TP mise en œuvre : Interprétation/visualisation des filtres et carte d'activation des classes (Grad-cam)

Djallel DILMI

I – Objectives

The aim of this practical work is to dispel the myth "CNN is a black box... it works but we don't know why" and to explore some techniques for visualizing filters to interpret: "what do CNNs really see?". This understanding could provide insights into questions raised by specialists in the application domain. The practical work is organized into several parts.

• The first part involves optimizing responses to filters in a convolutional neural network with an application in the field of facial recognition. We will study:

- The usefulness/impact of associating meaning with a filter during knowledge transfer.
- The importance/impact of the number of filters used in each layer.
- The importance/impact of the choice of learning rate during training.
- Identifying problems encountered during the learning phase.

• In the second part, we will test the validity of statements made in the field of facial recognition and photo anonymization. We will study:

- The technique of characterizing classes (Grad-CAM). How a CNN can help us characterize a class.
- The importance/impact of the choice of filter associated with Grad-CAM for proper characterization.

• In the third part, we will compare classification by k-nearest neighbors (kpp) to classification by a neural network. We will study:

- The importance/impact of choosing the classification algorithm based on constraints.

• In the fourth part, we will attempt to deploy the trained network in C++ or JavaScript. We will study:

- The importance/impact of the library used.
- The constraints and conditions of production (e.g., response time, platform portability).

It is worth noting that the goal of this practical work is discussion, analysis, and awareness, and technical solutions will be provided in case of difficulties.

II - Steps for the Practical session

For the implementation of this practical work, it is strongly recommended to use the standard Python libraries and avoid coding wherever possible. However, some libraries may have incompatibilities between different versions. To avoid wasting your time, create a virtual environment named "CVOUV" and install the correct versions of the following libraries:

- python==3.7.7
- jupyterlab
- tensorflow==2.1.0
- keras==2.3.1
- keras_vggface==0.6
- matplotlib
- pydot
- h5py==2.10.0
- python_graphviz 2.38(availible only conda)
- opencv-python

Face recognition

Facial recognition is a technique that utilizes facial features to:

- **Authenticate a person:** This involves verifying that a person is indeed who they claim to be, typically in the context of access control.

Or

- **Identify a person:** This entails locating a person within a group of individuals, in a particular location, image, or database.

According to the CNIL (French Data Protection Authority), in practice, facial recognition can be carried out using static images (photos) or dynamic images (video recordings) and involves two phases:

1. From the image, a feature vector or "template" representing the facial features is generated from a computer perspective. The data extracted to create this template qualifies as biometric data under the GDPR (General Data Protection Regulation) Article 4-14.
2. The recognition phase is then conducted by comparing these pre-generated feature vectors with models calculated in real-time for faces present in the candidate image.

In the last decade, in response to the widespread availability of face datasets (e.g., VGGFace2[1], etc.), models based on neural networks have emerged, such as VGG16[2] (2015), ResNet50 developed by Microsoft (2015[3], 2017), senet[4] (2017), and others.

[1] https://www.robots.ox.ac.uk/~vgg/data/vgg_face/
[2] https://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/parkhi15.pdf

For this session, we will take the example of the VGG16 network (2015), which is available in the Keras-VGGFace library. The first step would be to load the model, understand it, and familiarize ourselves with it.

III - First Part of the session (Interpretation of Filters):

1°) Explain the operating principle, nature, and different steps of the filter visualization algorithm through activation maximization. Then, if possible, identify the specific steps associated with this algorithm in the tutorial available at the following link: https://keras.io/examples/vision/visualizing_what_convnets_learn/.

2°) Consider the VGG16 neural network from the Keras-VGGFace library, trained for facial recognition. You are tasked with analyzing the network, describing it, and adapting the code from the first question to visualize the filters of different layers. Provide comments on the results, such as associating semantics with filters and assessing the quality of the training.

- Based on the visualization results, is the VGG16 network well-trained? Justify your answer.

IV - Second Part of the session (Grad-CAM and Occluding Parts of the Image):

1°) Now that you have a good understanding of the meaning of many filters in various layers, explain the operating principle, nature, and different steps of the Grad-CAM algorithm. Then, if possible, identify the specific steps associated with this algorithm in the tutorial available at the following link: https://keras.io/examples/vision/grad_cam/.

2°) For the VGG16 neural network from the Keras-VGGFace library, trained for facial recognition, adapt the Grad-CAM code for the class of Steven Soderbergh. Reflect on which layer(s) or filter(s) should be projected onto the person's photo.

- Is the interpretability of filters of any use in this context? Justify.
- The image "Steven-Soderbergh.jpg" is attached to the practical work.

3°) For companies using people's photos, the CNIL requires that photos used for clinical research purposes in the company be anonymized (respecting privacy). To meet this requirement, European and American (Western) pharmaceutical companies place a mask over the eyes. This practice is

[3] https://arxiv.org/abs/1512.03385
[4] https://arxiv.org/pdf/1709.01507.pdf

validated by the CNIL. In this part, examine this practice. i.e., does placing a mask over the eyes anonymize a person?

- Using the results from the first part on the visualization of filters, check whether the network focuses only on this region or not using Grad-CAM.
- Using the "occluding parts of the image" technique, analyze the evolution of scores (class probabilities).

Moreover, the article "Deep Learning for Face Recognition: Pride or Prejudiced?" (Shrutin[5] et al. 2019) claims that facial recognition is ethnicity-dependent, and CNIL recommendations are biased by the Caucasian population. This time, verify the accuracy of these claims (Figure 6 and 7 of the article).

Hint: Run Grad-CAM with the right filters from the right layers on two different populations.
Given the obtained results, what do you suggest for eye anonymization? Minimize the need for a mask as much as possible.

- To create your two groups, use the VGGFace2 database available at the following link: https://www.robots.ox.ac.uk/~vgg/data/vgg_face/

[5] https://arxiv.org/pdf/1904.01219.pdf