**Maman 12**

## **Q1**

### **Research/Training phase**

* I’ve split the training data into train/validation
* After extracting SIFT descriptors (128) I’ve tried reducing dimensionality for computing relief using PCA.
  + Total explained variance by 50 components: 0.86
  + Total explained variance by 80 components: 0.95
  + Total explained variance by 90 components: 0.97
* While this looks like a promising path to reduce computation, I’ve dropped it.
* Calculating SIFT features took some time. I’ve used the fact that cv2 doesn’t use pythons GIL and reading images is IO bound operation, to run in asyncio fashion to reduce run-time.
* I’ve trained kmeans over 20% of the training descriptors, with k=100 and n\_init = 4.
  + Increase # of descriptors to (maybe) get better means
  + Increase n\_init to (maybe) get better means
* After that, I’ve used my kmeans over the training data to produce a list of descriptor: cluster.
* Then I used histogram over all the descriptors clusters and normalized it.
* Why normalization is important here? Because some images might have 100 descriptors and other 1000. Normalization will get them to the same level. I’ve chosen l1 normalization here as I find it simply explaining the % a “feature” contributes. While there are many ways to normalize the data, I’ve decided not to dwell on it.
* Anyway, After this process, I’ve got / learnt procedures to transform a file\_path into a feature (1, 100) where sum(feature) = 1.
* I’ve used scikit learn encoder label to learn the labels and their transformations.
* Using all of the above, I now have train/validation features and labels.
* I’ve used xgboost classifier as its simply better than SVM both in results and computing speed.
* Xgboost hyper parameters were initially chosen by me from my own experience with the algorithm.
* The classifier (xgboost) outputs the roc auc (ovr) and mlogloss(multi) for both validation and training set during the training. This helps track for over/under fitting. Since xgboost builds many trees, I use an early stop parameter, to stop at the best iteration (to avoid overfitting on the training). The early stop uses the auc over the validation set to know when to stop.
* Eventually the training function yield clf, files\_to\_features, le
* Clf – the classifier (xgboost trained) – boosted trees (max depth = 6)
* Files\_to\_features (a delayed function, that uses the sift, semaphore and kmeans we’ve learnt/defined, and waiting for paths to execute)
* Le – label encoder, where we transform labels from strings to integers.

# **Testing phase**

At the beginning, we’ve split the data into Training and Testing sets. Later on we took this Training set and divided it into Training set and Validation set, and did all the work above.

Which means the testing set doesn’t know ANYTHING about what happened thus far.

* All data was split using stratification over the label.
* We’ve used the files\_to\_features function to get features from the testing set
* We used labelencoder to encode testing labels
* we’ve then used the classifier over the testing data to get probabilities.
* We’ve ran ROC AUC One VS Rest and averaged it.
* Both macro and micro averages were roughly the same with 0.938~ ROC AUC which is f$cking awesome!
* **Those results surprised me as I was 100% sure the descriptors will not represent the images well enough. And additional features like colour histogram and histogram of gradients will be needed.**
* I even thought about increasing my feature space using the 2 ideas above. But once I saw the results are very high, I’ve let it go.
* I also didn’t check the ROC curve / AUC score for each label individually because I see no merit in it.
* While it is possible one (or more) class performance is weak, we’re not going to do anything about it, so I simply don’t care.

## Optimal classifier hyperparameters:

1. While I choose very good hyper parameters, there are so many in xgboost algorithm. There is only one way to approximate the best hyper parameters and its through search.
2. This is np hard problem which called hyper parameter tuning.
3. To get such approximation one need to define a search space and use algorithms like grid search random grid search baysian hyper parameter optimization etc.
4. Just like any other machine learning algorithm, this is an approximation of the “real best values”
5. You can see in code the hyper parameters I’ve chose, including the objective function, loss metrics I’ve chosen to look at, boosting algorithm (trees of course!!!), etc.

### Q2