Projet Apprentissage Profond

CHEGGAF SAYAD Ahmed
CONTET Clément
DAI Guohao
DANTON LALOY Calliopé
DE ROCKER Tom

Département Sciences du Numérique
Deuxième année - Sciences et Ingénierie du Logiciel
2021-2022

Sujet

Le choix de l'affiche d'un film est un enjeux majeur dans sa commercialisation. Il est important de transmettre correctement le genre dans lequel le film se situe pour qu'il rencontre son public. Dans un problème proche, on sait que les miniatures utilisées par Netflix pour promouvoir ses films varient en fonction du profil de l'utilisateur et de ses préférences supposées. Lors de ce projet, notre objectif sera ainsi de chercher à déterminer les différents genres d'un film en fonction d'une de ses affiches.

Dans un premier temps, considérons ainsi les classes suivantes, correspondant aux genres possibles :

1. History
2. Animation
3. Family
4. Biography
5. Fantasy
6. Sci-Fi
7. Mystery
8. Horror
9. Romance
10. Crime
11. Adventure
12. Thriller
13. Action
14. Comedy
15. Drama

Ainsi, nous souhaiterions que notre système puisse attribuer à une affiche de film la ou les bonnes classes. Par exemple, nous voudrions que l'affiche du film Blade Runner (1982) soit classifiée "Action", "Drame", et "Science Fiction".

FIGURE 1 – Affiche du film Blade Runner de 1982

Base de données

Lien vers les fichiers sources et la base de données : https://github.com/KalyDL/ProjetAP

Composition de la base de données

Script d'extraction des informations

Nous avons réalisé un script qui :

1. récupère le nom, les genres et le lien vers l'affiche du film sur le site https://www.imdb.com et stocke ces informations dans un fichier csv
2. ne tient pas compte des films n'ayant pas d'affiche en triant par un critère d'extension
3. parcourt le fichier csv généré précedemment et télécharge pour chaque ligne l'image associée, en stockant dans son nom un id (correspondant à l'indice de sa ligne dans le fichiers csv) et l'ensemble de ses genres.
4. enlève les affiches ayant des labels n'ayant que peu de représentation (par exemple moins de 1/1000 de la base de donnée)

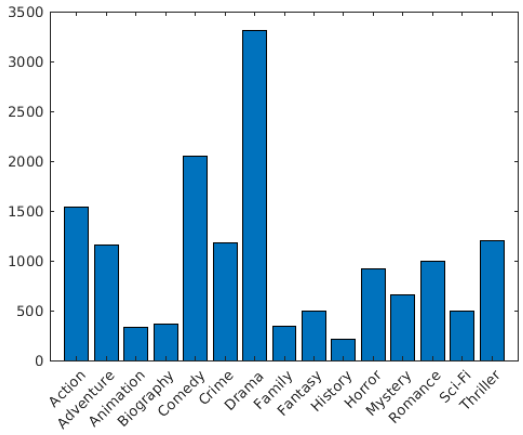
Partitionnement des images

On dispose d'une base de données d'un peu moins de 10 000 images. On la sépare de la manière suivante :

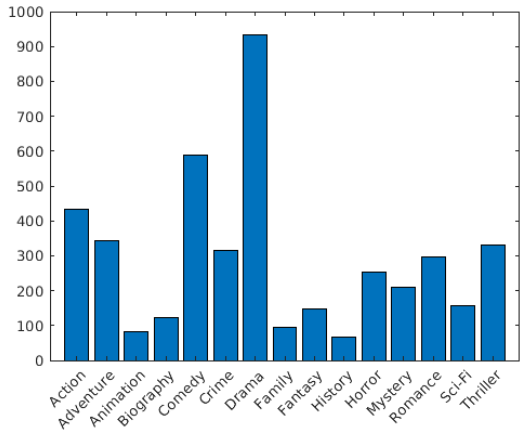
1. ensemble d'entraînement : 70%
2. ensemble de validation : 20%
3. ensemble de test : 10%

L'idéal serait de répartir dans nos trois ensembles les différentes classes de manière plutôt homogène. Cependant, cela n'est pas évident dans notre cas car chaque image peut possèder plusieurs classes différentes. De plus

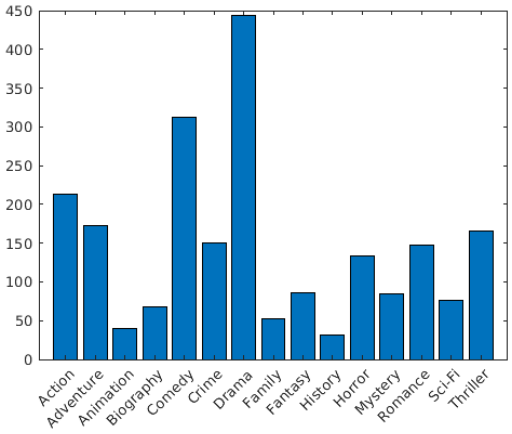
Pour s'assurer d'avoir une répartition la plus uniforme possible même sur les classes avec peu de représentants nous avons choisi de disperser aléatoirement les images associer à un genre en particulier en prenant les genres par ordre croissant de taille.



(a) *entraînement*



(b) *validation*



(c) *test*

FIGURE 2 – Répartitions des genres pour chaque ensemble de données

Script de chargement des données

Pour réaliser notre script de chargement, nous nous sommes largement inspiré du script fourni. La différence majeure réside dans le fait que comme nos images peuvent avoir jusqu'à 3 labels, les dimensions de y sont de $nbImage \times 3$ au lieu de $nbImage \times 1$.

Exemple de données

Au total, on dispose pour chaque label du nombre d'image suivant :

1. Documentary (10)
2. Film-Noir (40)
3. Short (58)
4. Musical(139)
5. Western (154)
6. Sport (202)
7. War (203)
8. Music (275)

9. History (356)
10. Animation (472)
11. Family (534)
12. Biography (679)
13. Fantasy (772)
14. Sci-Fi (776)
15. Mystery (994)
16. Horror (1344)
17. Romance (1649)
18. Crime (1732)
19. Adventure (1750)
20. Thriller (1778)
21. Action (2346)
22. Comedy (3296)
23. Drama (5419)

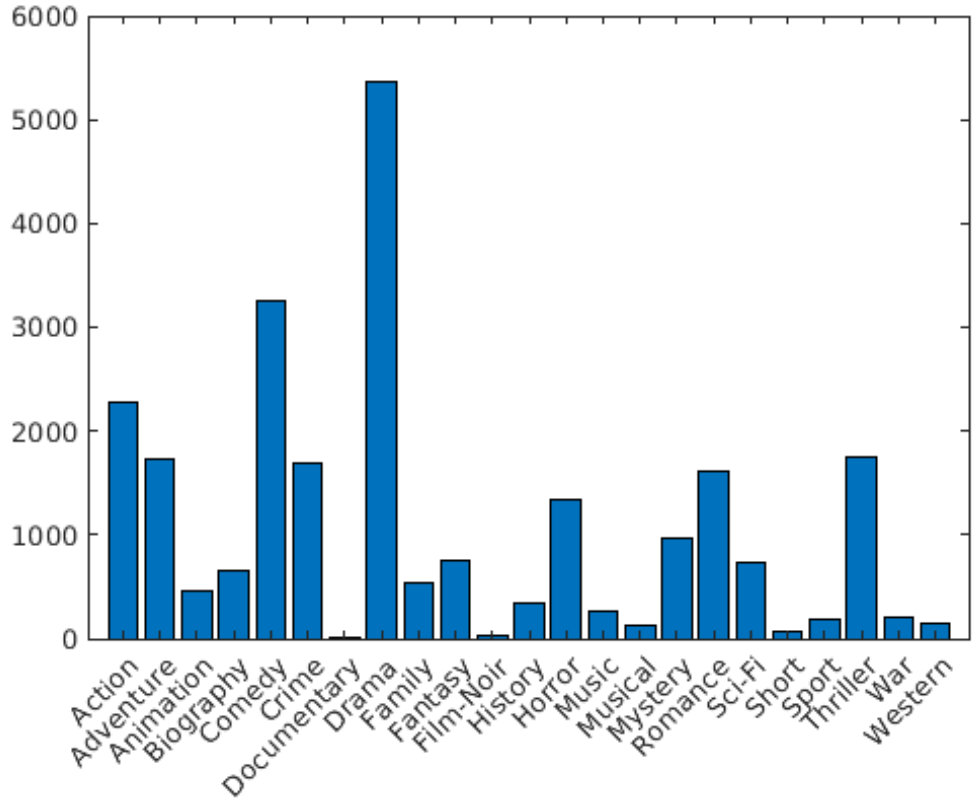


FIGURE 3 – Répartition des films par genre

La répartition des films selon leurs genres est loins d'être uniforme et certaines classes sont très peu représentées, parfois moins de 1% (cf histogramme). Nous avons donc fait le choix de retirer les films associés à ces classes trop petites (les 8 premières classes ont été retirées : Documentary

(10), Film-Noir (40), Short (58), Musical(139), Western (154), Sport (202), War (203) et Music (275)).

On peut visionner à quoi peut resemble notre base de données sur la capture d'écran suivante :

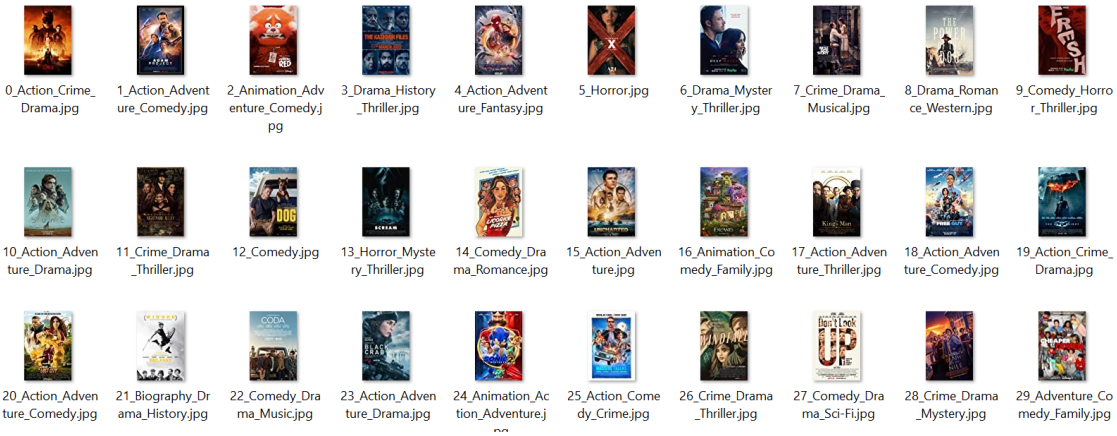


FIGURE 4 – Exemple de fichiers de notre base de données

Pronostic

Le problème semble plutôt compliqué, car il comporte un nombre de classes assez important, et que chaque image peut posséder jusqu'à 3 classes différentes. Cependant, nous pensons qu'il y a une certaine corrélation entre l'affiche d'un film et ses genres, et qu'il est possible d'obtenir des résultats assez satisfaisant.