Note méthodologique :
Preuve de concept d'un modèle de Convolutional visionTransformer (CvT)

Sommaire

Introduction

Cette note méthodologique a pour objectif de présenter une preuve de concept visant la classification automatique de produits à partir de leurs images, en utilisant une architecture de deep learning récente : le Convolutional Vision Transformer (CvT). Ce travail s'inscrit dans une volonté d'explorer une nouvelle technique récente et d'évaluer ses avantage et limites par rapport à une autre technique développée dans le passé pour executer la même tache de classification.

Dans les sections suivantes, la note décrira successivement les principes et fonctionnement du modèle CvT, le dataset utilisé pour tester l'approche, la méthodologie de modélisation et d'optimisation du modèle, ainsi que l'analyse comparative des performances de la nouvelle et de l'ancienne approche. Différentes techniques d'interprétabilités du modèle seront ensuite discutées. Enfin, les limites de l'approche ainsi que les pistes d'améliorations seront également explorées afin de proposer une vision critique de cette preuve de concept.

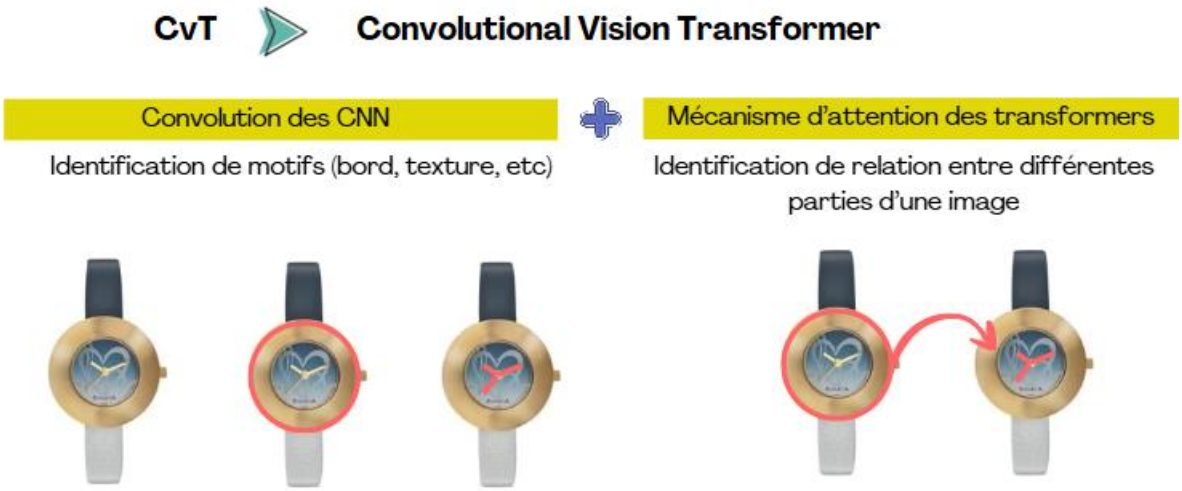
Les concepts de l'algorithme récent : le CvT

Principes généraux

Les modèles basés sur les transformateurs visuels convolutionnels (convolutionnal vision transformers ou CvT) correspondent à une famille de modèles basés sur la combinaison de deux approches majeures du deep learning :

- Les convolutions utilisées dans les réseaux de neurones convolutifs (convolutionnal neural network ou CNN) pour analyser les images
- Les mécanismes d'attentions qui sont au coeur des transformateurs et qui modélise des relations globales entre différentes parties de la séquence d'entrée.

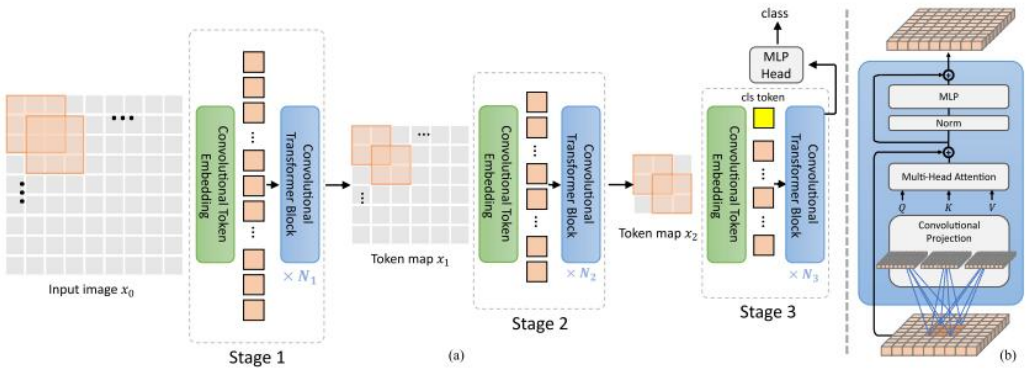
Concept clés du modèle CvT



L'idée est de profiter des forces de ces deux architectures pour la classfication d'images: la convolution pour détecter efficacement des motifs locaux comme des contours ou des textures dans une image et l'attention pour relier entre elles différentes région de l'image et comprenndre leur organisation globale.

L'approche à la base du modèle CvT a été présentée pour la première fois dans l'article de Wu *et al.* intitulé CvT: Introducing Convolutions to Vision Transformers publié en 2021.

Architecture du modèle CvT



Leur fonctionnement repose sur l'enchainement de 3 phases elles-même consituées de plusieurs blocs comprenant deux grandes étapes : **(1) l'encodage convolutionnel en token** et **(2) les blocs de transformations convolutionnels**.

Fonctionnement des modèles CvT

L'encodage convolutionnel en token

Au cours de cette étape, l'image (ou une carte de caractéristiques selon la phase) est transformée en une séquence appelée token utilisables par le transformateur de l'étape suivante. Concrètement, l'image est découpée en représentations appelées token à l'aide de filtres convolutionnels. Ces filtres parcourent l'image par petites zones, en produisant une carte de caractéristiques qui met en évidence certaines structures locales comme des contours ou des textures. Ces cartes sont ensuite normalisées afin de stabiliser les valeurs, puis transformées par une fonction d'activation, souvent la fonction ReLU, qui est là pour supprimer les valeurs négatives et conserver les valeurs positives. À la fin de cette étape, on obtient une séquence, prête à être traitée par la partie « transformateur ».

Formule de normalisation : $x_i = \frac{x_i - \mu_1}{\sigma_1}$

Les blocs de transformations convolutionnels.

Chaque bloc de transformations convolutionnels prend en entrée les tokens construits à l'étape précédente. Les token sont ensuite normalisés et transformés en trois ensembles de vecteurs appelés « requêtes » (Q), « clés » (K) et « valeurs » (V). Cette opération se fait grâce à des convolutions particulières : des convolutions par canal (qui traitent chaque canal séparément) puis des convolutions ponctuelles (qui mélangent les canaux et permettent d'augmenter la dimension des représentations). Ces trois ensembles de vecteurs servent ensuite à calculer les poids d'attention, c'est-à-dire à déterminer quelles parties de l'image doivent être mises en relation

Formule pour calculer les poids d'attention : $Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

Ces mécanismes d'attention permettent de modéliser les relations entre les différentes régions d'une image en attribuant à chacune un score reflétant son importance relative par rapport aux autres. Une fois l'attention calculée, les résultats passent par une projection linéaire, puis sont combiné avec l'entrée initiale grâce à une connexion résiduelle. Cela permet de conserver une partie de l'information brute et d'améliorer la stabilité du modèle. Enfin, un petit réseau supplémentaire, appelé bloc de propagation avant, affine encore les représentations : il élargit temporairement leur dimension, applique une fonction d'activation, puis les ramène à leur taille d'origine avant d'ajouter une nouvelle connexion résiduelle. Cette étape a pour but d'enrichir et de transformer les représentations apprises afin d'améliorer la capacité du modèle à capturer des relation complexes.

Ainsi, chaque bloc CvT construit progressivement une représentation de plus en plus riche de l'image, en combinant l'extraction locale de motifs par convolution et la compréhension globale de la structure grâce aux mécanismes d'attention. Cette approche permet d'obtenir des

modèles performants, capables de tirer le meilleur des deux techniques (convolution et mécanisme d'attention).

La classification

La représentation finale des tokens est agrégée via un pooling global, puis passée à une couche dense pour prédire la catégorie du produit parmi les sept possibles.

Présentation du dataset retenu

Le dataset retenu correspont au jeu de données de la société place de marché contenant un total de 1050 produits provenant de leur site de vente en ligne. Chacun des produits est annoté dans une des 7 catégories présentes dans le jeu de données (Baby Care, Cosmetic and personnal Care, Home Furnishing, Computer, Kitchen, Home decor and Festive needs, Watch) et est associé à une image fournie par les vendeurs. Les produits de ce jeu de données sont équitablement répartis dans chacune des catégories et il y a donc 150 produits par catégorie.

Présentation jeu de données utilisé

Présentation de la méthodologie de modélisation

Prétaitement des données

Une étape de prétraitement des images a été directement intégrée au modèle pour assurer une cohérence de l'aprrentissage lors de la phase de modélisation. Chaque image a été redimensionnée à une taille standard (224×224 pixels) et normalisée. Pour augmenter la robustesse du modèle et limiter le surapprentissage, des techniques d'augmentation de

données ont été appliquées, incluant des rotations aléatoires, des retournements horizontaux et des zooms. Ces transformations permettent au modèle de mieux généraliser sur des images légèrement différentes de celles présentes dans le jeu de données d'entrainement.

Exemple d'images générées par data augmentation

Modélisation

La modélisation a reposé sur l'utilisation du modèle CvT pré entraîné, qui a été adapté pour répondre aux besoins spécifiques de la tâche de classification automatique de produits. Afin de tirer parti des représentations déjà apprises, le squelette du CvT a été conservé tel quel et ses poids ont été figés, de sorte qu'ils ne soient pas modifiés durant l'entraînement. Ainsi, seule la partie finale du réseau a été modifiée : la couche de classification d'origine a été remplacée par une tête de classification personnalisée, composée d'un petit réseau entièrement connecté comprenant une couche linéaire, suivie d'une activation ReLU, d'un dropout pour la régularisation, puis d'une sortie linéaire dimensionnée au nombre de classes cibles.

L'entraînement a donc porté exclusivement sur cette nouvelle tête de classification, et a été optimisé à l'aide de l'optimiseur Adam qui a été appliqué aux paramètres entrainables des couches ajoutées au modèle CvT. Enfin, la fonction de perte utilisée pour optimiser le modèle lors de l'entrainement correspond à la CrossEntropyLoss, qui est une métrique adaptée aux problèmes de classification multi-classes.

Afin d'optimiser la convergence et d'éviter le surapprentissage, un mécanisme d'arrêt anticipé a été mis en place. Cela signifique que l'entraînement est interrompu si les performances de validation cessent de s'améliorer après un certain nombre d'époques. De plus, le modèle a été automatiquement enregistré au moment où il atteignait ses meilleures performances, de façon à garantir la conservation de la version la plus performante pour les étape ultérieures d'évaluation et d'analyse des résultats.

Métrique d'évaluation

La performance du modèle a été évaluée à l'aide de l'accuracy qui représente le pourcentage de produits correctement classés. La formule de l'accuracy est la suivante :

$$Accuracy = \frac{Nombre\ de\ prédictions\ correctes}{Nombre\ total\ de\ \text{prédictions}}$$

Le dataset étant équilibré (150 produits par catégorie), cette métrique reflète de manière fiable la capacité du modèle à classer correctement les produits. L'analyse a été complétée par une

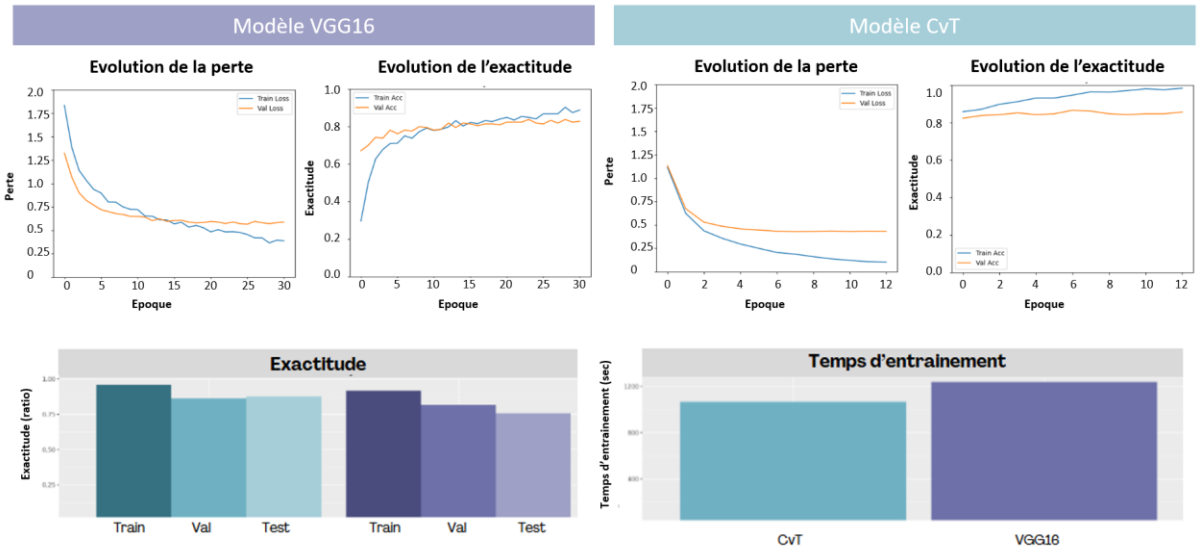
analyse des matrices de confusion et du rapport de classification qui apportent des informations plus précises sur les différents succès et erreurs du modèle.

Démarche d'optimisation

En complément de l'implémentation de l'arrêt anticipé de l'entraînement et de l'ajout d'une étape d'augmentation de données, l'optimisation du modèle CvT a reposé sur plusieurs aspects. En particulier, les hyperparamètres clés du modèle, tels que le gel du squelette du modèle CvT pré entrainé, le taux d'apprentissage, la taille de la couche dense ajoutée ou encore les paramètres de régularisation comme le dropout, ont été ajustés à l'aide de l'outil d'optimisation bayésienne Optuna, permettant d'identifier la combinaison la plus performante pour la tâche de classification.

Synthèse des résultats

Analyse comparatives des résultats – Performance globale



Les deux modèles, VGG16 et CvT, ont été entraînés dans des conditions similaires, avec le même nombre d'époques prévues et la même configuration d'arrêt prématuré. Toutefois, l'entraînement du modèle CvT s'est arrêté plus tôt que celui de VGG16. Dans les deux cas, la fonction de perte a diminué dès les premières époques, à la fois sur les données d'entraînement et de validation. Elle s'est ensuite stabilisée sur les données de validation autour de la 10^{ième} époque pour VGG16, et dès la 4^{ième} époque pour CvT. L'évolution de l'exactitude est apparue moins marquée sur les données de validation pour les deux modèles. Sur les données d'entraînement, on observe en revanche une augmentation plus nette en début d'apprentissage pour le modèle VGG16 que pour le modèle CvT.

Ces résultats suggèrent que, bien que le modèle CvT atteigne rapidement une stabilité sur les données de validation, le modèle VGG16 présente une progression plus classique sur les données d'entraînement

Concernant, les performances obtenues, le modèle CvT a attaint une accuracy globale de 88 % sur l'ensemble du jeu de test, ce qui représente une amélioration notable par rapport aux résultats du modèle VGG16 initialement développé (76 %). De plus, il est à noter que le temps

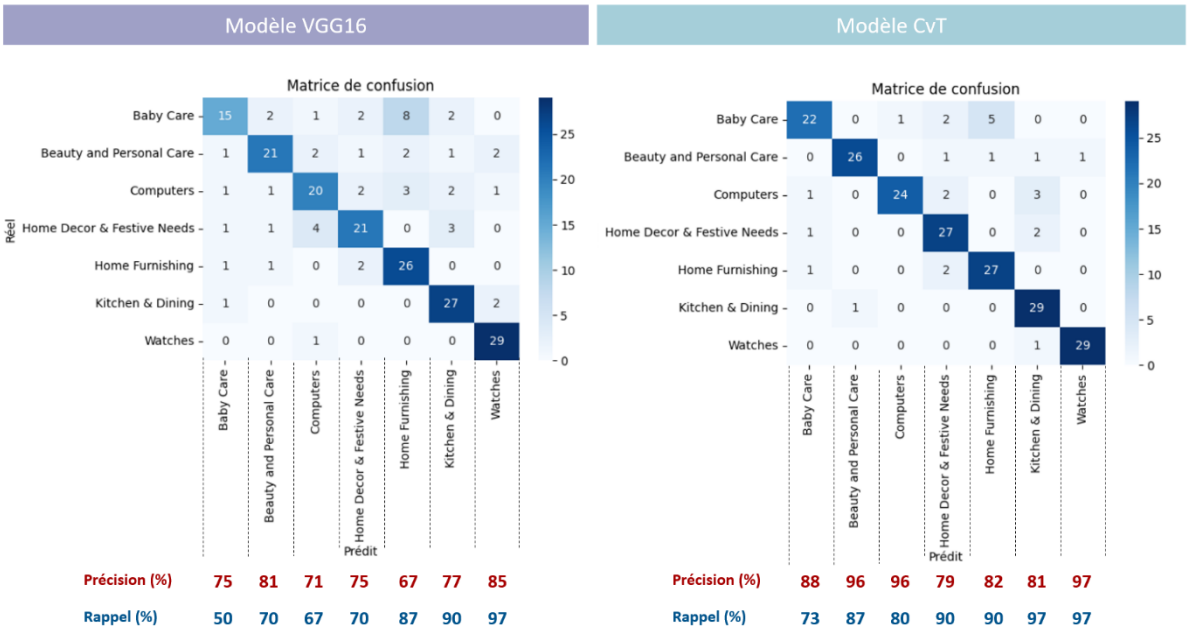
d’entrainement du modèle CvT est légèrement plus court ce qui rend cette approche plus efficace tant sur le plan des performances que sur le plan computationnel.

Pour aller plus dans le détail, les scores de rappel et de precision mesurés avec le modèle CvT sont globalement plus élevés et homogènes, avec des score de précisions et de rappel compris entre 76% et 97% et 77% et 97% respectivement.

Les catégories Watches et Kitchen & Dining ont obtenu des scores particulièrement élevés, avec un rappel supérieur à 95 %, traduisant une excellente capacité du modèle à identifier correctement ces classes. Les classes Beauty and Personal Care et Computers montrent également de très bonnes performances, combinant précision élevée et rappel satisfaisant.

Les performaces associées aux catégories Baby Care, Home Decor & Festive Needs, et Home Furnishing sont légèrement inférieures mais restent néanmoins très satisfaisantes.

Analyse comparatives des résultats – Analyse des erreurs



En comparaison avec le modèle VGG16, le modèle CvT présente donc un gain d’environ 12 points d’exactitude, ainsi qu’une nette progression de la performace sur la majorité des classes, en particulier pour la catégorie Beauty and Personnal Care et Home Decor. Cette amélioration suggère que la combinaison de la convolution et des mécanismes d'attention offre une meilleure généralisation et une réduction des erreurs de classification, tout en maintenant une bonne cohérence entre les classes.

Analyse de la feature importance globale et locale du nouveau modèle

L’analyse de la feature importance globale et locale du CvT s’est basée sur plusieurs approches dont: l’approche Grad-CAM permettant une analyse visuelle locale et l’approche SHAP permettant de quantifier l’importance des features à la fois globalement et localement.

SHAP : analyse de l'importance globale et locale

La méthode SHAP (SHapley Additive exPlanations) est une méthode basée sur la théorie des valeurs de Shapley en théorie des jeux, permettant d'attribuer à chaque caractéristique sa contribution à la prédiction d'un modèle. Pour les modèles de vision comme le CvT, SHAP peut être appliqué sur les pixels après prétraitement. Elle permet d'étudier l'importance de ces pixel à la fois localement et globalement

- Importance locale : pour une image donnée, SHAP calcule la contribution de chaque pixel à la prédiction finale, indiquant quelles régions ont influencé positivement ou négativement la classification de l'image en question.

- Importance globale : en agrégeant les valeurs SHAP sur un sous ensemble du dataset, il est possible de déterminer quelles caractéristiques visuelles sont les plus importantes pour la performance du modèle dans toutes les classes. Par exemple, il serait possible que certaines textures ou combinaisons de couleurs puissent apparaître systématiquement comme déterminantes pour une catégorie particulière.

Dans le cas des réseaux de neurones, la méthode Gradient Explainer de SHAP exploite les gradients du modèle pour approximer les valeurs de Shapley, ce qui permet d'expliquer efficacement les prédictions de modèles de deep learning tout en restant adapté aux données de grande dimension comme les images.

Grad-CAM : interprétation locale des activations

L'approche Grad-CAM (Gradient-weighted Class Activation Mapping) est une autre technique qui permet de visualiser les régions d'une image les plus influentes pour la prédiction d'une classe spécifique. La méthode repose sur différentes étapes :

1. Le calcul du gradient de la probabilité de la classe prédite par rapport aux cartes de caractéristiques (feature maps) d'une couche convolutionnelle du modèle.

2. Le moyennage spatiale de ces gradient pour obtenir des poids représentant l'importance de chaque feature map pour la prédiction.

3. La somme pondérée des cartes de caractéristiques permettant de générer une heatmap superposée à l'image originale, indiquant les zones les plus déterminantes pour la décision du modèle.

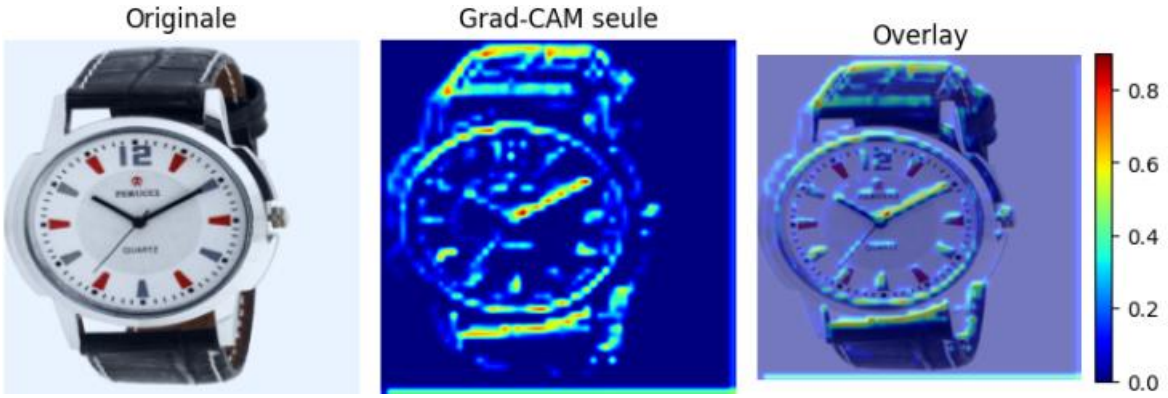
La méthode Grad-CAM permet ainsi de visualiser quelles parties des images des produits ont été utilisées par le CvT pour classer correctement chaque produit dans sa catégorie. Cette méthode fournit une interprétation locale, c'est-à-dire spécifique à chaque image individuelle.

Synthèse des résultats

L'analyse avec SHAP Gradient Explainer a produit des résultats difficilement interprétables, les cartes d'importance restant très homogènes entre les différents pixels de l'image. Cette limite s'explique par le fait que l'approche considère chaque pixel de façon isolée, alors qu'un réseau convolutionnel construit ses prédictions à partir de motifs et de structures plus complexes. En conséquence, SHAP ne reflète pas réellement la logique d'apprentissage du

modèle. À l'inverse, la méthode Grad-CAM s'est révélé plus pertinente, car cette méthode met en évidence les régions activées par les couches convolutionnelles, permettant ainsi d'identifier des zones visuellement cohérentes et directement reliées aux décisions du modèle.

Inteprétabilité des décisions prises par le modèle avec l'approche grad CAM



Limites et améliorations possibles

Bien que le modèle CvT ait montré des performances très satisfaisantes sur la tâche de classification automatique des produits, plusieurs limites inhérentes à l'architecture sont apparues ou pourraient apparaitre sur d'autres jeux de données et ouvrent la voie à des pistes d'amélioration.

La première limite concerne la capacité du modèle a gérer l'ambiguïté des catégories. Certains produits peuvent en effet appartenir à plusieurs classes (par exemple, des serviettes classées parfois dans la catégorie Baby Care et parfois dans la catégorie Home furnishing). Le modele CvT, comme la plupart des modeles de classification, est contraint de choisir une seule catégorie, ce qui peut entraîner des erreurs même lorsque la prédiction reste pertinente. Cette difficulté souligne une limite structurelle du modèle face à des problèmes multi-label ou à des taxonomies complexes. Des solutions pourraient inclure la mise en place d'une classification multi label ou l'utilisation d'une hiérarchie de classes, permettant au modèle de mieux capturer les relations entre catégories et de refléter plus fidèlement la réalité des produits.

Une autre limite est liée à la capacité du modèle à généraliser à partir d'un petit jeu de données. Avec seulement 630 images pour l'entraînement, le CvT peut avoir du mal à apprendre des représentations robustes, ce qui reflète une contrainte structurelle du modèle vis à vis de la quantité de données disponibles. Même si la collecte de nouvelles images ou l'utilisation de techniques plus avancées de data augmentation (transformations géométriques, photométriques, génération d'images synthétiques via GAN ou modèles de diffusion) pourrait pallier ce problème, cette limitation souligne la dépendance du CvT à un volume suffisant et varié de données pour atteindre son plein potentiel.

Un autre aspect critique est le coût computationnel du CvT. Les modèles de type Vision Transformer, et particulièrement le modèle CvT, nécessitent beaucoup de ressources (mémoire GPU et temps d'entraînement), ce qui peut limiter leur déploiement industriel à

grande échelle ou sur des infrastructures contraintes. Cette limitation du modèle pourrait être atténuée par des techniques d'optimisation comme la distillation de connaissance qui consiste à transferer les performances d'un modèles complexe vers un modèle plus léger ou encore le pruning qui consiste à supprimer certains paramètres d'un modèle pour gagner en temps de calcul tout en restant performant.

Bibliographie

Articles scientifiques :

Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L. 2021 CvT: Introducing convolutions to vision transformers, https://doi.org/10.48550/arXiv.2103.15808

Chollet F., 2017. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, https://doi.org/10.48550/arXiv.1610.02357

Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D., 2016. Grad-CAM : Visual Explanations from Deep Networks via Gradient-based Localization, https://doi.org/10.48550/arXiv.1610.02391

Site internet :

Raphael Kassel, Qu'est ce que la méthode Grad-CAM ? 2021. https://datascientest.com/grad-cam