Assignment: Deep Learning

13 de mayo, 2020

Data sets: NIH ChestXray14

The dataset we are going to use consists of 350 normal chest x-ray and 350 effusion x-ray, taken and selected from the public NIH ChestXray14 dataset:

https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-ch

Image pre-processing

Images are 512x512x3, but are in grayscale and the 3 chanels have the same values. Thus, we will keep only the first chanel. In addition, we have to reshape the images to 64x64x1.

Questions:

1. Normalize images.
2. Split the dataset into 500 train /100 validation /100 test. Try to balance the two classes.
3. Implement a Convolutional Neural Network (CNN) following the instructions below:
    - The number of convolutional layers should not be greater than 6.
    - Pooling layers should be included to reduce the number of parameters.
    - At the bottom of the network, the fully connected layers will have 128 and 32 nodes respectively.
    - Output layer with activation 'sigmoid'.
    - Trainable params should be at least than 60000.
4. Tune the hyperparameter batch_size checking the values in the set {25,35,50}
5. Assess the performance of the CNN predicting the categories of test images and obtain the confusion matrix.
6. Re-fit the CNN including data augmentation. Was the use of augmentation an improvement?
7. Compare these two CNN models.

8. Implement a convolutional autoencoder (CAE) network following the instructions below:
    - The full network should have two networks: a convolutional encode and a convolutional decode.
    - The total number of convolutional layers should not be greater than 8.
    - Pooling layers should be included to reduce the number of parameters.
    - Trainable params should be at least than 30000.
9. Tune the more compact layer (`z` layer) with three configurations(width x heigth x filters) what you can free choose. To evaluate the `z` layer performance use this flattened layer as input in a random forest(or boosting) algorithm to classify the images.
10. Once it was selected the best performing `z` layer configuration, to detect in which variables (nodes) there are significant differences between the two classes of images, a statistical test will be performed to determine those that are significant.
11. Visualize the results of the previous item using Volcano plot (see https://en.wikipedia.org/wiki/Volcano_plot_(statistics)).
12. Discusion and conclusion about CNN and CAE.

Important remarks

- You should do an R markdown (or R latex) report as dynamic as you can. Try to use r code into paragraph.
- Use relative paths instead of absolute paths to read / write files, to make it easier to run the code outside of your computer.

Delivery / Deadline

A zip file including the data set, the Rmd (or Rsw) file used as template for the report and output reports in pdf and html files.

Deadline: May 29th, 2020

Score of each question

- Questions 1, 2, 3, 4, 5: 30%
- Question 6, 7: 15%
- Questions 8, 9: 30%
- Questions 10, 11: 15%
- Questions 12: 5%
- Dynamic report quality: 5%