Classifiez automatiquement des biens de consommation

Soutenance du 1er août 2023

Yann Pham-Van

Classifiez automatiquement des biens de consommation

- Problématique
- Etude de faisabilité
- Classification supervisée
- Test de l'API
- Conclusion

Classifiez automatiquement des biens de consommation

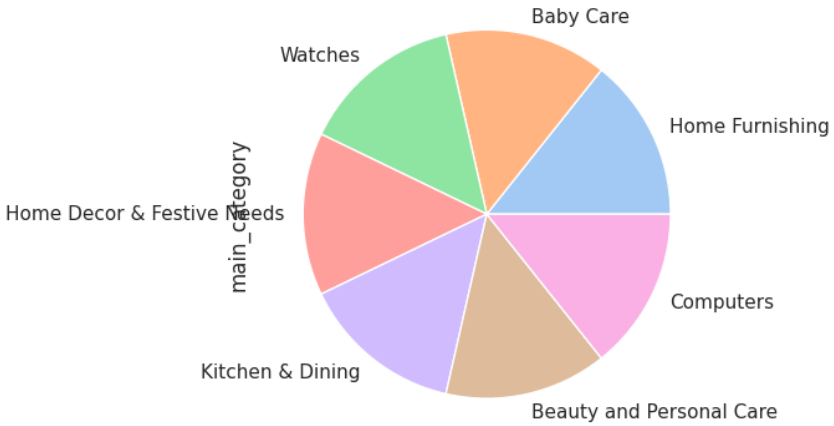
- **Problématique**
- Etude de faisabilité
- Classification supervisée
- Test de l'API
- Conclusion

Problématique

Qui fait quoi ?

QUOI	QUI			
	Aujourd'hui		Demain	
	Vendeur	Place de marché	Vendeur	Place de marché
Rédiger une description	X		X	
Poster une photo	X		X	
Choisir une catégorie	X			X
Objectifs : fiabilité, fluidité				

Problématique > jeu de données



- 1050 observations
- 15 variables
- 7 catégories principales
- 150 observations/catégorie
- 4 variables retenues
  - Nom
  - Description
  - Image
  - Catégorie

Classifiez automatiquement des biens de consommation

- Problématique
- **Etude de faisabilité**
- Classification supervisée
- Test de l'API
- Conclusion

Etude de faisabilité > prétraitement texte

Opération	Approche			Exemple : Franck Bella FB127A Analog Watch - For Boys,
	BoW lem	BoW stem Word2Vec	BERT USE	
Minuscules	X	X	X	franck bella fb127a analog watch - for boys,
RegexpTokeninzer	X	X	X	franck bella fb127a analog watch for boys
Stop words 'english'	X	X		been ours you're haven isn't being more himself
Mots rares < 10	X	X		solar sunlast casserole tip pulse wi fi offers
Mots < 3 caractères	X	X	X	
Mots avec caractère numérique	X	X	X	
WordNetLemmatizer	X			
PorterStemmer		X		
English words	X	X		photon pettifogging cordyceps
210 mots les + présents toutes catégories	X	X		pack product buy free deliveri cash

Etude de faisabilité > prétraitement image

Approche SIFT

- Réduction de la résolution d'une seule image bloquante
- Conversion en valeurs de gris
- Egalisation du contraste

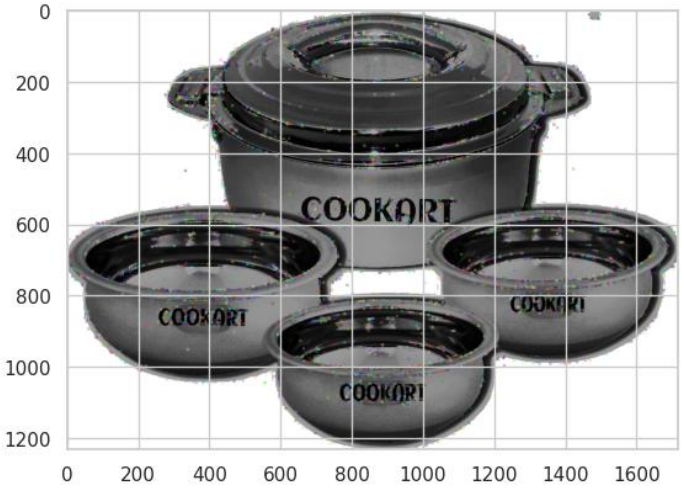
Etude de faisabilité > features textes

- Approches Bag of Words
  - CountVectorizer : fréquence d'apparition d'un mot dans un document
  - TfidVectorizer : fréquence ci-dessus rapportée à la fréquence dans le corpus
- Approches Word / Sentence embedding
  - Word2Vec
    - Dimension des vecteurs de mots = 400
    - Distance max entre mot actuel et mot prédit = 5
    - Nombre d'itérations sur le corpus = 100
    - Nombre de mots uniques = 643 → matrice embedding (643, 400)
  - BERT avec modèles types pré-entraînés

    bert-base-uncased, twitter-roberta, hub-TF-KerasLayer
  - USE

    Encode les sentences en un vecteur de 512 dimensions

Etude de faisabilité > features images



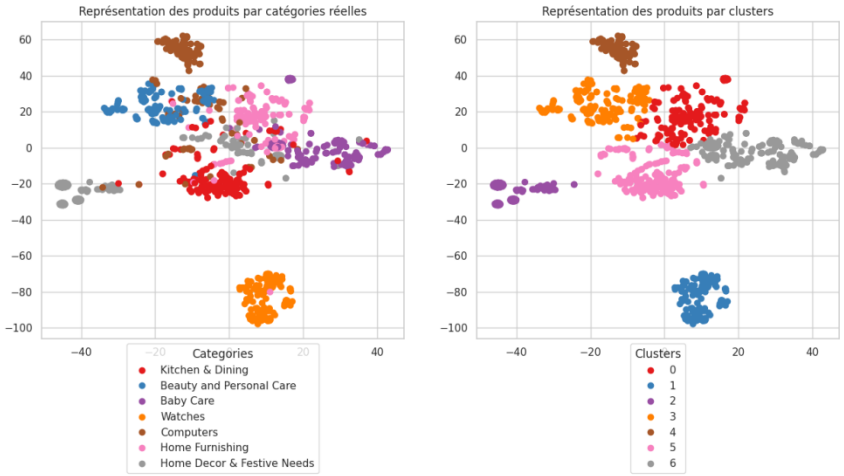
- Approche SIFT
  - 517 350 descripteurs sur vecteurs à 128 dimensions
  - 719 clusters de descripteurs → 719 features
  - 495 composantes après réduction ACP
- Approche CNN
  - 4096 features
  - 803 composantes après réduction ACP

Etude de faisabilité > résultats

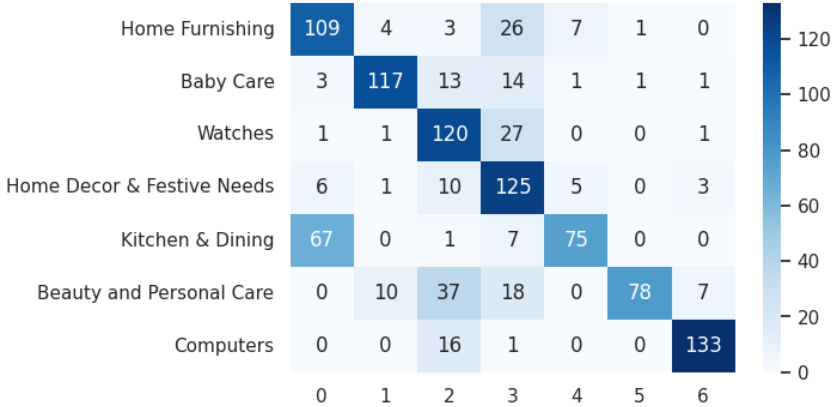
Score ARI des approches										
Textes									Images	
Bag of Words				Word embedding						
CountVectorizer		Tf-idf		Word 2Vec	BERT			USE	SIFT	CNN
lem	stem	lem	stem		base-uncased	Twitter-roberta	Hub-TF-Keras			
0.469	0.493	0.5	0.567	**0.604**	0.364	0.266	0.372	0.391	0.003	**0.483**

Etude de faisabilité > résultats

Word2Vec



CNN



Classifiez automatiquement des biens de consommation

- Problématique
- Etude de faisabilité
- **Classification supervisée**
- Test de l'API
- Conclusion

Classification supervisée

4 approches CNN avec VGG16

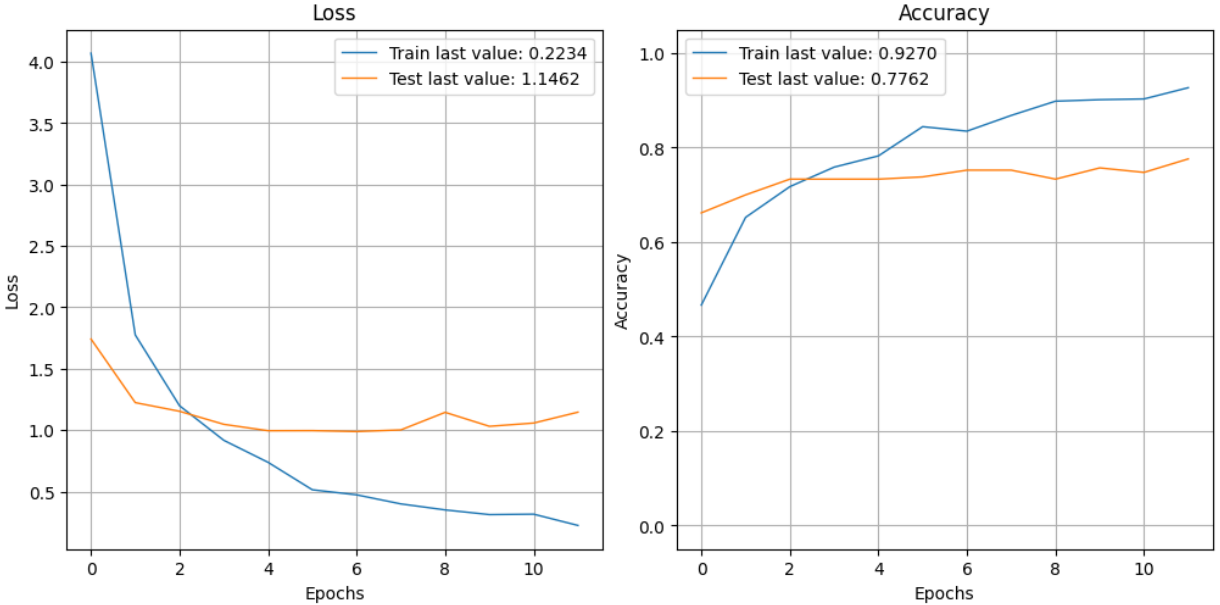
- préparation basique des images
- Image Data Generator avec data augmentation
- Image Dataset sans data augmentation
- Image Dataset avec data augmentation

Classification supervisée > résultats

Approches								
	Base		**Data Generator avec data augmentation**		**Image Dataset sans data augmentation**		**Image Dataset avec data augmentation**	
Batch-size	32	64	32	64	32	64	32	64
Best accuracy test	0.1476	0.1714	**0.819**	0.8048	0.8143	0.7857	0.7857	0.781
Temps de calcul (secondes)	296	359	556	746	237	**232**	367	626
Ratio temps/accuracy	2005	2095	679	927	**291**	295	467	802

Classification supervisée > résultats

Meilleure approche : Image Data Generator, batch-size = 32



Classifiez automatiquement des biens de consommation

- Problématique
- Etude de faisabilité
- Classification supervisée
- **Test de l'API**
- Conclusion

Test de l'API

- **Objectif**

  Collecte de données → nouvelle catégorie
- **Méthodologie**
  - Paramétrer code Python sur RapidAPI
  - Import JSON
  - Interpréter la structure des données
  - Transformer en DataFrame
  - Filtrer les champs pour respecter les 5 piliers du RGPD
  - Exporter en CSV

Test de l'API

10 produits « champagne »

	foodId	label	category	foodContentsLabel	image
0	food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods	NaN	https://www.edamam.com/food-img/a71/a718cf3c52...
1	food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	NaN
2	food_b3dyababjo54xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINE...	https://www.edamam.com/food-img/d88/d88b64d973...
3	food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS S...	NaN
4	food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	NaN
5	food_bmu5dmkazwuvpaa5prh1daa8jxs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFIT...	https://www.edamam.com/food-img/ab2/ab2459fc2a...
6	food_alpl44taoyv11ra0lic1qa8xculi	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne;...	NaN
7	food_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; Champagne; Peach	NaN
8	food_am5egz6aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	NaN
9	food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	NaN

Classifiez automatiquement des biens de consommation

- Problématique
- Etude de faisabilité
- Classification supervisée
- Test de l'API
- **Conclusion**

Conclusion

- Principe du moteur de classification validé
- Meilleur résultat faisabilité = Word2Vec
- Classification supervisée images satisfaisante
- Suggestion

  Ouvrir les prochaines catégories depuis un embedding dédié, associé à une approche CNN Image Dataset sans data augmentation pour le meilleur compromis performance et temps de calcul

Avez-vous des questions ?