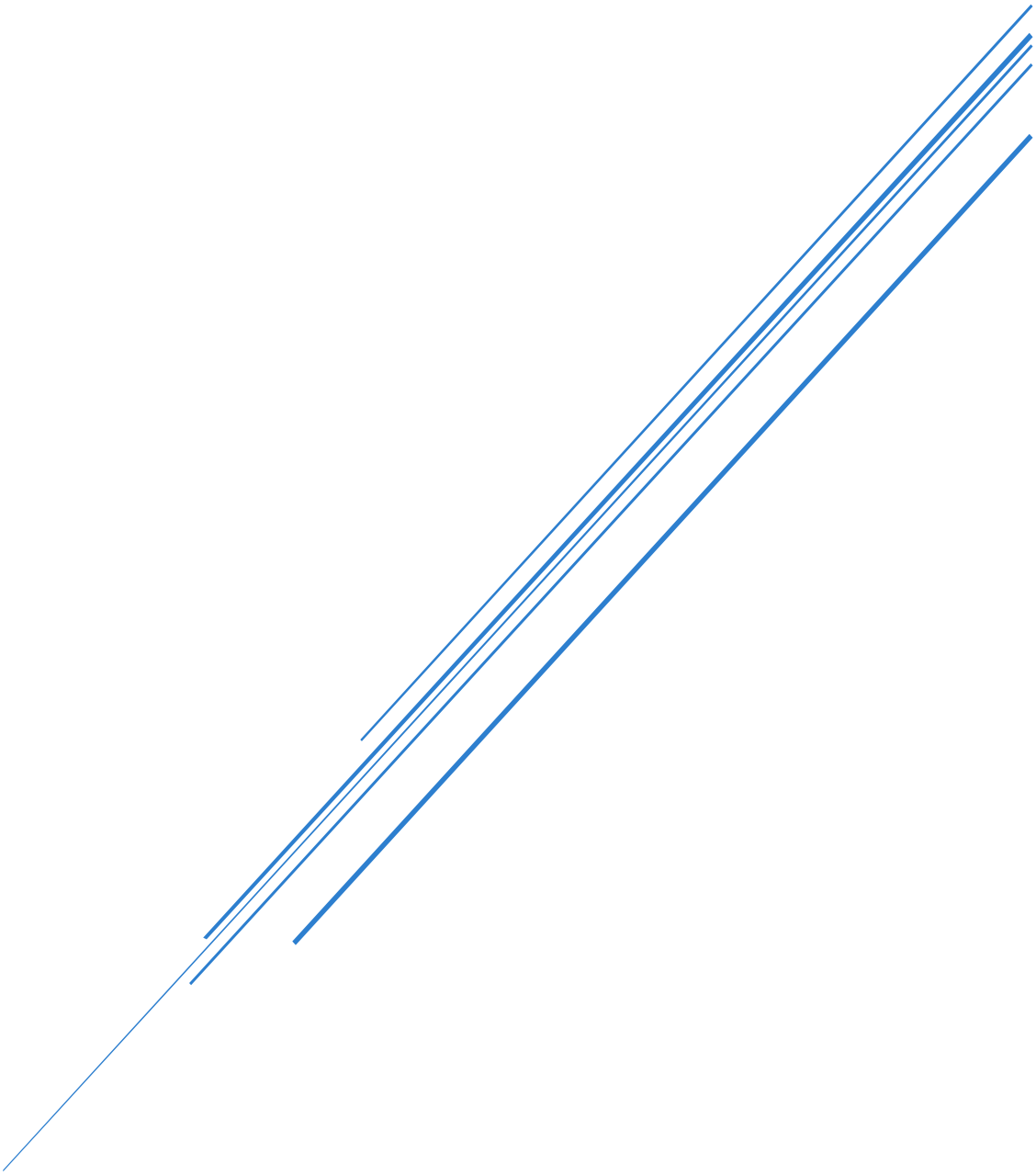
SUJET DE DEVELOPPEMENT DANS LE DOMAINE DES TRANSFORMERS

Cahier des charges



Encadrante : RAPHAELLE CHAINE

Etudiants : COTTIER ALEXANDRE et VU AHN DUY

1. Introduction

- **Présentation générale** : Les transformers sont des modèles de deep learning conçus pour traiter efficacement des séquences de données en capturant les relations entre les éléments, peu importe leur distance. Grâce à leur mécanisme d'attention, ils peuvent mettre en relation différents points d'une séquence, ce qui en fait un outil puissant pour des tâches de traduction, de génération de texte, et même d'analyse d'images. Leur flexibilité les rend également adaptés à des applications de reconstruction d'images, où l'objectif est de restaurer des données manquantes ou altérées en analysant les connexions entre les pixels restants.

2. Contexte et Objectifs du Projet

- **Contexte** : Ce projet s'inscrit dans le cadre des études sur les Transformers appliqués à la vision par ordinateur, en particulier avec l'utilisation des Vision Transformers pour la segmentation et la reconstruction d'images.
- **Objectifs** : Notre objectif principal est de trouver une application spécifique des vision transformers, comme la restauration d'images ou le redimensionnement d'images, en tant que sujet principal, plutôt qu'une étude générale des vision transformers, l'objectif de base étant d'étudier la capacité des Vision Transformers à reconstruire une image d'origine à partir de petits patchs (4x4 pixels) en utilisant le mécanisme d'attention. Nous évaluerons si l'image initiale peut être recréée avec précision à partir des variations de patchs obtenues.

3. Méthodologie, Description des Travaux Prévus :

- **Préparation des Données :** Choisir une image source, simple et de petite taille, pour faciliter l'analyse. Cette image sera segmentée en patchs 4x4.
- **Modélisation avec le Transformer :** Utiliser un modèle de Vision Transformer pour analyser et assigner une attention à chaque patch, en identifiant les parties les plus importantes pour la reconstruction.
- **Reconstruction de l'Image :** À partir des patchs pondérés par l'attention, tenter de reconstruire l'image d'origine et évaluer la fidélité de la reconstruction.
- **Analyse des Résultats :** Comparer l'image reconstruite à l'image originale pour déterminer l'efficacité de l'attention dans la préservation des informations.

4. Conception des Modèles et Choix Techniques

- **Structure du modèle** : Pour ce projet, nous allons utiliser un type de modèle de transformers adaptés à nos objectifs, le modèle Vision Transformer (ViT) pour la reconstruction d'images de petite taille.
    - **Vision Transformer (ViT)** : Le ViT est un modèle conçu pour le traitement des images en s'appuyant sur des principes similaires aux transformers utilisés en traitement du langage naturel. Contrairement aux réseaux de neurones convolutifs (CNN) qui se focalisent sur des régions spécifiques de l'image, le ViT divise l'image en une séquence de petits "patchs" ou blocs (par exemple, de 4x4 pixels) qui sont ensuite traités comme des tokens ou des mots dans une séquence. Chaque patch est linéarisé et projeté dans un espace de dimension fixe, puis un vecteur d'embedding est ajouté pour capturer des informations de position, permettant ainsi au modèle de conserver une idée de l'organisation spatiale.
- **Taille des couches (embeddings dimension) :** Comme les images sont de petite taille, une dimension d'embedding modeste (entre 32 et 64) peut suffire pour capturer l'information de chaque patch sans surcharger le modèle.
- **Nombre de têtes d'attention :** Un nombre de 4 à 8 têtes d'attention devrait être adéquat pour capturer les relations entre pixels sans complexifier excessivement le modèle.
- **Profondeur du modèle (nombre de couches) :** Pour cette tâche de reconstruction simple, 6 à 8 couches (transformer blocks) seront utilisées, équilibrant profondeur et temps d'entraînement.
- **Taille des patchs :** Chaque pixel de l'image pourrait être traité comme un "patch" de 1x1, mais pour des images un peu plus grandes, des patchs de 2x2 pourraient être envisagés.
- **Dropout :** Une régularisation avec un dropout de 0.1 à 0.2 est recommandée pour éviter le surapprentissage étant donné la taille limitée des données.5. Résultats Attendus et Critères d'Évaluation

5. Résultats Attendus

- Un ensemble de patchs 4x4 avec une carte d'attention montrant les zones prioritaires pour la reconstruction.
- Une image reconstruite qui permet d'analyser le degré de précision par rapport à l'image originale.
- Un rapport de synthèse détaillant les résultats, les conclusions et les perspectives.

6. Organisation et Planification

Durée : Plus d'un mois pour réaliser toutes les étapes, de la segmentation initiale à l'analyse finale.

Étapes Clés :

Préparation des données et segmentation en patchs (**Semaine 1**).

Implémentation du Vision Transformer et analyse par attention (**Semaine 2 et 3**).

Reconstruction de l'image et analyse des résultats (**Semaine 4**).

Production d'une vidéo a titre pédagogique (fin de **Semaine 4** voir **Semaine 5**).

Livrables : Rapport intermédiaire et final, vidéo de vulgarisation.

7. Ressources Nécessaires

Logiciels : Python, TensorFlow/PyTorch, Jupyter Notebook pour la modélisation.

Matériel : Accès à une machine avec GPU pour l'entraînement du modèle.

8. Conclusion et Perspectives

- **Conclusion attendue** : Synthèse des résultats possibles et de l'apport des transformers dans le traitement de séquences complexes.
- **Perspectives futures** : Exemples d'améliorations potentielles, comme l'utilisation de transformers pour des images de taille plus grande ou dans des tâches de restauration d'images plus complexes.