Benjamin Bouchard

Open Classroom
Plan de travail prévisionnel - Projet 7

Application du modèle ViT au Standford dogs dataset du projet 6

Vue d'ensemble

Les modèles, basés sur le mécanisme de l'Attention, sont devenus incontournables dans le domaine du NLP (exemples: BERT, GPT). Mais jusqu'à présent ces modèles, appelés Tranformers, n'avaient pas (ou alors sans succès) étaient appliqués au domaine de la vision par ordinateur.

Cependant un nouveau modèle, nommé ViT (Vision Transformer) vient d'être publié et s'annonce prometteur en comparaison des modèles à réseau de neurones convolutifs.

Le but de ce projet est donc d'appliquer ce modèle ViT au dataset Standford dogs (issu du projet 6) et de le comparer au modèle de référence actuel (ResNet …).

Il s'agira donc de comparer les performances de ces deux modèles. ***<to be completed>***

Sources et objectifs

1. Acquérir une compréhension minimum du mécanisme d'Attention dans le cadre de son utilisation dans la NLP

   Une revue des évolutions des récents modèles de NLP: A review of NLP

   Un article sur les réseaux LSTM

   Un livre sur la NLP: Modern Approaches in Natural Language Processing

Le mécanisme d'Attention a initialement été introduit en 2014 dans le cadre d'amélioration de modèles de transduction de séquences: Neural Machine Translation by Jointly Learning to Align and Translate

En 2017, l'article Attention Is All You Need qui a introduit le modèle de Transformer uniquement basé sur le mécanisme d'Attention.

2. Approfondir la compréhension de l'article introduisant le modèle ViT

L'article original: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Une vidéo sur l'article: vidéo 1 et une autre Vidéo 2

Un article de vulgarisation: Article 1 et un autre: Article 2

3. Différentes variations du modèle ont été proposées:

How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers

4. Différentes sources sur le même sujet:

-Github officiel : vision_transformer

-Le modèle est implémenté avec la nouvelle librairie FLAX de google (flax qui s'appuie sur elle-même sur JAX (jax, des vidéos d'introduction sur le sujet: Vidéo 1 Vidéo 2)) mais un "portage" pour l'utilisation vers TensorFlow est disponible dans ce github: ViT-jax2tf plus un collab qui explique comment "fine tuner" le modèle (fine_tune.ipynb).

-Certains des "checkpoints" sont directement disponibles sur Tensorflow hub: TFhub: vision_transformer. Il s'agira d'utiliser l'un deux pour faire du transfert learning.

- La description du modèle sur HuggingFace: HuggingFace ViT Model