

2.2 Pytorch Course

- Tensor basics
- Dataset Transforms
- Training Pipeline: Model, Loss and Optimizer
- Transfer Learning
- RNN and LSTM and GRU

3. Project: Face recognition project

3.1 Introduction about face recognition

Facial recognition is a way of identifying or confirming an individual's identity using their face. Facial recognition systems can be used to identify people in photos, videos, or in real-time.

Recently face recognition is attracting much attention in the society of network multimedia information access. Areas such as network security, content indexing and retrieval, and video compression benefits from face recognition technology because "people" are the center of attention in a lot of video. Network access control via face recognition not only makes hackers virtually impossible to steal one's "password", but also increases the user-friendliness in human-computer interaction. Indexing and/or retrieving video data based on the appearances of particular persons will be useful for users such as news reporters, political scientists, and moviegoers. For the applications of videophone and teleconferencing, the assistance of face recognition also provides a more efficient coding scheme

3.2 Dataset

Input data from folders by name of famous people. Each folder contains each people's photo by name and the folder name will be automatically used as class needed to be classified in our model.

Figure 1: Structure of data folder

As illustrated on the figure, there are six classes, which present six famous people to be the labels for in our classification project

In train folder there are a total of more than 1800 photos of 6 celebrities.

That's equivalent to having 300 pictures for each person, and the total Number of photos in the valaidate folder is 360 photos, each person has 60 photos.

3.3 Preprocess

3.3.1 Face detection

- MTCNN stands for Multi-task Cascaded Convolutional Networks.
- Composing of 3 stacked CNNs and works at the same time when detecting face. Each network has a different structure and plays a different role in the task like Figure 2

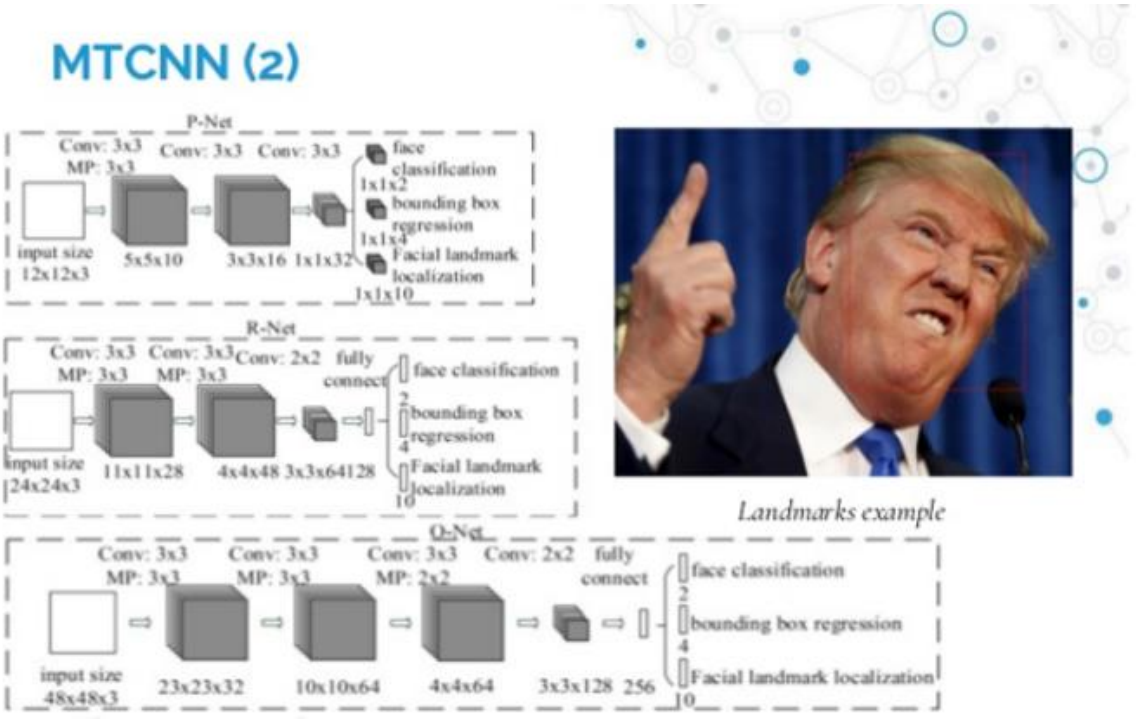
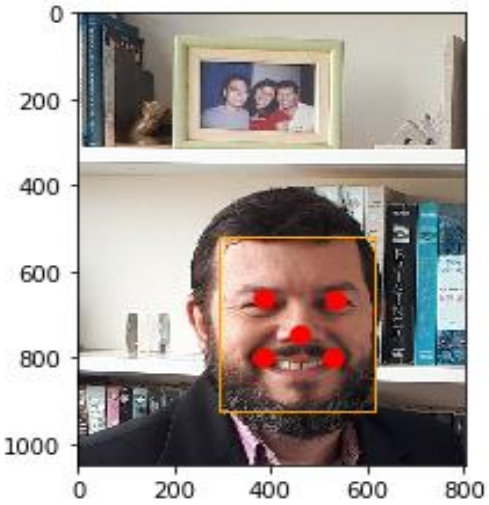


Figure 2: MTCNN architecture

- The output of MTCNN is the position of the face and points on the face such as eyes, nose, mouth like this:

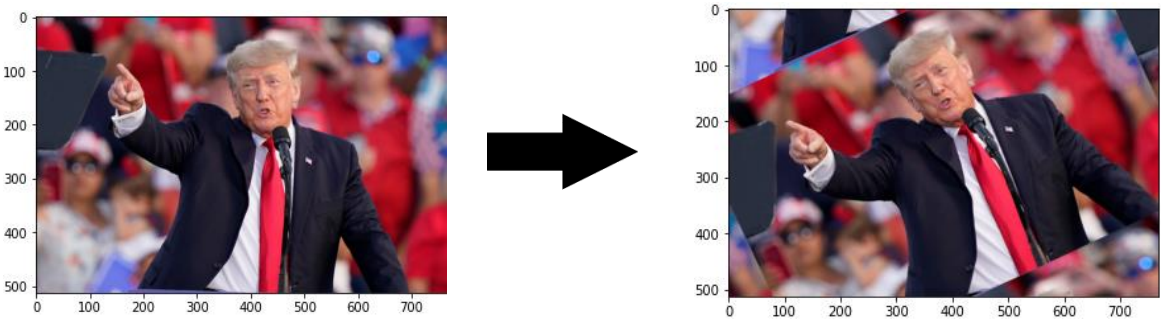


3.3.2 Augmentation

Data augmentation is the process of increasing the amount and diversity of data. Data augmentation is an integral process in deep learning, as in deep learning we need large amounts of data and in some cases it is not feasible to collect thousands or millions of images, so data augmentation comes to the rescue. In this project, I used transform form torchvision library to augment images.

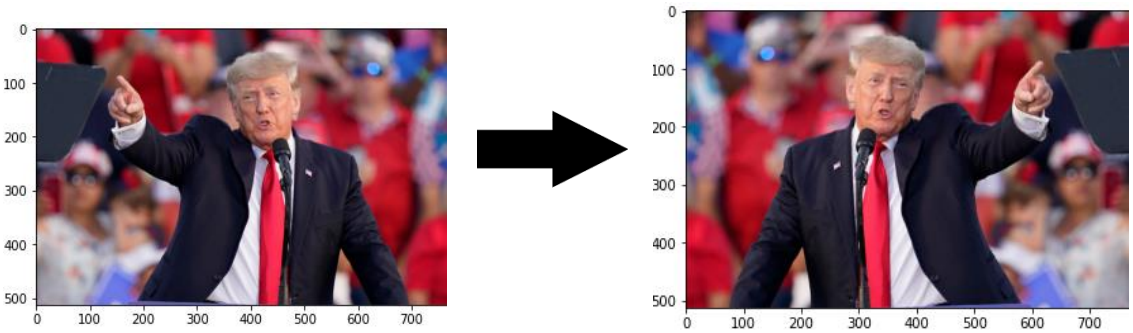
- *Rotating:*

Rotation operation as the name suggests, just rotates the image by a certain specified degree.



- *Flipping:*

Flipping allows us to flip the orientation of the image. We can use horizontal or vertical flip.



3.4 Model and training

3.4.1 VGG16 model

- *Introduction:*

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an annual computer vision competition. Each year, teams compete on two tasks. The first is to detect objects within an image coming from 200 classes, which is called object localization. The second is to classify images, each labeled with one of 1000 categories, which is called image classification. VGG 16 was proposed by Karen Simonyan and Andrew Zisserman of the Visual Geometry Group Lab of Oxford University in 2014 in the paper "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION". This model won the 1st and 2nd place on the above categories in 2014 ILSVRC challenge.

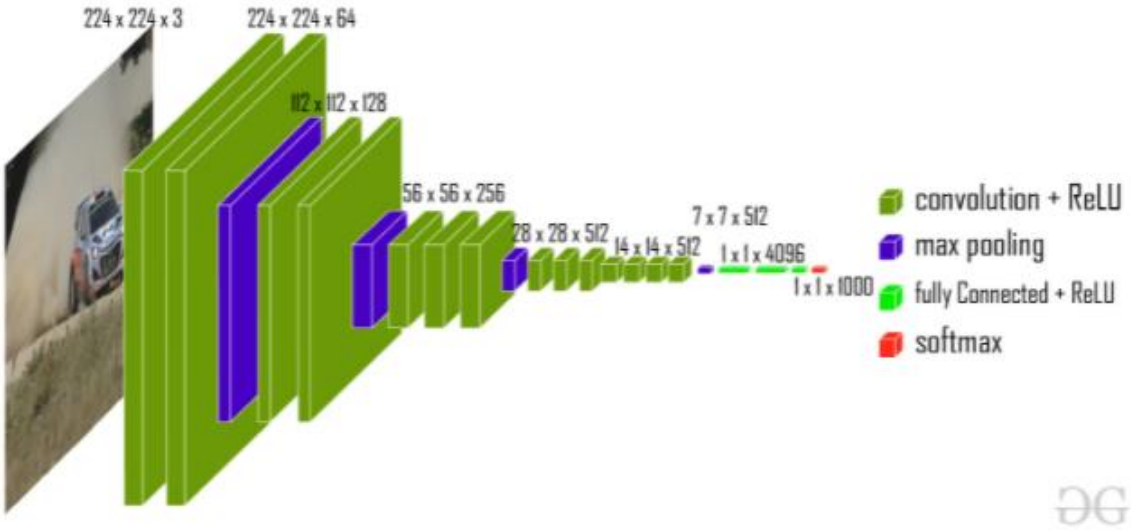


Figure 3: VGG-16 Architecture

This model achieves 92.7% top-5 test accuracy on ImageNet dataset which contains 14 million images belonging to 1000 classes.

- Architecture:

The input to the network is image of dimensions (224, 224, 3). The first two layers have 64 channels of 3*3 filter size and same padding. Then after a max pool layer of stride (2, 2), two layers which have convolution layers of 256 filter size and filter size (3, 3). This followed by a max pooling layer of stride (2, 2) which is same as previous layer. Then there are 2 convolution layers of filter size (3, 3) and 256 filter. After that there are 2 sets of 3 convolution layer and a max pool layer. Each have 512 filters of (3, 3) size with same padding. This image is then passed to the stack of two convolution layers. In these convolution and max pooling layers, the filters

we use is of the size 3*3 instead of 11*11 in AlexNet and 7*7 in ZF-Net. In some of the layers, it also uses 1*1 pixel which is used to manipulate the number of input channels. There is a padding of 1-pixel (same padding) done after each convolution layer to prevent the spatial feature of the image.

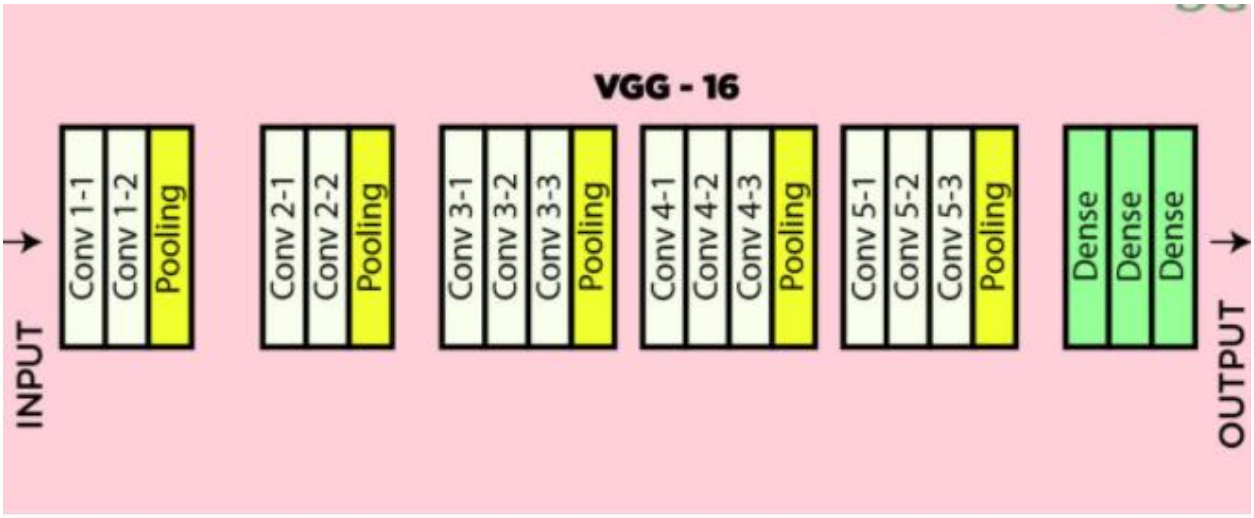


Figure 4: VGG-16 architecture map

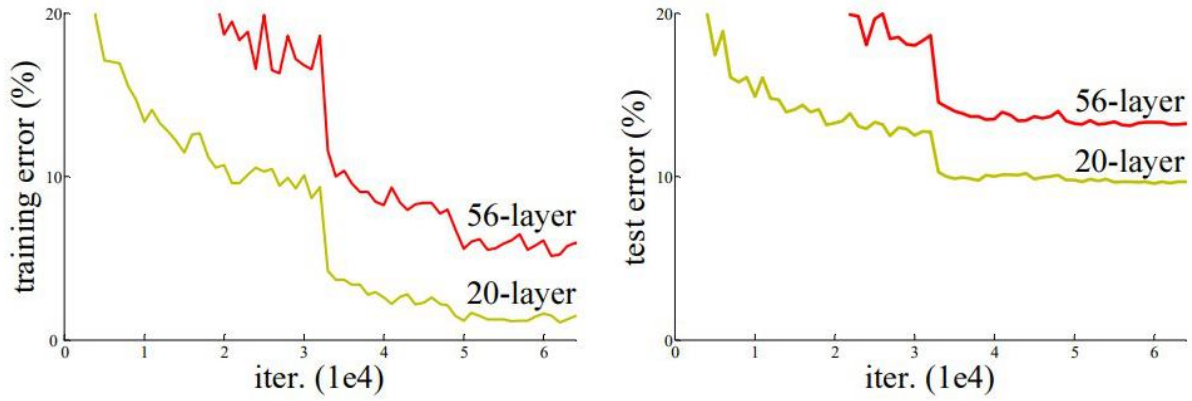
After the stack of convolution and max-pooling layer, we got a (7, 7, 512) feature map. We flatten this output to make it a (1, 25088) feature vector. After this there are 3 fully connected layer, the first layer takes input from the last feature vector and outputs a (1, 4096) vector, second layer also outputs a vector of size (1, 4096) but the third layer output 1000 channels for 1000 classes of ILSVRC challenge, then after the output of 3rd fully connected layer is passed to softmax layer in order to normalize the classification vector. After the output of classification vector top-5 categories for evaluation. All the hidden layers use ReLU as its activation function. ReLU is more computationally efficient because it results in faster learning and it also decreases the likelihood of vanishing gradient problem.

3.4.2 Model Resnet-18

- *Introduction:*

After the first CNN-based architecture (AlexNet) that win the ImageNet 2012 competition, every subsequent winning architecture uses more layers in a deep neural network to reduce the error rate. This works for less number of layers, but when we increase the number of layers, there is a common problem in deep

learning associated with that called Vanishing/Exploding gradient. This causes the gradient to become 0 or too large. Thus when we increases number of layers, the training and test error rate also increases.

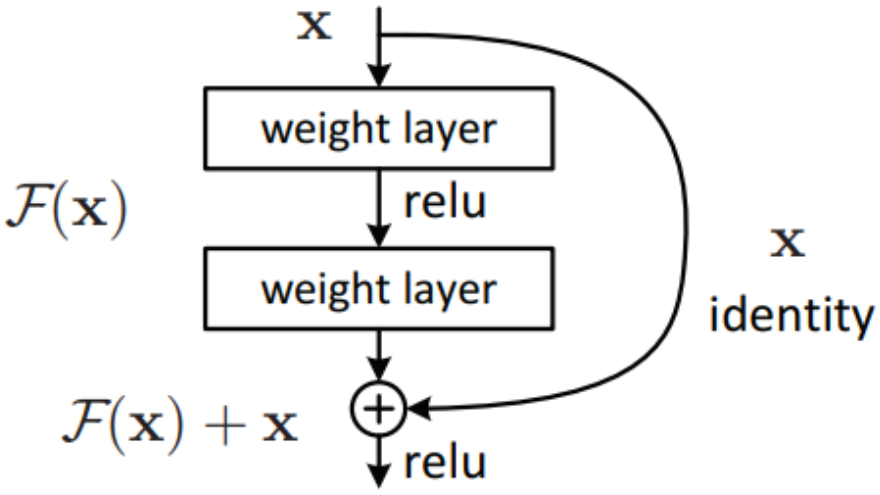


In the above plot, we can observe that a 56-layer CNN gives more error rate on both training and testing dataset than a 20-layer CNN architecture, if this was the result of over fitting, then we should have lower training error in 56-layer CNN but then it also has higher training error. After analyzing more on error rate the authors were able to reach conclusion that it is caused by vanishing/exploding gradient.

ResNet-18, which was proposed in 2015 by researchers at Microsoft Research introduced a new architecture called Residual Network.

- *Residual Block:*

In order to solve the problem of the vanishing/exploding gradient, this architecture introduced the concept called Residual Network. In this network we use a technique called skip connections. The skip connection skips training from a few layers and connects directly to the output. The approach behind this network is instead of layers learn the underlying mapping, we allow network fit the residual mapping. So, instead of say H(x), initial mapping, let the network fit, F(x) := H(x) – x which gives H(x) := F(x) + x.



The advantage of adding this type of skip connection is because if any layer hurt the performance of architecture then it will be skipped by regularization. So, this results in training very deep neural network without the problems caused by vanishing/exploding gradient. The authors of the paper experimented on 100-1000 layers on CIFAR-10 dataset.

3.5 Result

- *Training Datasets*: We train with images of 6 celebrities as introduced in part 3.2 Dataset

- *Model Resnet-18*: In Resnet-18 model, the model tries to learn image by image, which increase the score little by little. The initial score is not so high because the model's initialized parameters are not so good. This is reasonable because its initialized parameters are normally randomly generated. After some iterations, the model starts learning better and the score increase gradually

Our evaluation achieves an accuracy of 98% over 20 test images.

- *Model VGG-16*: In VGG-16 model, the model learns from the dataset gradually and it quickly achieves high accuracy score because of the simplicity of our dataset. The cross-validation score is a bit lower than the training score, which means that our model learns pretty well and does not over-fit the training dataset.

Our evaluation achieves an accuracy of 92.2% over 20 test images.

4. Conclusion

From the experimental results above, we can see that all two models learn pretty well on our dataset. Even though there are some little differences in validation accuracy of them, these distinctions are insignificant.

In order to obtain a result with clearer differences between more, we need to build a dataset with more images with variety of image-capturing condition. Furthermore, we need to adjust and choose some more appropriate parameters so that our models converge as expected in a time-and-memory-limited condition