Master's Thesis: Sequence-to-Sequence Models in Computer Vision

Contact: m.kwiatkowski@tu-berlin.de

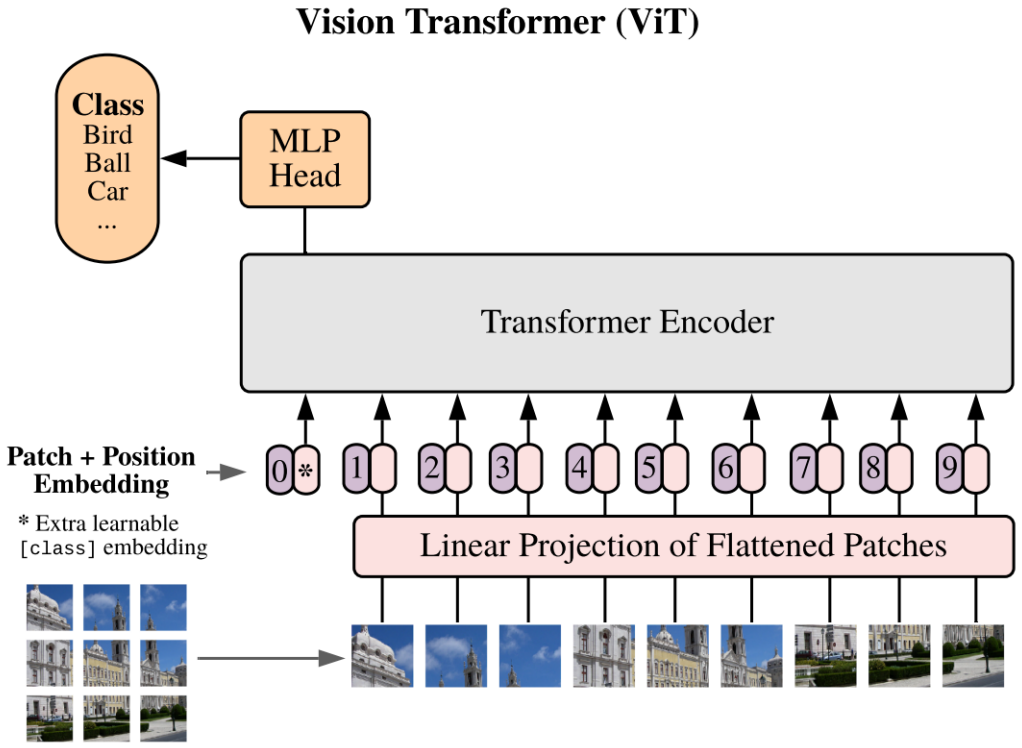
February 8, 2022

1 Introduction

Deep Learning models have been successfully used in many Computer Vision problems, such as classification, object detection, segmentation etc. Many of these models are designed to work on single images. However, when designing machine learning models for videos or sets of images, one is often faced with additional problems:

- Memory consumption
- Ordered/unordered data
- flexible sequence length/set size
- Various modes of sequential modelling:
  - One-to-One: most common case
  - One-to-Many: Image Captioning
  - Many-to-One: Video classification, Background Extraction, 3D Reconstruction (images to 3D model)
  - Many-to-Many: Video captioning, Video Segmentation, Object Tracking, Video prediction, Video Super Resolution

Several deep learning architectures have been introduced to process sequential data, such as Recurrent Neural Networks, LSTMs [3] and Transformers [4]. Applying these models directly on sequences of images is non-trivial. Many adjustments have been made to process videos, such as 3D-CNNs, Convolutional LSTMs and Vision Transformers [2, 1, 5].



2 Objective

The goal of this thesis is to do a comparative study of architectures for any of the following video-to-video tasks:

- Video Segmentation
- Object tracking
- Video Super-Resolution

The task is to train 3D-CNNs, LSTMs and Transformers on these tasks and evaluate their performances. The overall goal is to understand the advantages and limitations of each architecture in analyzing spatio-temporal data and image sets.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[3] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[5] Niccolò Zanichelli. Iaml distill blog: Transformers in vision, 2021.