Classification d'Images pour la Catégorisation de Produits sur un Site de E-Commerce

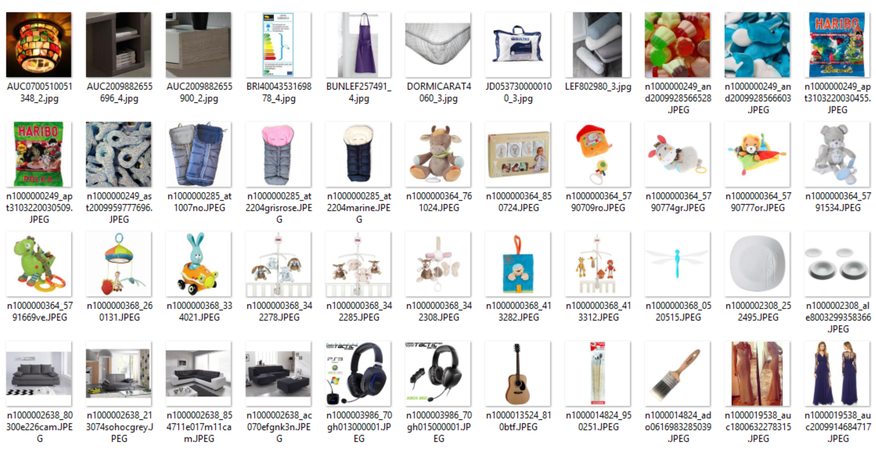
Z. Li, E. Guàrdia-Sebaoun, C. Grauer, M. Cornec, B. Goutorbe

Contexte

Cdiscount a généré plus de 3 milliards d'euros en 2016, ce qui en fait le deuxième acteur e-commerce en France. Avec l'ouverture de ses rayons à des vendeurs externes (via sa place de marché), Cdiscount voit la taille de son catalogue exploser, passant de 10 millions de produits fin 2015 à une prévision de plus de **30 millions de produits** à la fin de l'année. Face à un flux de produits arrivant sur la place de marché toujours plus important, la création manuelle de nouveaux produits est devenue impossible. Il est donc devenu nécessaire d'en automatiser certaines étapes, notamment la catégorisation.
Le catalogue de Cdiscount compte aujourd'hui plus de **6000 catégories**, dans lesquelles les produits sont répartis de manière fortement déséquilibrée (80

Objectifs

Dans un premier temps, nous avons attaqué le problème sous un angle sémantique. Si ces méthodes nous ont permis de traiter une bonne partie du flux d'entrée, il restait cependant une part non négligeable de produits non catégorisés. Le but de ces travaux est d'utiliser les images associées aux produits pour catégoriser ce reliquat.



Problématiques

Nous avons été confrontés à cinq défis majeurs :

- **Très grand nombre de classes** : les données contiennent plus de 6000 catégories, soit environ 6 fois plus que ImageNet.
- **Contraintes de temps et de budget** : le cadre de ces travaux nous a demandé de trouver le meilleur compromis entre les performances, le temps d'apprentissage et le coût de l'infrastructure.
- **Déséquilibre des données** : 80% des données représentent 20% des catégories. Cette répartition doit être prise en compte pour éviter le surapprentissage.
- **Données Bruitées** : les données contiennent environ 10% de données mal catégorisée.
- **Volatilité des Catégories** : l'arborescence des catégories évolue au cours du temps. Ceci accentue la contrainte de temps sur l'apprentissage du modèle.

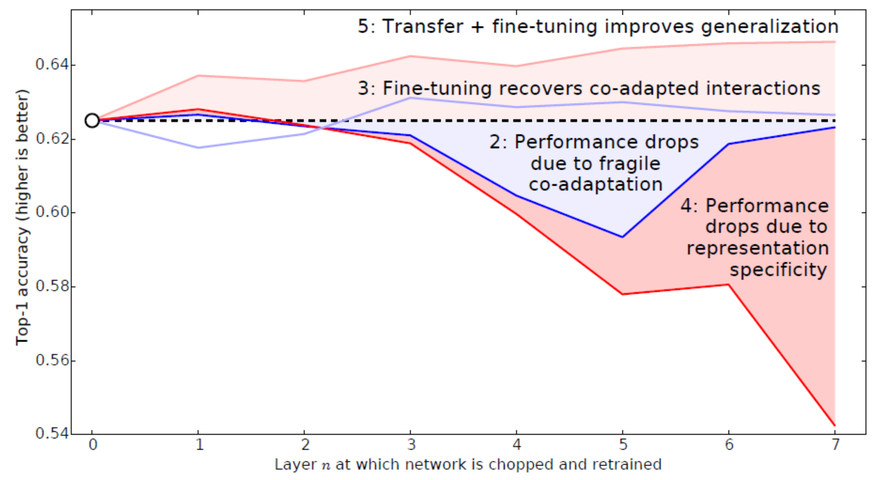
Données

Durant un stage de 6 mois, nous avons pu tester diverses approches sur un jeu de données comprenant plus de 15 millions d'images, associées à 9 millions de produits, eux même répartis dans plus de 5000 catégories.
Nous avons depuis rendu ce jeu de données public, lors de l'organisation d'un concours Kaggle.

Modèles

Mode	Matériel	Temps/Epoch	Images/sec	20M img 50 epochs
Full	K80, 12Go	33 000	30	350 jours
Bottlenecks 2layers NN	K80, 12Go	500	2000	5.8 jours

- **Solution choisie** : Inception v3 pré entraîné sur ImageNet.
- **Contraintes temporelles** : bottleneck features + réseaux de neurones à deux couches.



Mesures d'Évaluation

Nous avons utilisé **deux mesures d'évaluation** :

- Précision de classification
  $\tau_{prec} = \sum_{i=1}^{n} \frac{\mathbb{1}_{Vrai+}(i)}{n}$
- Taux de catégorisation à 90% de précision
  $\tau_{cat@90} = \sum_{i=1}^{n} \frac{\mathbb{1}_{Vrai+}(i,t,s)}{n}$ avec :
  $\mathbb{1}_{Vrai+}(i,t,s) = 1 \; si \;\; s \geq t$
  $\mathbb{1}_{Vrai+}(i,t,s) = 0 \; sinon$
  sc. $\tau_{\mathbf{prec}} \geq \mathbf{0.9}$
  s représente le score de confiance donné par le modèle et t le seuil au dessus duquel on accepte la catégorisation. Ici le modèle a la possibilité de ne pas répondre.

Résultats

Features	Classifieur	Fine Tuning	Précision
HOG	SVM	Non	0.1
ImageNet Specific Bottlenecks	SVM	Non	0.533
ImageNet Specific Bottlenecks	2NN - 4096	Non	0.598
Cdiscount Specific Bottlenecks	2NN - 4096	Non	0.614
Cdiscount Specific Bottlenecks	2NN - 4096	Oui	0.726