Instructions pour la compétition Kaggle 2024

IFT6390

November 15, 2024

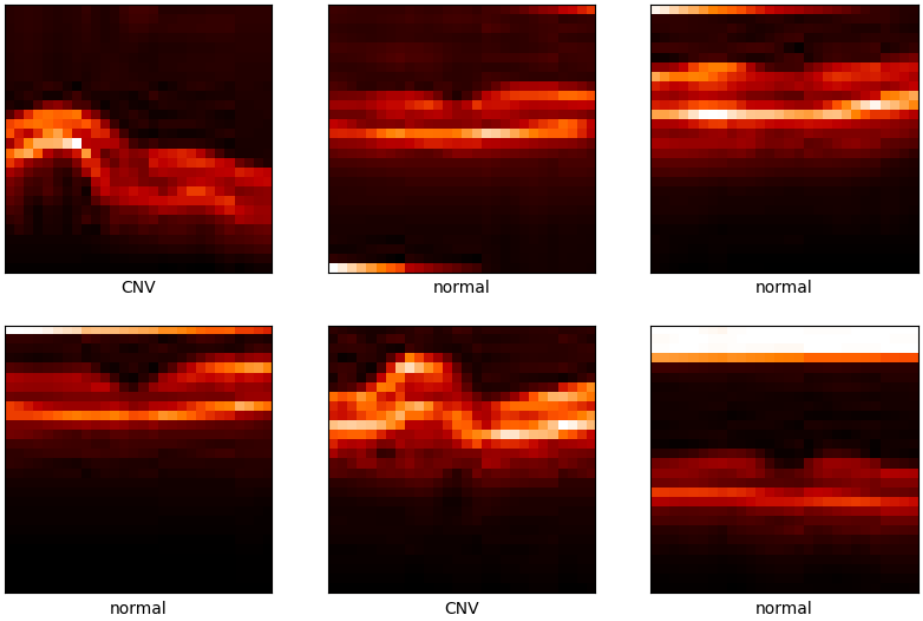

Figure 1: Images de l'ensemble d'entraînement

1 Description

Dans ce projet, vous participerez à une compétition Kaggle de classification d'images. L'objectif est de concevoir un algorithme d'apprentissage automatique capable d'identifier avec précision les maladies rétiniennes à partir d'une image de Tomographie par Cohérence Optique (OCT), un test d'imagerie pour obtenir des images en coupe transversale de la rétine. Afin de garantir l'anonymat du jeu de données et d'assurer une compétition équitable, nous ne divulguons pas le nom du jeu de données et appliquons des transformations de base sur les images

(miroirs, rotations). Votre objectif est de classer précisément les images selon les catégories suivantes :

- 0 : néovascularisation choroïdienne
- 1 : œdème maculaire diabétique
- 2 : drusen
- 3 : rétine saine

de manière à minimiser l'erreur de classification sur l'ensemble de test : $\sum_{i=1}^{n} \mathbb{1}_{\hat{y}_i \neq y_i}$
En résumé, les données fournies sont :

- `train_data.pkl` - Un fichier *pickle* contenant les images et les étiquettes de l'ensemble d'entraînement. Vous pouvez l'ouvrir en utilisant le script suivant en Python :

```
import pickle

# Charger le fichier pickle
with open('path_to_file.pkl', 'rb') as f:
    data = pickle.load(f)

# Acceder aux images et labels
images = data['images']
labels = data['labels']
```

  La variable "data" est ainsi un dictionnaire avec les clés "images" (une liste d'images **28x28** représentées sous forme de tableaux numpy) et "labels" (une liste d'étiquettes allant de 0 à 3).

- `test_data.pkl` - Un autre fichier pickle contenant les images de l'ensemble de test. Ce fichier peut être ouvert de la même manière, mais le dictionnaire ne contiendra pas la clé "labels" que vous devrez prédire.

2 Participation

Pour la section des étudiants de maîtrise (IFT6390), la tâche doit être réalisée **en équipes de 2.** Les étudiants de IFT3395 peuvent participer en **équipes de 3**. Pour participer à la compétition, vous devez :

- Créer un compte Kaggle si vous n'en avez pas déjà un.
- Rejoindre la compétition via le lien d'invitation suivant : `https://www.kaggle.com/t/1c523e26262b439fae68924234255054`.
- Dès lors, vous pouvez accéder à la compétition via `https://www.kaggle.com/competitions/ift3395-ift6390-identification-maladies-retine/`.

- Dans la section "Invite Others", entrez les noms de vos coéquipiers ou le nom de votre équipe.
- Votre coéquipier peut accepter la fusion de l'équipe.
- Remplissez le formulaire google `https://forms.gle/jYEHBWBsG7B6WQhC7` avec les informations de votre équipe avant le **22 novembre, 23:59**. Les équipes non inscrites ou inscrites en retard ne seront pas notées.

Note importante : Le nombre maximal de soumissions est de 2 par jour, par ÉQUIPE. Toute équipe dont les membres individuels dépassent le nombre de soumissions autorisé jusqu'à la date sera INCAPABLE de former une équipe.

Exemple : Aujourd'hui est le premier jour de la compétition. A, B et C sont trois coéquipiers qui n'ont pas encore formé d'équipe.

- A a soumis 0 fois.
- B a soumis 2 fois.
- C a soumis 1 fois.

Comme le nombre maximum de soumissions est de 2 par équipe et par jour, le total des soumissions possibles pour une équipe est de 2. Cependant, le nombre cumulé de soumissions pour A, B et C est de 3. Par conséquent, ils ne pourront pas former une équipe (Ils devront attendre demain et ne soumettre aucune soumission le jour suivant).

3 Premier jalon : Battre le score de référence (25 novembre)

Vous pouvez voir deux scores de référence sur le tableau de classement. Le premier score correspond à un classifieur attribuant des étiquettes aléatoires à chaque image. Le second score de référence correspond à un classificateur de régression logistique. Le classifieur de régression logistique a été entraîné en utilisant seulement 10% de l'ensemble d'entraînement. Pour ce premier jalon, vous devrez battre la régression logistique sur le tableau de classement public.

Voici quelques méthodes possibles :

- Classificateurs non linéaires tels que le SVM à noyau
- Caractéristiques faites à la main et régression logistique
- Arbres de décision et forêts aléatoires

Note importante : Pour battre le score de référence, vous n'êtes PAS autorisé à utiliser de bibliothèque d'apprentissage automatique, comme `scikit-learn`. Vous devez implémenter votre solution à partir de zéro en utilisant uniquement NumPy et les fonctionnalités de base de Python.

L'objectif est de concevoir la méthode la plus performante mesurée par la soumission des prédictions sur Kaggle. Votre performance finale sur Kaggle comptera pour l'évaluation, ainsi que le nombre de lignes de base battues.

4 Deuxième jalon : Compétition (3 décembre)

Vous avez jusqu'au **3 décembre à 23:59** pour obtenir les meilleures performances possibles. Dans cette phase, vous êtes libre d'implémenter toute méthode et d'utiliser toute bibliothèque, comme scikit-learn, Pytorch ou Tensorflow. Le tableau de classement de Kaggle comporte une composante publique et une composante privée pour éviter le "surapprentissage" sur le tableau de classement. Le tableau de classement public montre votre score calculé sur 50% de l'ensemble de test, tandis que le tableau de classement privé est basé sur votre score sur l'autre moitié. Vous ne pourrez voir que le tableau public pendant la compétition. Les points de cette phase seront attribués en fonction de votre classement sur le tableau privé.

Note importante : Vous devez soumettre deux solutions distinctes, une pour la première phase (battre le score de référence) et une pour la deuxième phase (votre modèle le plus performant). Vous devez nommer vos fichiers de soumission pour distinguer les deux.

5 Troisième jalon : Soumettre le code et le rapport (6 décembre)

Vous devez rédiger un rapport détaillant votre pipeline d'apprentissage, y compris le prétraitement, les algorithmes, l'optimisation et l'apprentissage, le réglage des hyperparamètres et la validation. Le rapport doit inclure :

- Titre du projet
- Nom de votre équipe sur Kaggle et liste des membres, noms complets et numéros d'étudiant.
- Introduction : description du problème et résumé de votre approche.
- Conception des caractéristiques : description et justification de vos méthodes de prétraitement.
- Algorithmes : aperçu des algorithmes d'apprentissage utilisés.
- Méthodologie : répartition entraînement/validation, régularisation, optimisation, etc.
- Résultats : analyse détaillée avec comparaisons de différentes valeurs d'hyperparamètres.
- Discussion : avantages/inconvénients de votre approche, idées d'amélioration.
- Références (obligatoire pour les idées empruntées).
- Annexe (optionnelle).

Le texte principal du rapport ne doit pas dépasser 6 pages. Les références et l'annexe peuvent être ajoutées.

Instructions de soumission

- Code séparé pour les jalons 1 et 2.
- Prédictions de test à soumettre sur Kaggle.
- Rapport en pdf à soumettre sur Gradescope.

6 Critères d'évaluation

1. Points minimum si vous battez le score de référence.
2. Qualité et solidité technique du rapport.
3. Points **bonus** selon votre classement final.

7 Dates limites

Les dates limites pour ce projet sont fermes, et chaque soumission doit inclure tous les éléments requis pour chaque jalon. Tout retard peut entraîner une perte de points.

- La date limite pour former les équipes et remplir le formulaire Google Form est le **22 Novembre, 23:59**
- La date limite pour battre le score de référence est le **25 novembre, 23:59**.
- La compétition Kaggle se termine le **3 décembre, 23:59**.
- Vous devez télécharger votre rapport et code sur Gradescope avant le **6 décembre, 23:59**.