Point de mi-parcours : Toutes les comédies françaises sont-elles bleues ?

"Warner Bros."
Maxime Poli Nicolas Schlegel Alex Fauduet **Virginie Loison**

30 mars 2020

1 Position de la problématique

Le choix de l'affiche d'un film est considéré comme un élément déterminant du marketing pour la sortie d'un film. Il faut qu'en un coup d'oeil à l'affiche, tout individu puisse deviner le genre du film associé.

Le but de ce projet est de construire un algorithme qui devine les genres principaux d'un film à partir de son affiche. Par exemple, l'algorithme doit associer les genres Comédie et Policier au film Ocean's 8. Il s'agit donc d'un problème de classification multilabel.

Nous construisons le projet à partir de zéro, et ne reprenons pas de projet déjà existant. Néanmoins, nous verrons plus bas que nous ne sommes pas les premiers à nous confronter à ce problème.

Ce projet représente un double intérêt pour le client.
Premièrement, l'équipe Data Science de Warner Bros. travaille actuellement sur un algorithme qui prédit la catégorie de téléspectateurs auprès de qui un film va marcher. Le genre d'un film est une donnée cruciale pour cette prédiction. A l'heure actuelle, Warner Bros. récupère la base de données d'Allociné pour connaître le genre d'un film. Mais cette base de données est imparfaite. Notre algorithme permettrait de vérifier facilement les informations qui sont dans la base de données d'Allociné.
Deuxièmement, cet algorithme pourra être utilisé par le département marketing au moment de la conception d'une affiche pour un film. S'il y a plusieurs affiches candidates, on pourra les passer en argument de l'algorithme que nous aurons créé. On pourra alors vérifier que l'affiche véhicule bien les bonnes informations sur les genres du film.

2 Cadre fixé pour le sujet d'étude

Rappelons tout d'abord que nous traitons ici un problème de classification multi-label.

Warner Bros. nous a mis à disposition la base de données d'Allociné. C'est sur cette base de données que nous entraînons et testons nos modèles. Dans cette base de donnée, un film peut avoir au maximum trois genres. Les genres sont classés par ordre décroissant d'importance. Nous traitons donc ici un problème de classification multi-label, avec trois classes maximum.

En accord avec le client, nous avons choisi une méthodologie ascendante. Nous avons commencé par coder des algorithmes "simples" (kNN, CNN). Nous n'attendions pas de bons résultats de la part de ces algorithmes. Mais, comme l'a souligné Lise Regnier, ils permettront d'avoir un point de comparaison lorsque nous coderons des algorithmes plus élaborés (modèles de Deep Learning haut niveau). Nous augmentons progressivement la complexité de nos algorithmes.

Nous nous sommes pour l'instant placés dans le cadre d'une classification mono-label. Cela veut dire que nous ne considérons que le genre le plus important de chaque film. Cela nous permet de prendre en main la base de donnée, et de commencer à coder des algorithmes avec mons de difficulté. Nous migrerons ensuite vers le multi-label.

Nous n'avons pas d'objectif quantitafif précis. Au vu de l'état de l'art, tout algorithme ayant une accuracy supérieure à 20% serait cohérent. Nous ne chercherons pas seulement à avoir une bonne accuracy. Lise Regnier, notre contact chez Warner, accorde aussi de l'importance à l'interprétabilité du résultat. Il faut que le programme soit clair, et qu'on puisse connaître les facteurs qui l'ont poussé à faire son choix.

3 Verrous scientifiques, points bloquants

Comme c'est souvent le cas en Data Science, un premier problème rencontré concerne la base de données qui nous a été fournie. Voici une liste non exhaustive des imperfections de la base de donnée:

- Certains films de la base de données n'ont pas encore d'affiche officielle. Il s'agit majoritairement de films qui ne sont pas encore sortis. L'image qui leur est associée dans la base de données ne doit donc pas être prise en compte dans notre analyse
- Certains genres (qui sont pour nous des labels) sont redondants, ou peu intéressants pour notre analyse. Il y a par exemple dans la base de donnéesle genre "Divers" (inexploitable pour nous). Et on y retrouve les genres "Animation" et "Dessin animé".

- La base de données n'est pas uniforme, c'est-à-dire que certains genres sont beaucoup plus représentés que d'autres. Il s'agit là du plus gros problème pour nous. Par exemple, sur les 37 genres référencés, 3 genres (Comédie, Drame, et Comédie Dramatique) sont les genres principaux de 50% des films. On sait qu'une répartition non uniforme des labels n'est pas optimale pour avoir des algorithmes de classification performants.

Nous expliquerons dans l'annexe ce que nous avons commencé à implémenter pour contourner ces problèmes.

D'autre part, nous avons rapidement été confrontés à un frein technologique. Nos algorithmes sont rapidement devenus trop gourmands en calculs pour tourner en temps raisonnable sur nos ordinateurs personnels. Pour pallier à cela, Lise Regnier nous a mis à disposition une machine virtuelle.

4 État de l'art

D'autres personnes se sont déjà penchées sur la classification de genres de films. Il existe de la bibliographie sur la classification de genres sur les posters, sur les bandes-annonces, ainsi que sur les synopsis de films. Citons quelques sources qui nous ont particulièrement inspirés.

Certains articles utilisent des méthodes purement statistiques sur des features "bas niveau" des posters. C'est par exemple le cas de la référence [3]. Cet article utilise des histogrammes de couleurs sur des zones délimitées des posters. Plusieurs approches sont testées, qui sont des combinations d'approches Bayésiennes et d'algorithmes k-NN (k-Nearest Neighbors). Les méthodes bayésiennes aboutissent à une accuracy de 0.8, alors que les méthodes k-NN aboutissent à une accuracy de 0.2. Cetta accuracy de 0.8 semble très élevée par rapport au reste de l'état de l'art, et est donc à manier avec précaution. Cet article développe aussi beaucoup l'importance du preprocessing des données, et montre son impact sur la performance des algorithmes.

On peut aussi trouver à ce sujet des algorithmes dits "force brute". Le Github [2] en offre un exemple d'implémentation. Les posters sont traités "tels quels" par un modèle de deep learning, CNN. Il revendique une accuracy de 0.65. Mais la manière dont il calcule son accuracy est peu orthodoxe, c'est pourquoi l'accuracy réelle est probablement plus basse.

On peut enfin trouver des idées d'algorithmes de deep learning qui combinent des features "bas niveau" et des features "haut niveau". Citons par exemple l'article [1]. Cet article utilise en parallèle un CNN (voir paragraphe précédent) et un modèle de reconnaissance d'object : YOLO. Il revendique une accuracy de 0.2.

5 Organisation des prochaines séances

5.1 D'un point de vue humain

5.1.1 Au sein de l'équipe étudiante

Nous faisons des points hebdomadaires avec toute l'équipe par visioconférence. A chaque point hebdomadaire, nous définissons les objectifs d'avancement pour la semaine, et répartissons les tâches entre nous quatre. Nous faisons en sorte que chacun puisse avancer de la manière la plus autonome possible au cours de la semaine. Cette répartition est ensuite notée sur le fichier TODO.md de notre github. Voici par exemple les tâches de chacun pour cette semaine :

- Maxime : installer la nouvelle machine virtuelle. Importer et pré-traiter la nouvelle base de données que Warner Bros. nous a envoyée.
- Alex : implémenter et tester un réseau CNN pour notre cas. Se documenter sur une bonne accuracy à utiliser.
- Nicolas : chercher (sur Github) et tester un réseau YOLO pour notre cas.
- Virginie : créer des ensembles d'entraînement équirépartis en genre, monolabels et multilabels. Se documenter sur les notions de bootstrapping et de bagging.

5.1.2 Au niveau des relations avec les clients/tuteurs

Notre contact chez Warner Bros., Lise Regnier, est disponible. Nous faisons un point en visioconférence tous les lundi avec elle. Nous lui expliquons à chaque point notre avancement, et elle nous donne des conseils et des guidelines. Lise Regnier prend aussi le temps de contextualiser notre travail dans le milieu de la production cinématographique, ce qui nous donne une meilleure compréhension de ce qui est attendu. Elle a également à coeur de mettre à disposition pour nous les ressources nécessaires au bon déroulement du projet, comme une base de données complète ou une machine virtuelle avec une puissance de calcul adaptée.

5.2 D'un point de vue technique

En plus des rendez-vous hebdomadaires avec Lise Regnier, nous faisons un point entre nous une fois par semaine.

Nous utilisont githbub pour notre projet. A la demande de Warner Bros, nous avons mis le git en privé. Mais si le département le souhaite, il peut nous faire parvenir un nom d'utilisateur github afin que nous lui accordions l'accès.

Si le département souhaite avoir des nouvelles régulières du projet, il peut le demander à Virginie (virginie.loison@eleves.enpc.fr) en précisant la fréquence des points attendue.

5.3 D'un point de vue des objectifs

Notre projet continue d'avancer malgré l'impossibilité de nous retrouver physiquement. Nous avons bon espoir de pouvoir fournir des résultats probants.

A Annexe : détails de la méthode déjà implémentée

A.1 Preprocessing de la base de donnée

Comme mentionné dans la partie 3, la base de données que nous avons n'est pas optimale. Voici les traitements que nous lui avons appliqués pour corriger cela.

- Nous avons supprimé de la base de données les films qui ne sont pas encore sortis. Cela permet de supprimer les films qui n'ont pas encore d'affiche officielle.
- Nous avons fusionné certains genres redondants. Nous avons par exemple regroupé les genres "Animation" et "Dessin animé".
- Nous avons supprimé les films avec des genres inexploitables. Nous avons par exemple supprimé le genre "Divers".

Ces premières manipulations nous ont permis de passer de 33 à 17 genres. Les genres sont maintenant répartis de manière un peu plus homogène.

Pour pallier au manque d'uniformité de la base de données, nous avons créé des ensemble d'entraînement et de test de nos algorithmes qui sont le plus homogènes possibles.

A.2 Méthode K Nearest Neighbors

Le premier algorithme que nous avons implémenté est une classification k-NN. Lorsqu'on souhaite attribuer un genre à un poster, cette méthode cherche dans l'ensemble d'entraînement les posters qui en sont le plus proches, ses voisins. Elle détermine ensuite le genre prédominant parmi les voisins du poster. C'est ce genre qui lui est assigné.

Nous avons cohoisi de calculer la distance entre 2 posters en faisant la somme des différences des pixels entre les posters.

Sur le problème monolabel (où on ne considère que le genre principal de chaque film), nous obtenons une accuracy de 15%. Cette accuracy est assez basse.

Comme attendu, cette première méthode n'est pas très performante. Elle nous a néanmoins permis de mettre la main à la pâte, et de nous familiariser avec la base de données. Une amélioration possible de cette méthode serait de calculer la distance entre les posters d'une manière différente, en calculant des histogrammes de couleurs par exemple. Mais nous n'avons pas approfondi cette méthode pour l'instant. Nous nous sommes concentrés sur d'autres algorithmes dont la bibliographie est plus prometteuse.

A.3 Convolutional Neural Network

Nous avons ensuite implémenté une première méthode de deep-learning : un CNN. Il s'agit d'une méthode classique pour le traitement di'amages : on applique tour à tour des convolutions et des alinéarité (les fonctions d'activation, des "relu" ici) puis une fonction affine (couche dite dense) pour peu à peu transformer l'image en un vecteur associant une probabilité à chaque genre.

Pour le moment, c'est une approche très "force brute" parce qu'on ne recherche pas à extraire une quelconque information de l'image avant de la fournir au réseau : on lui donne l'image telle quelle, et toute l'extraction et le traiement se fait par les couches du réseau.

On essaiera par la suite de fournir au réseau plus précise et qui contient plus de sens, par exemple la présencce de personnages ou d'objets, et de combiner ces deux approches pour affiner nos résultats.

A.4 La méthode Yolo

Pour la suite de notre travail, nous avons prévu d'utiliser un algorithme capable de reconnaître des éléments dans une image. Nous avons opté pour le méthode YOLO v3, qui, selon la bibliographie, donne les résultats les plus probants.

Nous avons décidé d'utiliser un algorithme déjà entraîné et de l'adapter à notre projet. Nous avons choisi [4] car il est populaire sur Github et sous licence du MIT. Pour trouver les poids du réseau pré-entraîné, nous avons suivi les suggestions du créateur du Git. Il est entraîné à reconnaître 80 concepts tels que les personnes, les cravates, les couteaux, les voitures, les bouteilles ... [5]

Voici quelques exemples de résultats:

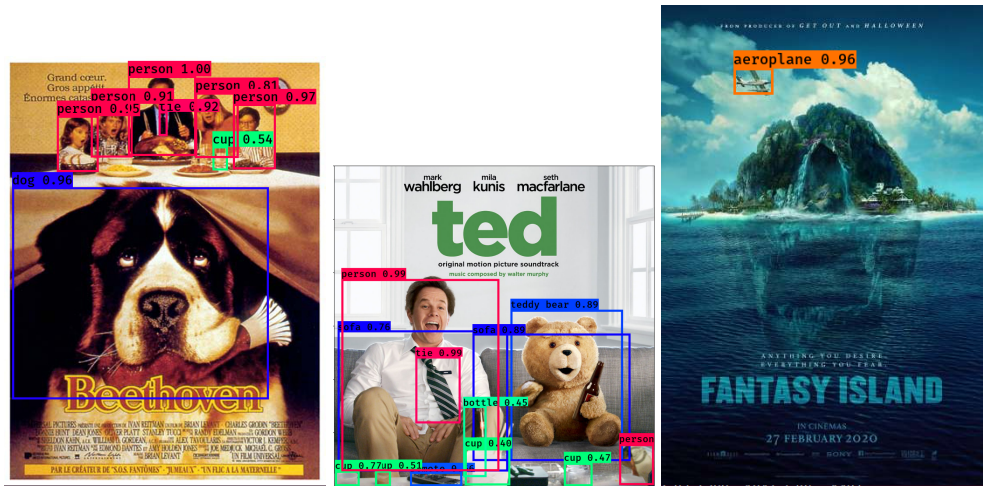

Figure 1: Résultat de la méthode YoloV3

Aujourd'hui, nous avons incorporée la méthode dans notre projet et elle est utilisable. Nos interrogations se portent maintenant sur la structure que nous devons donner aux résultats que nous donne l'algorithme YOLOv3 pour qu'ils puissent être ensuite intégrés à un réseau de neurones.

References

[1] Wei-Ta Chu and Hung-Jui Guo. Movie Genre Classification based on Poster Images with Deep Neural Networks. In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes - MUSA2 '17*, pages 39–45, Mountain View, California, USA, 2017. ACM Press. Meeting Name: the Workshop Reporter: Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes - MUSA2 '17.

[2] davideiacobs. davideiacobs/-Movie-Genres-Classification-from-their-Poster-Image-using-CNNs, February 2020. original-date: 2019-02-24T11:40:40Z.

[3] Marina Ivašić-Kos, Miran Pobar, and Ivo Ipsic. Automatic Movie Posters Classification into Genres. *Advances in Intelligent Systems and Computing*, 311:319–328, January 2015. Reporter: Advances in Intelligent Systems and Computing.

[4] qqwweee. qqwweee/keras-yolo3, March 2020. original-date: 2018-04-03T03:36:21Z.

[5] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.