Capstone Memo

Quality/Quantity tradeoff

Status:

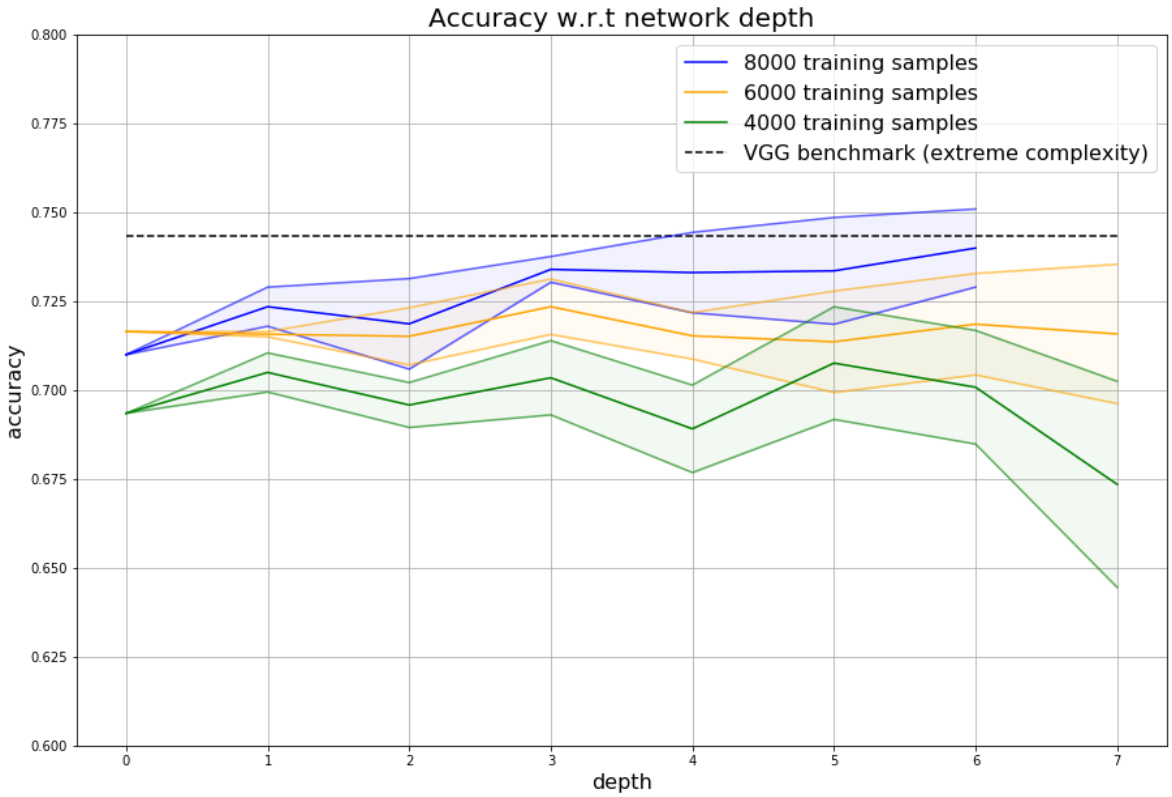
The project is progressing well, we are not facing major issues.

What we did:

As a reminder, the main objective is to provide recommendations to better approach the following problem: given limited resources to label data, what is the optimal strategy to allocate these ressources ? What is the best trade-off between poorly labeling important amounts of data and labeling with high precision smaller amounts of data ?

We want to provide a visual interpretation of this trade-off and hence plan on plotting iso-error curves, with axis being data quality on one hand and data quantity on the other hand. But, as we explained during the presentation, each point on this plot would correspond to one experiment, which actually corresponds to 1) training models with the corresponding dataset size and with the corresponding noise in the labels, 2) finding the best model under this setting and 3) reporting the error found. This technically means that for each point, we would need to test various architectures with different complexities to find the actual best model. This task of plotting the iso-error being already experiments heavy, our first attempt was to prove that the same architecture with some chosen complexity could be used for all the different data sizes and label noise rates tested.

We conducted this first step of experiments and obtained curves tracking the performance (accuracy) of the models do not vary too much with their complexity. We used CIFAR-10 dataset and trained simple CNNs with various depth. We also added an horizontal line showing the performance of a VGG architecture (far greater complexity). Here are the curves that we obtained:



Thanks to this graph, we can confirm that accuracies remain close for different network depths. Indeed, there is not a general trend showing that increasing the depth would improve the accuracy. Thus, for the following experiments, we decided to train only one model for each quality and quantity values.

What we are currently working on:

Now that this first step is done, we are currently working on producing the iso-error plot for different quality and quantities of the dataset, keeping the same model architecture and complexity for all experiments. For different dataset sizes, and different error rate values (introduced by modifying the label of a percentage of the data), we are evaluating the performance of a convolutional neural network for the classification of these datasets. This will then enable us to plot iso-error curves with respect to the quality and the quantity of the dataset used. As it not possible to perform an infinite number of experiments, we are doing a fixed number of experiments and then we will use interpolation techniques in order to create our iso-error plots.

Then, we are planning to perform these experiments on different datasets (2 other classes from CIFAR-10, MNIST, …) so that we could try to find more consolidated results.