Machine Learning

Practical work 11 - Understanding Deep Neural Networks

Teachers: A. Perez-Uribe (Email: andres.perez-uribe@heig-vd.ch) & J. Hennebert
Assistants: Benoît Hohl (Email: benoit.hohl@heig-vd.ch) and Yasaman Izadmehr (Email: yasaman.izadmehr@heig-vd.ch)

Summary for the organization:

- Submit a report before Monday 6.12.22 11h00 via Moodle.
- Modality: PDF report (max. 8 pages)
- The file name must contain the number of the practical work, followed by the names of the team members by alphabetical order, for example 11_dupont_muller_smith.pdf.
- Put also the name of the team members in the body report.
- Only one submission per team.

0. Notebooks and libraries

Download the notebook material from the Moodle platform or run those in the servers.

1. Understanding Convolutional neural networks by using filter activation statistics

The objective of this exercise is to train a deep convolutional neural network capable of determining the features (e.g., the convolutional filters) that allow it to properly recognize the digits 0 to 9.

The first part of the notebook CNN_filter_visualization.ipynb corresponds to the training of the CNN. We also show the learning evolution and its performance.

The second part of the notebook shows the activation of the filters at each convolutional layer and the activation of the dense layers.

We then compute the average output of every filter of each layer for a given input class (e.g., select all the input images corresponding to the class ‘zero’, compute the output of

the filters for each of them and then average those outputs) and visualize that average output value.
Finally, we present an algorithm to find the index of the filter with the highest activation, and then compute the mode (i.e., majority) through the whole dataset, with the aim of visualizing which filter is activated most of the time for each pixel of the input images, and for each input class (e.g., for every digit 0 to 9).

Run the notebook in order to solve the digit recognition task using a CNN and observe the visualization of the filter activation statistics. If there are no clear patterns that appeared to be exploited by the CNN to classify the digits, run the training of the CNN once again and check. Hopefully, you will get something as nice as the example we provided in the slides of the course.

2. Activation maximization as a means for understanding a CNN model

The idea behind activation maximization is simple: generate an input image that maximizes the output activations of a given unit in the network. It can be an output of the network (e.g., a class) or any output of a convolution or a hidden unit of the dense part of the network. The `keras-vis` package computes the derivative of the ActivationMaximization loss with respect to the input, and uses this gradient to update the input image. In this way, we can compute an image that shows us what is being detected by a given filter. For example, if we find that a filter is activated by a vertical line, that can be a feature that allows the network to detect ones, sevens, fours and maybe nines.

Follow the instructions of the notebook “CNN_activation_maximization” and perform the experiments indicated at the end in the blue box.

3. Class Activation Maps

Class activation maps or grad-CAM is another way of visualizing attention over input. The intuition is to use the outputs of the last convolution layer to exploit the spatial information that gets completely lost in the subsequent dense layers. To compute the CAM’s, we replace the final fully-connected layers by a Global Average Pooling (GAP) layer that computes the mean of the activations of the filters of a given layer. We fine-tune the resulting network and generate “heat-maps” that indicate the parts of the image that appear to be more relevant for the network that is attempting to classify the given image.

Follow the instructions of the notebook “CNN_CAM_keras_vis” and perform the experiments indicated at the end.

Summary of work to include in the report

- Present the results obtained in point 1 and comment those results. E.g., describe the patterns that are shared by different digits and that are being used by the CNN to solve the recognition task. Can you infer what do the activation maps of the L2 and L3 layers represent ?

- Present the results of the experiments proposed at the end of the notebook corresponding to “activity maximization” and answer the question concerning the use of such a technique.

- Present the results of the experiments proposed at the end of the notebook corresponding to “Class Activation Maps” and answer the question concerning the use of such a technique.