Soutenance Projet 6

Classifiez automatiquement des biens de consommation

Étudiant: Hang ZHONG
Évaluateur: Kezhan SHI
Mentor : Arthur MELLO
Date: 18/05/2021

OPENCLASSROOMS

- 5 min
- 15 min
- 5 min
- 5-10 min

I. Problématique et Présentation du jeu de données

II. Prétraitements et Clustering

III. Conclusion et Recommandation

IV. Questions-réponses

Partie I

Problématique

Contexte

- 'Place de Marché' souhaite lancer une Marketplace e-commerce.
- Vendeur -> Acheteur :
  - Photo
  - Description

Problématique

- Attribuer la catégorie manuellement ❌
- Pour les vendeurs: faciliter la mise en ligne de nouveaux articles
- Pour les acheteurs: faciliter la recherche de produits
- Automatiser la classification par un moteur ✅

Etude de faisabilité

- Description de produit (**NLP**):
  - Prétraitement
  - tfidf, bag of words, NMF, LDA
- Photo de produit (**Traitement d'image**):
  - Prétraitement
  - SIFT, ORB, CNN
- Combinaisons et clustering

Partie I

Présentation du jeu de données

- 1 fichier csv + 1 dossier des images :
  - 1.7MB + 368MB
  - 1050 articles: catégorie, description
  - photo de produit
  - 7 catégorie x 150 articles:

    'Home Furnishing ', 'Baby Care ', 'Watches ', 'Home Decor & Festive Needs ', 'Kitchen & Dining ', 'Beauty and Personal Care ', 'Computers '

Ic4718ae90f2889...a99ee1cc106c.jpg
2 800×1 488

0ca8e323551dd71...57ef7c3e77aee.jpg
1 400×1 440

1e8741b5ae27a51...c94b3f3312aee.jpg
719×1 145

1eda39f01d0a8a2...4b32fc7da1027.jpg
220×464

Partie II Texte

1. Prétraiter des données texte

2. Feature engineering

Nettoyage(A->a, !?, 123) Tokenization

Normalisation: Lemma + stemming

Stopwords, mot court, doc_freq

Entrée: Jack klein BlackLed Digital Watch - For Boys - Buy Jack klein BlackLed Digital Watch - For Boys BlackLed Online at Rs.150 in India Only at Flipkart.com. - Great Discounts, Only Genuine Products, 30 Day Replacement Guarantee, Free Shipping. Cash On Delivery!

Sortie: ['digit', 'watch', 'boy', 'digit', 'watch', 'boy', 'discount']

Entrée: Buy Binatone WR3000N only for Rs. 1800 from Flipkart.com. Only Genuine Products. 30 Day Replacement Guarantee. Free Shipping. Cash On Delivery!

Sortie: []

Partie II Texte

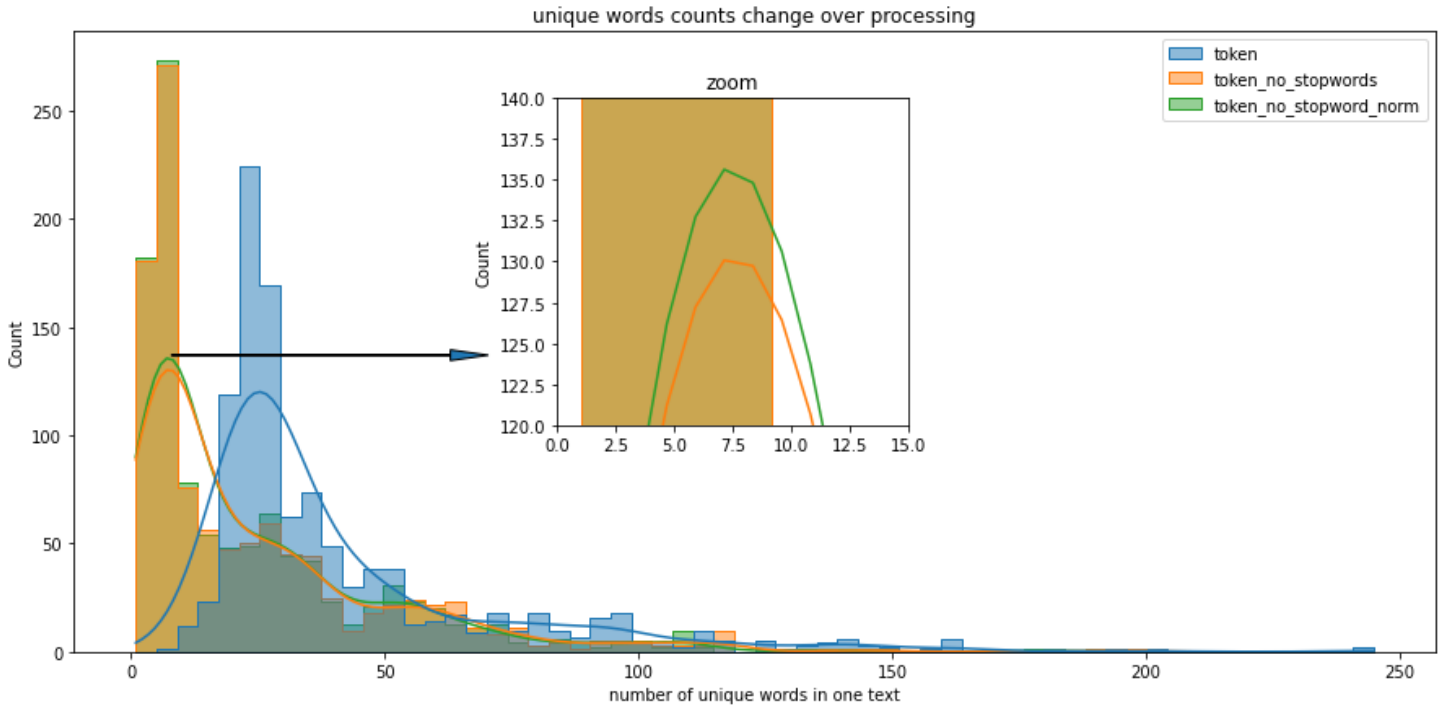
1. Prétraiter des données texte

2. Feature engineering

Nettoyage(A->a, !?, 123)
Tokenization

Normalisation:
Lemma + stemming

Stopwords, mot court, doc_freq

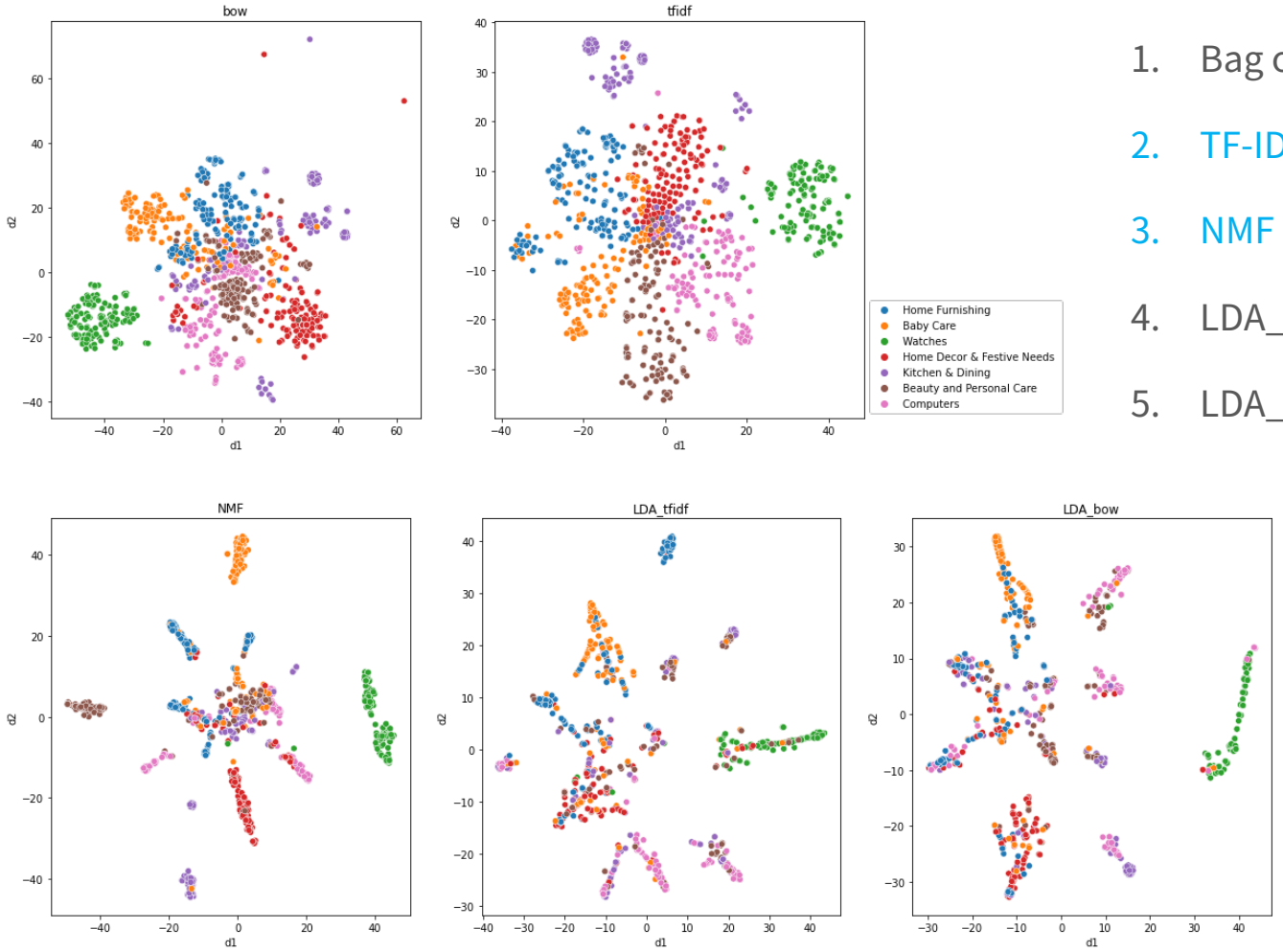


Réduiction de nombre de mots uniques:
5023(original)->4770(stopword)->3754(normalisation)->**1016**(filtre par doc_freq)

Partie II Texte

1. Prétraiter des données texte

2. Feature extraction



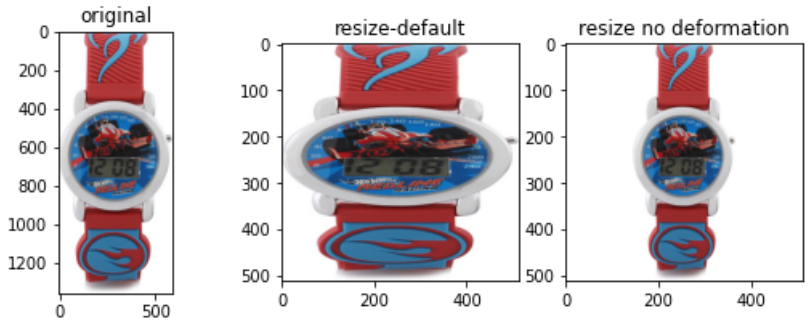
1. Bag of Words
2. TF-IDF
3. NMF
4. LDA_tfidf
5. LDA_BoW

Résultat de la comparaison: TF-IDF > Bag of Words
NMF > LDA

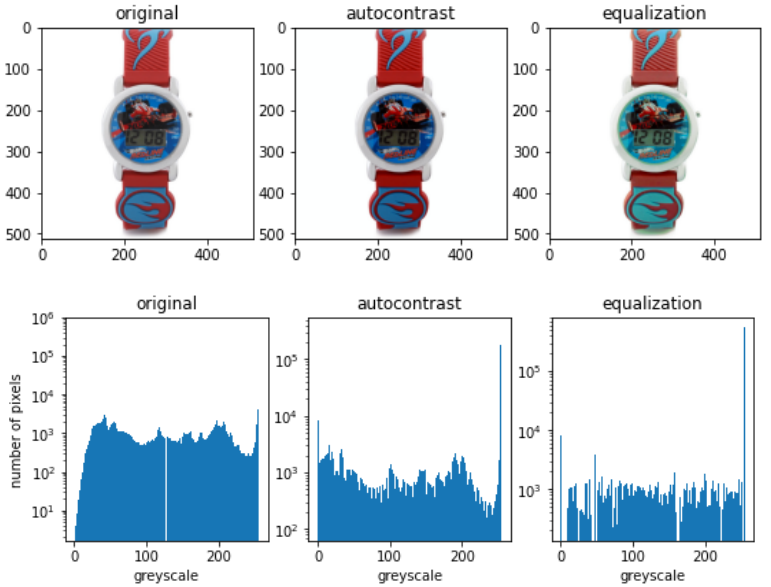
Partie II Photo

1. Prétraiter des images

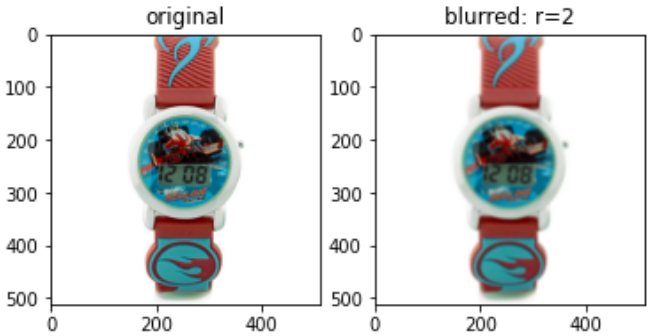
2. Feature extraction



Redimensionner



Auto-contrast et Equilisation



Débruitage

Partie II Photo

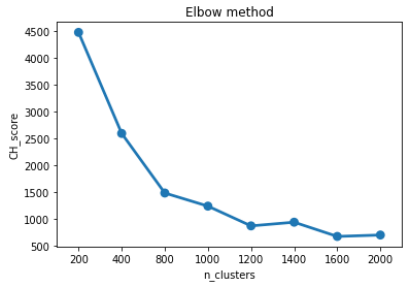
1. Prétraiter des images

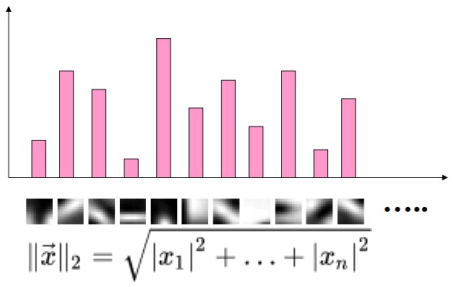
2. Feature extraction

Bag of features (SIFT & ORB) :

1. Calcul des descripteurs
2. Clustering des descripteurs
3. Histogramme des visual words, normaliser
4. Output: (1050, 1200)

SIFT: (1104377, 128)
ORB: (510018, 32)





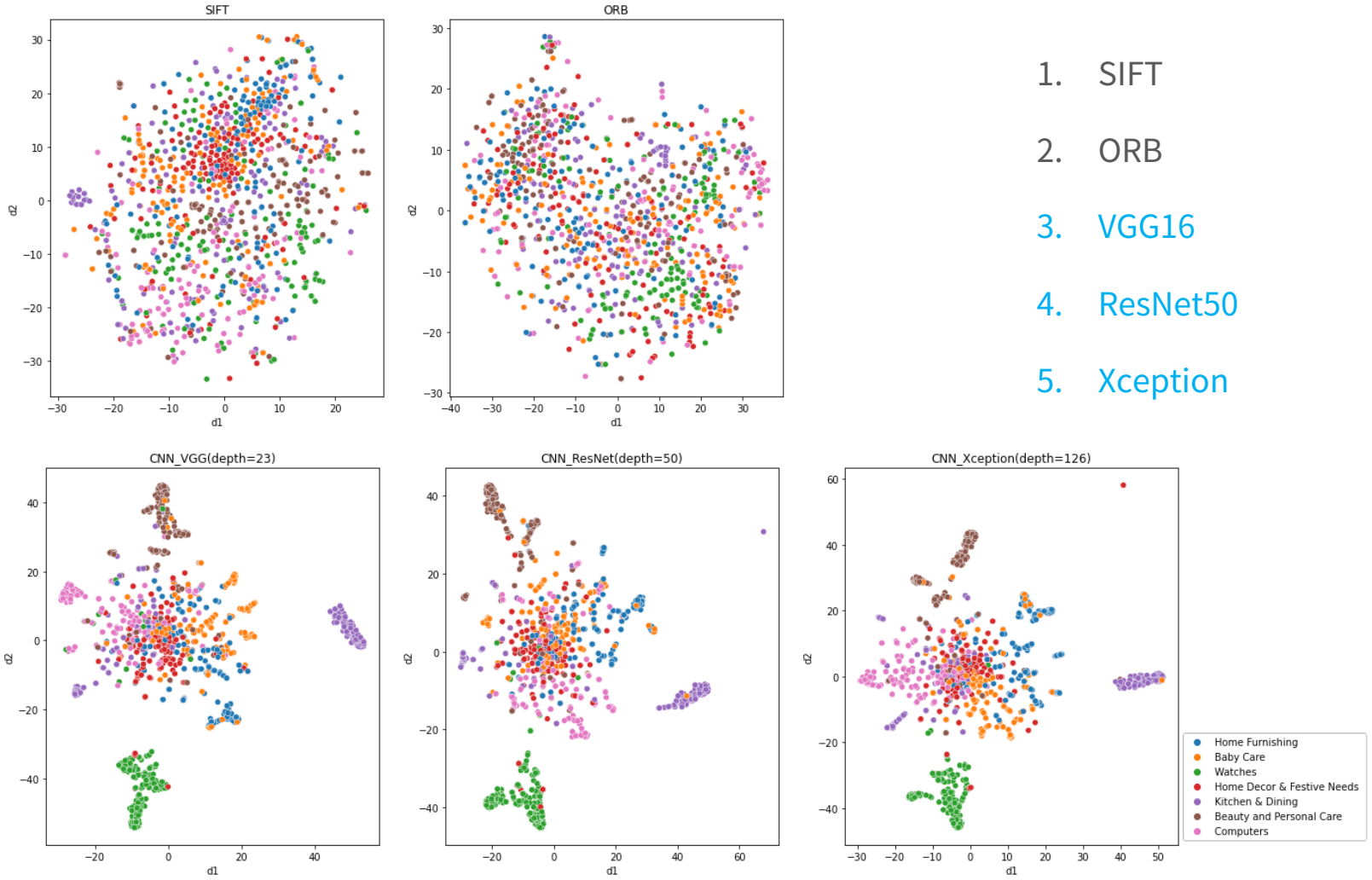
Bag of features (CNN) :

1. VGG16, ResNet50, Xception
2. Prétraiter Redimensionner,
3. Prédire directement
4. Output: (1050, 1000)

Partie II Photo

1. Prétraiter des images

2. Feature extraction



1. SIFT
2. ORB
3. VGG16
4. ResNet50
5. Xception

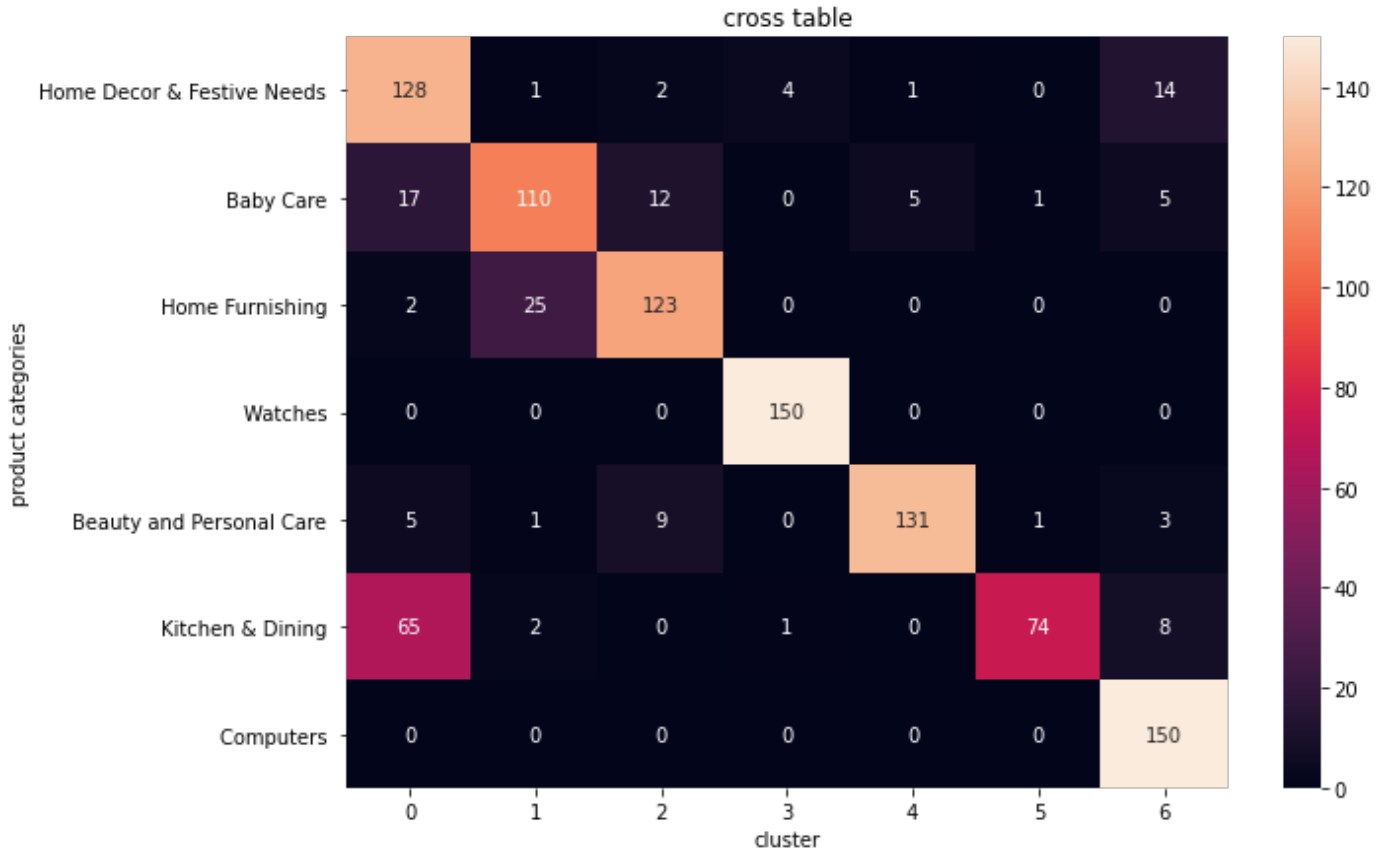
Résultat de la comparaison: CNN > SIFT et ORB

Partie II Combi

1. Best combi
2. Classification erronée
3. Amélioration

Clustering Hiérarchique –> hyper paramètres -> coefficient de Gini

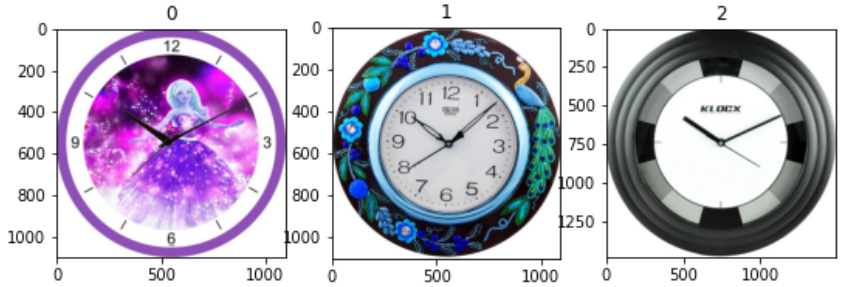
gini_score	VGG16_PCA	ResNet50_PCA	Xception_PCA
NMF	0.32	0.46	0.37
TF-IDF_PCA	0.21	0.16	0.14 ✅



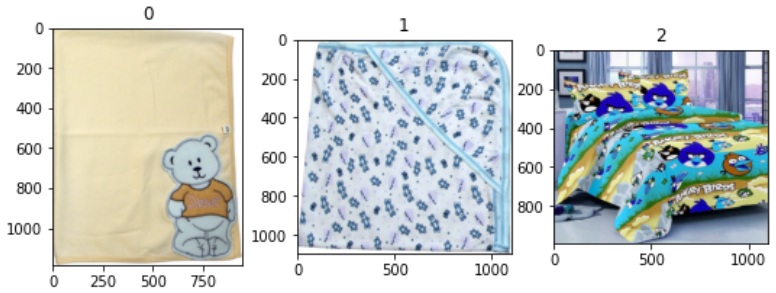
Partie II Combi

1. Best combi
2. Classification erronée
3. Amélioration

Pseudo-mauvaise classification

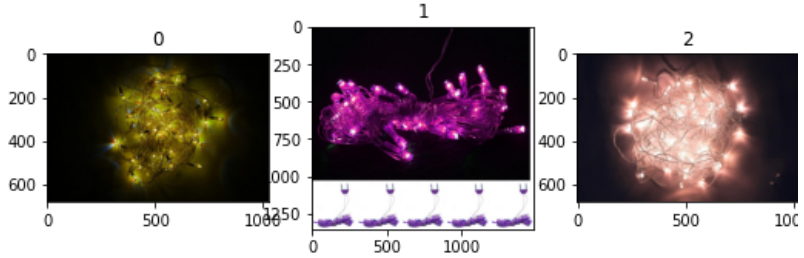


«home decoration» classées à tort dans «montres»

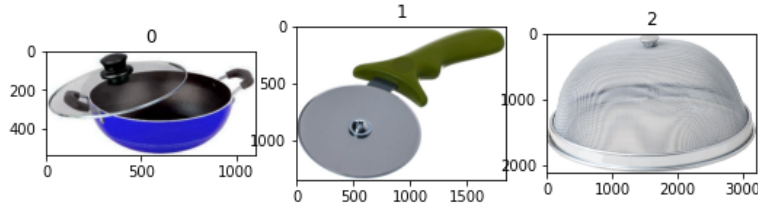


«Babycare» classées à tort dans «home furnishing»

Réelle-mauvaise classification



«home decoration» classées à tort dans «ordinateur»



«kitchen» classées à tort dans «home decoration»

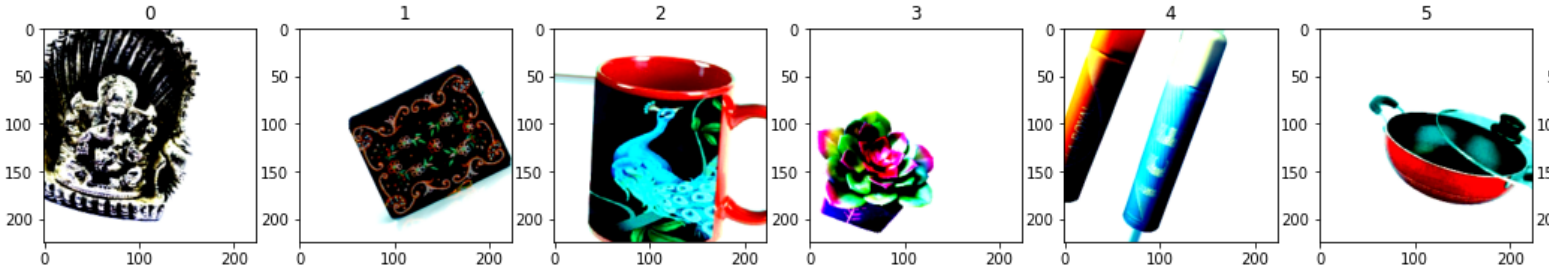
Partie II Combi

1. Best combi
2. Classification erronée
3. Amélioration

Transfer-Learning: fine-tuning partiel

1. Remplace la dernière couche fullly-connected (1000 classes -> 7 classes)
2. Fixe tous les autres paramètres pré-entraîné
3. Entraîner la couche non-fixée sur nos propres images
   (70% images originales + augmentés)

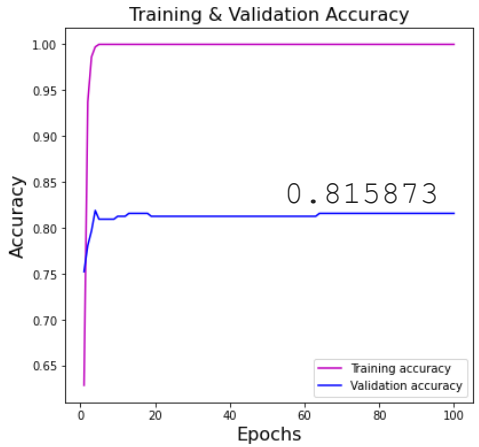
Data augmentation



Partie II Combi

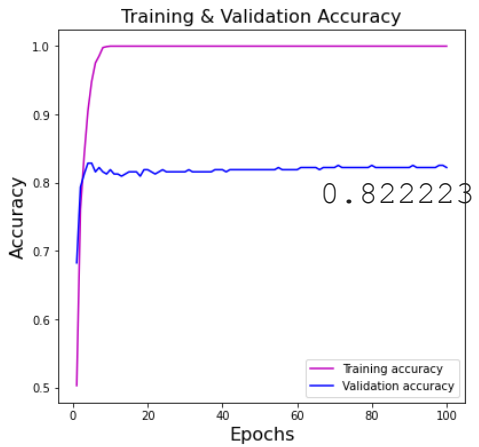
1. Best combi
2. Classification erronée
3. Amélioration

Amélioration Data augmentation



Entraîné sur les images originales

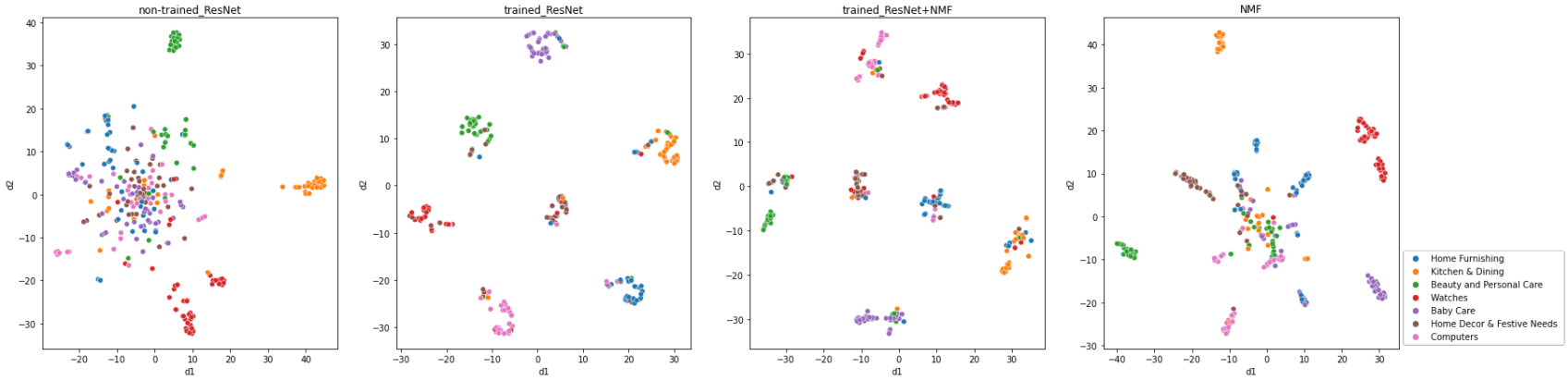
1%



Entraîné sur les images originales + augmentés

1. Underfitting
2. 1% d'amélioration

Amélioration Transfer-Learning (sur 30% test données)



Partie III

Conclusion et Recommandation

Conclusion:

Monteur de classification avec une description et une photo est tout à fait faisable

Recommandations:

1. CNN Transfer-Learning: Entraîner le model avec plus d'images
2. Essayer les autres méthodes pour représenter les donnés textuelles
3. Pondération sur des features textuelles ou des features en vision

Partie IV

Q&A

MERCI