

OPENCLASSROOMS

CLASSIFICATION AUTOMATIQUE DE BIENS DE CONSOMMATION

POUR L'ENTREPRISE PLACE DE MARCHÉ

Présenté par Cédric Vachaudez



Déroulé

Problématique
Découverte
Visualisation

Dataset



Etude de faisabilité

Prétraitement
Extraction de features
Visualisation



Choix du modèle
Ajustement des hyperparamètres
Quelles catégories ?

Classification



Data Augmentation

Méthode d'augmentation
Traitements appliqués
Résultats finaux



Fonctionnement
Respect du RGPD
Utilisation d'API



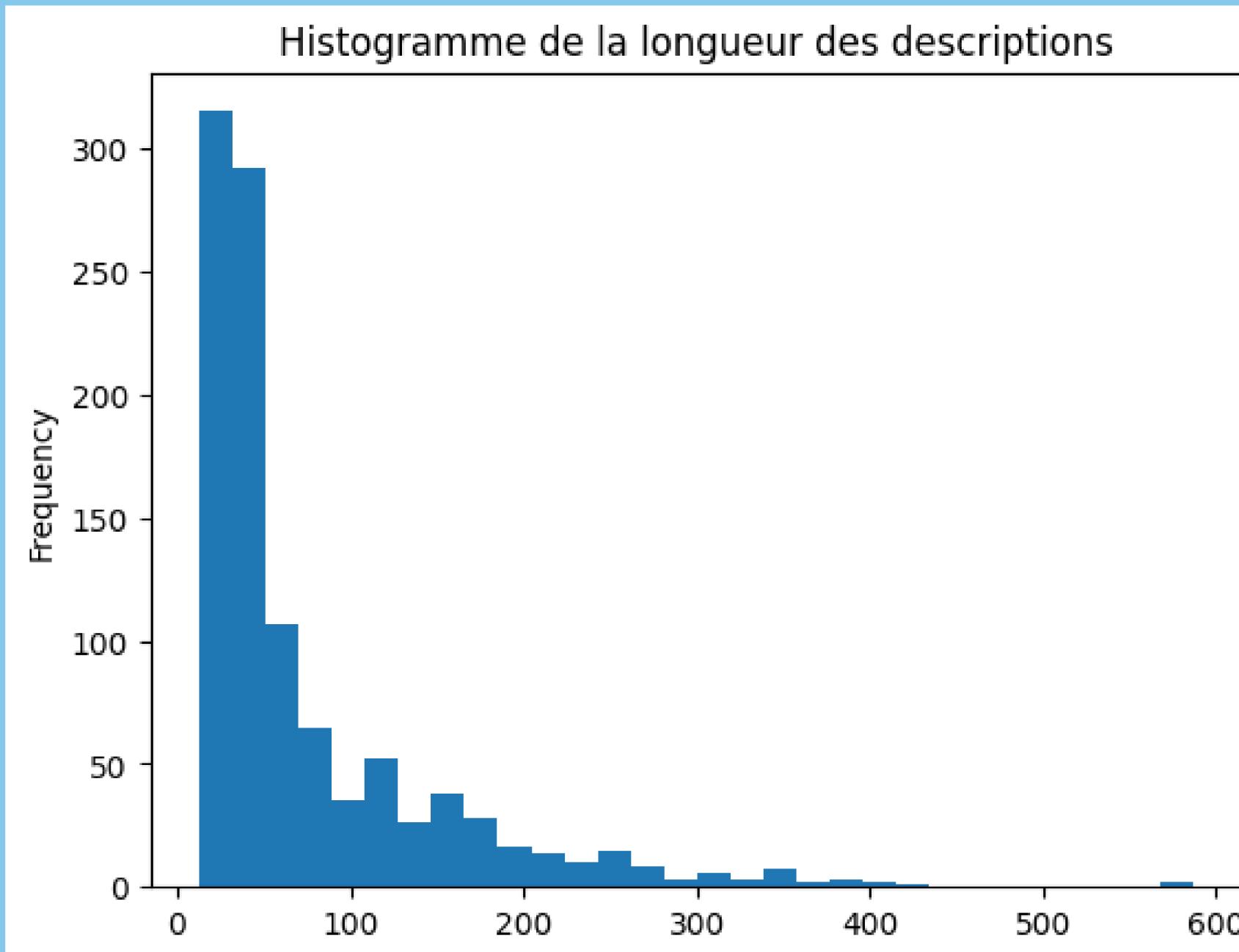
Dataset

découverte et visualisation

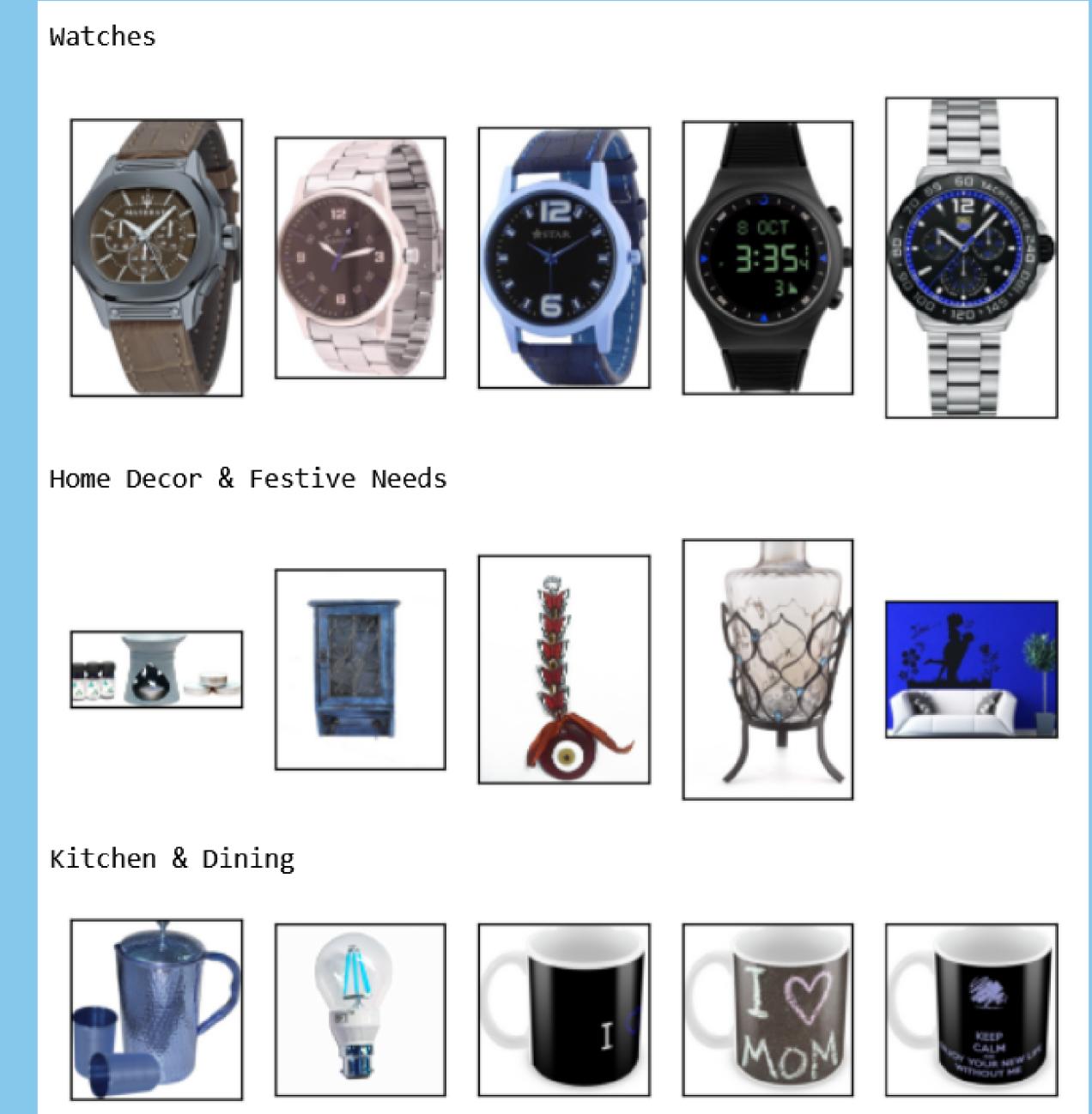
Aperçu du dataset

		description	product_category_tree	image
0	55b85ea15a1536d46b7190ad6fff8ce7	Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain, Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This c...	["Home Furnishing >> Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Eyelet Do..."]	
1	7b72c92c2f6c40268628ec5f14c6d590	Specifications of Sathiya's Cotton Bath Towel (3 Bath Towel, Red, Yellow, Blue) Bath Towel Features Machine Washable Yes Material Cotton Design Self Design General Brand Sathiya's Type Bath Towel GSM 50...	["Baby Care >> Baby Bath & Skin >> Baby Bath Towels >> Sathiya's Baby Bath Towels >> Sathiya's Cotton Bath Towel (3 Bath Towel, Red, Y..."]	
2	64d5d4a258243731dc7bbb1eef49ad74	Key Features of Eurospa Cotton Terry Face Towel Set Size: small Height: 9 inch GSM: 360, Eurospa Cotton Terry Face Towel Set (20 PIECE FACE TOWEL SET, Assorted) Price: Rs. 299 Eurospa brings to you an ...	["Baby Care >> Baby Bath & Skin >> Baby Bath Towels >> Eurospa Baby Bath Towels >> Eurospa Cotton Terry Face Towel Set (20 PIECE FA..."]	
3	d4684dcdc759dd9cdf41504698d737d8	Key Features of SANTOSH ROYAL FASHION Cotton Printed King sized Double Bedsheet Royal Bedsheet Perfect for Wedding & Gifting, Specifications of SANTOSH ROYAL FASHION Cotton Printed King sized Double Be...	["Home Furnishing >> Bed Linen >> Bedsheets >> SANTOSH ROYAL FASHION Bedsheets >> SANTOSH ROYAL FASHION Cotton Printed King sized ..."]	
4	6325b6870c54cd47be6ebfbffa620ec7	Key Features of Jaipur Print Cotton Floral King sized Double Bedsheet 100% cotton, Jaipur Print Cotton Floral King sized Double Bedsheet (1 bed sheet 2 pillow cover, White) Price: Rs. 998 This nice bed...	["Home Furnishing >> Bed Linen >> Bedsheets >> Jaipur Print Bedsheets >> Jaipur Print Cotton Floral King sized Double Bed..."]	
5	893aa5ed55f7cff2eccea7758d7a86bd	Maserati Time R8851116001 Analog Watch - For Boys - Buy Maserati Time R8851116001 Analog Watch - For Boys R8851116001 Online at Rs.24400 in India Only at Flipkart.com. - Great Discounts, Only Genu...	["Watches >> Wrist Watches >> Maserati Time Wrist Watches"]	

Visualisation



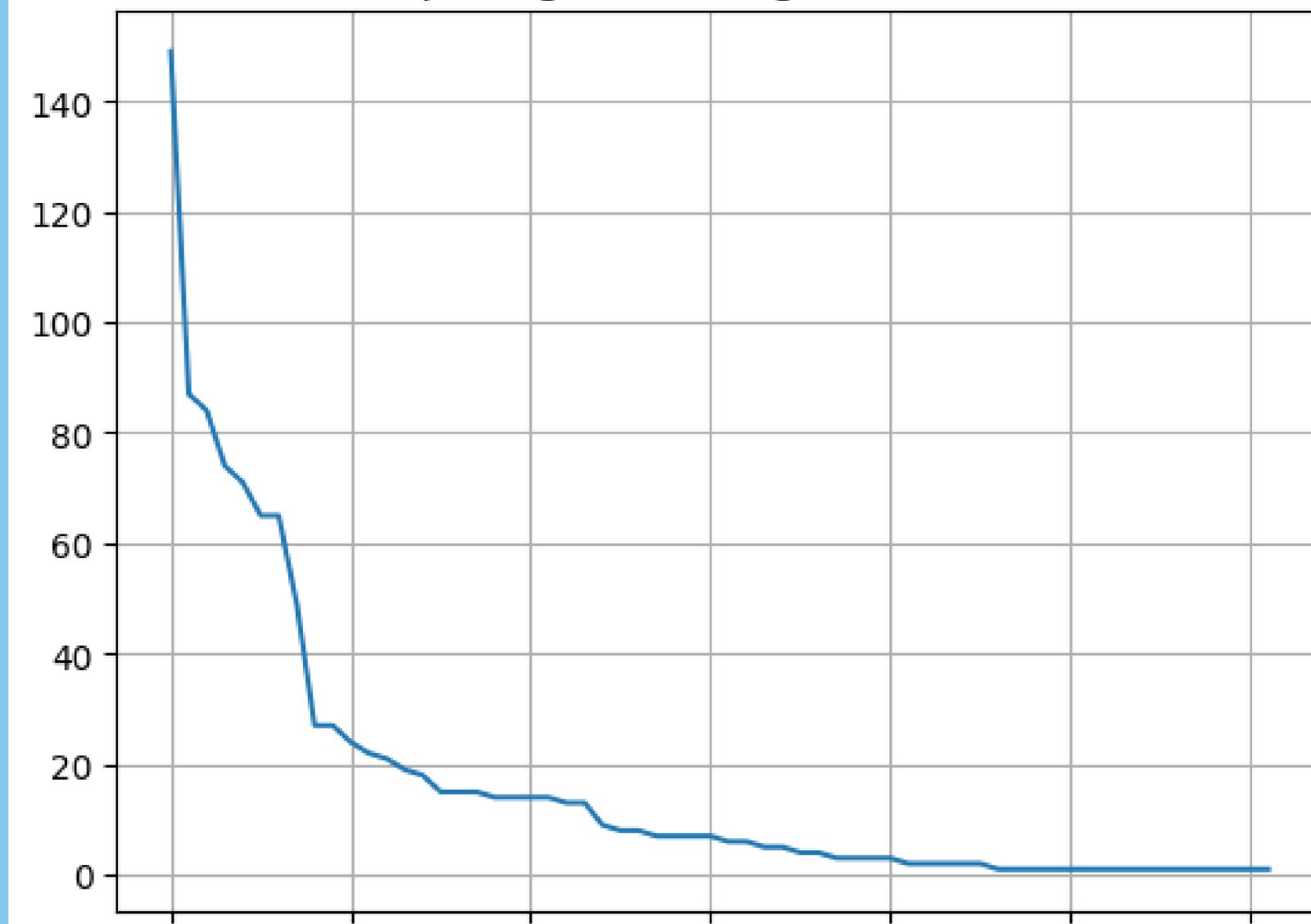
Taille des descriptions



Aperçu des images

Visualisation des catégories

Remplissage des catégories au level 1



Etude de faisabilité

en utilisant les descriptions

Traitements du texte

avec spaCy et NLTK

01 Nettoyage

Passage en minuscule

Homogénéisation des espaces, séparateurs et ponctuation

02 Tokenisation

Division de la phrase en une liste de "symboles", de tokens individuels

03 Lemmatisation

Transformation des mots en leur racine

04 Suppression des stopwords

Utilisation d'une liste générique de spaCy en y ajoutant les termes les moins représentatifs du dataset

05 Suppression de la fin des textes

On ne garde que la première moitié (le reste étant souvent redondant ou porteur de confusion)

I don't like those 3 black cats, they are all three sitting on my single mat —————→ not like black cat sit single mat

Bag-of-words et TF-IDF

Représentation des occurrences des mots dans une matrice.

Pour le Bag-of-words on compte le nombre de mots.

Pour TF-IDF on calcule un score représentant son importance au sein du document.

BERT

Vectorisation mot par mot d'une phrase.

Modèle de transformer entraîné à trouver un mot masqué (MLM), et à deviner la phrase suivante (NSP). Prend en compte la position et le contexte des mots.

Utilise un système d'attention.

USE

Vectorisation d'une phrase complète.

Modèle entraîné à prédire les phrases suivantes et précédentes, choisir la meilleure réponse possible, comparer l'intention de phrases deux à deux.

Agglomère les tokens au départ plutôt qu'à la fin.

Word2Vect

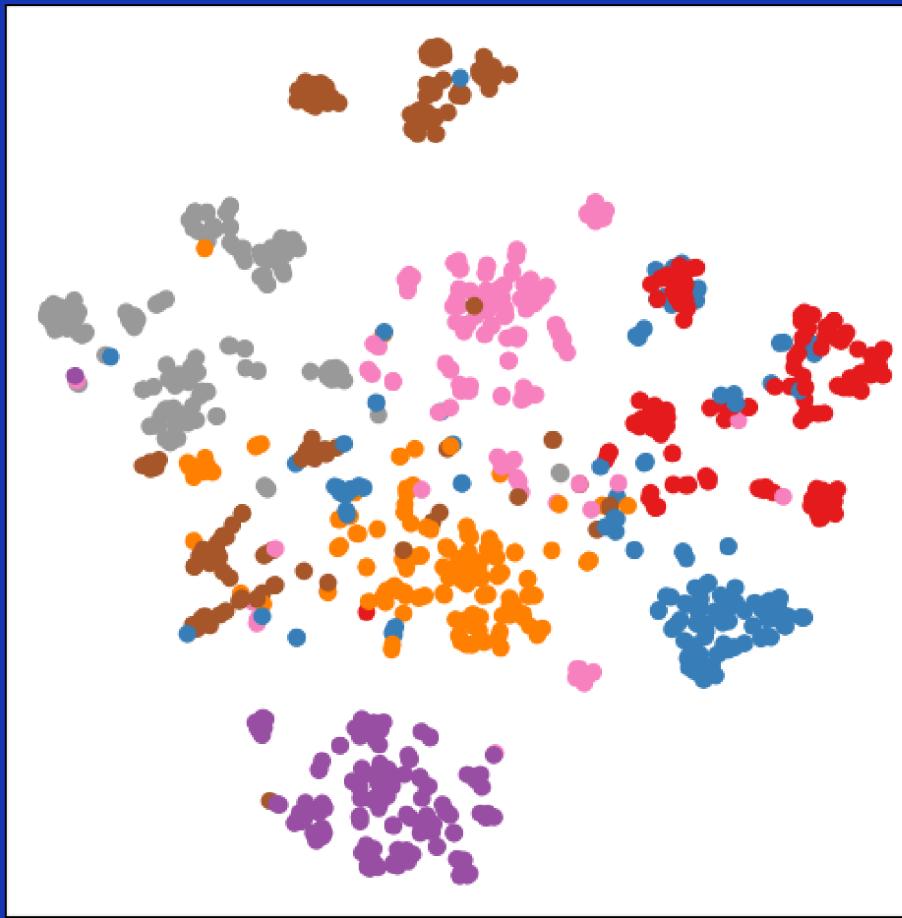
Vectorisation mot par mot.

Initialement entraîné à prédire un mot selon son contexte proche (MLM avec une fenêtre) sur un large corpus.

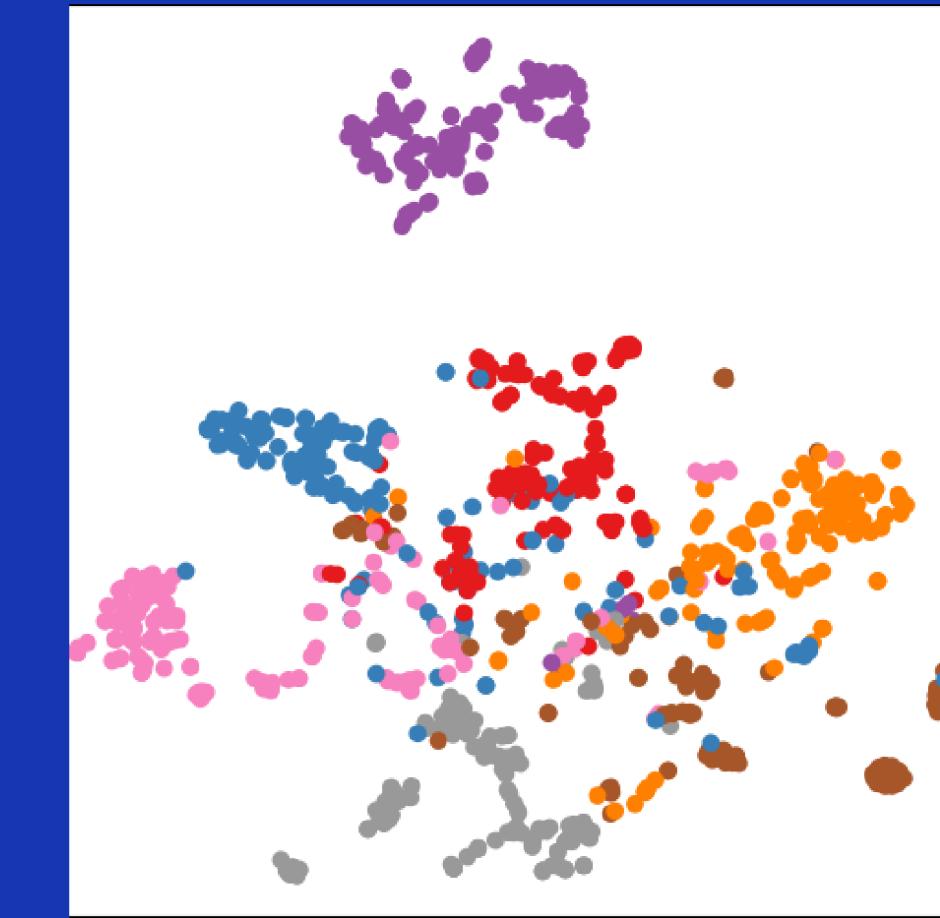
Des mots au sens proche ont des vecteurs proches.

Algorithmes utilisés

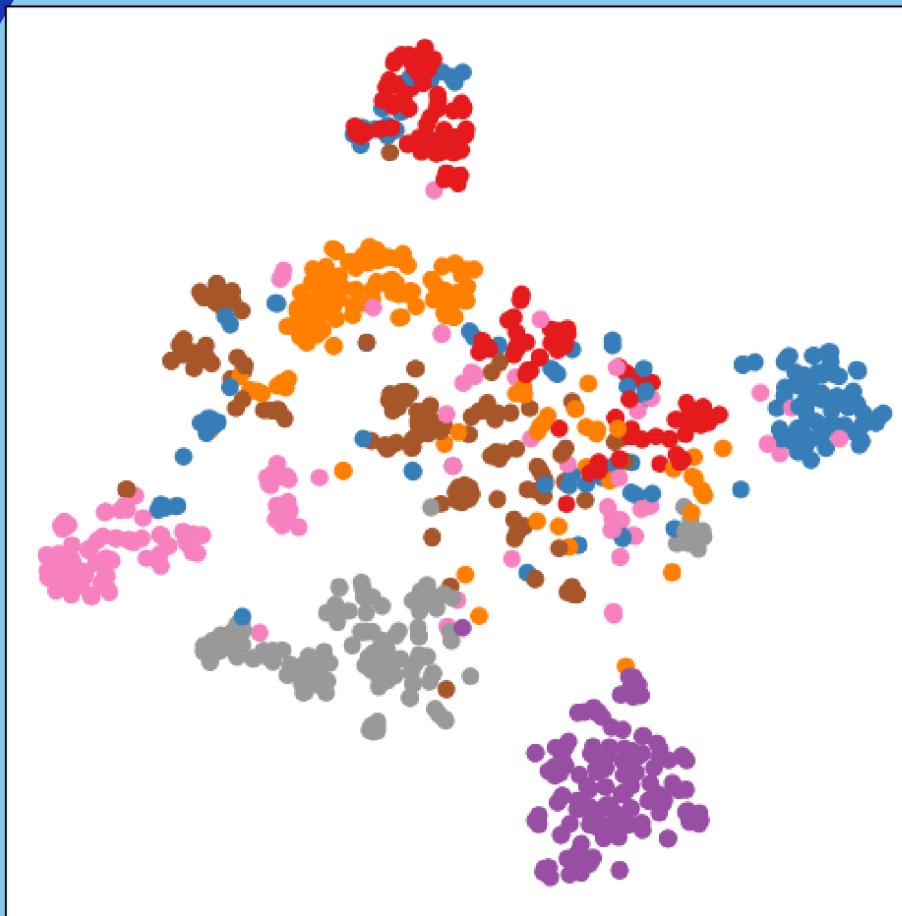
Projections



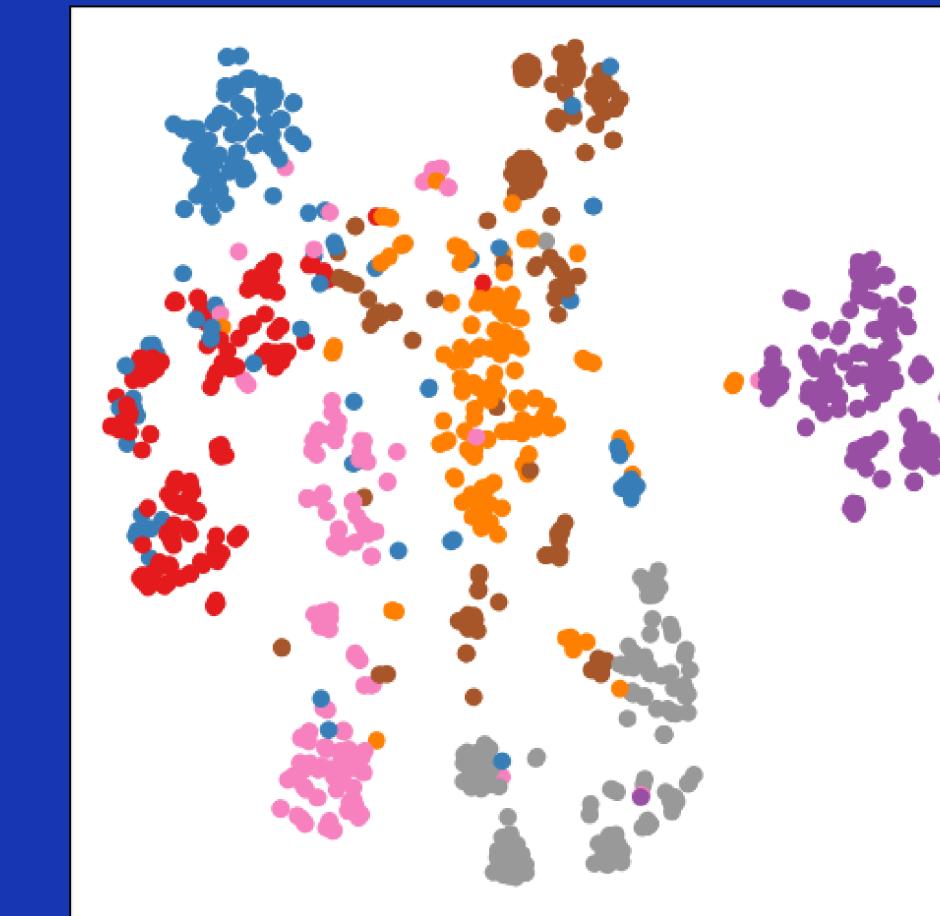
Tf-idf
ARI : 0.59



USE
ARI : 0.58



BERT
ARI : 0.45



Word2Vect
ARI : 0.52

Etude de faisabilité

en utilisant les photos

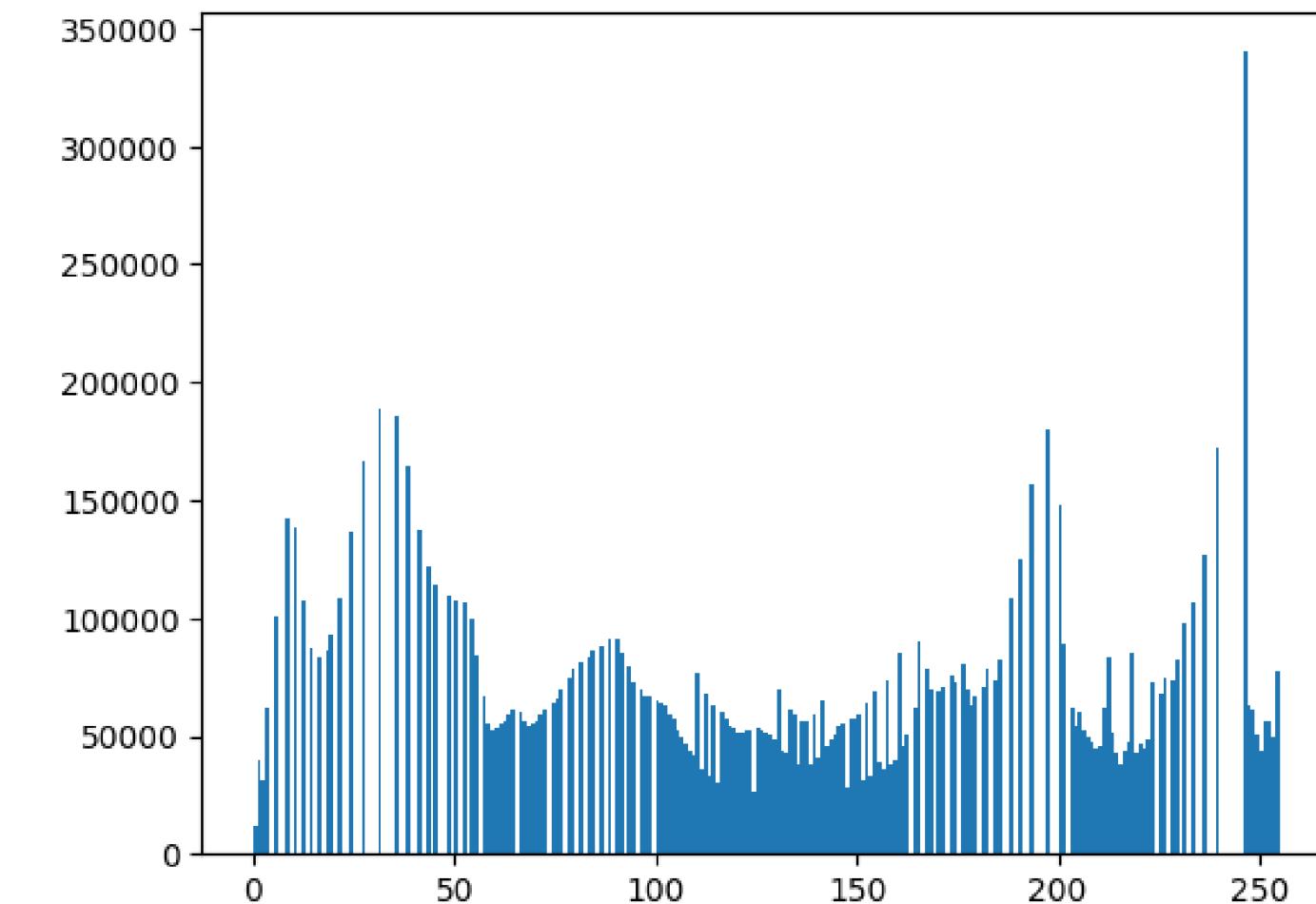
Prétraitement pour ORB

avec OpenCV

01 Passage en niveaux de gris

02 Egalisation

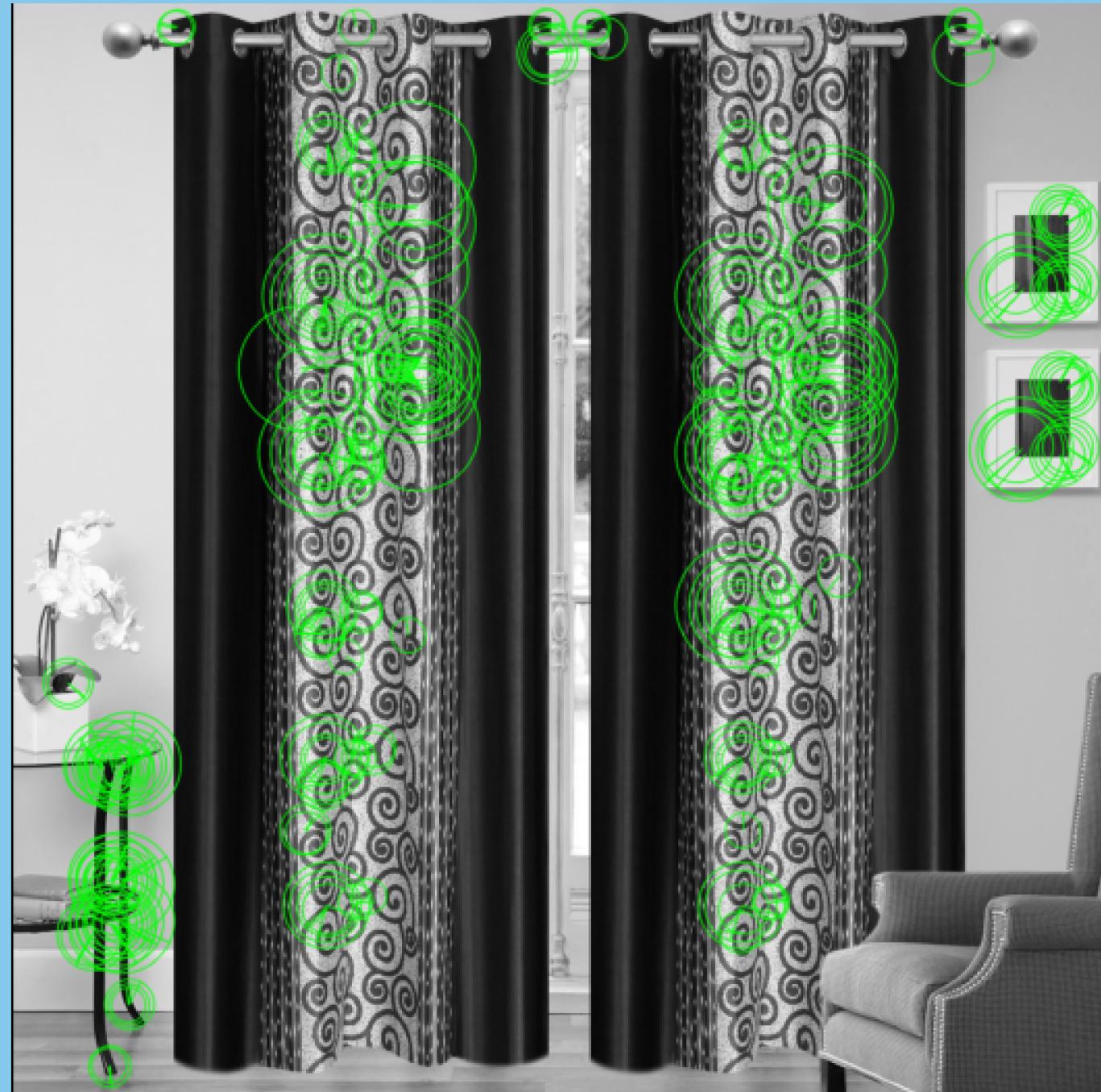
Pour corriger les images peu contrastés



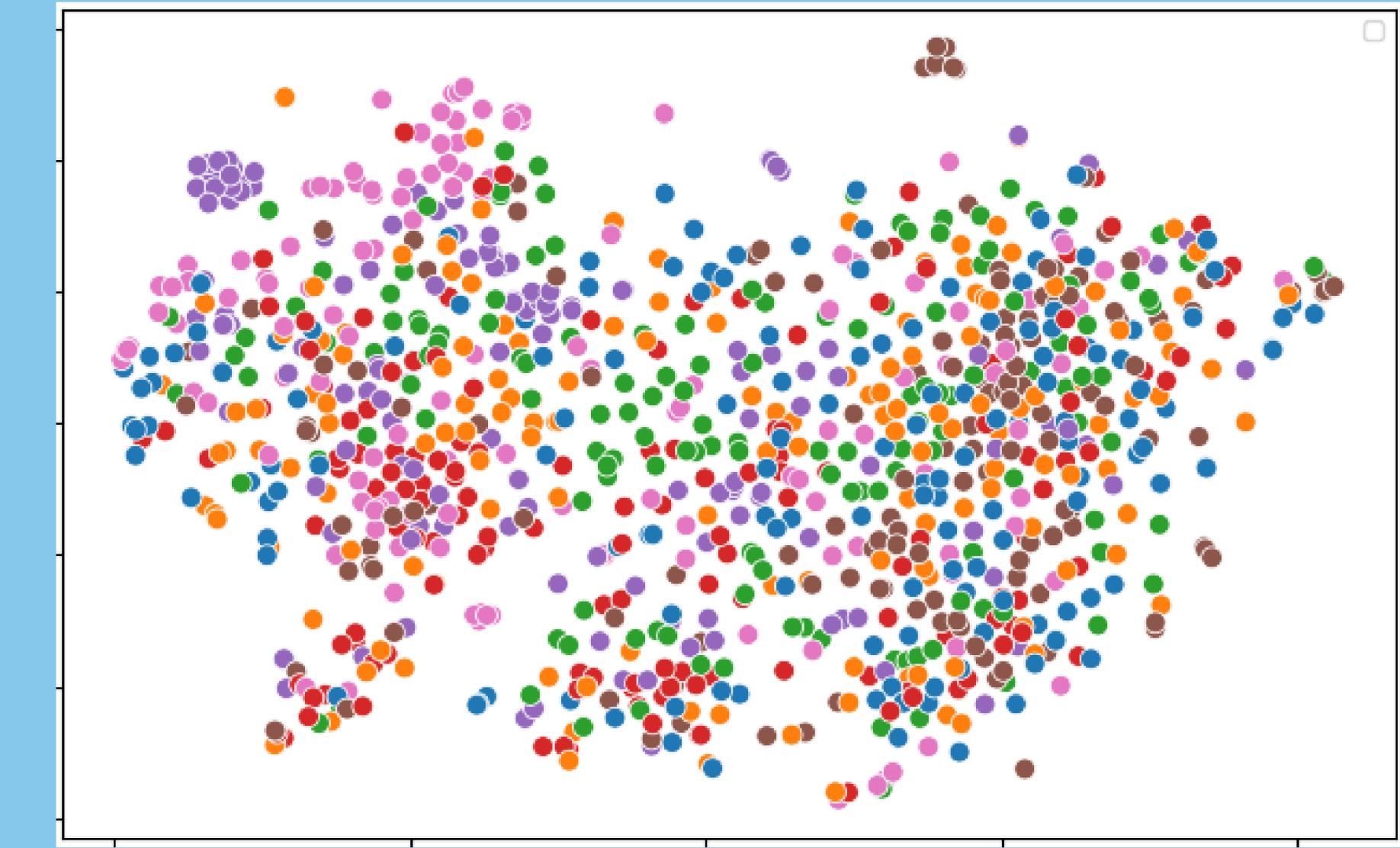
03 Floutage

Pour un éventail de features trouvées plus variées

Features extraction



Exemple des features trouvées



Projection 2D

Score ARI : 0.03

Utilisation du transfert learning

01 Prétraitement

Redimensionnement et utilisation de preprocess_input

02 Utilisation de modèles pré-entraînés

DenseNet121, MobileNetV2, VGG16 et VGG19

Suppression des dernières couches de classification

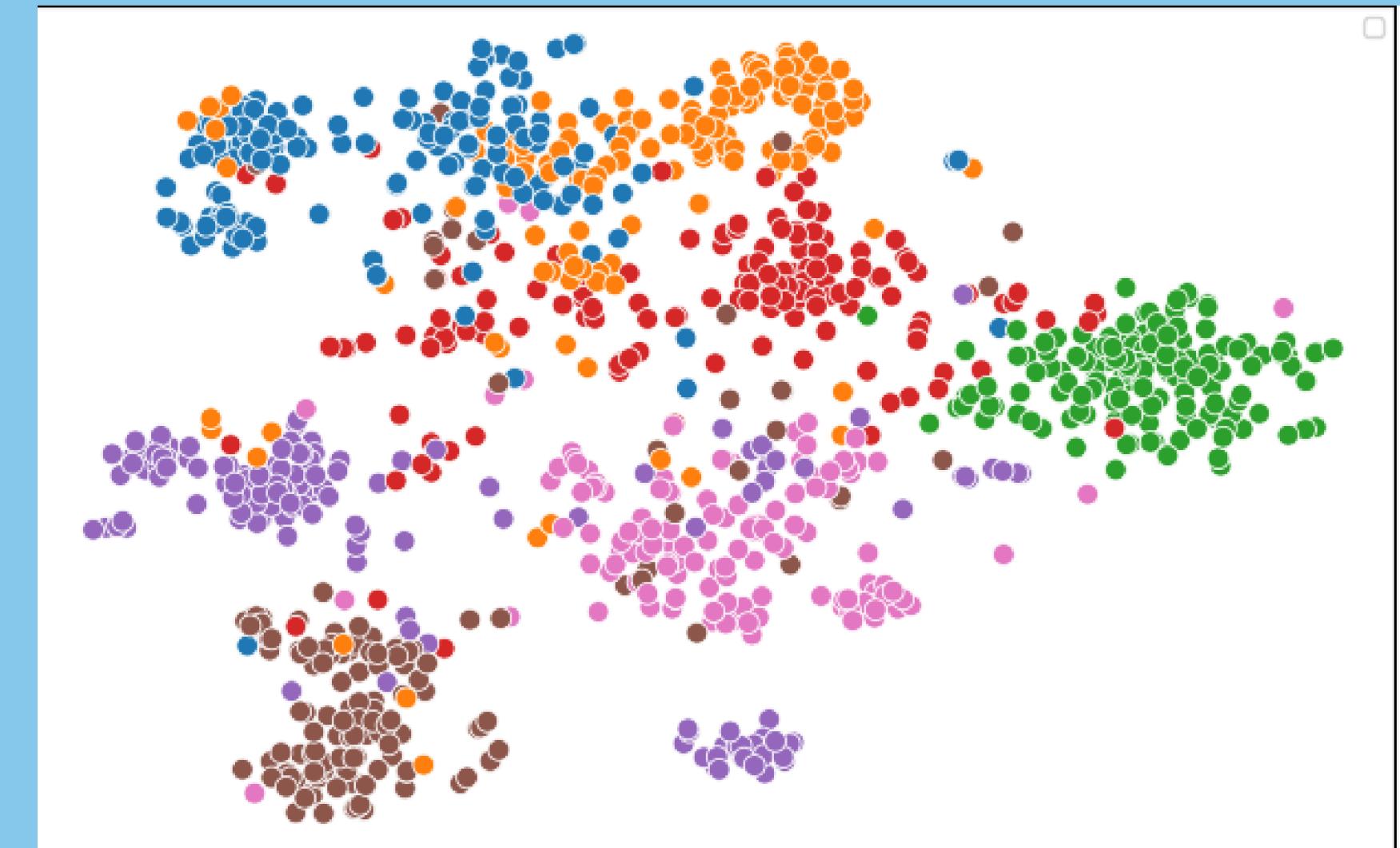
03 Réduction de dimension

Application d'une PCA pour pouvoir utiliser KNN ensuite

Transfert learning

cat	pred_0	pred_3	pred_4	pred_1	pred_5	pred_6	pred_2
Computers	140	1	6	2	0	1	0
Watches	6	139	5	0	0	0	0
Home Decor & Festive Needs	1	5	136	2	2	1	3
Beauty and Personal Care	11	0	16	116	1	5	1
Kitchen & Dining	18	1	14	5	112	0	0
Baby Care	4	0	20	3	4	109	10
Home Furnishing	1	0	11	1	0	74	63

Matrice de confusion



Projection 2D

Score ARI : 0.57

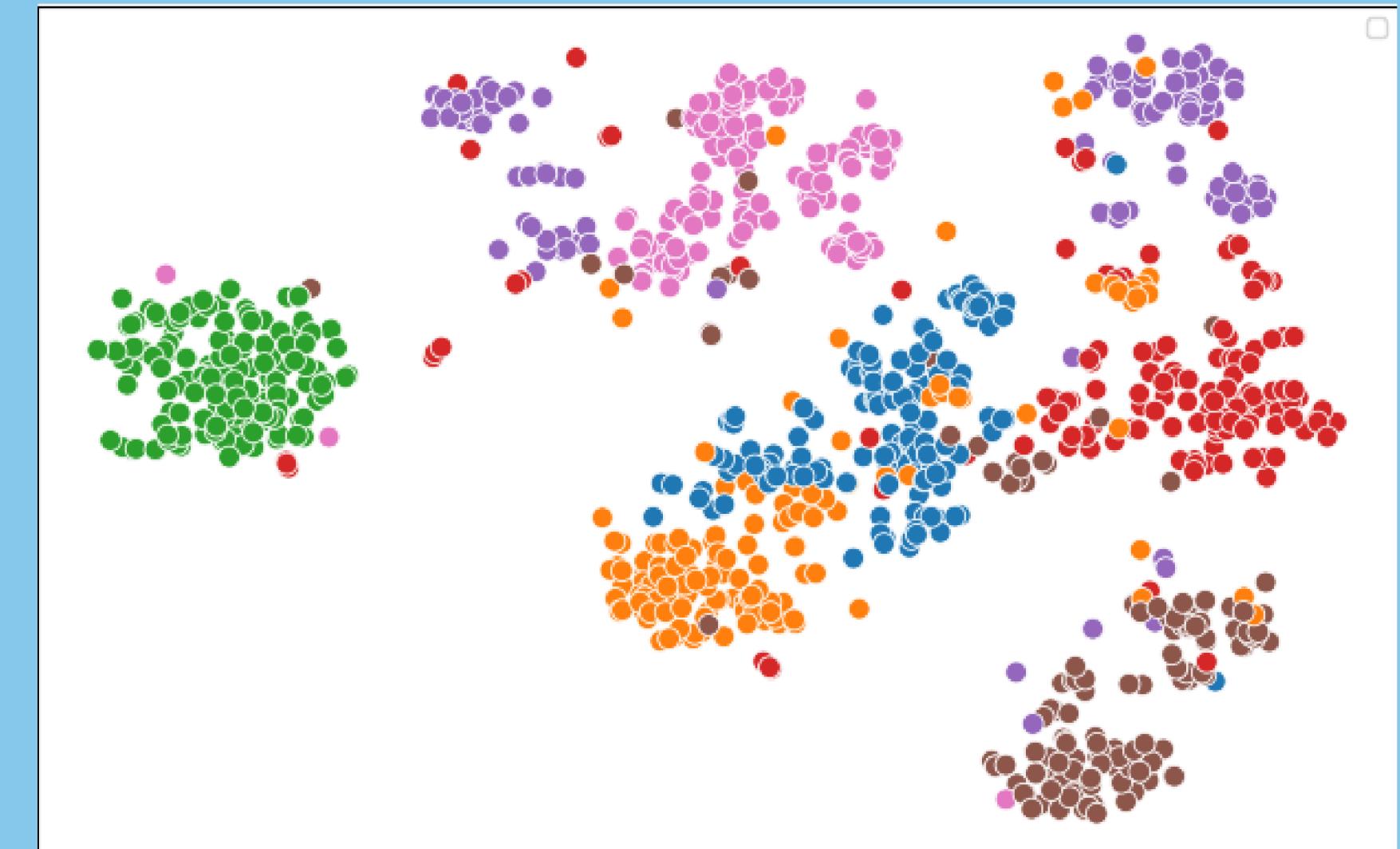
Etude de faisabilité

en combinant photo et description

USE + DenseNet121

	pred_2	pred_1	pred_4	pred_0	pred_3	pred_5	pred_6
cat							
Watches	150	0	0	0	0	0	0
Computers	2	134	1	0	0	13	0
Beauty and Personal Care	1	7	123	4	1	13	1
Home Furnishing	0	0	1	121	1	3	24
Kitchen & Dining	0	27	3	0	119	1	0
Home Decor & Festive Needs	7	3	1	9	23	107	0
Baby Care	0	6	3	21	9	10	101

Matrice de confusion



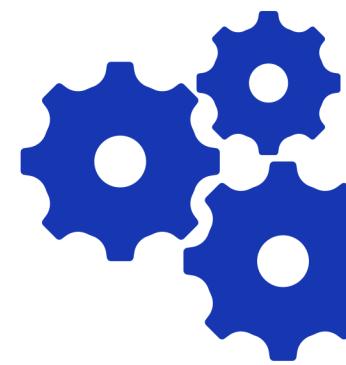
Projection 2D

Score ARI : 0.63

Classification

en utilisant les photos

Mise en place de l'algorithme



Choix des métriques

- Accuracy score
- Score de Cohen kappa
- Temps d'exécution

Prétraitement

- Redimensionnement
- Prétraitement automatique via Tensorflow

Choix du type d'entraînement

- Fine-tuning partiel
- Fine-tuning total

Sélection du modèle

- EfficientNet, MobileNetV2, DenseNet121, VGG16

Ajustement des hyperparamètres

- Calque Dropout
- EarlyStopping
- Batch size

Résultats avec un CNN from scratch

Watches	39	1	1	1	1	1	1	6
Kitchen & Dining	2	34	4	2	1	3	4	
Computers	2	10	27	3	2	3	2	
Home Furnishing	3	1	2	24	3	6	11	
Beauty and Personal Care	0	8	4	4	21	4	8	
Baby Care	1	7	1	10	3	16	11	
Home Decor & Festive Needs	1	13	5	6	4	5	16	

Accuracy score : 0.51

Score de Cohen kappa : 0.43

Temps d'exécution par image : 3 ms

Résultats avec Xception

Watches	49	0	0	0	0	0	0	1
Computers	0	45	3	0	1	0	0	0
Kitchen & Dining	1	2	44	0	1	0	0	2
Beauty and Personal Care	0	2	3	37	4	2	2	1
Home Furnishing	0	0	0	0	36	11	3	3
Baby Care	0	3	2	1	6	34	3	3
Home Decor & Festive Needs	0	2	2	0	3	10	33	

Accuracy score : 0.80

Score de Cohen kappa : 0.77

Temps d'exécution par image : 100 ms

Résultats avec MobileNetV2

Watches	48	2	0	0	0	0	0	0
Computers	1	41	1	0	3	2	1	
Home Furnishing	0	0	41	0	0	3	6	
Kitchen & Dining	3	3	0	41	2	0	1	
Beauty and Personal Care	0	3	2	0	40	3	1	
Home Decor & Festive Needs	1	1	4	1	3	36	4	
Baby Care	0	1	10	1	2	5	30	

Accuracy score : 0.80

Score de Cohen kappa : 0.76

Temps d'exécution par image : 28 ms

Modification des catégories



Home Furnishing	150	Watches	150
Baby Care	150	Baby Care	110
Watches	150	Home Furnishing	93
Home Decor & Festive Needs	150	Beauty and Personal Care	86
Kitchen & Dining	150	Kitchen & Dining	76
Beauty and Personal Care	150	Coffee Mugs	74
Computers	150	Combos	64
		Computers	63
		Showpieces	60
		Blankets, Quilts & Dohars	57
		Home Decor & Festive Needs	56
		Routers	49
		Baby Girls' Clothes	40
		USB Gadgets	38
		Ethnic	34

Résultats avec MobileNetV2

Watches	48	0	0	0	1	0	0	0	0	0	1	0	0	0	0
Home Furnishing	0	25	0	0	0	1	0	2	3	0	0	0	0	0	0
Coffee Mugs	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0
Combos	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0
Beauty and Personal Care	0	2	0	2	19	0	0	3	1	0	0	0	1	0	0
Computers	0	0	0	0	2	16	0	0	1	2	0	0	0	0	0
Kitchen & Dining	0	2	0	0	2	3	16	1	0	0	0	0	0	0	1
Baby Care	0	14	0	0	1	0	1	15	1	0	1	2	0	0	1
Blankets, Quilts & Dohars	0	3	0	0	0	0	0	2	13	0	1	0	0	0	0
Routers	0	0	0	0	0	3	0	0	0	13	0	0	0	0	0
Showpieces	0	4	0	0	0	1	0	0	0	0	12	0	0	2	1
Baby Girls' Clothes	0	0	0	0	0	0	0	3	1	0	0	9	0	0	0
USB Gadgets	0	0	1	0	1	2	0	1	0	0	0	0	8	0	0
Ethnic	0	0	0	0	2	1	0	0	0	0	5	0	0	3	0
Home Decor & Festive Needs	2	8	0	0	1	2	0	0	1	0	5	0	0	0	0

Accuracy score : 0.67

Score de Cohen kappa : 0.65

Temps d'exécution par image : 28 ms

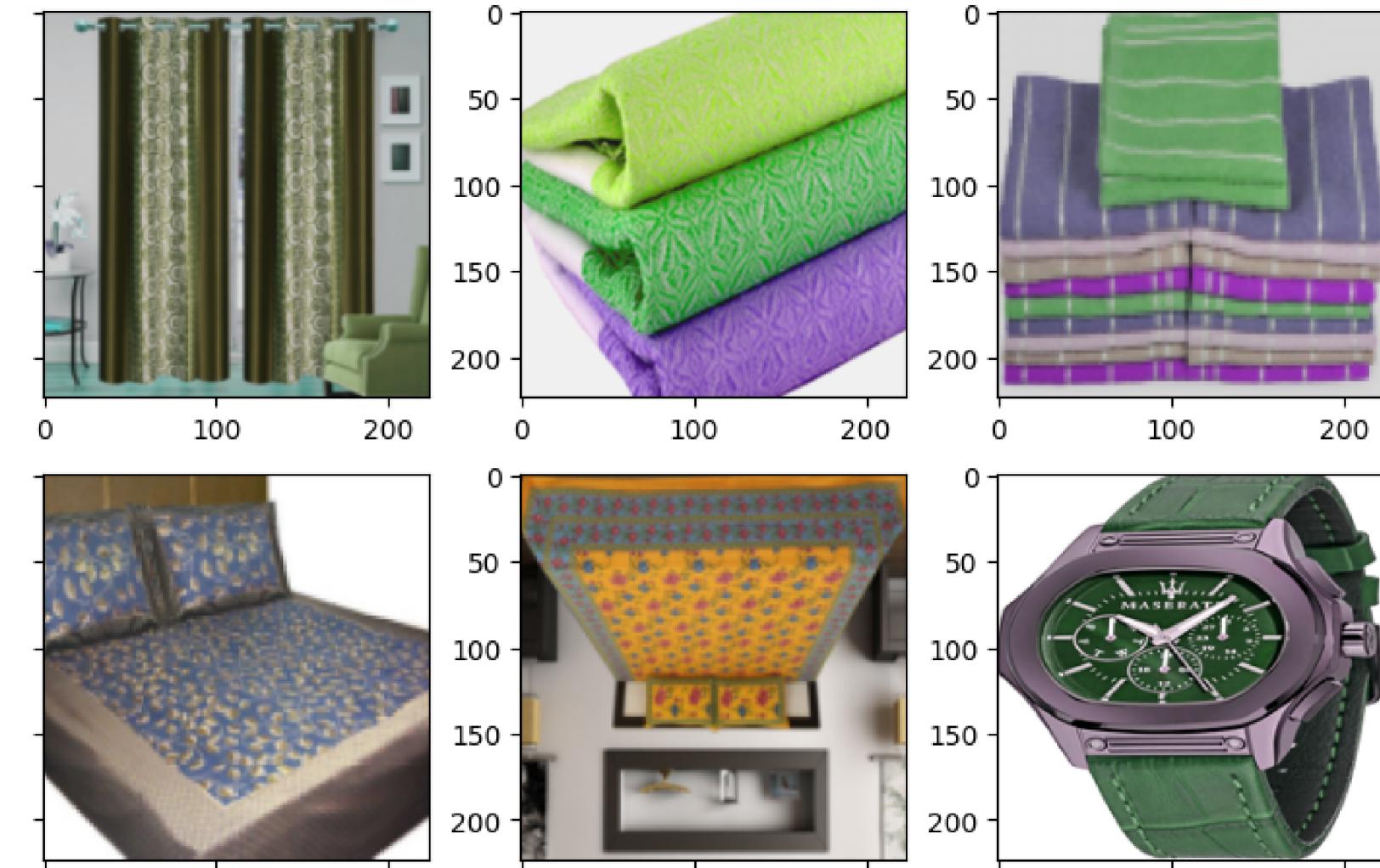
Data Augmentation

Augmentation

avec Albulmentations

01 Application de transformations

Recadrage, rotation, pixellisation, décalage couleur, fliping



02 Choix de la proportion d'images augmentées

Amener chaque catégorie à 58 image minimum

Varier les images en en ajoutant 58 systématiquement

03 Ajustement d'hyperparamètres

Batch size et Validation split à modifier

Résultats avec MobileNetV2 augmenté

Watches	49	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Coffee Mugs	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0
Home Furnishing	0	0	24	0	0	3	0	1	0	2	0	1	0	0	0
Beauty and Personal Care	2	0	0	21	1	1	1	0	0	0	0	1	0	1	0
Combos	0	0	0	0	20	0	0	1	0	0	0	0	0	0	0
Baby Care	0	0	5	2	0	19	0	0	0	1	2	4	2	1	0
Kitchen & Dining	0	1	1	0	1	0	18	0	0	0	0	1	0	2	1
Computers	1	0	1	1	0	0	1	12	2	0	0	1	0	2	0
Routers	0	0	0	0	0	0	0	1	12	0	0	0	0	3	0
Blankets, Quilts & Dohars	0	0	6	1	0	1	0	0	0	10	1	0	0	0	0
Showpieces	0	0	2	1	0	0	2	2	0	0	10	2	0	0	1
Home Decor & Festive Needs	0	0	2	1	0	0	2	0	0	1	2	8	0	1	2
Baby Girls' Clothes	0	0	0	0	0	6	0	0	0	0	0	0	7	0	0
USB Gadgets	0	0	0	0	0	0	2	1	1	0	0	2	0	7	0
Ethnic	1	0	0	0	0	0	1	0	0	0	2	0	0	2	5

Accuracy score : 0.71

Score de Cohen kappa : 0.68

Temps d'exécution par image : 27 ms

Utilisation d'une API

pour récupérer plus de produits

Agrandir le dataset

avec requests

01 API

Edamam Food and Grocery Database
Rapid API

02 Librairie

Utilisation de requests (plus haut niveau et moderne que http.client)

03 Respect du RGPD

- Respecter le consentement des personnes
- Respecter le principe de minimisation
- Respecter les durées de conservations
- Respecter les demandes d'exercice des droits des personnes concernées
- Respecter la sécurité et la confidentialité

Conclusion