



Rakuten Institute of Technology



DataScientest

Participants :

Olga TOLSTOLUTSKA

Mohamed BACHKAT

Charly LAGRESLE

Mentor : Manu POTREL

Promotion: DST Bootcamp DEC22



白いクリーマー

商品仕様 ★サイズ
8.3×7×H7.8cm ★重量
160g ★容量 240cc ★色
ほんの少しグレーが
かかった白 ★素材 日
本製磁器 ★その他 電
子レンジ,食器洗浄機
OK! オープン不可 こ
ちらはアウトレット
商品です。細かなキ
ズ、ピンホール(針
を刺したような穴) ...

Rakuten Data Challenge: extracting information from texts and images with Deep Learning to enrich the catalog of product

Material *Porcelain*
Color *White*
...

Filters

<https://www.rakuten.co.jp>, Jan. 2021

- **84 916** observations
- **27** catégories à déterminer
- **0** données dupliquées
- **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

Exploration des données

- Cible déséquilibrée
- 27 catégories dont certaines sont délicates à dissociées

Cat.	Code et libellé	Cat.	Code et libellé	Cat.	Code et libellé
	10 - Livre d'occasion		1300 - Jouet Tech		2280 - Affiche
	40 - Jeu Console		1301 - Chaussette		2403 - Revue
	50 - Accessoire Console		1302 - Gadget		2462 - Jeu oldschool
	60 - Tech		1320 - Bébé		2522 - Bureautique
	1140 - Figurine		1560 - Salon		2582 - Décoration
	1160 - Carte collect.		1920 - Chambre		2583 - Aquatique
	1180 - Jeu Plateau		1940 - Cuisine		2585 - Soin et Bricolage
	1280 - Déguisement		2060 - Chambre enfant		2705 - Livre neuf
	1281 - Boite de jeu		2220 - Animaux		2905 - Jeu PC

Exploration des données / Text

- Détection de la langue
 - 81% français
 - 14% anglais
 - Traduction

Exploration des données / Images

Préparation des données / Text

L'exemple de transformations appliquées :

- designation : Une table très jolie!
- description :
 - Dimensions : 60 x 33 cm

Etape	Résultat
Fusion de deux colonnes	Une table très jolie! <ul style="list-style-type: none">▪ Dimensions : 60 x 33 cm
Détection de la langue et traduction en français	Une table très jolie! <ul style="list-style-type: none">▪ Dimensions : 60 x 33 cm
Suppression les balises html	Une table très jolie! Dimensions : 60 x 33 cm
Suppression des caractères non alpha-numériques	Une table très jolie Dimensions x cm
Passage en minuscules	une table très jolie dimensions x cm
Supression des accènts	une table tres jolie dimensions x cm
Les mots d'un caractère	une table tres jolie dimensions cm
Suppression des <i>stopwords</i>	table tres jolie dimensions cm
Extraction de la racine des mots	tabl tres jol dimens cm
Vectorisation du texte via un Tokenizer	[6, 1, 2, 4, 5]

Générateur d'images:

- Streaming per batch : les images sont transmises sous de batchs ce qui évite de traiter l'ensemble des données d'un coup
- Redimensionnement en taille 224x224 px
- Application de la fonction preprocess_input spécifique à chaque modèle

Un schéma simplifié du fonctionnement de concaténation.

- concaténation est faite sur les avant-dernières couches de deux modèles.
- les autres couches sont *freezées*.
- couches denses complètent la fusion pour obtenir une classification sur 27 classes.

Analyse du meilleur modèle

Pas d'impacte sur les performances réduites du modèle d'image.

- Toutes les catégories dépassent le score de 54% et
- Une catégorie sur trois dépasse le score de 90%

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

- La catégorie 1080 (Jeu Plateau) gagne 25 points
- La catégorie 2705 (Livre neuf) gagne 23 points

- Le traitement des 84916 images nécessite d'utilisation de générateurs.
- Disponibilité limité de ressources de calcul de type GPU ou TPU via Google Colab.
- Coupures de lien entre Google Drive et Google Colab ont entraîné une grande perte de temps
- La création d'un modèle de fusion a été une tâche ardue, principalement pour la gestion des entrées sous forme de générateurs.

“

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



Les modifications globaux :

- Uniformisation des données dans le code. Actuellement, des dataframes Pandas, des tableaux Numpy, des générateurs d'images fonctionnent ensemble. Tout pourrait être géré autour d'un seul type de données, comme les tf.data.Dataset.

Le modèle de texte:

- une couche d'embedding pré-entraînée, par exemple celle issue de CamemBERT.



Le modèle d'image :

- évolution traitement et preprocessing des images
 - cropping d'image
 - augmentation des données via transformation
- évolution de modèles testés :
 - implémenter *Batch Normalization*,
 - entraîner des couches de model issue de transfer learning
 - configurer différemment les hyperparamètres
 - entraînement des couches de model issue de transfer learning
- analyse de patterns générés par les couches
- test autres modèles avec autre taille des images en entrées

Fusion

- ajout d'autres modèles au modèle de fusion
- test un autre approche de la fusion : utiliser un modèle pour identifier un group global et ensuite sous-group précis. Par exemple première model prédit un group "Livre" et deuxième model prédit "Nouveau" ou "Ancien".





Le projet **Rakuten** a été très intéressant, car complexe et faisant appel à des notions avancées mêlant le traitement de textes et le traitement d'images.

L'exploration de données, le travail de groupe, les différentes implémentations et sprints ont fait de ce projet un projet répondant, nous l'espérons, aux besoins d'une entreprise.

- 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)

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 / Text

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

Annexe Exploration des données / Target

- Connaissance du métier : une erreur de classification n'est pas fatale
 - 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*

