


# Développez une preuve de concept

**Morgan May**

Projet n°7 du parcours Ingénieur Machine Learning



# Sommaire

1. Thématique
  2. Etat de l'art
    - a. Mise en contexte dans l'état de l'art de la vision par ordinateur
    - b. Etat de l'art antérieur des Vision Transformers
  3. Analyse de l'article de référence
    - a. Présentation de l'article
    - b. Architecture proposée d'un Vision Transformer
    - c. Principaux résultats obtenus
  4. Expérimentations & conclusion
    - a. Jeu de données utilisé
    - b. Modèle utilisé et analyse comparative des résultats
    - c. Perspectives & conclusion
  5. Bibliographie et sources
- 



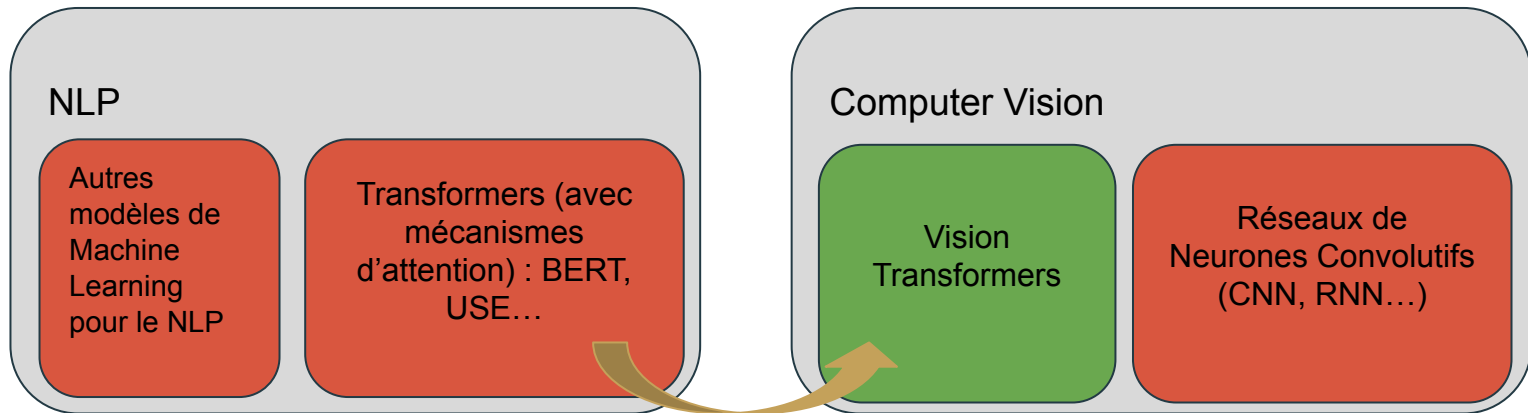
# 1. Thématique



# 1. Thématique

La thématique de veille choisie est celle des Vision Transformers.

Les Vision Transformers sont des algorithmes de Deep Learning relativement innovants car ils sont la transposition dans le domaine de la vision par ordinateur du principe des Transformers; des modèles utilisés en Traitement Automatique du Langage (NLP).

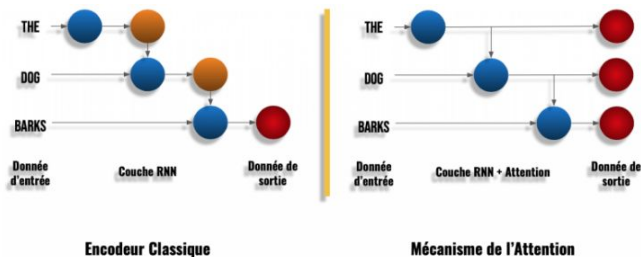


# 1. Thématique

Les Transformers sont des architectures de réseaux de neurones particulièrement adaptés pour le traitement de données séquencées comme du texte.

Développés notamment depuis 2017 et la publication *Attention Is All You Need* (Vaswani, 2017) par une équipe de Google, leur principe de fonctionnement repose sur un mécanisme d'attention dont voici une explication brève du principe pour un problème de NLP :

Les Transformers sont une forme améliorée des encodeurs-décodeurs classiques uniquement basés sur des RNN. Pour l'encodeur, dans la forme classique, une couche du RNN ne produit qu'une seule sortie pour une séquence : chaque sortie correspondante à chaque mot ne sert qu'à calculer la sortie du mot suivant.

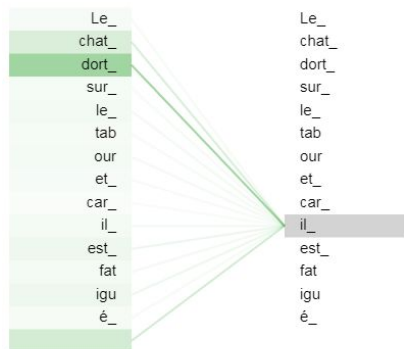


Avec un mécanisme d'attention, chaque sortie de récurrence utilisée par le RNN pour le mot suivant et correspondante à chaque mot forme également une information transmise à la couche suivante.

# 1. Thématique

Le décodeur est également un réseau de neurones récurrent dont la principale fonction est l'interprétation de la relation entre les mots.

**Le mécanisme d'attention (ici de self-attention pour une application sur une même séquence) permet ainsi au modèle de capter des notions de contexte afin de proposer une interprétation plus juste d'une même séquence.**



Visualisation d'une couche de self-attention

Sur l'exemple ci-contre de l'application d'un mécanisme d'attention sur la phrase *"Le chat dort sur le tabouret car il est fatigué"*, la visualisation proposée montre que le mot "il", une fois décodé, est fortement lié à "chat" et "dort" mais il n'est pas très lié à "tabouret" : le modèle a compris que "il" ne faisait pas référence à "tabouret" mais bien à "chat".

**C'est ce principe d'attention que les Vision Transformers tentent de transposer au traitement d'image.**



## 2. Etat de l'art

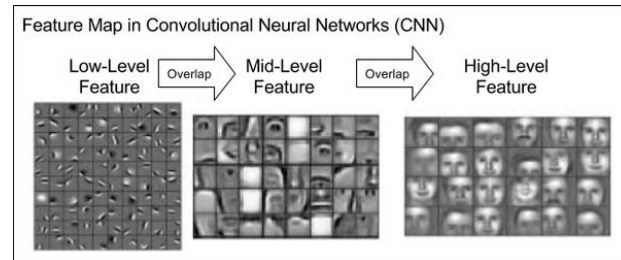
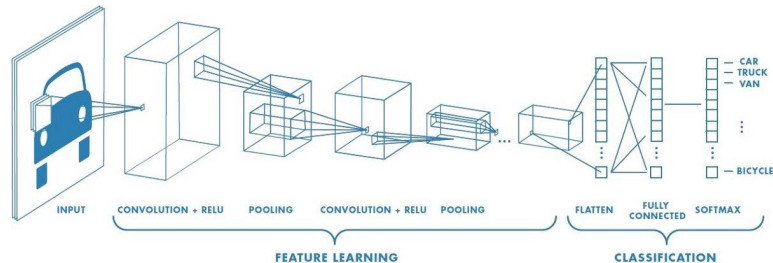


## 2. a. Mise en contexte dans l'état de l'art de la vision par ordinateur

Dans le domaine de la vision par ordinateur (classification d'image, segmentation, détection et reconnaissance d'objets...), les réseaux de neurones convolutifs sont la référence et ont permis d'énormes progrès depuis le début des années 2010 en particulier.

La convolution, au cœur de ces réseaux de neurones est une opération mathématique permettant d'extraire de l'information de plusieurs pixels répétée de multiples fois sur l'ensemble d'une image.

Le principe de ces modèles est donc d'extraire des motifs des images à l'aide de couches de convolution puis de réaliser une prise de décision (de classification par exemple) à l'aide de couches denses classiques.

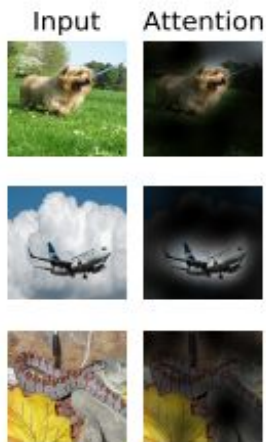




## 2. b. Etat de l'art antérieur des Vision Transformers

Les Vision Transformers (ou “ViT”) étant une transposition du principe des Transformers qui est devenu la référence pour le traitement de séquences (de textes) notamment, les essais de transposition au traitement de l'image ont rapidement suivi l'article *Attention Is All You Need* (Vaswani et al., 2017).

Le premier principal axe de recherche a donc été l'adaptation d'une image à une forme de “séquence” afin d'appliquer le mécanisme d'attention.



L'objectif visé à travers la transposition du mécanisme d'attention est de permettre à un modèle de mieux interpréter la relation entre les pixels et ainsi de se focaliser sur certains pixels plutôt que d'autres comme illustré ci-contre.

Contrairement aux réseaux de neurones convolutifs, les Vision Transformers n'apprennent pas à repérer directement des motifs mais plutôt à sélectionner les pixels utiles.



### 3. Analyse de l'article de référence



## 3. a. Présentation de l'article

L'article principalement étudié ici, sur lequel s'est fortement basée notre expérimentation, est dénommé ***"An image is worth 16x16 words : transformers for image recognition at scale"*** et a été publié en 2021 à travers la conférence internationale "International Conference on Learning Representations" par une équipe de Google.

Ce papier de recherche suggère que ces Vision Transformers atteignent des performances remarquables sur des tâches de reconnaissance d'image; parfois même meilleures que les performances de l'état de l'art réalisées par des réseaux de neurones convolutifs. Il démontre ainsi que les Vision Transformers peuvent être de bonnes alternatives aux réseaux exploitant des convolutions.

Published as a conference paper at ICLR 2021

### AN IMAGE IS WORTH 16x16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy<sup>1</sup>, Lucas Beyer<sup>2</sup>, Alexander Kolesnikov<sup>2</sup>, Dirk Weissenborn<sup>2</sup>,  
Xiaohua Zhai<sup>2</sup>, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,  
Georg Heigold, Sylvain Gelb, Jakob Uszkoreit, Neil Houlsby<sup>1</sup>

<sup>1</sup>equal technical contribution, <sup>2</sup>equal advising  
Google Research, Brain Team  
{adosovitskiy, neilhoulshy}@google.com

#### ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.

#### 1 INTRODUCTION

Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

In computer vision, however, convolutional architectures remain dominant (Lecun et al., 1989; Krizhevsky et al., 2012; He et al., 2016). Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention (Wang et al., 2018; Carion et al., 2020), some replacing the convolutions entirely (Ramachandran et al., 2019; Wang et al., 2020a). The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Therefore, in large-scale image recognition, classic ResNet-like architectures are still state of the art (Mahajan et al., 2018; Xie et al., 2020; Kolesnikov et al., 2020).

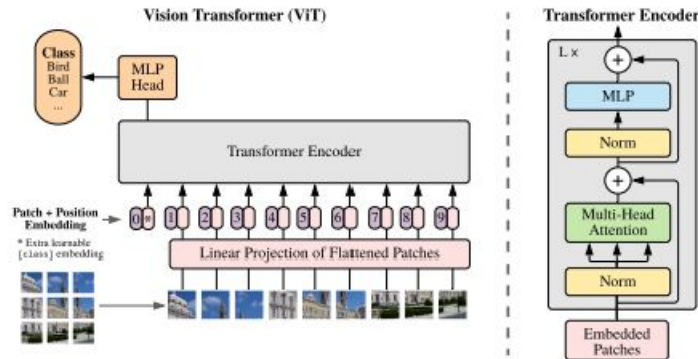
Inspired by the Transformer scaling successes in NLP we experiment with applying a standard Transformer directly to images, with the fewest possible modifications. To do so, we split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image classification in supervised fashion.

When trained on mid-sized datasets such as ImageNet without strong regularization, these models yield model accuracies of a few percentage points below ResNets of comparable size. This seemingly discouraging outcome may be expected: Transformers lack some of the inductive biases

<sup>1</sup>Pre-training code and pre-trained models are available at [https://github.com/google-research/vision\\_transformer](https://github.com/google-research/vision_transformer)

### 3. b. Architecture proposée d'un Vision Transformer

A la problématique de l'adaptation des images sous la forme de séquences, les auteurs ont repris le principe évoqué antérieurement par une équipe de recherche de l'Ecole Polytechnique Fédérale de Lausanne "On the Relationship Between Self-Attention and Convolutional Layers" (Cordonnier et al. 2020) : découper chaque image en patches de dimensions égales et inclure un indice de localisation du patch dans la projection.



La séquence de vecteurs issue de ces projections est ensuite utilisée comme entrée d'un encodeur Transformer qui met en œuvre le mécanisme d'attention.

La sortie du Transformer est ensuite exploitée par des couches denses afin de procéder à la résolution du problème de classification comme le ferait un réseau convolutif.

### 3. c. Principaux résultats obtenus

Voici les principaux constats établis dans cet article en testant un pré-entraînement sur différents datasets connus comme ImageNet et des architectures diverses et en les comparant avec des ResNet :

- Si pré-entraînés sur des datasets très volumineux comme JFT-300M (300 millions d'images), alors les Vision Transformers sont capables de meilleurs résultats que des ResNet sur des tâches de classification d'image (sur ImageNet par exemple).
- Les Vision Transformers sont plus efficaces que les ResNet une fois pré-entraînés : ils nécessitent 2 à 4 fois moins de ressources de calcul pour atteindre des performances similaires.
- Par effet contraire, ils ne sont donc pas pertinents face à des ResNet s'ils ne sont pas pré-entraînés sur de très gros datasets ; ce qui peut être un frein à leur application dans des domaines où les données sont compliquées à agréger en nombre (comme l'imagerie médicale) et les techniques de transfer learning relativement limitées depuis des images plus classiques.



## 4. Expérimentations & conclusion



## 4. a. Jeu de données utilisé

Afin de mener des expérimentations sur un Vision Transformer, nous choisissons de réutiliser le jeu de données du projet précédent de classification d'images afin d'obtenir rapidement une bonne idée des performances d'entraînement et de résultats, notamment en le comparant avec un réseau convolutif que l'on a déjà optimisé.

Le jeu de données est donc **un ensemble de 20580 photos de chiens, classées en fonction de leur race parmi 120 races différentes**. Ce dataset est mis à disposition par l'Université de Stanford et disponible à l'adresse <http://vision.stanford.edu/aditya86/ImageNetDogs/>.

L'objectif sur lequel nous avons testé notre Vision Transformer est la classification d'une image de chien en fonction de sa race.



## 4. b. Modèle utilisé et analyse comparative des résultats

Pour réaliser le test d'un Vision Transformer sur ce dataset, **étant donné la nécessité d'être pré-entraîné sur de larges datasets, nous avons choisi d'opérer par transfer learning** en récupérant sur le dépôt Tensorflow Hub un modèle ViT pré-entraîné.

Nous avons ainsi trouvé un dépôt particulièrement pertinent car proposant différentes implémentations Tensorflow de Vision Transformers qui sont en fait **les modèles pré-entraînés et mis à disposition par l'équipe de recherche de Google sur GitHub et convertis depuis la librairie JAX.**

Le modèle que nous avons choisi est un Vision Transformer "S/16" qui est un modèle composé de **22 millions de paramètres et performant une décomposition des images en 16x16 patches. Il a été pré-entraîné sur ImageNet-21k et optimisé sur ImageNet-1k.**

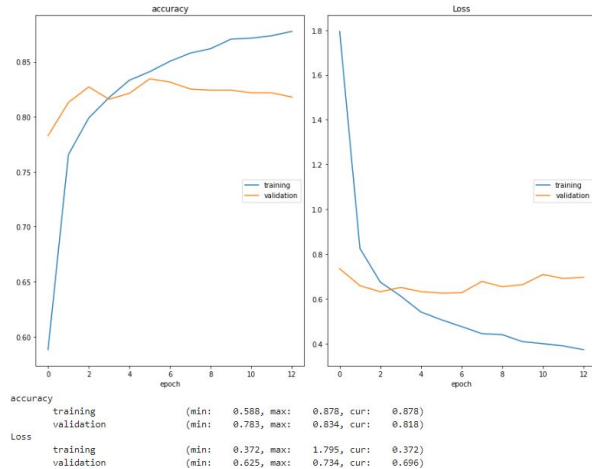
Pour le transfer learning, les paramètres correspondants aux premières couches sont figés, seuls les 46200 paramètres correspondants à la dernière couche softmax que l'on a rajoutée sont entraînés.



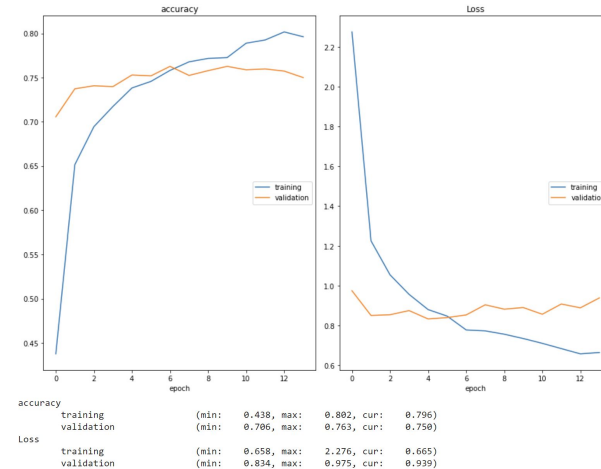
## 4. b. Modèle utilisé et analyse comparative des résultats

Nous entraînons sur le jeu de données (avec Data augmentation) notre ViT (ou plutôt sa couche softmax) en le comparant avec un ResNet-50, réseau de neurones convolutif, que nous avons optimisé légèrement lors du projet précédent :

Courbes d'apprentissage du Vision Transformer pré-entraîné :



Courbes d'apprentissage d'un ResNet-50 pré-entraîné et optimisé :



Ces courbes d'apprentissage, bien que très similaires (l'entraînement ne concernant ici que les dernières couches de toute manière), nous permet de faire principalement le constat que **les performances obtenues par le ViT sont meilleures que celles obtenues par le ResNet pourtant légèrement optimisé (83.4% d'accuracy sur le validation set contre 76.3%).**

Une première analyse de cette différence de performance pourrait conclure à une efficacité supérieure du mécanisme d'attention pour sélectionner les formes pertinentes pour la classification sur ce jeu de données dans lequel un certain nombre d'éléments (plusieurs chiens sur une photo, présence d'humains, chiens portant des vêtements...) peuvent nuire directement à la recherche de motifs effectuée par des réseaux convolutifs.

## 4. c. Perspectives & conclusion

Cet exercice de veille technologique a essentiellement permis de comprendre le principe des Vision Transformers et de vérifier ses performances sur un cas d'application simple en le comparant avec un ResNet-50.

Les performances obtenues avec ce Vision Transformer pré-entraîné sur ImageNet sont assez prometteuses et corroborent l'idée que ce type d'architecture puisse représenter une vraie alternative aux réseaux convolutifs.

Parmi les perspectives de développement de ces Vision Transformers, on peut notamment se poser la question de l'adaptation des Vision Transformers à l'exploitation non pas d'une image mais d'une séquence d'images (flux vidéo, images 3D). On pourrait ainsi imaginer recycler le principe de décomposition en patches + embedding en décomposant non pas une image mais une série d'images : chaque patch étant lui-même une image de la séquence d'image.

Enfin, la combinaison des capacités de convolution et d'attention semble théoriquement très prometteuse d'un point de vue de l'efficacité notamment en recherchant des motifs (grâce aux convolutions) sur des zones particulières de l'image (grâce à l'attention).



## 5. Bibliographie & sources

## 5. Bibliographie & sources

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, [Attention Is All You Need](#), 2017, in arXiv:1706.03762
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. [An image is worth 16x16 words: transformers for image recognition at scale](#), 2021, in ICLR 2021
- Jean-Baptiste Cordonnier, Andreas Loukas, Martin Jaggi, [On the Relationship Between Self-Attention and Convolutional Layers](#), 2020, in ICLR 2020
- Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, Alexey Dosovitskiy. [Do Vision Transformers See Like Convolutional Neural Networks ?](#) In arXiv:2108.08810
- Le mécanisme de l'attention en Deep Learning - Comprendre rapidement, inside-machinelearning.com disponible à l'adresse <https://inside-machinelearning.com/mecanisme-attention/>
- A la découverte du Transformer, ledatascientist.com, disponible à l'adresse <https://ledatascientist.com/a-la-decouverte-du-transformer/>



**Merci pour votre attention**

