

ConvNeXt-Base

Dataset retenu

Le dataset se concentre sur la classification automatique des articles à partir des images produits, dans le but de pallier les limites de l'attribution manuelle actuellement réalisée par les vendeurs. Chaque exemple du jeu de données comprend une image illustrant un article ainsi qu'un label correspondant à sa catégorie (montres, accessoires, ordinateur, etc.). Ces visuels permettent d'extraire des caractéristiques visuelles pertinentes (forme, couleur, texture, style) qui peuvent être exploitées par des modèles de vision par ordinateur pour apprendre à distinguer les différentes classes. Cette approche exclusivement basée sur l'image est particulièrement utile dans les cas où les descriptions textuelles sont absentes, imprécises ou peu informatives. Elle constitue une première étape vers un système de catégorisation automatique plus robuste, visant à améliorer la qualité des données, à fluidifier la mise en ligne des articles pour les vendeurs et à offrir une navigation plus cohérente pour les acheteurs.



Les concepts de l'algorithme récent

Le modèle ConvNeXt-Base est un réseau de neurones conçu pour analyser des images, en combinant des idées des anciens modèles de convolution avec des techniques récentes inspirées des Transformers. Il a été développé pour moderniser les architectures classiques comme ResNet, sans pour autant les abandonner. ConvNeXt-Base applique plusieurs améliorations, comme une normalisation plus efficace et une structure simplifiée mais performante. Avec environ 88 millions de paramètres, il obtient d'excellents résultats sur des ensembles d'images comme ImageNet, tout en restant rapide et efficace. Ce modèle a été proposé par Liu et al. dans leur article "A ConvNet for the 2020s" (Liu et al., 2022).

ConvNeXtBase fait partie d'une famille d'algorithmes appelés réseaux de neurones convolutifs (ou CNN en abrégé). Ce sont des systèmes qui imitent, de façon très simplifiée, le fonctionnement du cerveau humain pour traiter des images. Ces réseaux sont utilisés depuis longtemps en vision par ordinateur, c'est-à-dire dans tous les domaines où un ordinateur doit "voir" et analyser une image.

Ces dernières années, une nouvelle famille d'algorithmes, les Transformers, a beaucoup fait parler d'elle car elle donne d'excellents résultats, notamment dans le traitement des images et des textes. Cependant, ces modèles sont souvent très lourds à faire fonctionner. Ils demandent beaucoup de puissance, de mémoire et de temps.

ConvNeXtBase a été créé pour combiner le meilleur des deux mondes. Il garde la structure simple et rapide des CNN traditionnels, mais il emprunte aussi plusieurs idées efficaces aux Transformers, comme une meilleure façon de normaliser les données et d'analyser le contexte d'une image. Cela le rend à la fois puissant, moderne et plus efficace que les anciens modèles.

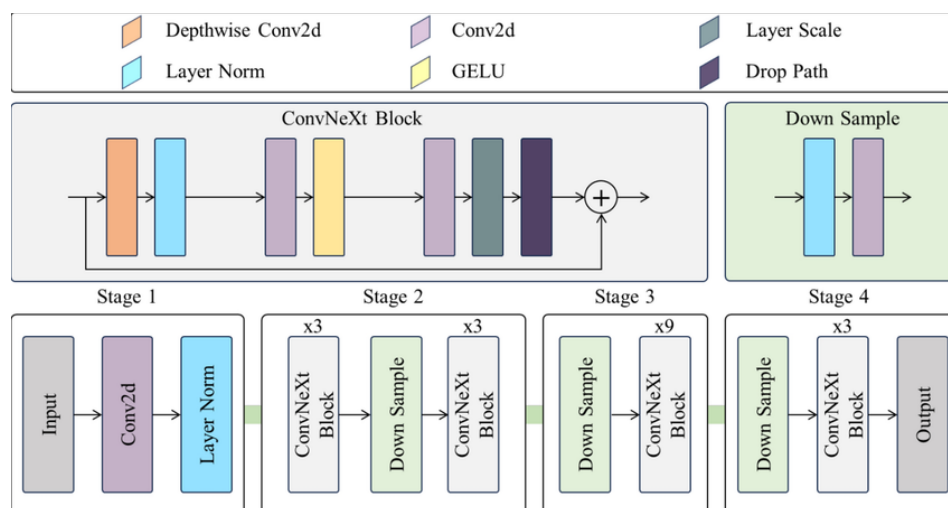


Figure : Architecture de ConvNeXtBase

L'architecture est composée de 4 étapes (stages) successives, comprenant des blocs ConvNeXt empilés. Chaque bloc inclut une convolution depthwise, une normalisation LayerNorm, une activation GELU, une convolution linéaire, et une connexion résiduelle avec Drop Path. Des modules de Down Sampling entre les stages réduisent la résolution spatiale tout en augmentant la profondeur des features. Cette structure permet d'extraire progressivement des représentations visuelles globales tout en conservant une architecture purement convolutionnelle, optimisée pour la classification d'images

La modélisation

La modélisation dans le cadre de ce projet consiste à construire un pipeline de reconnaissance visuelle permettant à un algorithme d'intelligence artificielle d'analyser automatiquement des images et de les regrouper selon leur similarité visuelle. Contrairement aux CNN classiques comme VGG ou ResNet, ConvNeXt a été repensé pour intégrer certaines techniques modernes inspirées des

Transformers (comme ceux utilisés dans ChatGPT), tout en conservant les avantages des architectures purement convolutives. L'objectif est d'exploiter les capacités d'un modèle préentraîné sur le jeu de données ImageNet, riche de millions d'images, et de le réutiliser comme extracteur de caractéristiques visuelles (ou "features"). Cela permet de bénéficier de sa capacité à détecter des motifs, formes, textures et structures globales, sans avoir à tout réentraîner depuis zéro.

Dans notre approche, chaque image est d'abord transformée et normalisée, puis passée dans ConvNeXtBase (avec la couche de classification supprimée) afin d'en extraire une représentation vectorielle. Ce vecteur contient des centaines de dimensions numériques qui encodent les éléments visuels essentiels de l'image. Ces représentations sont ensuite utilisées pour faire du regroupement non supervisé avec des algorithmes de clustering comme KMeans. L'idée est que des images ayant des caractéristiques similaires devraient, dans un espace mathématique abstrait, être proches les unes des autres. Par conséquent, elles seront affectées au même groupe ou "cluster".

Sur le plan mathématique, plusieurs notions entrent en jeu. Le réseau de neurones convolutionnel repose sur des opérations de convolution, où des filtres (ou noyaux) glissent sur l'image pour extraire des motifs locaux. Ces convolutions sont suivies d'activations non linéaires (comme GELU), de normalisations (LayerNorm), et de résidus, qui permettent d'entraîner des réseaux profonds sans perte d'information (voir figure). À la sortie du modèle, les vecteurs produits vivent dans un espace de grande dimension, où la distance euclidienne ou des métriques comme la cosinus similarity servent à comparer les images entre elles. Le clustering vient alors segmenter cet espace selon des centres calculés mathématiquement, souvent via minimisation de variance intra-cluster.

Pour évaluer la qualité du regroupement des images par le modèle, nous utilisons une métrique appelée Adjusted Rand Index (ARI). L'ARI mesure le degré de correspondance entre les clusters prédits par le modèle et les classes réelles des images (si elles sont connues). Contrairement à la simple accuracy, l'ARI prend en compte toutes les paires d'éléments et vérifie si elles sont placées ensemble ou séparées dans les deux groupements (le vrai et le prédit). Il corrige également le score attendu par hasard, ce qui le rend plus robuste et plus juste, notamment quand les classes sont déséquilibrées ou nombreuses. L'ARI varie entre -1 et 1 : un score de 1 signifie que les clusters correspondent parfaitement aux classes, 0 signifie que le groupement est équivalent au hasard, et des valeurs négatives indiquent une mauvaise correspondance.

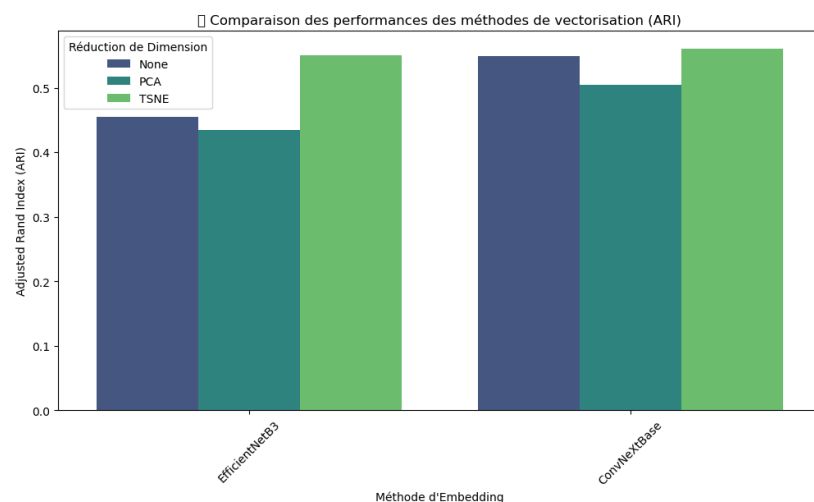
L'intérêt de l'ARI est qu'il permet de juger la qualité du regroupement visuel indépendamment des étiquettes exactes. Par exemple, peu importe si le modèle

appelle le cluster "1" ou "2" tant que les éléments similaires sont bien regroupés. Cela est très utile dans les cas non supervisés ou lorsqu'on cherche à vérifier si les embeddings extraits par un modèle capturent bien la structure sous-jacente des données. En résumé, cette modélisation mêle des techniques avancées de vision par ordinateur, des fondements mathématiques solides et une évaluation fine via l'ARI pour analyser la qualité des regroupements d'images produits à partir des features visuelles apprises.

Une synthèse des résultats

Afin d'évaluer la qualité des représentations visuelles extraites par le modèle ConvNeXtBase, nous avons comparé EfficientNetB3 (qui avait donné les meilleurs résultats dans la classification d'image précédemment) et ConvNeXtBase à travers une analyse de clustering basée sur l'indice ARI.

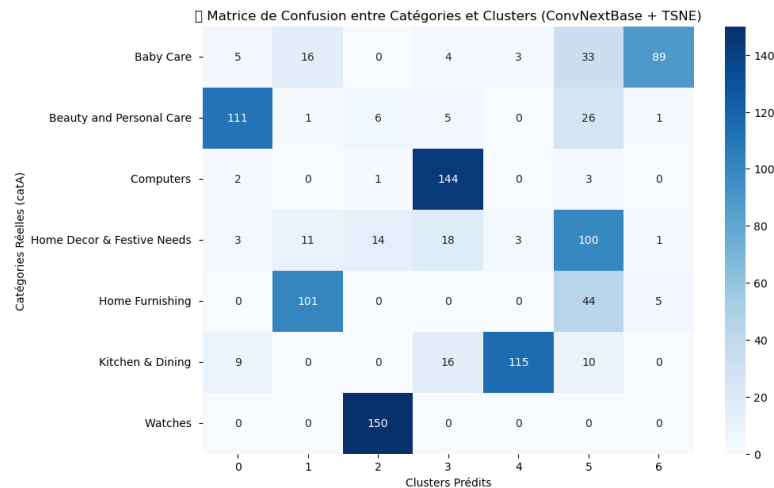
Cette figure met en évidence la supériorité de ConvNeXtBase par rapport à



EfficientNetB3 pour la tâche de vectorisation d'images, évaluée via l'Adjusted Rand Index (ARI). On observe que, quelle que soit la méthode de réduction de dimension utilisée (aucune, PCA ou t-SNE), ConvNeXtBase obtient des scores ARI systématiquement plus élevés. Cela s'explique par le fait que ConvNeXtBase est une architecture plus récente et mieux optimisée pour capter des caractéristiques visuelles à la fois locales et globales. Contrairement à EfficientNetB3, qui est très compact mais parfois trop compressé, ConvNeXtBase adopte des blocs de traitement plus profonds, avec des mécanismes inspirés des Transformers (comme la normalisation LayerNorm ou les grandes convolutions), ce qui lui permet d'extraire des représentations plus riches, structurées et discriminantes. Ces embeddings facilitent ensuite le regroupement d'images similaires dans l'espace des features, ce qui se traduit par un meilleur score ARI. En résumé, ConvNeXtBase

offre une meilleure qualité de représentation, ce qui le rend plus adapté aux tâches de clustering ou d'analyse non supervisée.

La matrice de confusion montre les correspondances entre les catégories réelles des produits et les clusters prédits après application de l'algorithme de clustering sur les vecteurs extraits par le modèle ConvNeXtBase, suivie d'une réduction de dimension par t-SNE. L'objectif de cette figure est d'évaluer dans quelle mesure les représentations générées par le modèle permettent de regrouper correctement les images par similarité visuelle, même sans supervision directe.

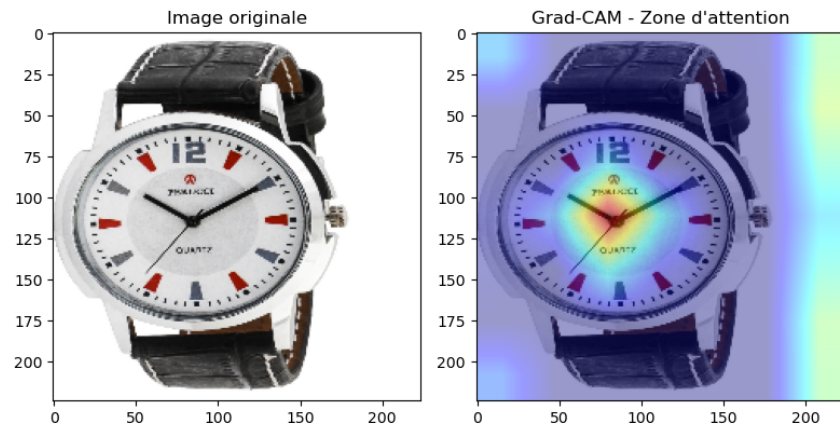


On constate tout d'abord que certaines catégories sont très bien isolées. Par exemple, les catégories "Watches" et "Computers" présentent une forte concentration dans un seul cluster (respectivement le cluster 2 et le cluster 3), ce qui indique que le modèle a su extraire des caractéristiques visuelles suffisamment distinctives pour permettre un regroupement cohérent. Ces catégories ont probablement des éléments visuels très spécifiques (formes, textures, structures) que le modèle identifie facilement.

En revanche, d'autres catégories comme "Baby Care" ou "Home Decor & Festive Needs" sont réparties sur plusieurs clusters. Cela traduit une plus grande hétérogénéité visuelle au sein de ces catégories, ou une similarité visuelle avec d'autres classes. Par exemple, "Baby Care" est confondu en partie avec les clusters de "Kitchen & Dining" et "Beauty and Personal Care", ce qui peut s'expliquer par la présence d'emballages similaires, de couleurs proches ou d'objets difficilement distinguables par des critères purement visuels. De même, la catégorie "Home Decor & Festive Needs" est très dispersée, ce qui peut indiquer que les images de cette classe sont variées en termes de contenu, rendant leur regroupement plus difficile.

Un autre élément important à noter est que certaines catégories comme "Home Furnishing" ou "Beauty and Personal Care" se chevauchent partiellement dans les clusters, notamment dans le cluster 0 et 1. Cela peut être le signe d'une frontière

floue entre ces classes, ou d'un manque de diversité visuelle dans les données d'entraînement du modèle.



En résumé, cette matrice confirme que ConvNeXtBase combiné à t-SNE est capable d'encoder efficacement certaines classes très visuelles et spécifiques, mais montre aussi les limites du clustering non supervisé lorsque les catégories sont visuellement proches ou internes variables. Cela souligne l'intérêt d'un ajustement fin du modèle ou l'utilisation de méthodes supervisées pour mieux séparer les catégories ambiguës.

L'analyse de la feature importance globale et locale du nouveau modèle

La méthode Grad-CAM est particulièrement utile pour analyser les "features" (caractéristiques) qu'un modèle de vision par ordinateur, comme ConvNeXtBase, utilise pour prendre une décision. Elle permet de visualiser les zones de l'image qui ont le plus influencé la prédiction du modèle, en générant une carte de chaleur superposée à l'image d'origine. Sur l'exemple ci-dessus, l'image montre une montre-bracelet. Sur la partie droite, la carte Grad-CAM révèle que le modèle se concentre principalement sur le centre du cadran, notamment les aiguilles, les chiffres et le logo. Ce sont des éléments visuellement distinctifs qui aident le modèle à comprendre qu'il s'agit d'une montre. Le reste de l'image (le bracelet ou le fond blanc) est beaucoup moins mis en valeur, ce qui est logique, car ces zones n'apportent pas d'information unique pour la classification. Grad-CAM offre donc une explication visuelle intuitive qui permet de vérifier que le modèle s'appuie sur les bonnes parties de l'image et pas sur des détails sans rapport, ce qui est essentiel pour la confiance et la compréhension des modèles de deep learning.

Les limites et les améliorations possibles

Malgré ses performances impressionnantes, le modèle ConvNeXtBase présente plusieurs limites. Tout d'abord, il dépend fortement de son pré-entraînement sur des jeux de données génériques comme ImageNet, ce qui peut réduire sa pertinence lorsqu'il est appliqué à des domaines spécifiques (santé, industrie, etc.) sans phase d'adaptation. Sa taille importante et le nombre élevé de paramètres en font un modèle relativement lourd, avec un coût de calcul élevé, notamment en termes de mémoire GPU et de temps d'inférence. De plus, comme beaucoup de modèles profonds, ConvNeXt reste une boîte noire dont les décisions sont difficiles à interpréter. Enfin, le modèle peut se montrer sensible à des variations visuelles comme le bruit, les changements de luminosité ou les rotations, ce qui peut impacter ses performances si ces cas ne sont pas bien représentés dans les données d'entraînement.

Pour surmonter ces limites, plusieurs pistes d'amélioration peuvent être envisagées. La plus directe consiste à réaliser un fine-tuning ciblé sur le domaine d'application, afin d'adapter le modèle aux spécificités des nouvelles données. Pour alléger le modèle et le rendre plus rapide, on peut utiliser des techniques de compression comme la quantification ou le pruning (suppression de certaine partie pour alléger le modèle) , qui réduisent la taille du réseau sans compromettre significativement sa précision. Une autre piste prometteuse est l'intégration de mécanismes d'attention ou la fusion avec des architectures hybrides combinant convolutions et Transformers, pour améliorer la prise en compte du contexte global. Enfin, des méthodes d'explicabilité avancées comme SHAP ou Grad-CAM, ainsi qu'une stratégie d'augmentation de données robuste, peuvent contribuer à rendre le modèle plus transparent, plus fiable et mieux généralisé à des situations variées.