Classifiez automatiquement des biens de consommation

. . .

Soutenance Projet 6
OpenClassrooms
Formation Data Scientist
17 Mai 2023

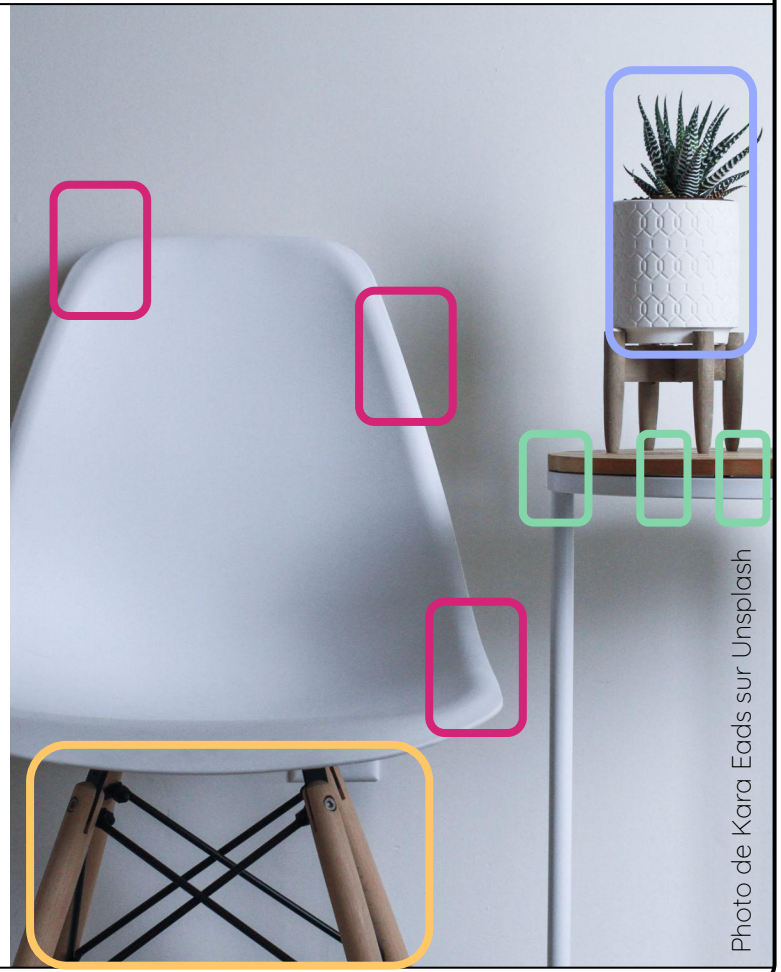


Photo de Kara Eads sur Unsplash

Marketplace e-commerce

Vendeur

Photo

Description

Catégorie

Est-il possible d'automatiser la catégorisation des articles ?

Plan soutenance : 3 missions

01 Étude de faisabilité d'un moteur de classification

02 Classification supervisée (images)

03 Test API

1050 articles - 15 colonnes

Descriptions produits

Noms produits

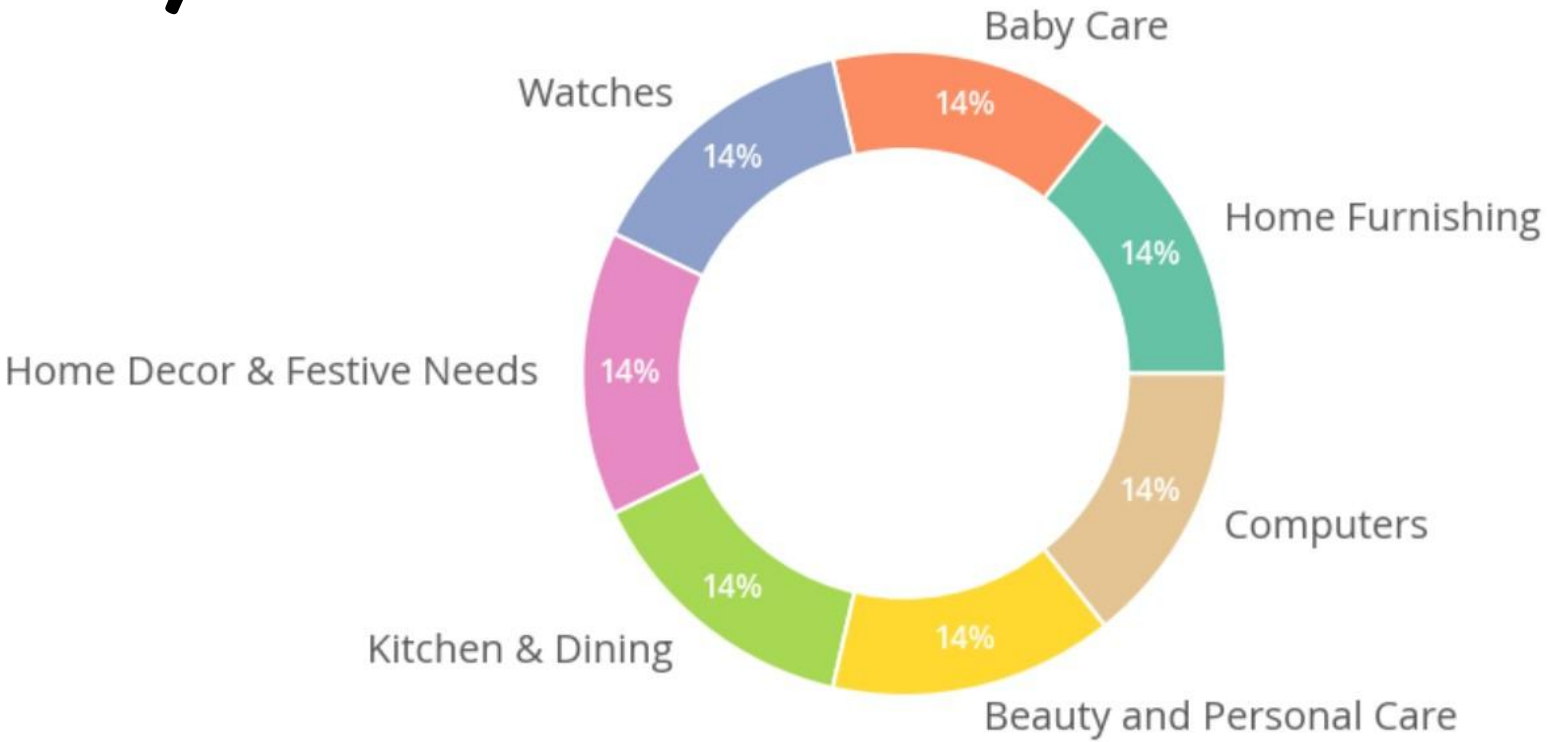
Noms fichiers image

Arbre catégories

* 0 contraintes de propriété intellectuelle

* 0 valeurs manquantes, 0 doublons

Analyse de la cible



*parfaitement équilibrée

Étude de faisabilité

Méthodologie

01 Prétraitement des données

02 Extraction des features

03 Réduction dimensionnelle

04 Clustering

05 Analyse graphique + ARI

Réduction de dimension x2

Features extraites

Réduction linéaire (ACP ou SVD)

Features pertinentes

Réduction non linéaire (t-SNE)

2D

Méthodologie

01 Prétraitement des données

02 Extraction des features

03 Réduction dimensionnelle

04 Clustering

05 Analyse graphique + ARI

Indice de rand ajusté

Évalue la concordance de deux partitions du jeu de données

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)}$$

0 — Clustering aléatoire

1 — Clustering correspond exactement à la partition originale

− Faisabilité du moteur de classification +

01 Prétraitement données texte

“Great Discounts and FREE shipping !!”

Minuscules, Ponctuation, Tokenisation	‘great’	‘discounts’	‘and’	‘free’	‘shipping’	✕
Suppression stop-words	‘great’	‘discounts’	✕	‘free’	‘shipping’	
Lemmatisation	‘great’	‘discount’		‘free’	‘shipping’	
Racinisation	‘great’	‘discount’		‘free’	‘ship’	

02 Extraction features : bag-of-words

Document -> Vecteur de la taille du vocabulaire

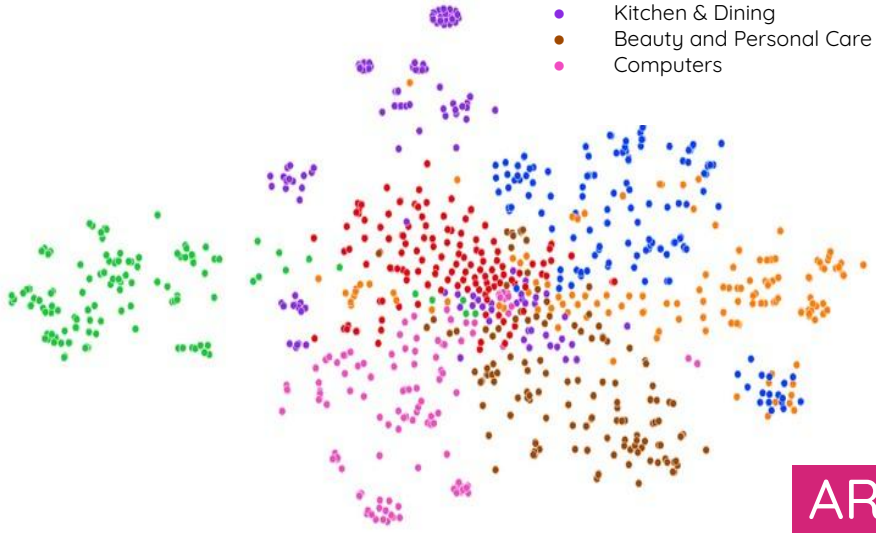
- TF : composante = $\frac{\text{nombre d'occurrences du token dans le doc}}{\text{nombre de tokens dans le doc}}$

- TF-IDF : composante = TF x $\frac{\text{nombre de documents}}{\text{nombre de docs contenant le token}}$

05 Descriptions en 2D (t-SNE) : tf.idf

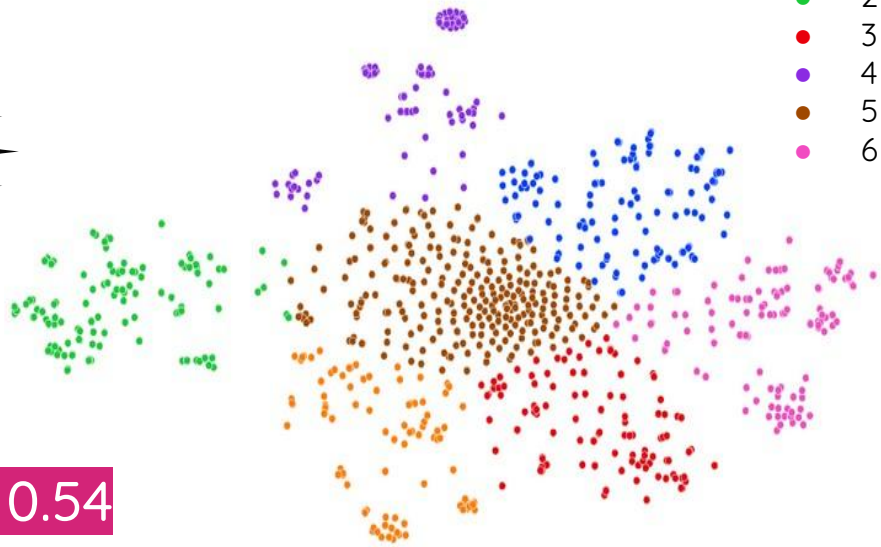
Vraies catégories

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



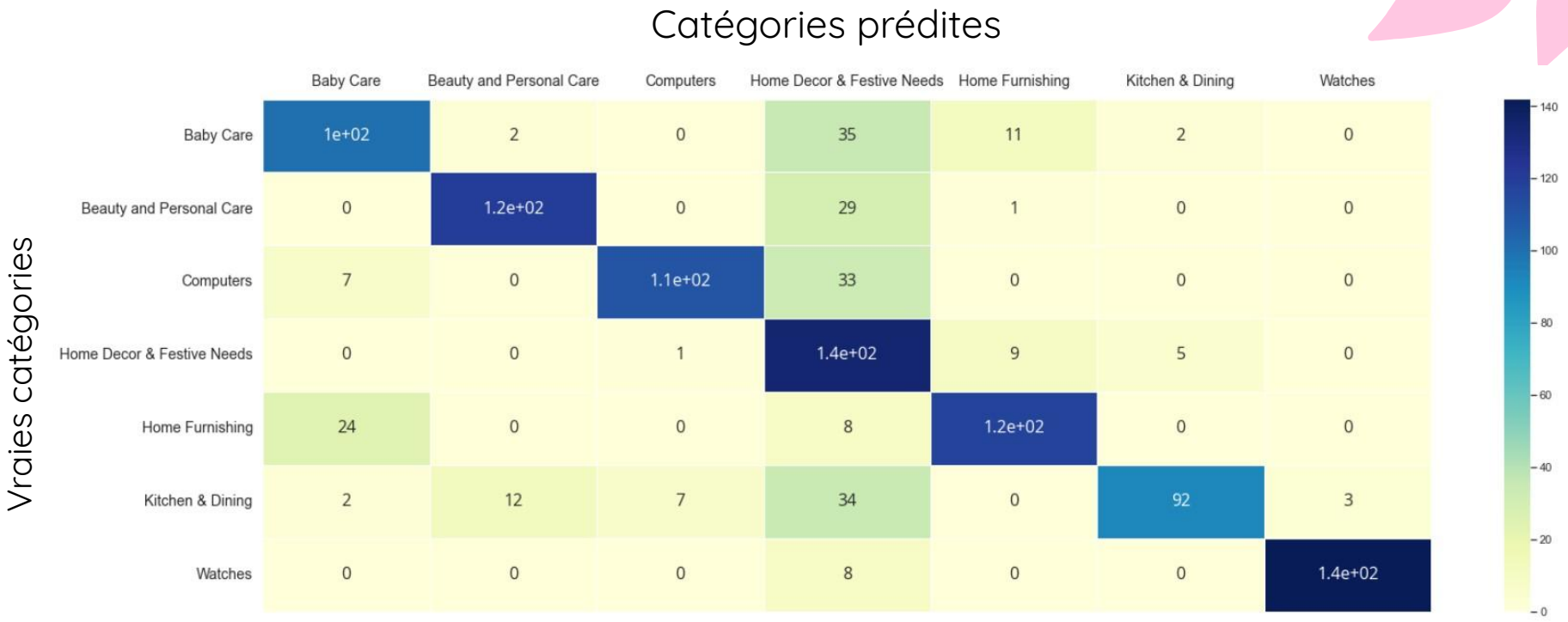
Clusters (K-means)

- 0
- 1
- 2
- 3
- 4
- 5
- 6



ARI = 0.54

Matrice de confusion - tf.idf



02 Extraction features : word embeddings

Token-> Vecteur de nombres réels qui conservent similarité et contexte

w2v

Word2Vec
(dim: 300)

Bert

Bidirectional Encoder Representation from Transformers
(dim: 768)

USE

Universal Sentence Encoder
(dim : 512)

Entraine un réseau de neurones

Utilise des modèles **pré-entraînés**

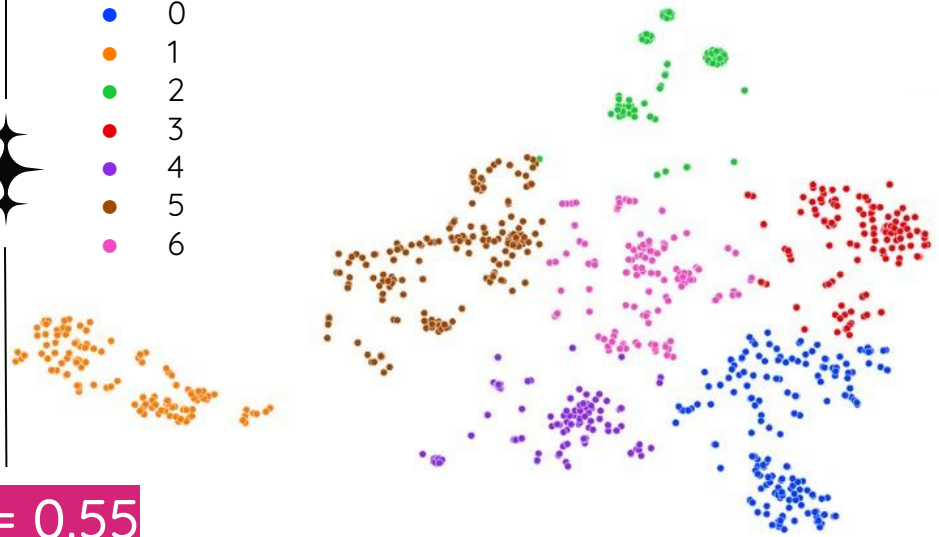
05 Descriptions en 2D (t-SNE) : w2v

Vraies catégories

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

Clusters (K-means)

- 0
- 1
- 2
- 3
- 4
- 5
- 6

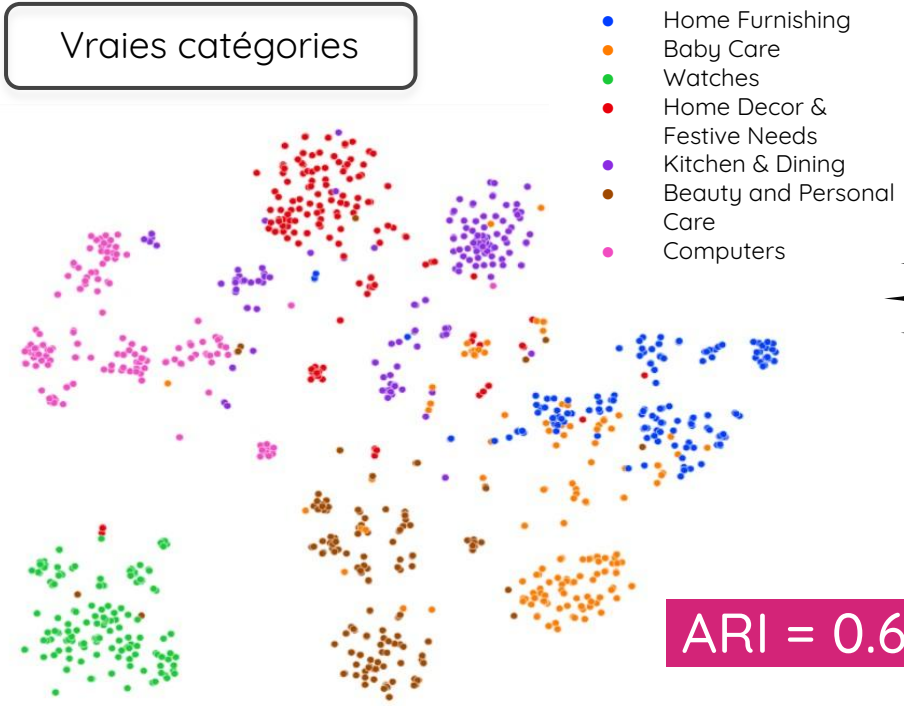


ARI = 0.55

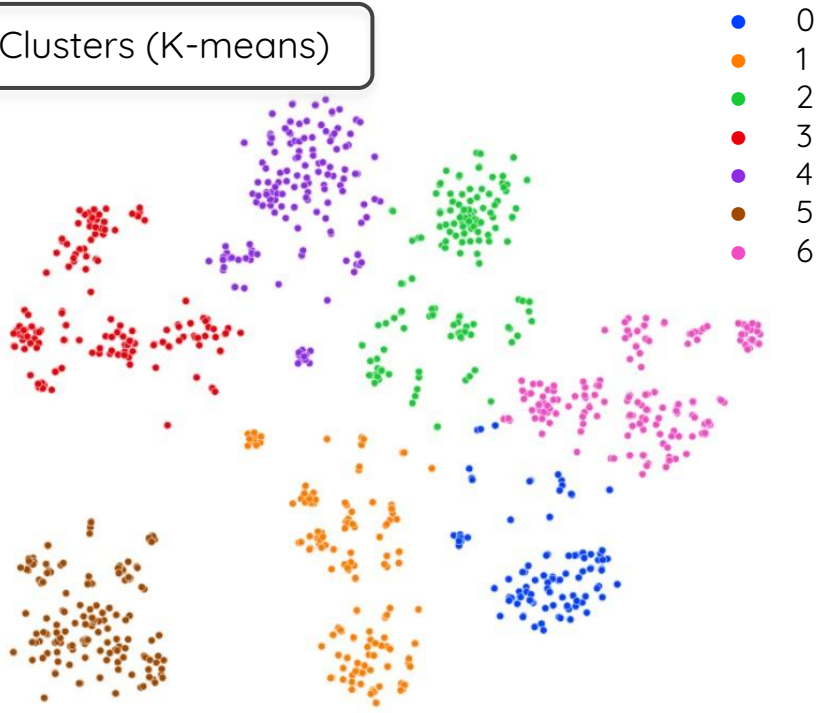
Matrice de confusion - w2v

05 Noms d'articles en 2D (t-SNE) : use

Vraies catégories

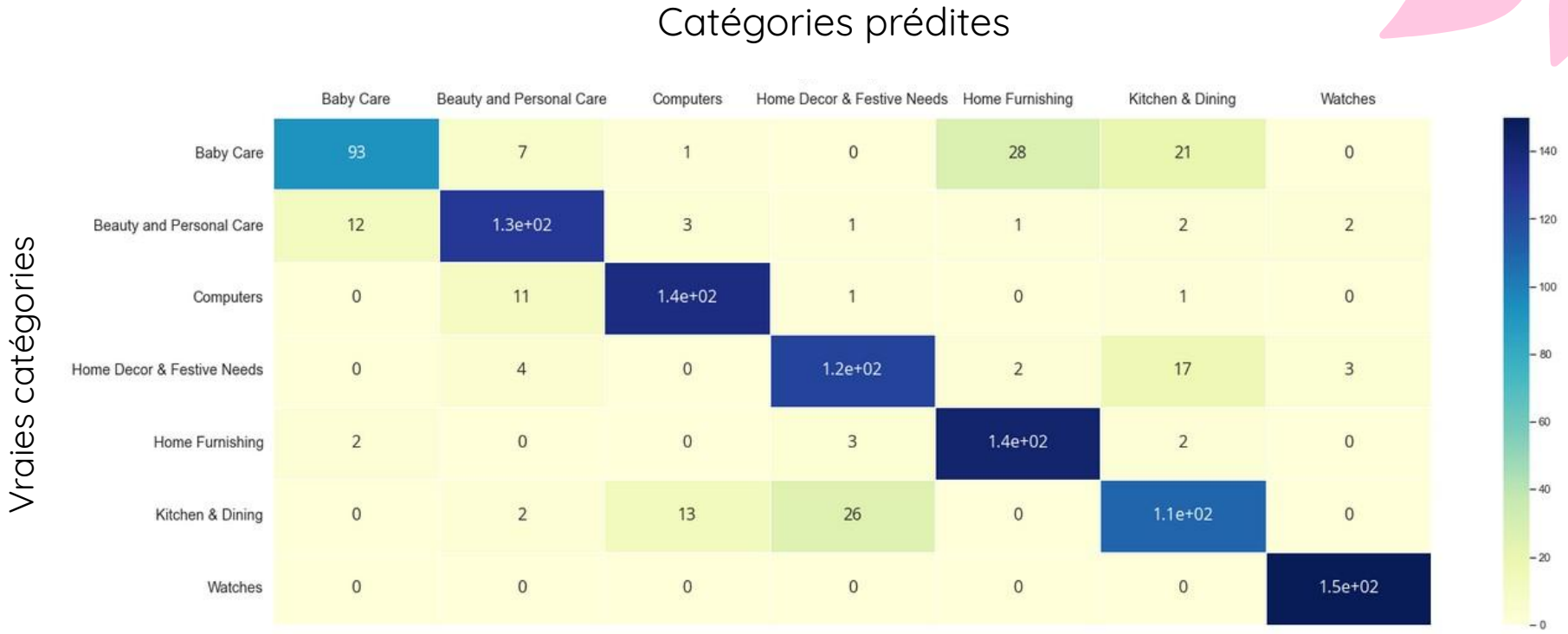


Clusters (K-means)



ARI = 0.69

Matrice de confusion - use



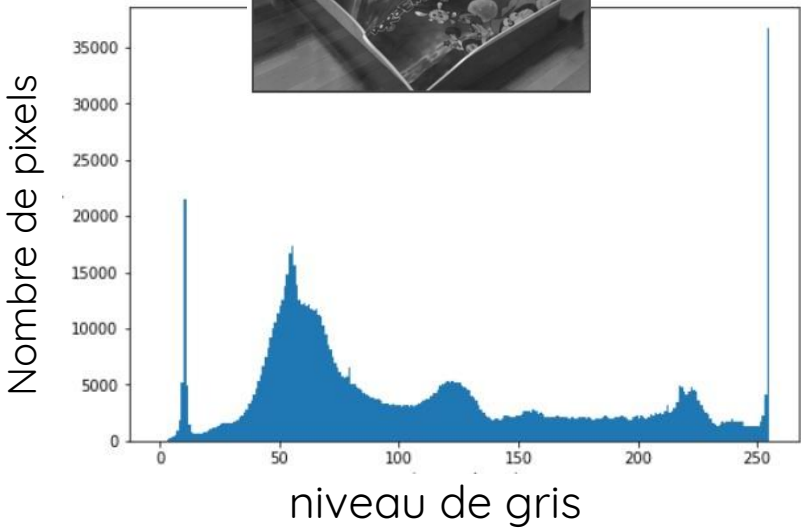
01 Prétraitement données image

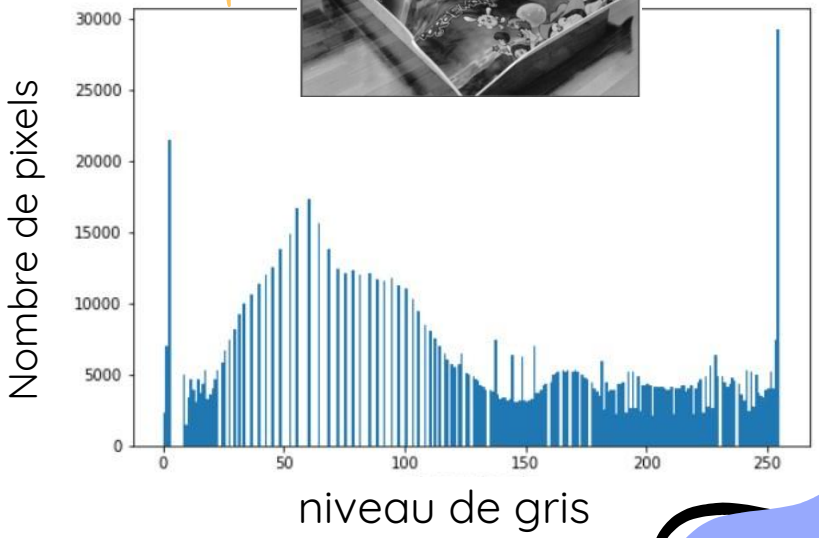
Passage au gris

Re-dimension

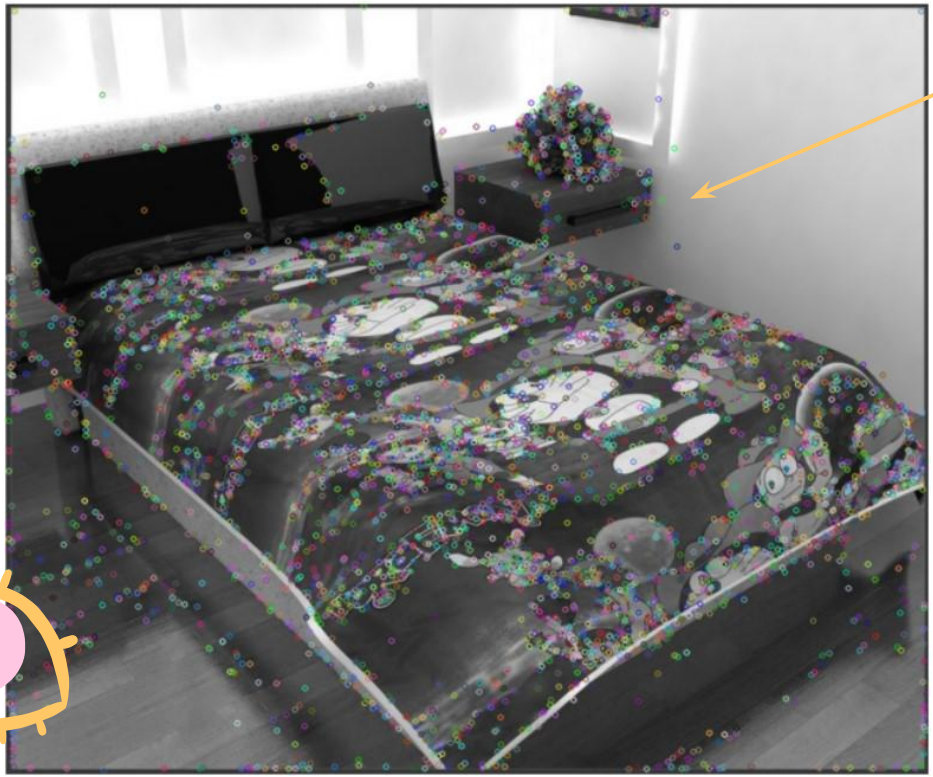
Amélioration contraste

Égalisation histogramme :





02 Extraction features : SIFT



01. Détection de **points d'intérêts**

02. Création de **descripteurs**

```
[[  0.  15. 132. ...   0.   0.   0.]
 [130.  21.   0. ...   0.   0.   0.]
 [131.  53.   1. ...   0.   0.   0.]
 ...
 [ 49.   3.   2. ...   0.   1.   8.]
 [  0.   0.   0. ...   0.   2.  12.]
 [  0.   0.   1. ...   0.   1.   3.]]
```

dim: nombre de points d'intérêts

dim: 128

Extraction features :

01.

1050 images

SIFT

dim: 128

```
[[  0.  15. 132. ...   0.   0.   0.]
 [130.  21.   0. ...   0.   0.   0.]
 [131.  53.   1. ...   0.   0.   0.]
 ...
 [ 49.   3.   2. ...   0.   1.   8.]
 [  0.   0.   0. ...   0.   2.  12.]
 [  0.   0.   1. ...   0.   1.   3.]]
```

1050 * 300 descripteurs

K-Means (train)

dim: 128

K visual words

K-Means (predict)

02.

K visual words

fréq d'apparition visual word dans image

normalise

Bag-of-features

02 Extraction features : transfer learning

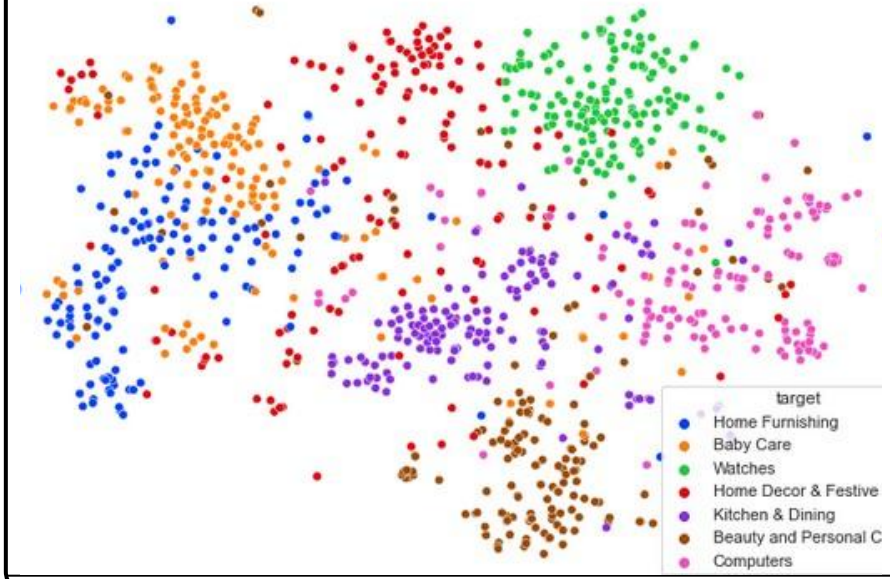
Utilise les features d'un modèle pré-entraîné pour représenter nos images

Vgg-16 ✦ ResNet-50

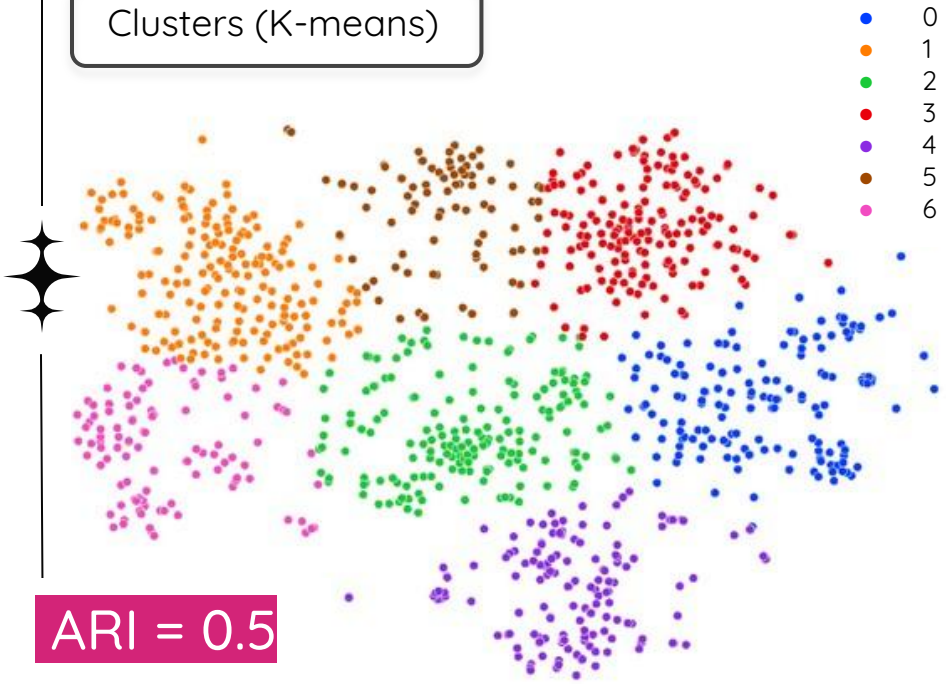
*Réseaux de neurones convolutifs

05 Images en 2D (t-SNE) : resnet.50

Vraies catégories

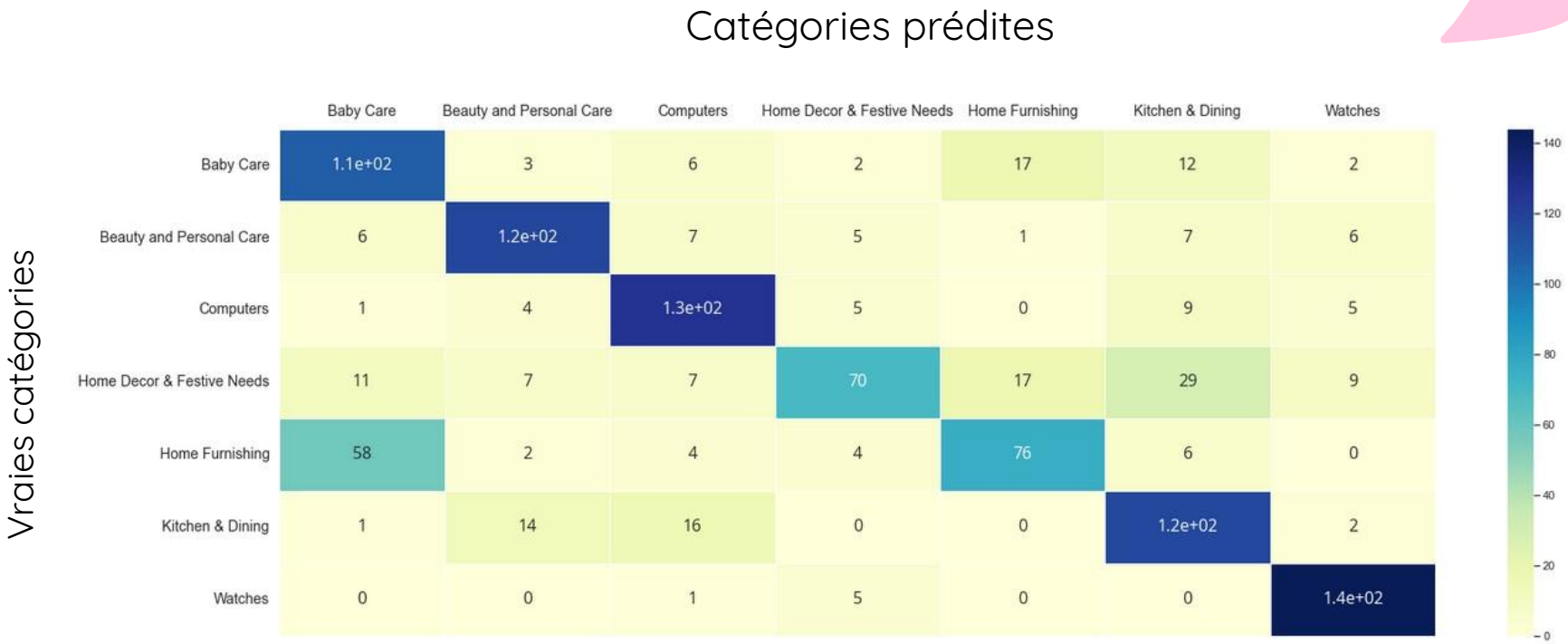


Clusters (K-means)



ARI = 0.5

Matrice de confusion - resnet.50



Articles “Home Furnishing” classés en “Baby Care”

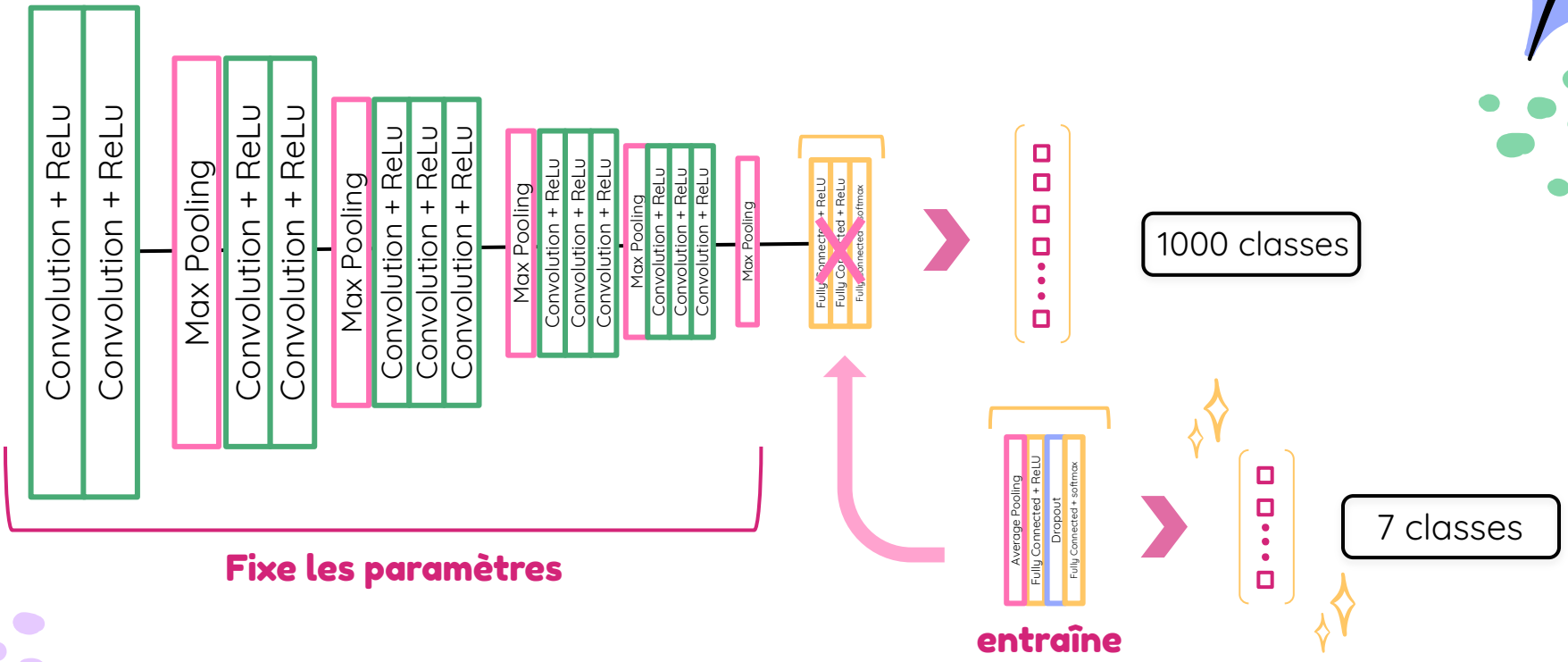




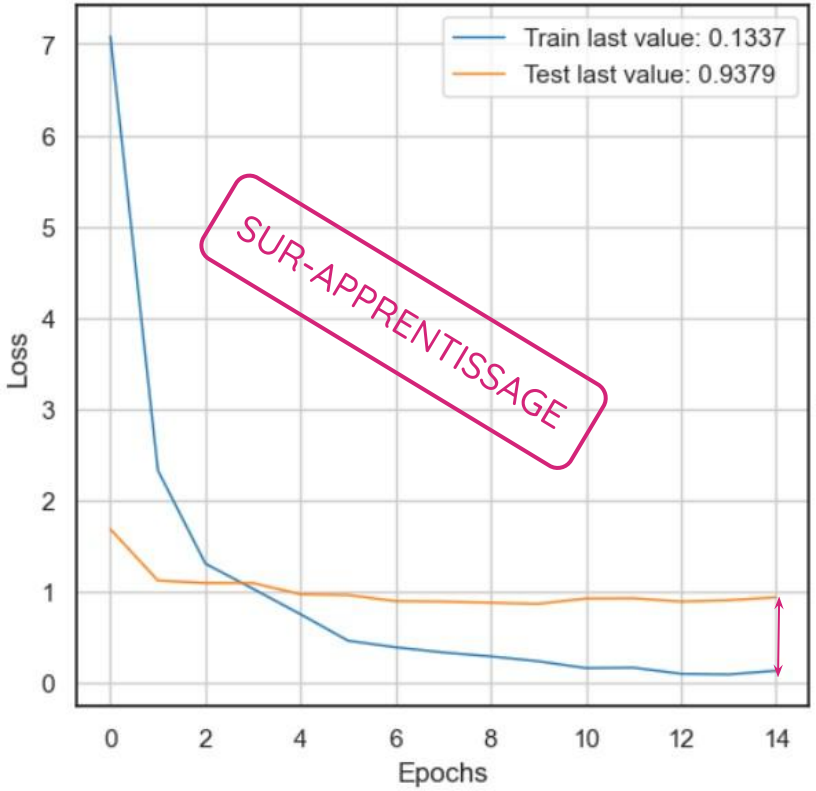



Classification supervisée

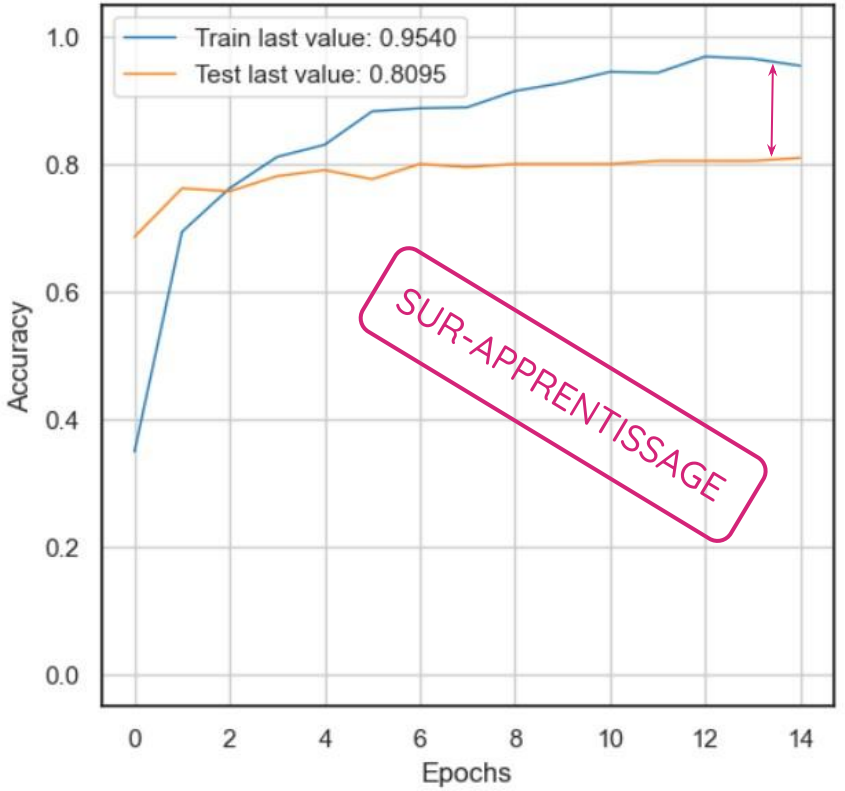
Transfer learning - vgg16



Evolution de l'erreur du modèle (entropie croisée)



Evolution de la proportion de classes correctement prédites

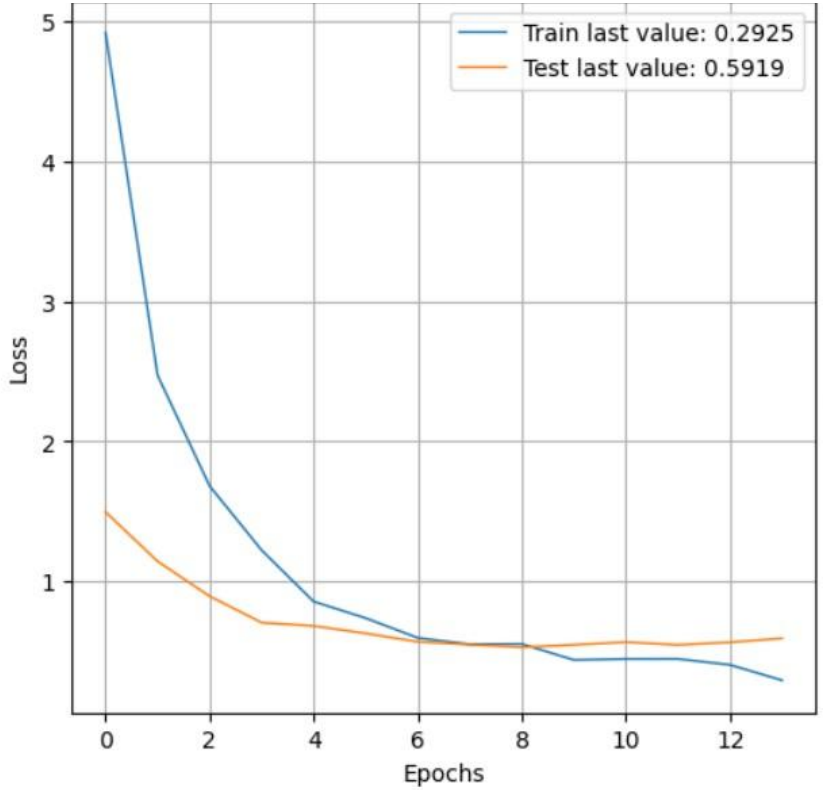


Data Augmentation - vgg16

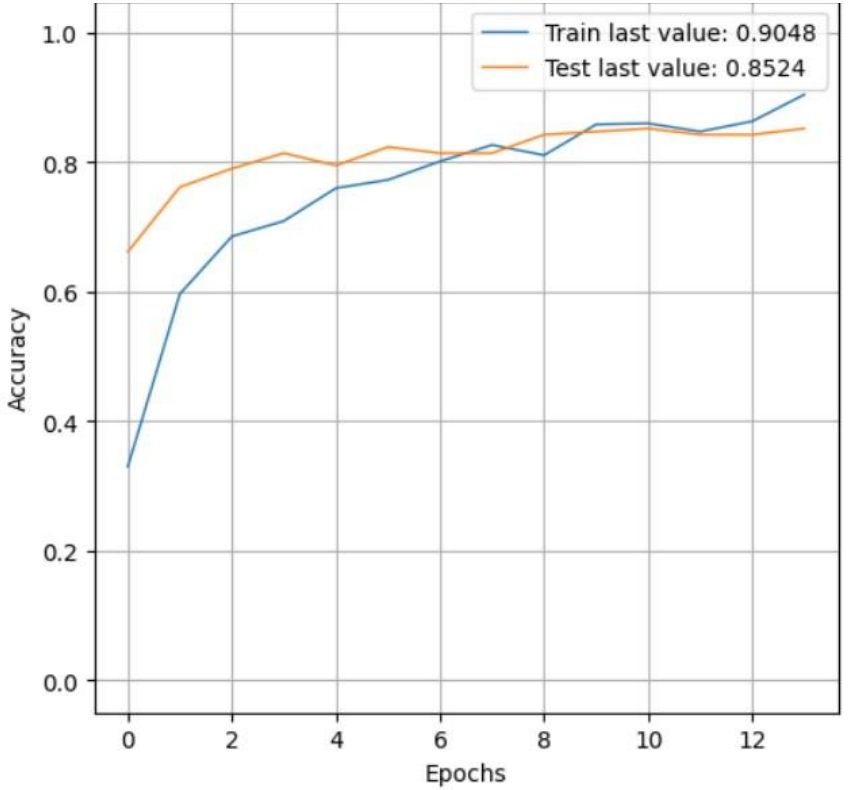
Augmente la diversité du jeu de données en lui appliquant des transformations

Retourner	Pivoter	Zoomer

Evolution de l'erreur du modèle (entropie croisée)



Evolution de la proportion de classes correctement prédites



Optimisation - vgg16

80%

TRAIN

75%

TRAIN

VALIDATION

25%

Recherche Hyperparamètres (hyperopt)

Minimise l'erreur

- loss
- epochs
- batch_size

20%

TEST

Mesure la performance

Meilleurs hyperparamètres

Mesures de performance

Rappel

$$\frac{\text{Vrais positifs}}{\text{Vrais positifs + Faux négatifs}}$$

Précision

$$\frac{\text{Vrais positifs}}{\text{Vrais positifs + Faux positifs}}$$

F1-score

$$2 \times \frac{\text{Précision x Rappel}}{\text{Précision + Rappel}}$$

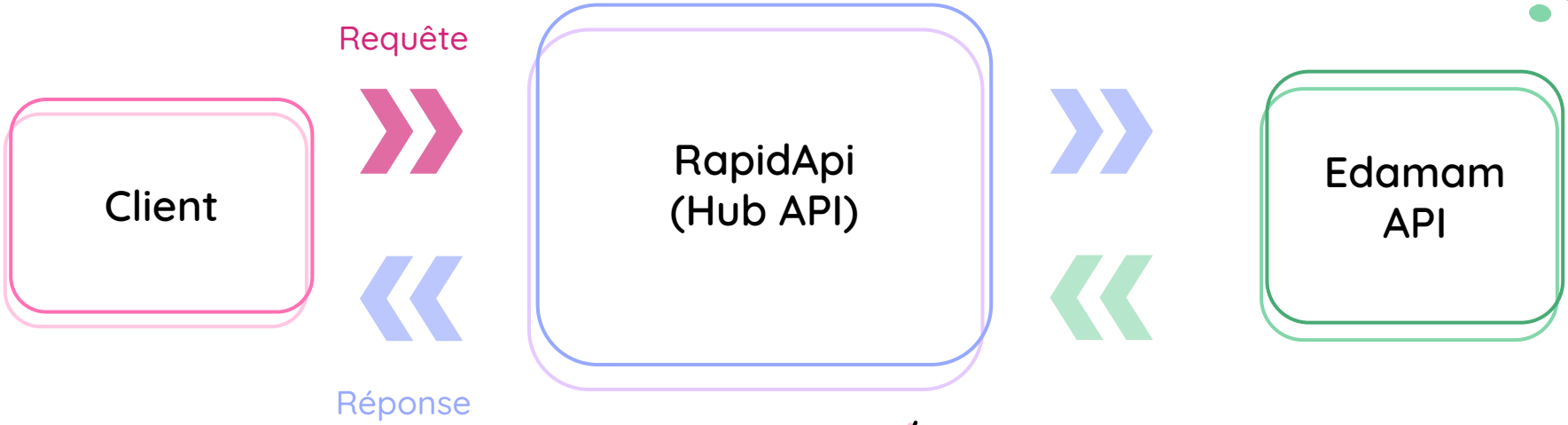
F1-scores (jeu de test)

Classes	Vgg16	+ data augmentation	+ data augmentation + optimisation
Baby Care	0.71	0.62	0.74
Beauty and Personal Care	0.80	0.82	0.86
Computers	0.84	0.87	0.84
Home Decor and Festive Needs	0.67	0.68	0.75
Home Furnishing	0.83	0.72	0.82
Kitchen and Dining	0.81	0.81	0.85
Watches	0.95	0.90	0.94
Accuracy	0.80	0.77	0.83

Test API

Application Programming Interface

Réduire le temps de développement en ré-utilisant des services déjà développés



Conclusion

Oui, il est possible d'automatiser la catégorisation des articles.

Annexes

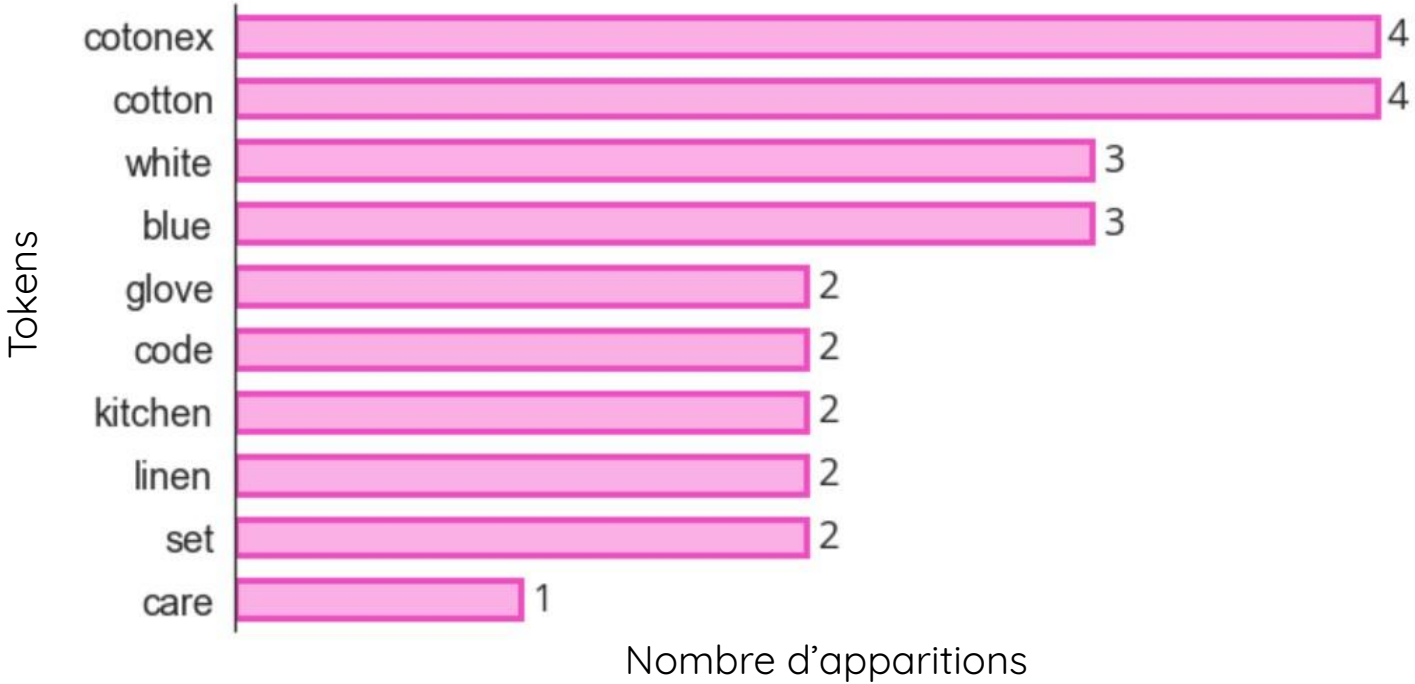
Home Furnishing

Baby Care

*Tokens les plus utilisés par classe

Exemple - count terms

10 tokens les plus utilisés dans la description d'une paire de gants de cuisine



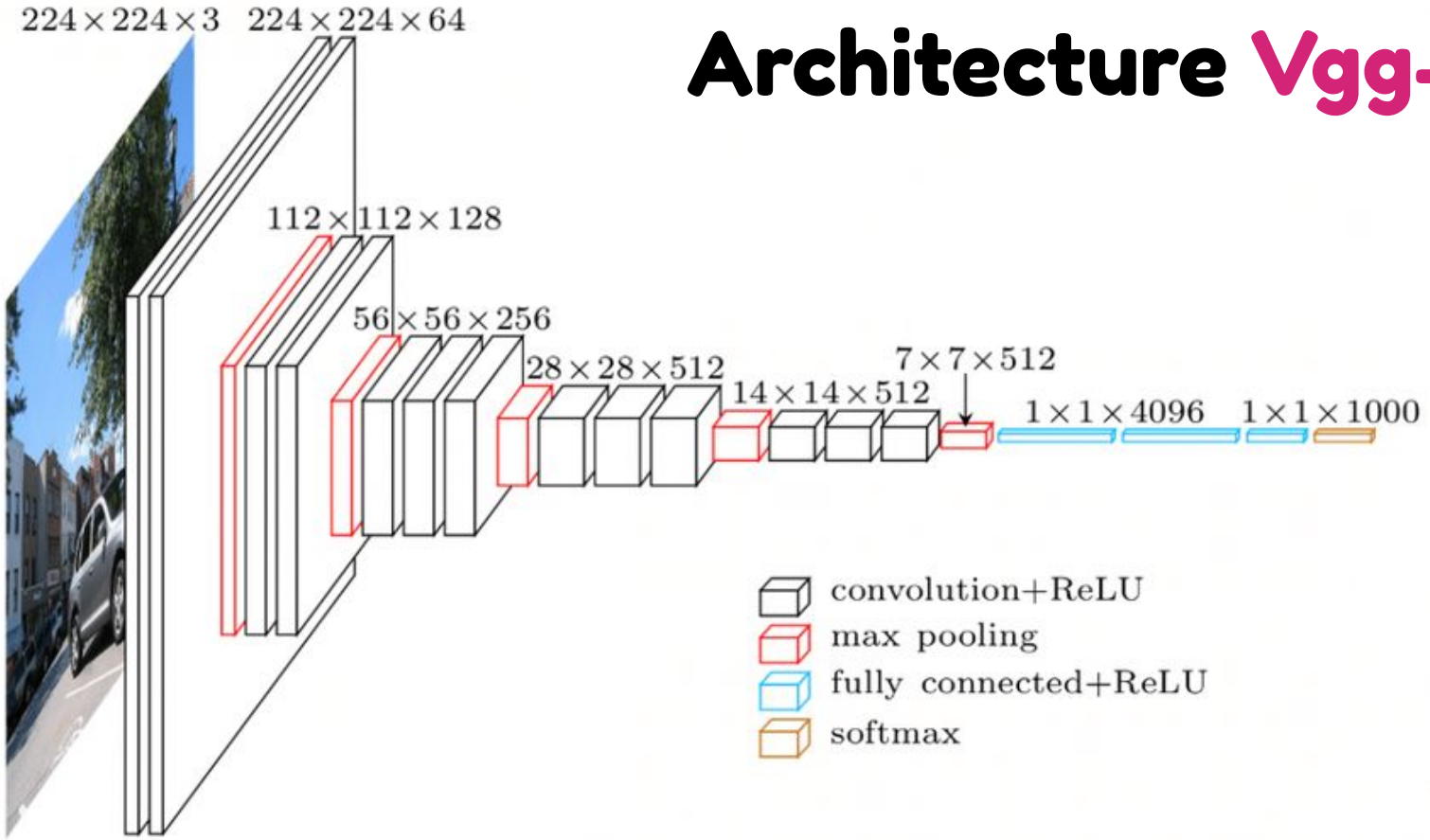
Entropie croisée

Pour une observation i :

$$\text{Entropie Croisée}_{(i)} = -\sum_{c}\left[\text{Proba observée}_{(c)} \times \log(\text{Proba prédite}_{(c)})\right]$$

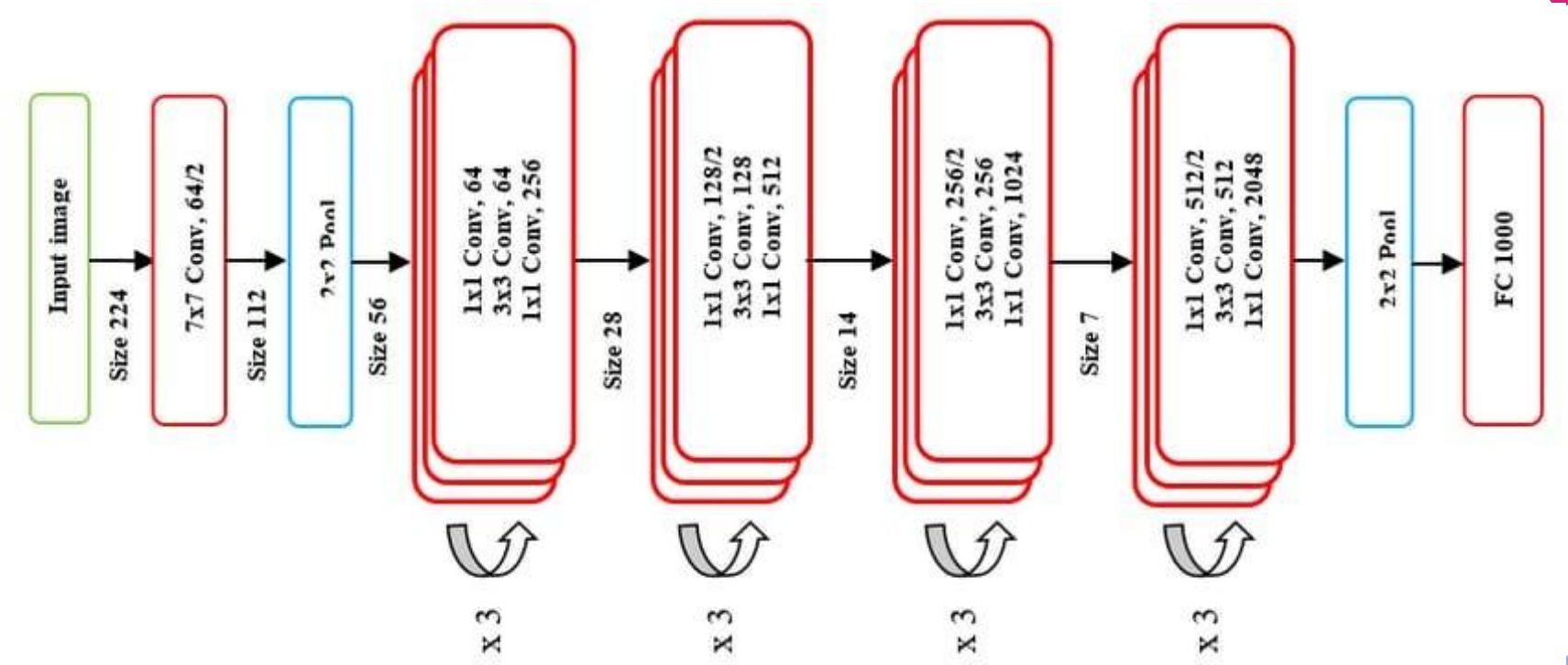
*proba d'appartenir à la catégorie c

Architecture Vgg-16



Source : https://www.researchgate.net/publication/327070011_Accelerating_Deep_Neural_Networks_on_Low_Power_Heterogeneous_Architectures

Architecture ResNet-50



Source : https://ijeecs.iaescore.com/index.php/IJEECS/article/view/20562

Règles de protection des données personnelles

01 Finalité

02 Proportionnalité et pertinence

03 Durée de conservation limitée

04 Sécurité et confidentialité

05 Droit des personnes

Credits

This template has been created by **Slidesgo**, and includes icons by **Flaticon**, infographics & images by **Freepik** and content by Eliana Delacour


slidesgo