Exercice de classification

# Contexte

Il s’agit d’un problème de classification d’images de visages. Le critère de classification n’est pas décrit ici, il est défini de façon implicite par les labels de la base d’apprentissage. Néanmoins, je vous précise qu’il s’agit d’une combinaison de caractéristiques « softbio » (barbe, moustache ou non, forme du menton, couleur des yeux, chapeau ou pas, … )

En plus d’une bonne performance de classification, on cherche à équilibrer les performances entre femmes et hommes. Pour cela, pour chaque image de la base d’apprentissage, on fournit les labels de genre.

La métrique finale, retenue pour évaluer chaque modèle, sera la moyenne des performances de classification entre hommes et femmes, à laquelle on soustrait le double de la différence des performances entre hommes et femmes (en valeur absolue). Pour obtenir une bonne métrique finale, il faut donc obtenir une bonne performance de classification sur la population totale (hommes + femmes) et, en même temps, avoir des performances de classification similaires entre hommes et femmes.

# Data

Les données sont les suivantes :

* Une base d’apprentissage : un dossier « train » contenant 90.000 images de visage, numérotées de 000000 à 89999. Les fichiers sont au format jpg, les images font 80x80 et sont en couleur.
* Les labels d’apprentissage : un fichier nommé « train.txt », contenant trois colonnes. La première liste les noms des 90000 images de la base d’apprentissage. La deuxième colonne liste les labels à prédire (-1 ou 1) de la base d’apprentissage, tandis que la troisième liste les genres binaires (-1 : femme, 1 : homme) correspondants.
* Une base de test : un dossier « test » contenant 9.999 images de visage, au même format que la base d’apprentissage.

Le rendu devra être organisé comme le fichier « train.txt », sans la colonne du label de genre. Ce sera donc un fichier texte à deux colonnes : la première indiquant le nom de l’image de la base de test, la deuxième indiquant la prédiction du label à prédire.

# Organisation

* Le travail s’effectue en binôme. Si une personne reste seule, elle me contacte et on trouvera un trinôme.
* La labélisation n’est pas parfaite. Une fonction de coût robuste pourrait aider lors de l’apprentissage.
* Il est demandé aux étudiants de fournir les prédictions du jeu de test pour l’évaluation, sous forme de fichier texte, nommé « predictions\_nom1\_nom2.txt », d’exactement 9999 lignes en respectant l’ordre des images de la liste « test.txt ».
* Il est demandé aux étudiants de fournir le modèle appris, le code python/tensorflow /pytorch/keras ayant servi à apprendre le modèle ainsi qu’un code python permettant de générer « predictions\_nom1\_nom2.txt » à partir du modèle.
* Le rendu définitif est attendu pour le 4 décembre à 8h44
* Il est demandé aux étudiants de fournir un document répondant synthétiquement aux questions suivantes :

1. Quel type d’architecture avez-vous utilisé, pourquoi ?
2. Combien de convolutions contient votre architecture, pourquoi ?
3. Comment vous êtes-vous préoccupé de l’over-fitting ?
4. Est-ce que l’asymétrie de le base d’apprentissage est un problème ? si oui, comment avez traité ?
5. Est-ce que le fait que les labels soient bruités a posé problème ? Comment l’avez-vous géré ?
6. Pour les images de la base de test, à partir de la sortie de votre réseau, comment attribuez-vous le label ?
7. Comment avec vous pris en compte les problèmes d’équité de l’algorithme ?
8. J’ai précisé que le label correspondait à une combinaison de caractéristiques « softbio » et non image. Qu’est ce que cela change pour vous ?
9. Avez-vous tenté d’utiliser un modèle pré-appris ? Que vous l’ayez fait ou non, comment procéder, choisir le modèle, qu’apprendre… ?
10. Avec un mois supplémentaire pour travailler, que feriez-vous pour améliorer vos résultats ?
11. Question non notée : A quoi correspond le label recherché ?