**Note méthodologique : preuve de concept**

**1. Introduction**

Dans le cadre de notre démarche de veille technologique, nous avons choisi d’évaluer l’architecture Vision Transformer (ViT), proposée par Dosovitskiy et al. en 2020, pour la classification d’images. Ce choix s’inscrit dans une volonté d’explorer les architectures récentes qui dépassent les performances des CNNs classiques tels que ResNet50.

**2. Présentation de la méthode**

Le Vision Transformer (ViT) applique le principe des Transformers à la vision par ordinateur. L’image d’entrée est divisée en patchs (par exemple, 16x16 pixels), chaque patch étant linéarisé et projeté dans un espace de dimension fixe. Ces vecteurs sont ensuite traités comme une séquence par un modèle Transformer standard, sans aucune opération de convolution. L’architecture permet de capturer des dépendances globales dans l’image, contrairement aux CNNs qui fonctionnent principalement sur des voisinages locaux.

**3. Justification du choix**

* **Récence et impact** : ViT est une architecture de moins de 5 ans, ayant suscité un grand intérêt dans la communauté scientifique et de nombreux travaux dérivés (DeiT, Swin Transformer, etc.).
* **Performance** : Sur des benchmarks comme ImageNet, ViT atteint voire dépasse les performances des meilleurs CNNs, notamment lorsque l’entraînement est réalisé sur de larges jeux de données.
* **Flexibilité** : ViT peut être adapté à différents types de tâches en vision (classification, détection, segmentation) et bénéficie d’une forte compatibilité avec les frameworks modernes (PyTorch, TensorFlow).

**4. Sources et documentation**

* Dosovitskiy, A., et al. (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv:2010.11929.
* Touvron, H., et al. (2021). "Training data-efficient image transformers & distillation through attention." arXiv:2012.12877.
* Synthèses et tutoriels sur Papers With Code, Machine Learning Mastery, Towards Data Science.

**5. Implémentation et expérimentation**

L’implémentation de ViT a été réalisée à l’aide de la bibliothèque timm (PyTorch Image Models) qui propose des modèles pré-entraînés et facilement adaptables. Les étapes principales sont :

* Prétraitement des images (redimensionnement, normalisation)
* Découpage en patchs et vectorisation
* Passage dans le modèle ViT pré-entraîné ou fine-tuné
* Évaluation sur le jeu de données utilisé précédemment avec ResNet50

**6. Comparaison avec ResNet50**

Les résultats expérimentaux montrent que ViT, même en fine-tuning sur un jeu de données de taille modérée, offre des performances comparables ou supérieures à ResNet50, notamment en termes de précision top-1. L’interprétabilité des décisions peut être renforcée par l’analyse des cartes d’attention générées par le modèle.

**7. Limites et perspectives**

* ViT nécessite généralement plus de données pour atteindre son plein potentiel, mais des variantes comme DeiT rendent la méthode plus accessible pour des jeux de données de taille moyenne.
* Les ressources de calcul requises sont plus importantes qu’avec un CNN classique, mais restent compatibles avec une expérimentation sur GPU standard.

**8. Conclusion**

L’architecture Vision Transformer représente une avancée majeure en vision par ordinateur et constitue une alternative pertinente et performante aux architectures CNN traditionnelles pour la classification d’images. Son adoption croissante dans l’industrie et la recherche confirme sa robustesse et son potentiel d’innovation.