

Projet 6

Classifiez automatiquement des biens de consommation

Développement d'un moteur de classification automatique d'articles basé sur les images et les descriptions textuelles.



David Scanu | Mai 2025



Parcours **AI Engineer**



Sommaire

- 🚀 Contexte & problématique du projet
- 🔍 Jeu de données
- 💡 Étude de faisabilité
 - 💬 Features texte
 - 🖼️ Features image
- 🤖 Classification supervisée d'images avec CNN et Data Augmentation
- 🍾 Collecte de produits à base de “champagne” via l'API Openfood Facts
- 🏁 Conclusion

★ Contexte et problématique métier

L'entreprise "Place de marché"

- Lancement d'une **marketplace e-commerce anglophone**
- Vendeurs proposent des **articles avec photos et descriptions**
- Actuellement : **attribution manuelle des catégories** par les vendeurs

⚠ Problématiques identifiées

- Attribution des catégories peu fiable
- Volume d'articles limité mais expansion prévue
- Besoin d'automatisation pour le passage à l'échelle

🎯 Objectifs du projet

- **Automatiser la tâche d'attribution des catégories**
- Étudier la faisabilité d'un **moteur de classification automatique**
 - Exploiter à la fois le texte (en anglais) et l'image des produits
- **Classification supervisée à partir des images**
 - **Réseau de neurones profond (CNN)**
 - Avec **data augmentation**





Présentation du jeu de données

Jeu de données Flipkart

- **1050 produits e-commerce**
- **7 catégories principales** avec **150 produits chacune** :
 - 🏠 Home Furnishing
 - 🧒 Baby Care
 - ⌚ Watches
 - 🎀 Home Decor & Festive Needs
 - 🍴 Kitchen & Dining
 - 💄 Beauty and Personal Care
 - 💻 Computers

Caractéristiques des données

- **Descriptions textuelles en anglais**
- **Images des produits**
- **Métadonnées** (prix, marque, spécifications)

Points forts du dataset

- **Répartition parfaitement équilibrée**
- Diversité des types de produits
- Données textuelles et visuelles exploitables





Étude de faisabilité : Démarche globale



Objectif de l'étude : Évaluer la capacité des différentes approches à **regrouper automatiquement les produits similaires**.



Méthodologie générale :

- Exploration des modalités texte et image
- Comparaison systématique des techniques classiques et avancées
- Évaluation quantitative des performances



Évaluation des performances :

- **Visualisation** par réduction dimensionnelle (**PCA + t-SNE**)
- **Clustering** non-supervisé (**K-means**)
- **Mesures quantitatives (ARI)** entre clusters et catégories réelles



Résultat attendu : Recommandation argumentée sur la faisabilité technique et les méthodes à privilégier pour un système de classification automatique performant.



Approches textuelles explorées

- **Méthodes statistiques :**
 - Bag-of-Words, TF-IDF
- **Word embeddings classiques :**
 - Word2Vec
- **Embeddings contextuels avancés :**
 - BERT, Universal Sentence Encoder



Approches images explorées

- **Vision par ordinateur classique :**
 - SIFT
- **Deep Learning :**
 - Transfer Learning (architectures pré-entraînées)
 - VGG16
 - ResNet50
 - InceptionV3



Prétraitement des données textuelles

1. Normalisation et nettoyage

- Conversion en minuscules
- Suppression des caractères spéciaux et chiffres via expressions régulières
- Élimination des espaces superflus

2. Tokenization

- Découpage en mots individuels avec NLTK
- Préservation des termes significatifs

3. Filtrage des stop words

- Utilisation du dictionnaire anglais de NLTK
- Élimination des mots non informatifs (the, and, of...)

4. Lemmatisation

- Réduction à la forme canonique via WordNetLemmatizer
- Conservation du sens sémantique (vs stemming)



Données combinées

- **Fusion du nom du produit et sa description**
- Fonction de prétraitement (***processed_text***)



Exemple concret

- **Avant** : "Buy Clues Abstract Single Quilts & Comforters Gold at Rs. 1349 at Flipkart.com. Only Genuine Products. Free Shipping. Cash On Delivery!"
- **Après** : "buy clue abstract single quilt comforter gold r flipkart com genuine product free shipping cash delivery"



Avantages

- Réduction du bruit
- Normalisation du vocabulaire
- Préparation optimale pour les différentes techniques d'extraction de features.

Extraction des features textuelles

Objectif : Transformer les textes prétraités en représentations vectorielles exploitables.

1. Approches statistiques

- **Bag-of-Words (CountVectorizer)**

- Dimensions : 1000
- Chaque document est un **vecteur contenant la fréquence d'apparition de chaque mot** du vocabulaire
- *Classifications basées sur la présence de mots-clés spécifiques*

- **TF-IDF (TfidfVectorizer)**

- Dimensions : 1000
- Extension du BoW qui **pondère les termes selon leur fréquence** dans le document et leur rareté dans le corpus
- *Réduit l'importance des mots trop communs*

2. Word/Sentence Embeddings

- **Word2Vec (embeddings classiques)**

- Dimensions : 100
- **Représentation vectorielle dense des mots**, moyennée pour obtenir un vecteur par document
- *Capture des relations sémantiques entre mots*

- **BERT (DistilBERT via Sentence Transformers)**

- Dimensions : 768
- **Embeddings contextuels issus d'un modèle de langage bidirectionnel pré-entraîné**
- *97% des performances de BERT pour 60% moins de paramètres*

- **Universal Sentence Encoder**

- Dimensions : 512
- Modèle spécialement conçu pour **encoder des phrases ou paragraphes entiers en vecteurs denses**



Extraction des features d'images

Objectif : Transformer les images produits en représentations vectorielles exploitables

Prétraitement appliqué

- **Redimensionnement aux dimensions requises** par chaque modèle
- **Conversion RGB et normalisation adaptée** à chaque architecture
- **Pooling global** pour obtenir des **vecteurs à taille fixe**

1. 🔍 Approche classique de vision par ordinateur :

SIFT (Scale-Invariant Feature Transform)

- Dimensions : 500
- Principe : Détection de points-clés invariants à l'échelle et à la rotation

2. 🧠 Approches par transfer learning (CNN) :



Keras

Les **applications Keras** sont des **modèles d'apprentissage profond** disponibles avec les **poids pré-entraînés** sur ImageNet.

Ces modèles peuvent être utilisés pour la **prédiction**, l'**extraction de caractéristiques** et le **transfer learning**.

VGG16

- Dimensions : 512
- Principe : **Réseau à 16 couches** avec architecture simple et homogène

ResNet50

- Dimensions : 2048
- Principe : **Réseau à 50 couches** avec connexions résiduelles

InceptionV3

- Dimensions : 2048
- Principe : Architecture avec modules "**Inception**" combinant convolutions multi-échelles



Réduction de dimension et visualisation

Objectif : Transformer nos représentations vectorielles haute dimension en visualisations interprétables en 2D



Prétraitement des données

- Standardisation (moyenne 0, écart-type 1)
- Équilibrage des échelles entre différentes approches



Réduction préliminaire par PCA

- Appliquée aux **vecteurs de grande dimension** (> 50)
- **Conservation d'environ 70-90% de la variance** selon la méthode
- Réduction du bruit et **accélération de t-SNE**



Transformation t-SNE

- Algorithme de réduction non-linéaire préservant les relations de proximité
- Transformation des vecteurs en points 2D

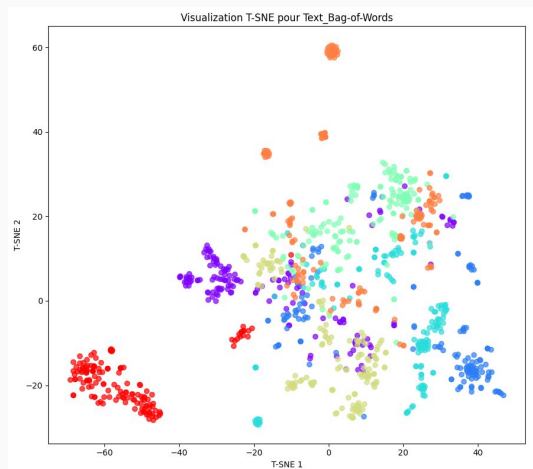


Visualisation des clusters

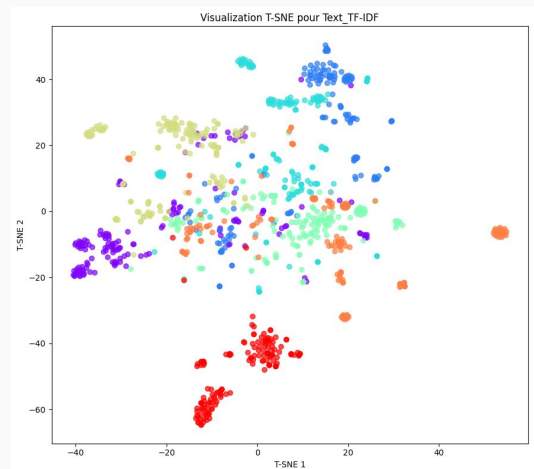
- Nuage de points colorés par catégorie réelle
- Clusters naturels révélant les proximités sémantiques et visuelles
- Interprétation de la séparation des classes

Visualisation des features textuelles

Bag of words

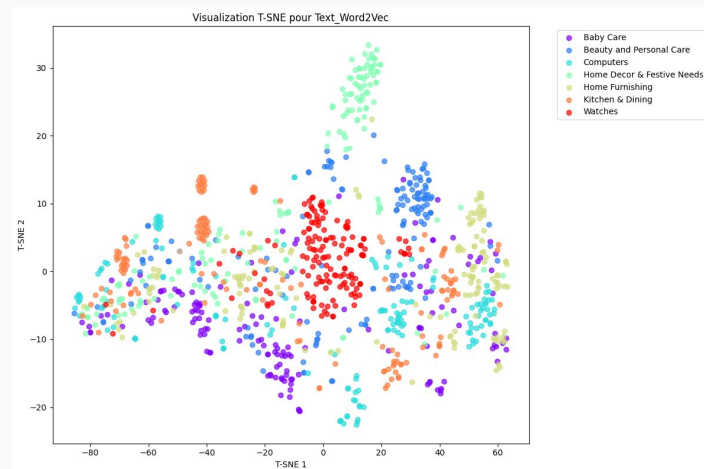


TF-IDF



Séparation claire pour les montres. Fort chevauchement des autres catégories au centre

Word2Vec

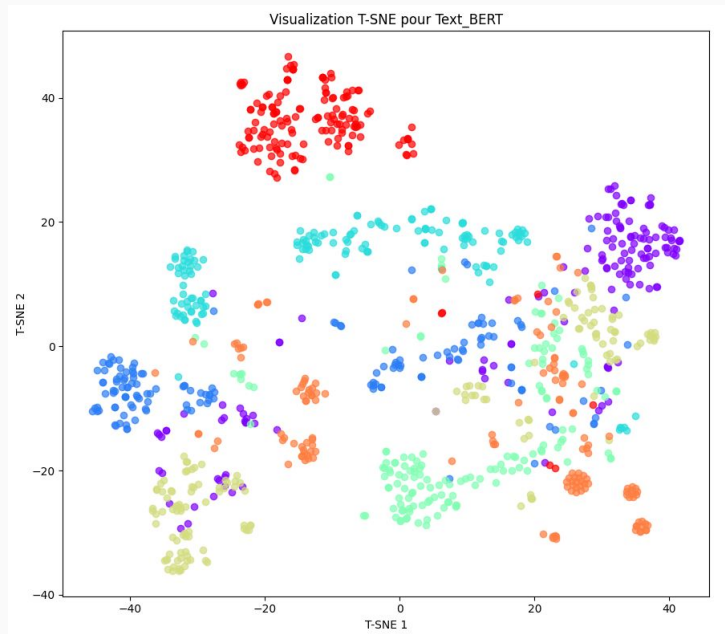


Distribution très dispersée avec chevauchement important

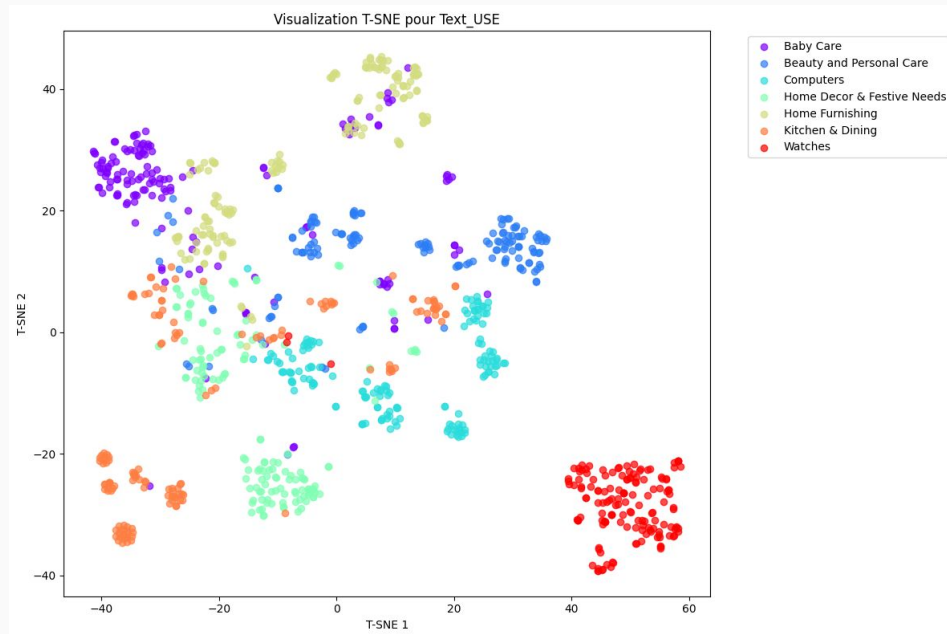


Visualisation des features textuelles

BERT



USE

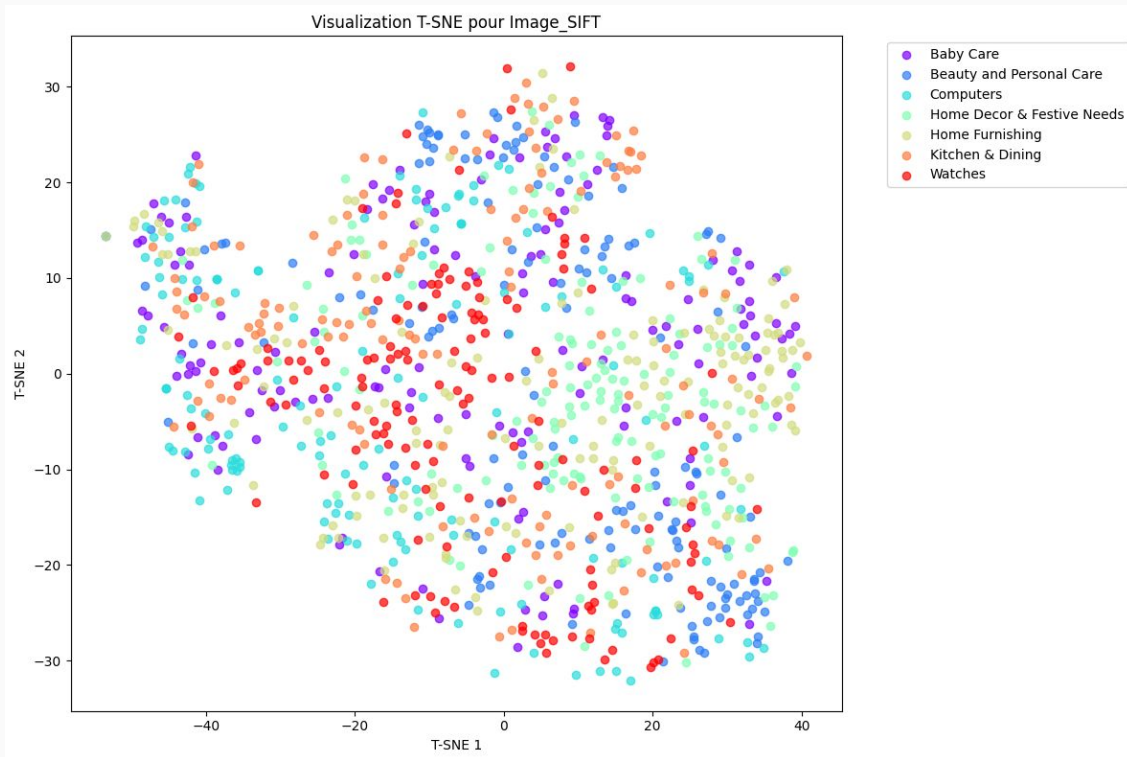




Visualisation des features d'images

Interprétation

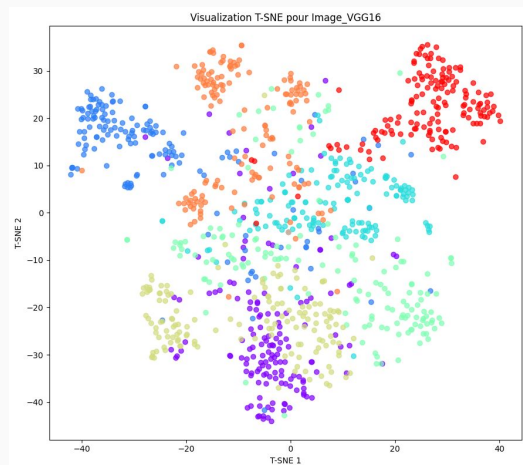
- Distribution très mélangée
- Absence presque totale de clusters distincts par couleur
- Nombreux chevauchements entre toutes les catégories de produits





Visualisation des features d'images

VGG16



ResNet50



InceptionV3



- Niveau élevé de séparation entre certaines catégories
- Cohésion interne de ces groupes

Cette visualisation t-SNE démontre clairement la **faisabilité d'un système de classification automatique** des produits basé sur les **caractéristiques visuelles** extraites.

Mesure de similarité par clustering K-means

Objectif : Évaluer quantitativement la capacité des différentes méthodes à regrouper les produits par catégorie.

Clustering K-means

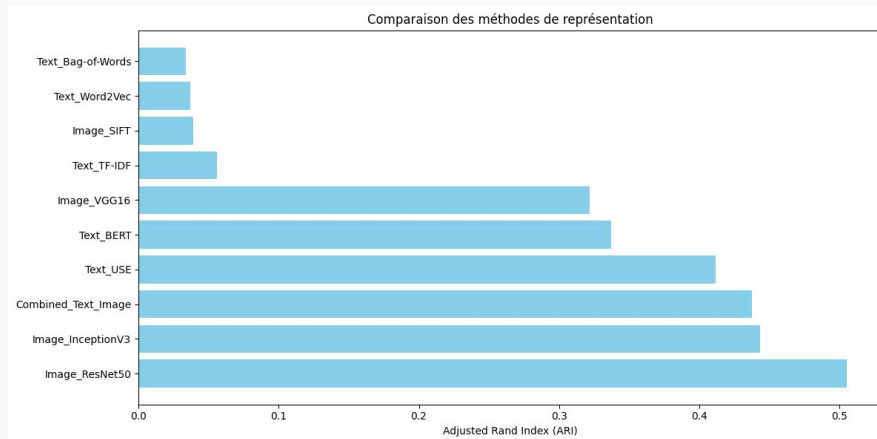
- Standardisation préalable des caractéristiques
- **Application de l'algorithme K-means avec k=7** (nombre de catégories réelles)

Calcul de l'Adjusted Rand Index (ARI)

- **Mesure de concordance entre les clusters générés et les catégories réelles**
- Valeurs entre -1 et 1 :
 - 1 = correspondance parfaite
 - 0 = correspondance aléatoire
 - Négative = pire que le hasard
- **Évalue objectivement** la pertinence des représentations extraites
- **Confirmation quantitative** des observations visuelles

Résultats par méthode (ARI)

- **ResNet50** : 0.50 - Performance très bonne
- **InceptionV3** : 0.44 - Bonne performance
- **USE** : 0.41 - Meilleure approche textuelle
- **BERT** : 0.34 - Bonne performance textuelle
- **VGG16** : 0.32 - Performance moyenne
- **Approches classiques** < 0.06 - Performances faibles



Conclusions de l'étude de faisabilité



Hiérarchie claire des approches

- 🏆 **ResNet50** : ARI exceptionnel de 0.505
- 🥈 **InceptionV3** : ARI ≈ 0.44
- 🥉 **Embeddings avancés** (USE, BERT) : ARI > 0.33
- ⚠️ **Méthodes traditionnelles** (BoW, SIFT) : ARI < 0.06



Prédominance visuelle

- **Les caractéristiques visuelles surpassent les caractéristiques textuelles**
- Les images contiennent plus d'information discriminante pour la catégorisation
- Signal visuel plus fort et plus cohérent que le signal textuel



Faisabilité démontrée

- L'étude **confirme la possibilité de créer un moteur de classification automatique performant** basé sur **ResNet50**, avec un ARI dépassant 0.5, ce qui constitue une base solide pour la suite du projet.





Classification supervisée d'images avec data augmentation

Objectif : Développer un modèle performant de classification d'images par catégorie de produits.



Organisation de l'environnement

- Structure de dossiers pour les **artefacts d'entraînement**
- Séparation en phases : **standard** et **fine-tuning**
- Conservation des **métriques** et **visualisations**



Préparation des données

- **Encodage des catégories** (LabelEncoder)
- **Division des données** : Train (64%) / Validation (16%) / Test (20%)
- **Normalisation et redimensionnement** (224×224 pixels)



Data augmentation

- **Transformations géométriques :**
 - rotations ($\pm 5.5^\circ$), retournements horizontaux
- **Modifications colorimétriques :**
 - luminosité ($\pm 15\%$), contraste ($\pm 10\%$)
- **Variations de saturation** ($\pm 15\%$) et **teinte** ($\pm 5\%$)



Architecture du modèle

- **Base : ResNet50 pré-entraîné sur ImageNet**
- **Adaptation à nos 7 catégories spécifiques**
- Stratégie d'entraînement en deux phases :
 - **Phase 1** : Couches de base gelées
 - **Phase 2** : Fine-tuning des blocs supérieurs



Évaluation complète

- **Métriques de performance** par catégorie
- **Visualisation des résultats** et des erreurs
- Analyse de la **matrice de confusion**

Approche justifiée par : L'étude de faisabilité qui a démontré l'excellent potentiel de **ResNet50** (ARI = 0.50) pour la classification de nos produits.





Data augmentation - Diversifier pour mieux généraliser

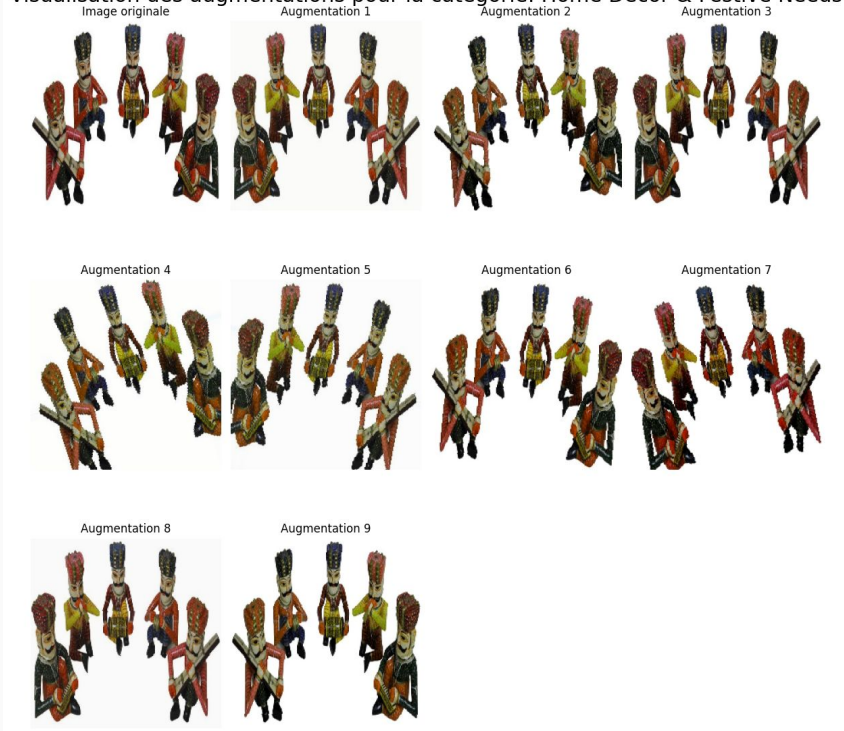
Pourquoi la data augmentation ?

- **Dataset limité** (670 images d'entraînement)
- **Prévention du surapprentissage**
- **Amélioration de la robustesse du modèle**
- **Simulation de variations naturelles des produits**

Transformations implémentées

-  **Transformations géométriques**
 - Retournements horizontaux aléatoires
 - Rotations légères ($\pm 5.5^\circ$)
 - Conservation des proportions originales
-  **Transformations colorimétriques**
 - Ajustement de luminosité ($\pm 15\%$)
 - Modification du contraste ($\pm 10\%$)
 - Variation de saturation ($\pm 15\%$)
 - Légère modification de teinte ($\pm 5\%$)

Visualisation des augmentations pour la catégorie: Home Decor & Festive Needs





Architecture du modèle de classification

Base : ResNet50 pré-entraîné sur ImageNet

- Réseau profond de 50 couches avec connexions résiduelles
- Extraction de caractéristiques visuelles hiérarchiques et robustes
- Poids pré-entraînés sur 1,2 million d'images (1000 catégories)

Optimisation

- Optimiseur Adam avec learning rate de 0.0005
- Loss : **Sparse Categorical Crossentropy**
- Métrique principale : **Accuracy**



Couches d'adaptation

1. **GlobalAveragePooling2D**
 - a. Transformation des cartes de caractéristiques en vecteur 1D
 - b. Réduction significative du nombre de paramètres
2. **Dense (512 neurones) + ReLU**
 - a. Première couche entièrement connectée
 - b. Régularisation L2 ($\lambda=0.001$) pour limiter le surapprentissage
3. **Dropout (60%)**
 - a. Désactivation aléatoire de 60% des neurones pendant l'entraînement
 - b. Robustesse et généralisation améliorées
4. **Dense (256 neurones) + ReLU**
 - a. Seconde couche entièrement connectée
 - b. Régularisation L2 ($\lambda=0.001$)
5. **Dropout (40%)**
 - a. Seconde couche de dropout pour renforcer la généralisation
6. **Dense (7 neurones) + Softmax**
 - a. Couche de sortie avec **activation softmax**
 - b. **Distribution de probabilités sur les 7 catégories**



Phase 1 : Entraînement avec couches de base gelées

★ Stratégie d'entraînement

- Poids de ResNet50 gelés
- Apprentissage ciblé sur les couches personnalisées

⚙️ Configuration de l'entraînement

- Optimiseur Adam (learning rate initial : 0.001)
- Batch size : 16 (optimisé pour GPU)
- Entraînement sur 20 époques avec data augmentation
- Pipeline TensorFlow optimisé (prefetch, parallel calls)

🔧 Système de callbacks

- ModelCheckpoint : Sauvegarde automatique du meilleur modèle
- EarlyStopping : Arrêt si aucune amélioration après 10 époques
- ReduceLROnPlateau : Division du taux d'apprentissage par 5 après 5 époques sans progrès



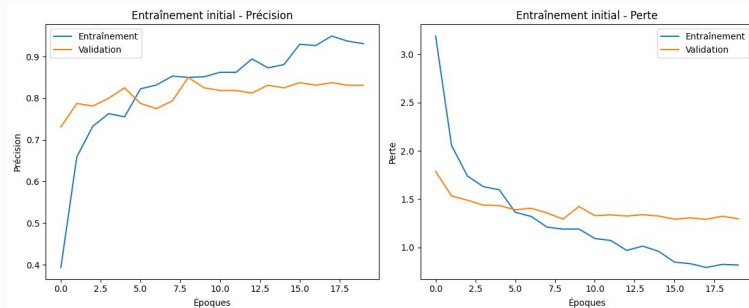
Résultats obtenus

- Précision maximale sur validation : 85.0% (époque 9)
- Précision sur test : 86.1%
- Comportement : Apprentissage rapide initial suivi d'un plateau
- **Observation** : Écart croissant entre entraînement (94.3%) et validation (83.1%) indiquant un début de surapprentissage



Points clés

- Convergence rapide vers une bonne performance
- Régularisation efficace limitant le surapprentissage
- Base solide pour la phase de fine-tuning



Phase 2 : Fine-tuning avec dégel progressif

Approche de fine-tuning

- **Dégel stratégique de couches profondes de ResNet50**
- **22.80% des paramètres rendus entraînaibles** (5,6 millions sur 24,7 millions)
- **Ciblage des 10 dernières couches du réseau de base + couches personnalisées**

Configuration spécifique

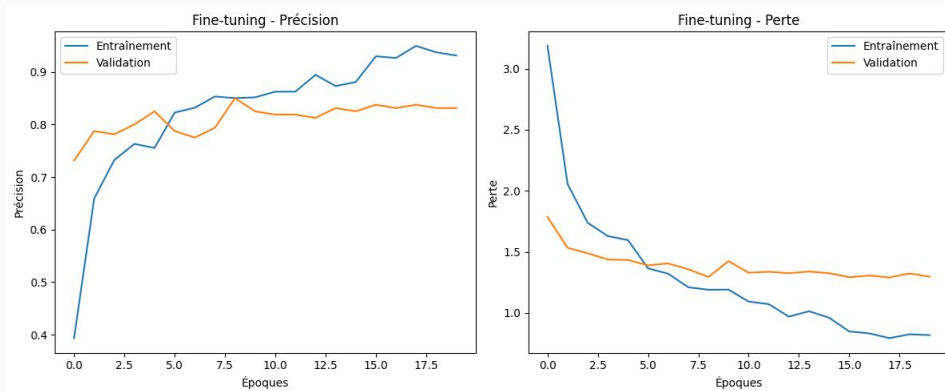
- **Taux d'apprentissage très faible** (0.00002) pour modifications subtiles
- **Batch size augmenté à 24 images**
- **Continuation de la data augmentation**
- **Entraînement sur 30 époques maximum** (avec early stopping)

Résultats obtenus

- Meilleure précision validation : **85.1%** (époque 2)
- Accuracy finale sur test : **86.5%**

Bilan

Légère amélioration (+0.4%) par rapport à la phase 1, suggérant que le modèle avait déjà bien capturé les caractéristiques importantes des produits dans la première phase.





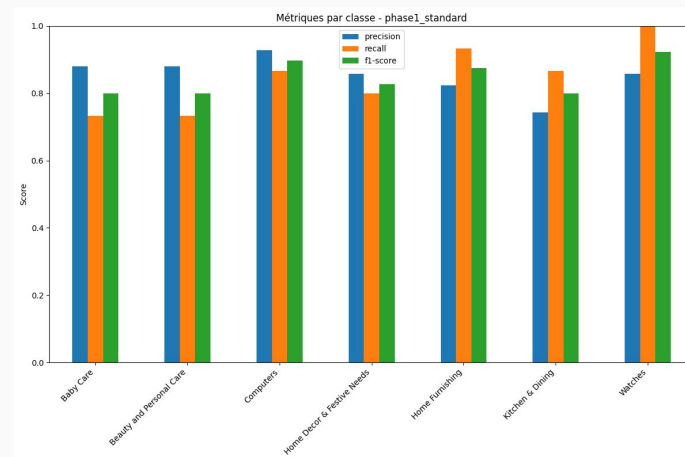
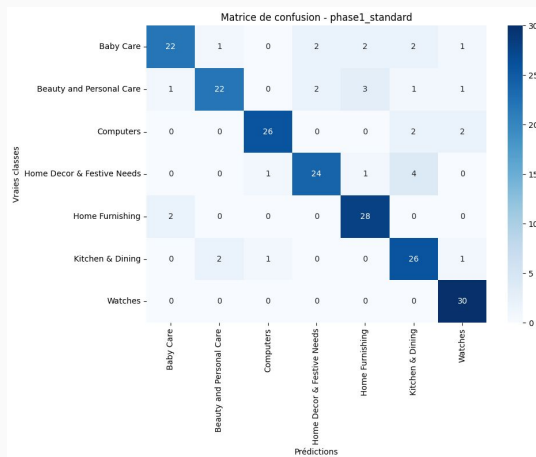
Évaluation des performances du modèle (Phase 1)

Performance globale

- **Accuracy** : 84.8%
- **Precision** : 85.3%
- **Recall** : 84.8%
- **F1-score** : 84.6%

Conclusions

- **Performance globale solide pour l'ensemble des catégories**
- Variations attendues dans la précision entre catégories
- Base solide pour un système de classification automatique
- Potentiel d'amélioration par enrichissement des données d'entraînement



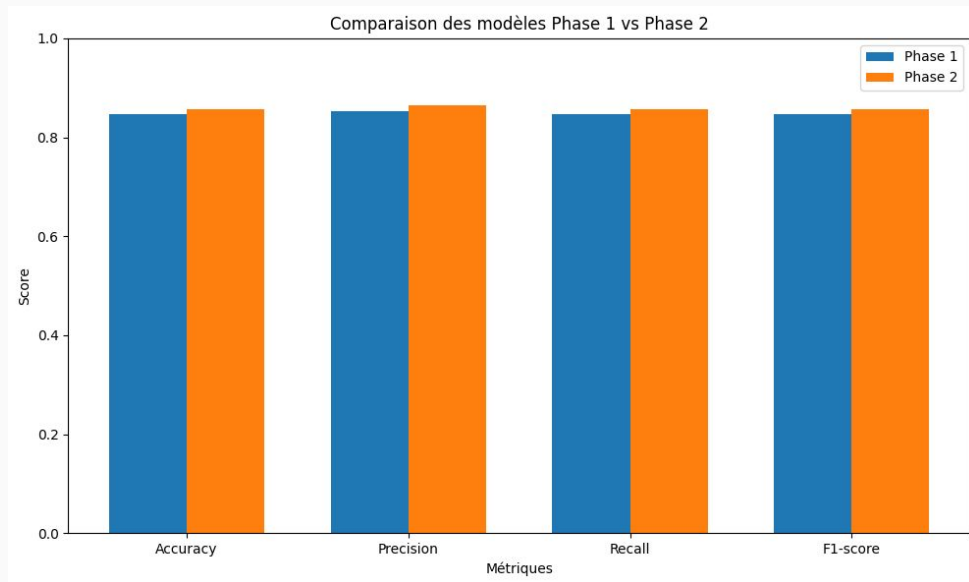


Comparaison des performances Phase 1 vs Phase 2

Conclusion

- Le fine-tuning a apporté une **amélioration modeste** mais réelle
- Performance finale très satisfaisante** (85.7% d'accuracy)
- Base solide pour un déploiement en production**
- Potentiel d'amélioration future par augmentation du volume de données

	Phase 1 (Base)	Phase 2 (Fine-tuning)	Amélioration
Accuracy	84.8%	85.7%	+0.9%
Precision	85.3%	86.4%	+1.1%
Recall	84.8%	85.7%	+0.9%
F1-score	84.6%	85.6%	+1.0%



Visualisation des prédictions du modèle (phase 2)

Vrai: Home Decor & Festive Needs
Préd: Home Decor & Festive Needs
Conf: 0.62



Vrai: Baby Care
Préd: Home Decor & Festive Needs
Conf: 0.78



Vrai: Computers
Préd: Home Decor & Festive Needs
Conf: 0.25



Vrai: Computers
Préd: Computers
Conf: 0.58



Vrai: Home Furnishing
Préd: Home Furnishing
Conf: 0.62



Vrai: Kitchen & Dining
Préd: Baby Care
Conf: 0.29



Vrai: Beauty and Personal Care
Préd: Beauty and Personal Care
Conf: 0.89



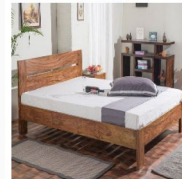
Vrai: Home Furnishing
Préd: Kitchen & Dining
Conf: 0.36



Vrai: Beauty and Personal Care
Préd: Beauty and Personal Care
Conf: 0.97



Vrai: Beauty and Personal Care
Préd: Home Furnishing
Conf: 0.94



Vrai: Home Decor & Festive Needs
Préd: Kitchen & Dining
Conf: 0.93



Vrai: Beauty and Personal Care
Préd: Computers
Conf: 0.41



Vrai: Home Decor & Festive Needs
Préd: Kitchen & Dining
Conf: 0.71



Vrai: Beauty and Personal Care
Préd: Computers
Conf: 0.75



Vrai: Baby Care
Préd: Beauty and Personal Care
Conf: 0.50



Vrai: Home Decor & Festive Needs
Préd: Kitchen & Dining
Conf: 0.48



Vrai: Kitchen & Dining
Préd: Baby Care
Conf: 0.36



Vrai: Baby Care
Préd: Home Decor & Festive Needs
Conf: 0.65










API OpenFood Facts

Objectif : Explorer l'intégration d'une nouvelle catégorie "Épicerie fine" avec des produits à base de champagne.

Mise en œuvre technique

- **Script Python**
- **Requête REST** vers l'API OpenFood Facts
- Sauvegarde au **format CSV**

Champs extraits pour chaque produit

-  **foodId** : identifiant unique du produit
-  **label** : nom commercial du produit
-  **category** : classifications hiérarchiques
-  **foodContentsLabel** : composition et ingrédients
-  **image** : URL de l'image du produit

foodId	label	category	foodContentsLabel	image
31139340 04147	Canard Duchêne	Boissons, Boissons alcoolisées, Vins, Vins effervescents, Champagnes, Liquide	Pinots et de Chardonnay	https://images.openfoodfacts.org/images/products/31139340/04147/front_fr.4.400.jpg
32584312 20000		Boissons, Boissons alcoolisées, Vins, Vins effervescents, Champagnes, Champagnes bruts	Champagne	https://images.openfoodfacts.org/images/products/32584312/20000/front_en.4.400.jpg
31140800 34057	Champagne rosé	Bebidas, Bebidas alcohólicas, Vinos, Vinos espumosos, Champán		https://images.openfoodfacts.org/images/products/31140800/34057/front_es.3.400.jpg
31853707 29960	Champagne Impérial Brut	Getränke und Getränkezubereitungen, Getränke, Alkoholische Getränke, Weine, Schaumweine, Champagner		https://images.openfoodfacts.org/images/products/31853707/29960/front_de.3.400.jpg
3049610 004104	Veuve Clicquot Champagne Ponsardin Brut	Boissons et préparations de boissons, Boissons, Boissons alcoolisées, Vins, Vins effervescents, Champagnes	Champagne	https://images.openfoodfacts.org/images/products/3049610/004104/front_fr.39.400.jpg
3282946 015837	Nicolas Feuillatte	Boissons, Boissons alcoolisées, Vins, Vins français, Vins effervescents, Champagnes, Champagnes français, Champagnes bruts	Champagne, Contient des _sulfites_	https://images.openfoodfacts.org/images/products/3282946/015837/front_fr.7.400.jpg
311391031 2013	Champagne Alfred Rothschild et Cie brut	Boissons, Boissons alcoolisées, Vins, Vins français, Vins effervescents, Champagnes, Champagnes français, Champagnes bruts	Champagne brut (contient _sulfites_)	https://images.openfoodfacts.org/images/products/311391031/2013/front_fr.3.400.jpg
20712907	Champagne	Boissons, Boissons alcoolisées, Vins, Vins effervescents, Champagnes		https://images.openfoodfacts.org/images/products/000/007/2907/front_fr.6.400.jpg
31853702 83905	Champagne Ruinart	Boissons, Boissons alcoolisées, Vins, Vins effervescents, Champagnes	champagne	https://images.openfoodfacts.org/images/products/31853702/83905/front_en.5.400.jpg
341618101 7169	Champagne AOP, brut	Boissons, Boissons alcoolisées, Vins, Vins effervescents, Champagnes, Champagnes bruts	Champagne	https://images.openfoodfacts.org/images/products/341618101/7169/front_fr.7.400.jpg



Conclusion globale du projet

🎯 Réalisation d'un moteur de classification automatique pour "Place de marché"



Étude de faisabilité (Partie 1)

- Exploration complète de **9 approches d'extraction de caractéristiques** (5 textuelles, 4 visuelles)
- Démonstration de la **capacité des algorithmes à regrouper naturellement les produits** (ARI max = 0.505)
- Identification de **ResNet50** et **Universal Sentence Encoder** comme méthodes optimales



Classification supervisée (Partie 2)

- **Développement d'un modèle de classification d'images** atteignant **85.7% d'accuracy**
- **Implémentation efficace de data augmentation et fine-tuning**
- Amélioration progressive des performances en **deux phases d'entraînement**



Bénéfices métier

1. **Automatisation** : Réduction de la charge manuelle pour les vendeurs
2. **Qualité** : Catégorisation cohérente et fiable des produits
3. **Évolutivité** : Base solide pour le passage à l'échelle
4. **Flexibilité** : Capacité d'intégration de nouvelles catégories via l'API testée



Conclusion

Un **système intelligent de classification** qui permettra à "Place de marché" d'**optimiser son expérience utilisateur** tout en se préparant à une **croissance du volume de produits**.



Merci 🙏

Avez-vous des questions ?



Annexes

Dépôt GitHub

- <https://github.com/DavidScanu/oc-ai-engineer-p06-classifier-biens-consommation>

Notebooks

- [david-scanu-p06-notebook-01-feasibility-images-texts.ipynb](#)
- [david-scanu-p06-notebook-02-classification-images.ipynb](#)

A propos

- [David Scanu Développeur en intelligence artificielle Carrefour | LinkedIn](#)
- [Formation AI Engineer OpenClassrooms](#)

