

Rakuten Institute of Technology



DataScientest

Participants :

Olga TOLSTOLUTSKA

Mohamed BACHKAT

Charly LAGRESLE

Mentor : Manu POTREL

Promotion: DST Bootcamp DEC22

Rakuten

- Site de e-commerce avec 1.3 milliards d'utilisateurs
- Suggestions de recherche et recommandations pour l'utilisateur
- Classification des produits nécessaire
- Manuellement impossible

Objectifs

Prédire la catégorie d'un produit sur la base de son **titre**, sa **description** et de son **image**

1 + 1 = 3 ... Un **modèle de texte**, un **modèle d'image** et un **modèle de fusion**

The screenshot shows the Rakuten search interface for the query "Console". The left sidebar includes a "CATEGORIES" section with "Maison" expanded, showing sub-categories like "Mobilier" (+10 000), "Luminaires" (454), "Accessoires de rangement" (345), and "Plus de 1000". It also has sections for "FILTRES", "ETAT" (Neuf, Occasion, Reconditionné), "PRIX" (Min, Max, Ok button), and "OPTION D'EXPÉDITION" (Livraison gratuite, Livraison rapide, Expédié par Rakuten). The main area displays two sponsored results: "Pied De Table Diall - 30 X H. 350 Mm Noir - Diall" at 5,50 € Neuf and "Cockpit Bravo Throttle Quadrant Honeycomb Aeronautica Noir" at 267,48 € Neuf.

Présentation des données

- **84 916 observations**
- **27 catégories à déterminer**
- **0 donnée dupliquée**

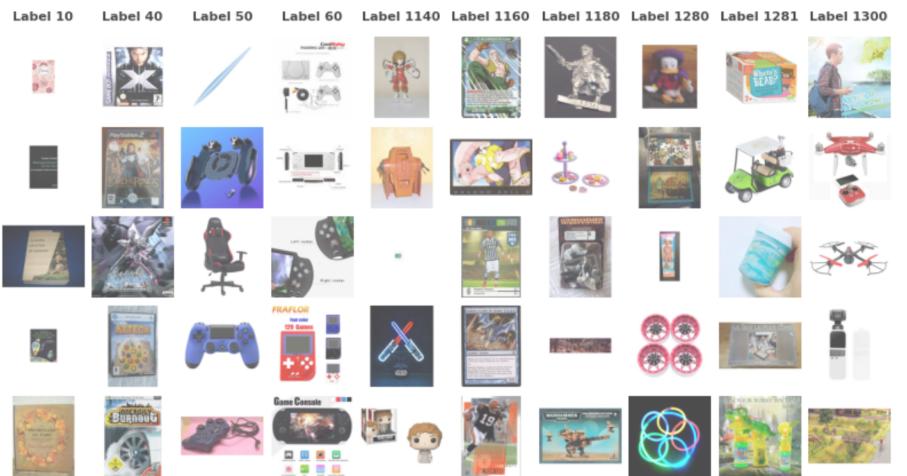
Textes

- Un produit est désigné par : `designation` et `description` soit un titre et sa description
- 35% de NaNs pour `description`

Images

- Une image couleur par produit
- Peut comporter un support ou une mise en scène
- Taille `500x500px` en JPG

	designation	description	productid	imageid
0	Olivia: Personalisiertes Notizbuch / 150 Seite...	NaN	3804725264	1263597046
1	Journal Des Arts (Le) N° 133 Du 28/09/2001 - L...	NaN	436067568	1008141237
2	Grand Stylet Ergonomique Bleu Gamepad	PILOT STYLE Touch Pen de marque Speedlink est ...	201115110	938777978
3	Peluche Donald - Europe - Disneyland 2000	(Mar...	50418756	457047496
4	La Guerre Des Tuques	Luc a des idées de grandeur. Il veut or...	278535884	1077757786



Données déséquilibrées

- 27 catégories (codes fournis)
- 7 domaines différents trouvés (non labelisés, non utilisés)
- Sur-représentation de la classe 2583
- Sous-représentation des classes 60 , 1320 et 2220

Challenge probable

→ Les modèles auront probablement (comme nous) du mal à distinguer les catégories de produits appartenant même domaine

Numéro de catégorie	Description
10	Livre occasion
40	Jeu vidéo, accessoire tech.
50	Accessoire Console
60	Console de jeu
1140	Figurine
1160	Carte Collection
1180	Jeu Plateau
1280	Jouet enfant, déguisement
1281	Jeu de société
1300	Jouet tech
1301	Paire de chaussettes
1302	Jeu extérieur, vêtement
1320	Autour du bébé
1560	Mobilier intérieur
1920	Chambre
1940	Cuisine
2060	Décoration intérieure
2220	Animal
2280	Revues et journaux
2403	Magazines, livres et BDs
2462	Jeu occasion
2522	Bureautique et papeterie
2582	Mobilier extérieur
2583	Autour de la piscine
2585	Bricolage
2705	Livre neuf
2905	Jeu PC

Nombres de mots

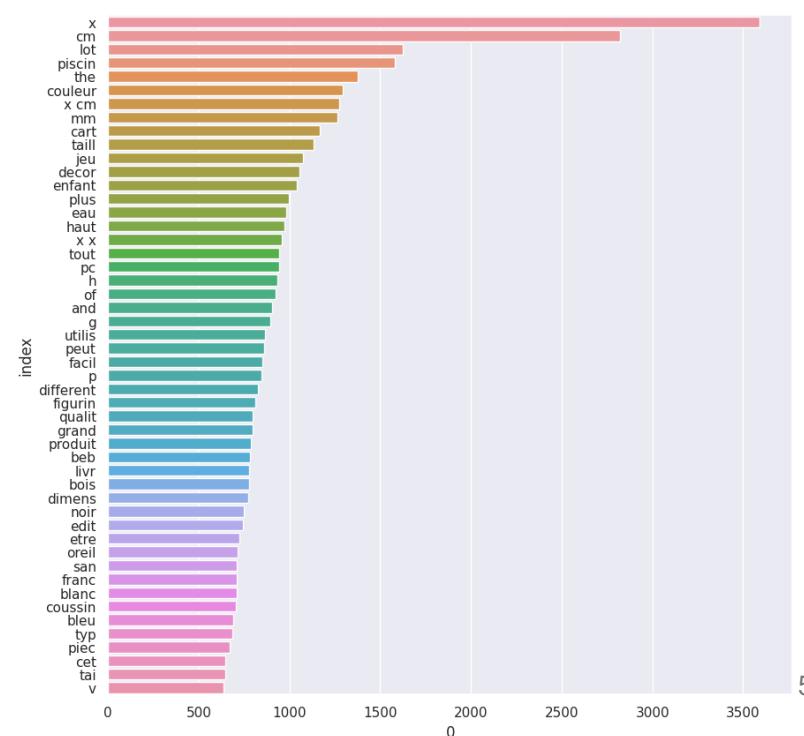
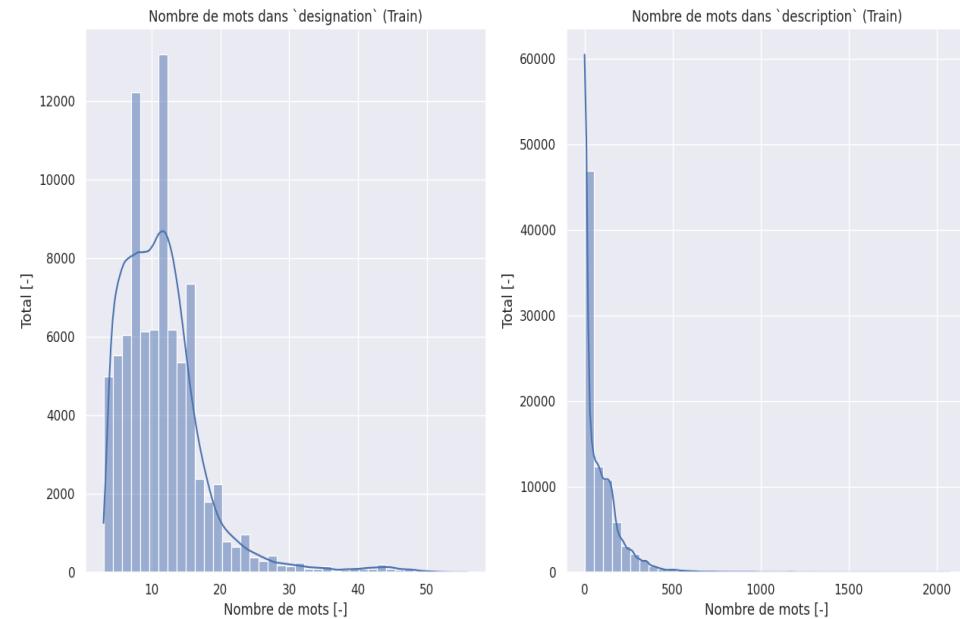
- Variable (rappel : 35% de `descriptcion` ne compte pas de mots)
- Limite à 500 mots

Langues

- Détection de la langue pour traduction à effectuer
 - 81% français
 - 14% anglais et autres langues

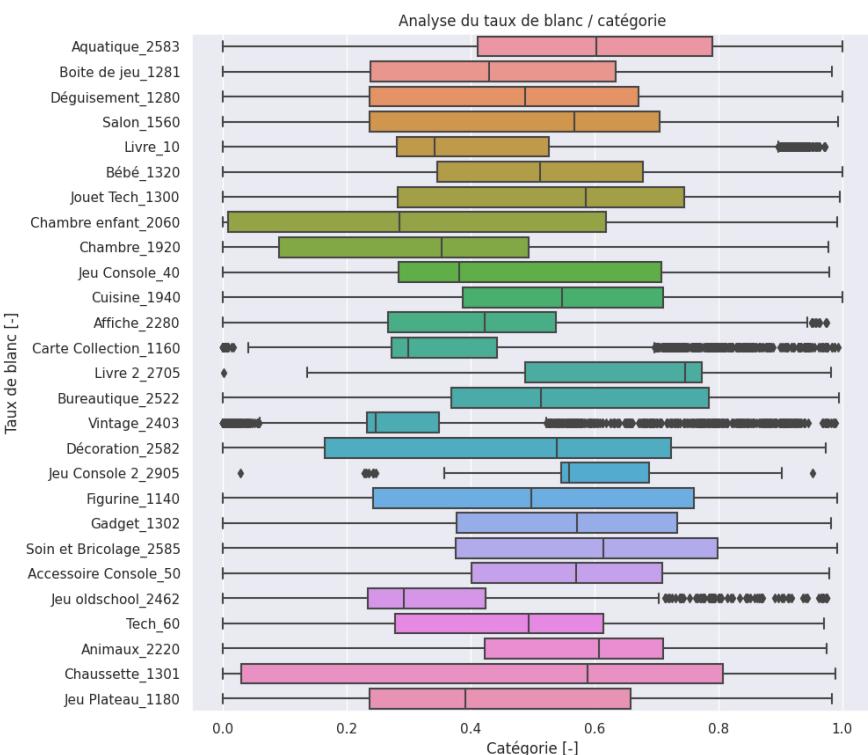
Fréquences des mots

- Grande disparité d'apparition
- Quelques domaines/catégories apparaissent :
 - Dimensions (`cm` , `x` , `mm` , `taill` , `lot`)
 - Autour de la piscine (`eau` , `piscin`)



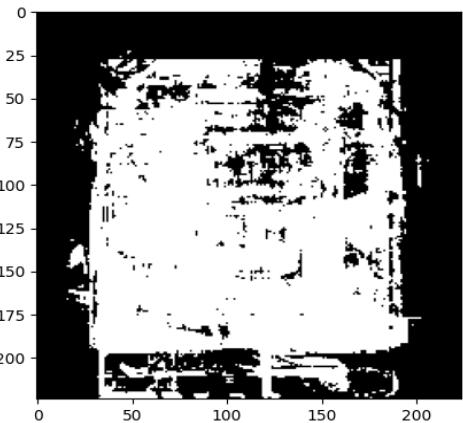
Analyse de canaux

- Fortes disparités dans le taux de blanc de images
 - Catégorie 1301 très étendue
 - Catégorie 2403 , 1160 et 2462 plus restreintes
 - Nombreux *outliers* pour 2403 et 1160



Analayse de la variance

- Masque de variance très net sur les bordures des images
- Possibilité de rogner les images de 20% sans trop de perte de données



Préparation des données / Textes

L'exemple de transformations appliquées :

- colonne `designation` : `Une table très jolie!`
- colonne `description` : `\+Dimensions : 60 x 33 cm`

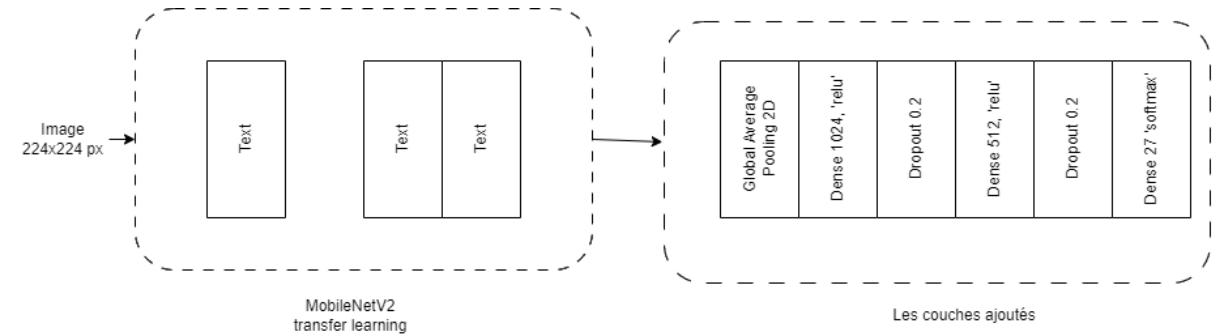
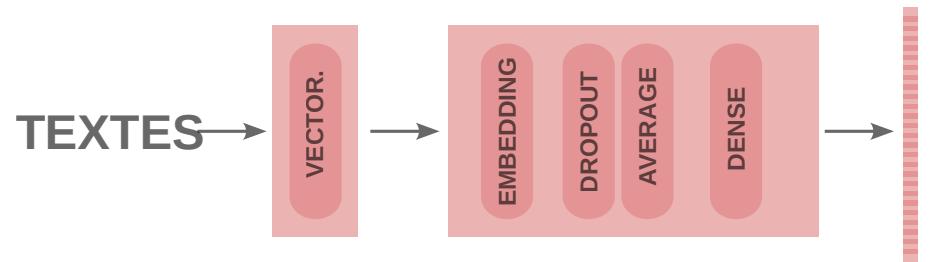
Étape	Résultat
Fusion de <code>description</code> et <code>designation</code>	<code>Une table très jolie! \&#43;Dimensions : 60 x 33 cm</code>
Détection de la langue et traduction en français	<code>Une table très jolie! \&#43;Dimensions : 60 x 33 cm</code>
Suppression les balises html	<code>Une table très jolie! Dimensions : 60 x 33 cm</code>
Suppression des caractères non alpha-numériques	<code>Une table très jolie Dimensions x cm</code>
Passage en minuscules	<code>une table très jolie dimensions x cm</code>
Suppression des accents	<code>une table tres jolie dimensions x cm</code>
Suppression des mots d'un caractère	<code>une table tres jolie dimensions cm</code>
Suppression des <i>stopwords</i>	<code>table tres jolie dimensions cm</code>
Extraction de la racine des mots	<code>tabl tres jol dimens cm</code>
Vectorisation TF-IDF du texte via un <code>Tokenizer</code>	<code>[6, 1, 2, 4, 5]</code>

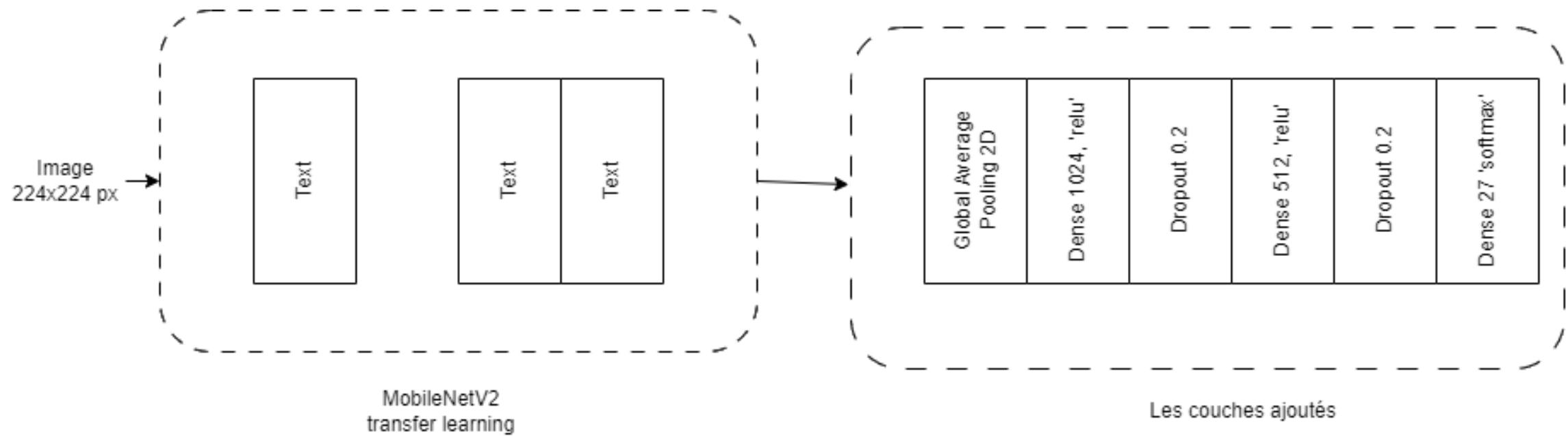
Générateur d'images:

- Streaming per *batch*: images transmises forme de *batchs* ce qui évite de traiter l'ensemble des données (limite RAM + CPU)
- Rognage des images de 20%
- Redimensionnement en taille `224x224 px`
- Application de la fonction `preprocess_input` spécifique à chaque modèle de *deep-learning*



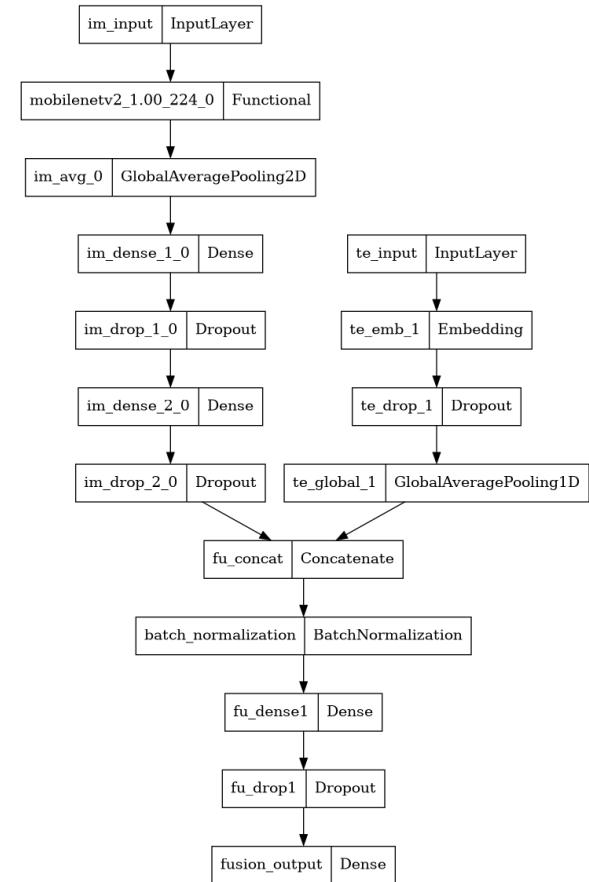
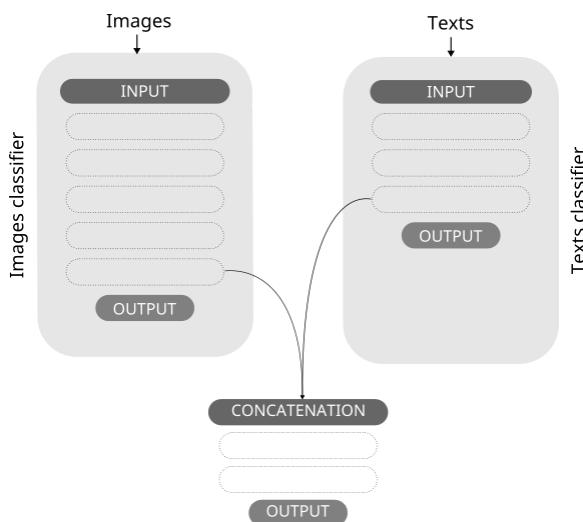
```
<div class="container"> <div class="col"> <h2>Modèle Texte</h2> <ul> <li> Couche de TextVectorization <li> Embedding <li> Couches denses + Dropout </ul> </div> <div class="col"> <h2>Modèle Image</h2> <ul> <li> Couche de TextVectorization <li> Embedding <li> Couches denses + Dropout </ul> </div> </div>
```





Un schéma simplifié du fonctionnement de concaténation.

- La concaténation est faite sur les avant-dernières couches de deux modèles.
- Les autres couches des modèles sont *freezées*.
- Une couche de *BatchNormalization*
- Deux couches denses complètent la fusion pour obtenir une classification sur 27 classes.



Analyse du meilleur modèle 1/2

Analyse des *weighted f1-scores*

- Toutes les catégories dépassent le score de 55%
- Une catégorie sur trois dépasse le score de 90%
- Au final : *weighted f1-score* 82.2 %

Le modèle concaténé s'aide du modèle d'image pour catégoriser les produits où le modèle de texte sous-performait :

- Les catégories 10 et 2705 sont très impactées par la fusion
 - La catégorie 10 Livre neuf gagne 15 points
 - La catégorie 2705 Livre occasion gagne 18 points

	Text f1-score	Image f1-score	Fusion f1-score
10	0.464	0.413	0.610
40	0.612	0.417	0.717
50	0.842	0.223	0.821
60	0.914	0.446	0.902
1140	0.728	0.349	0.730
1160	0.930	0.778	0.954
1180	0.605	0.119	0.551
1280	0.668	0.284	0.654
1281	0.530	0.167	0.572
1300	0.928	0.500	0.930
1301	0.950	0.343	0.937
1302	0.856	0.215	0.844
1320	0.811	0.297	0.800
1560	0.817	0.475	0.809
1920	0.912	0.642	0.905
1940	0.890	0.341	0.866
2060	0.786	0.346	0.780
2220	0.896	0.243	0.899
2280	0.722	0.596	0.767
2403	0.755	0.504	0.775
2462	0.772	0.291	0.803
2522	0.926	0.544	0.914
2582	0.720	0.311	0.709
2583	0.977	0.686	0.972
2585	0.799	0.267	0.775
2705	0.684	0.675	0.859
2905	0.953	0.676	0.965
weighted avg	0.808	0.469	0.822

Analyse des erreurs > 10%

- Livres : 10 , 2080 , 2403 et 2280
- Jouets : 1080 , 1280 et 1281
- Mobilier: 2582 et 1560

Nous nous attendions à avoir des erreurs au sein de produits du même domaine.

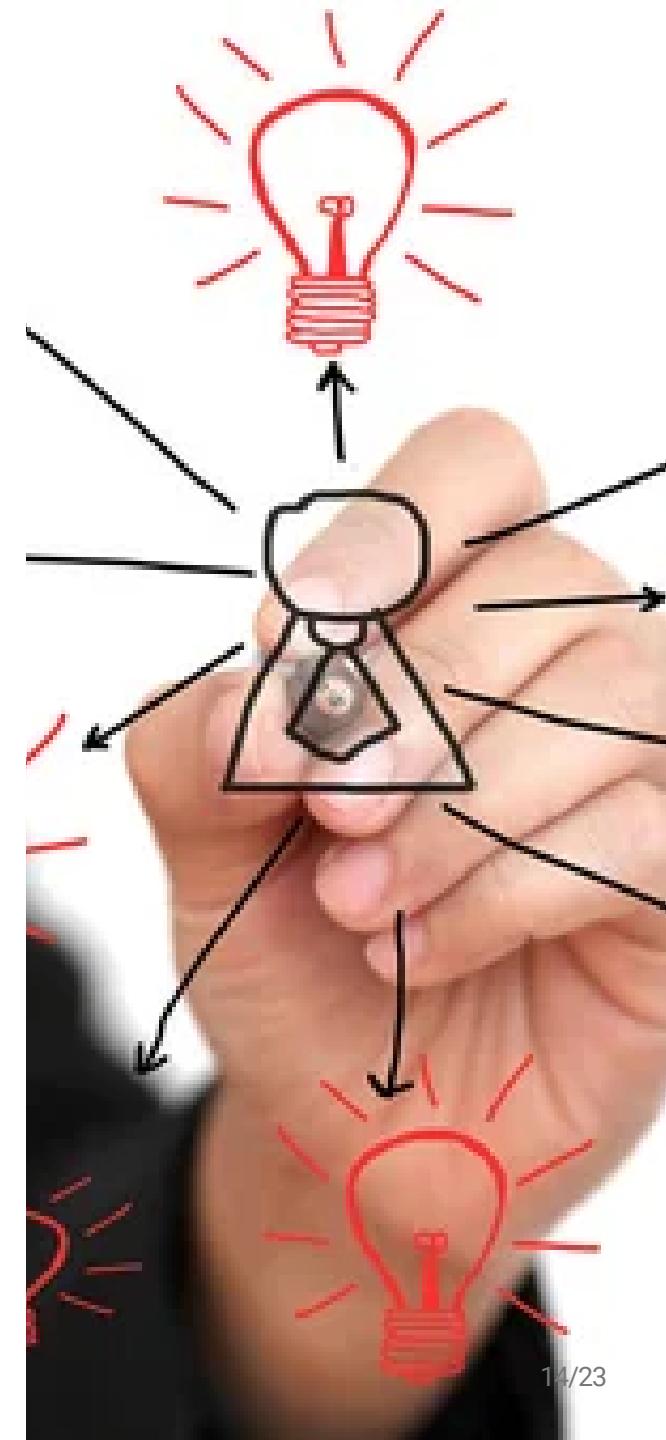
réalité	10	40	50	60	1140	1160	1180	1280	1281	1300	1301	1302	1320	1560	1920	1940	2060	2220	2280	2403	2462	2522	2582	2583	2585	2705	2905
10	0.61	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.12	0.15	0.00	0.02	0.00	0.00	0.06	0.00	
40	0.08	0.71	0.04	0.00	0.02	0.01	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.04	0.02	0.00	0.00	0.00	0.01	0.00	
50	0.00	0.04	0.87	0.01	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	
60	0.00	0.01	0.10	0.84	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.01	
1140	0.02	0.02	0.00	0.00	0.75	0.00	0.02	0.08	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.02	0.00	0.00	0.00	0.00	
1160	0.00	0.00	0.00	0.00	0.01	0.95	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	
1180	0.01	0.05	0.02	0.00	0.11	0.03	0.48	0.07	0.07	0.00	0.00	0.00	0.03	0.00	0.00	0.04	0.00	0.01	0.08	0.00	0.00	0.00	0.00	0.00	0.01	0.00	
1280	0.00	0.01	0.00	0.00	0.08	0.00	0.00	0.62	0.11	0.09	0.00	0.01	0.02	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	
1281	0.01	0.04	0.01	0.00	0.02	0.02	0.02	0.18	0.56	0.00	0.01	0.02	0.00	0.00	0.02	0.00	0.03	0.01	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.00	
1300	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
1301	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.02	0.00	0.90	0.00	0.02	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	
1302	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.01	0.00	0.00	0.00	0.81	0.03	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.02	0.01	0.01	0.00	0.00
1320	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00	0.00	0.01	0.83	0.01	0.02	0.01	0.03	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.00
1560	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.83	0.02	0.00	0.05	0.00	0.01	0.00	0.01	0.04	0.00	0.01	0.00	0.00	0.00
1920	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.92	0.00	0.02	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00
1940	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.05	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.04	0.03	0.01	0.00	0.00	0.00	0.00
2060	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.07	0.05	0.00	0.00	0.75	0.00	0.00	0.01	0.02	0.03	0.00	0.02	0.00	0.00
2220	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.01	0.90	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.00
2280	0.07	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.76	0.13	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
2403	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
2462	0.01	0.07	0.08	0.00	0.02	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2522	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.02	0.00	0.93	0.00	0.00	0.01	0.00	0.00	
2582	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.10	0.01	0.00	0.05	0.01	0.00	0.00	0.01	0.73	0.02	0.04	0.00	0.00	0.00	
2583	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.08	0.00	0.00	0.02	0.00	0.00	0.01	0.09	0.02	0.73	0.00	0.00	0.00	0.00	0.00
2585	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.01	0.08	0.00	0.00	0.02	0.00	0.00	0.03	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00
2705	0.09	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86	0.00	
2905	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Informations complémentaires

- Pourquoi ces produits et ces catégories à classer en particulier?
- Comment la classification initiale des targets a-t-elle été faite?

Ressources et techniques

- Disponibilité limitée de ressources de calcul de type GPU ou TPU via Google Colab.
- Pertes d'accès fréquentes entre Google Drive et Google Colab, perte de temps
- Arrivées tardives des notions de générateurs et de deep-learning dans les modules



Perspectives

Le modèle de textes:

- Utiliser des modèles pré-entraînés
 - Word2Vec
 - CamemBERT

Le modèle d'images :

- Étapes de pré-processing
 - Augmentation des données via transformations
- Modèles
 - Implémenter *Batch Normalization*,
 - Entrainer des couches de modèles issues de *transfer-learning *
 - Configurer différemment les hyperparamètres
 - Tester les *Vision Transformers*
- Analyse de patterns générés par les couches
- Test autres modèles avec autre taille des images en entrée

Fusion

- Ajouter d'autres modèles, plus performants sur les classes délicates à prédire



Conclusion

Merci à tous pour ces 3 mois très intenses et riches !

<a style="color: #f44336"
href="localhost:8080">Streamlit + FastAPI +
Docker = ❤

“

Nous continuons de croire que le monde numérique a le potentiel d'améliorer la vie de chacun d'entre nous. Oubliez la peur.
Adoptez l'optimisme.

Hiroshi Mikitani – Fondateur et CEO de Rakuten



Annexes

Notions

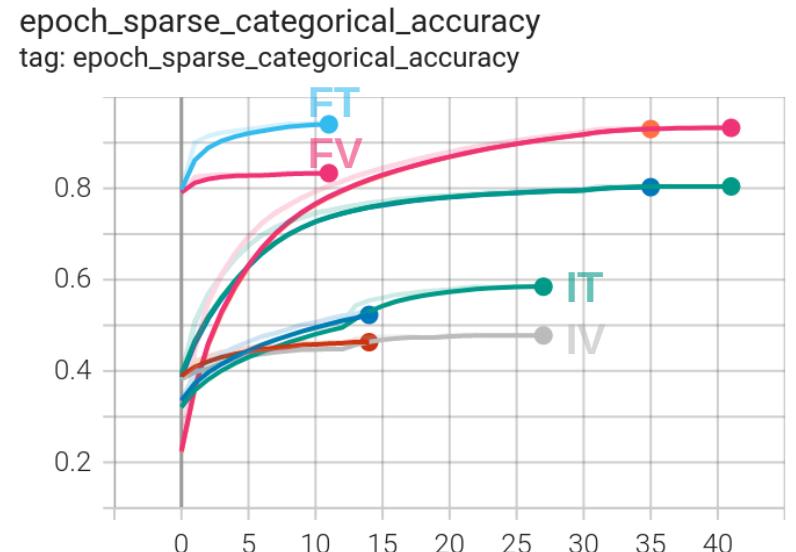
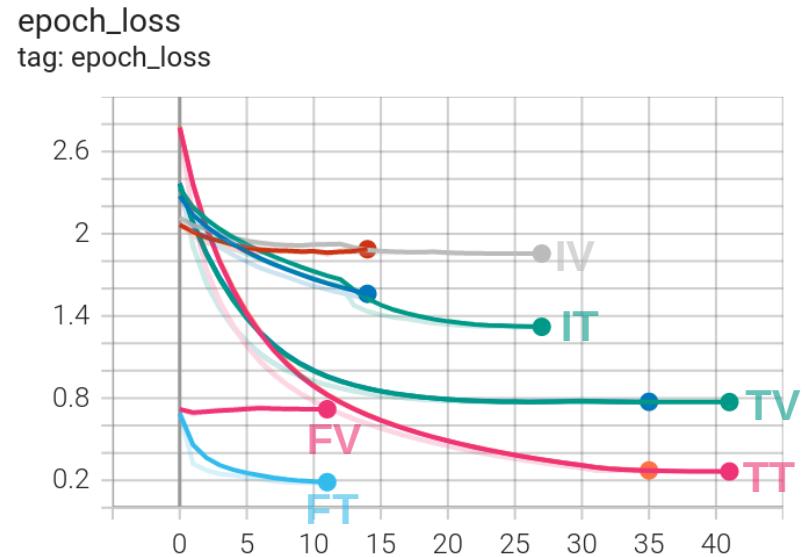
- Connaissance du métier : une erreur de classification n'est pas fatale
 - Labelisation : comment a-t-elle été effectuée
 - Jeu de données déséquilibré : dû à une survente ou à des difficultés à classer ces produits
 - Forte tendance à l'*overfitting*
- Choix de la métrique : *f1 weighted score* pour un bon équilibre entre *accuracy* et *recall*

Remarques

- Modèle aléatoire : score de 3.7% en moyenne
- Une métrique personnalisée aurait pu être créée

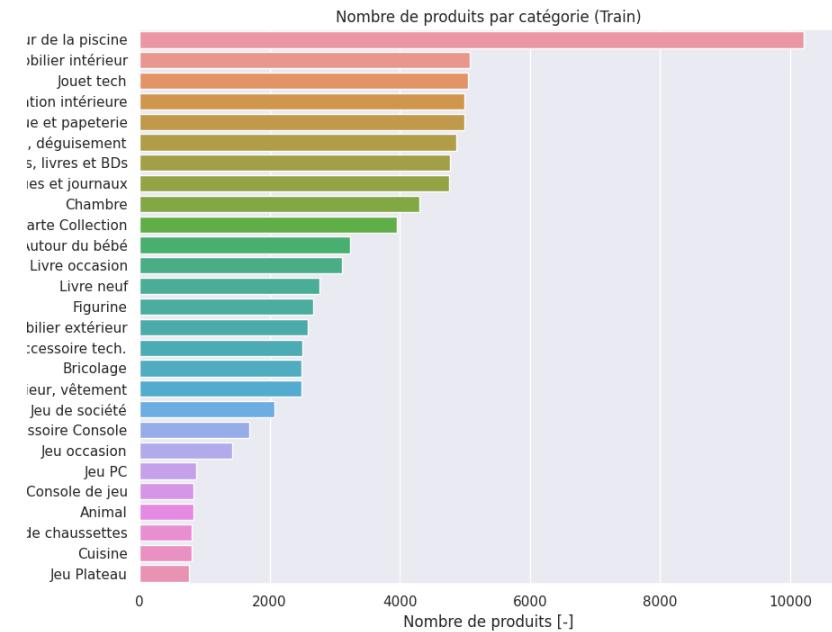
Nécessaires au contrôle des modèles lors de l'apprentissage

- Suivi via *TensorBoard*
- **EarlyStopping** : met fin à l'apprentissage si val_loss augmente pendant plus de 5 périodes à partir de la 8ème période
- **ReduceLROnPlateau** : réduit le taux d'apprentissage si val_loss stagne sur un plateau pendant plus de 5 périodes



Déséquilibre des targets

- Non homogénéité de la répartition des classes
- Environ 7 classes sur-représentées
- 3 classes sous-représentées



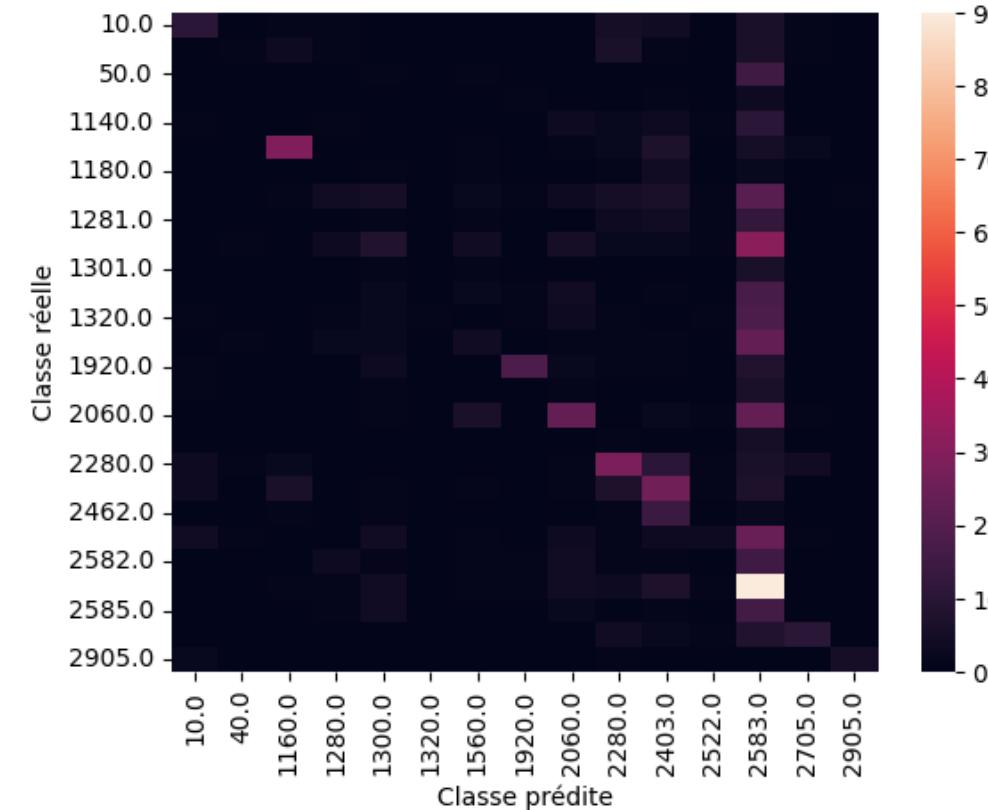
Annexe : Machine Learning / Text

- Catégorie 10 (Livre d'occasion) souvent confondue avec 2705 (Livre neuf) et 2403 (Revue)
- Catégorie 40 (Jeu console) souvent confondue avec 10 (Livre occasion) et 2462 (Jeu oldschool)
- Catégorie 1280 (Déguisement) souvent confondue avec 1281 (Boîte de jeu) et 1140 (Figurine)

	logistic_regression	random_forest	kneighbours	decision_tree
10	0.449	0.472	0.260	0.376
40	0.543	0.595	0.320	0.499
50	0.730	0.770	0.518	0.618
60	0.875	0.891	0.814	0.823
1140	0.670	0.685	0.530	0.586
1160	0.877	0.865	0.731	0.836
1180	0.391	0.475	0.381	0.420
1280	0.638	0.603	0.449	0.534
1281	0.507	0.482	0.322	0.410
1300	0.899	0.866	0.729	0.890
1301	0.880	0.849	0.847	0.793
1302	0.754	0.758	0.603	0.671
1320	0.699	0.674	0.579	0.576
1560	0.793	0.746	0.636	0.660
1920	0.893	0.905	0.856	0.828
1940	0.821	0.791	0.648	0.669
2060	0.731	0.739	0.620	0.664
2220	0.744	0.751	0.488	0.632
2280	0.790	0.812	0.577	0.764
2403	0.720	0.734	0.604	0.671
2462	0.722	0.768	0.604	0.709
2522	0.891	0.846	0.763	0.755
2582	0.697	0.671	0.525	0.571
2583	0.964	0.926	0.900	0.919
2585	0.728	0.676	0.522	0.540
2705	0.643	0.644	0.321	0.545
2905	0.950	0.977	0.045	0.963
weighted F1-Score	0.771	0.760	0.619	0.695

Machine Learning / Image

Classifier	Acc.	Precision weighted	Recall weighted	F1 weighted
LogReg	0.18	0.16	0.18	0.16
RF	0.12	0.04	0.12	0.04
KNN	0.18	0.16	0.18	0.16
SVC	0.18	0.17	0.18	0.17
GradBoost	0.09	0.08	0.09	0.06



Annexe : Les modèles / Deep learning / Image

Model	Accuracy	Val accuracy
VGG16	0.50	0.49
ResNet	0.16	0.18
MobileNet	0.87	0.47

