Vision Transformers

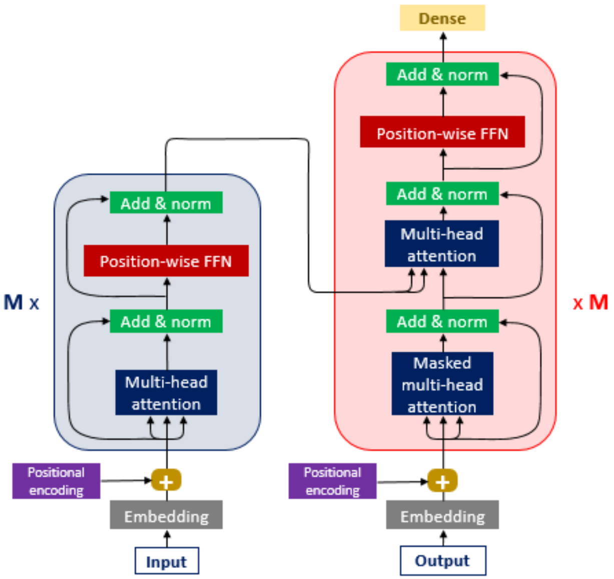
Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, India

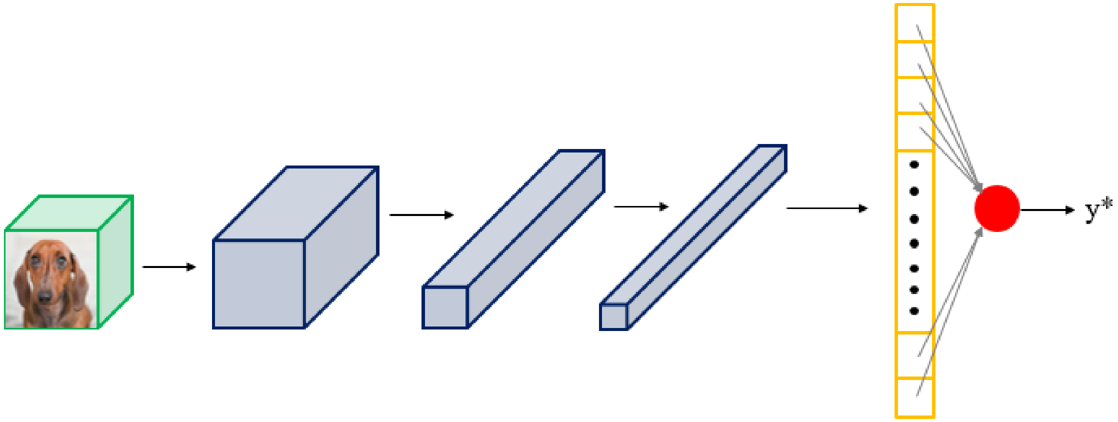
RAMAKRISHNA MISSION VIVEKANANDA UNIVERSITY
SCHOOL OF MATHEMATICAL SCIENCES
PRAJNA BHAVAN
SCHOOL OF INDIAN HERITAGE

Introduction

- CNNs did away with the need for hand-crafted visual features
  - It can learn to perform tasks directly from data
- But the design of CNN architectures are specific to images
- Vision Transformers are motivated by the need for developing task-agnostic yet computationally efficient models
- An input image is represented as a sequence of image patches

Transformer vs CNN

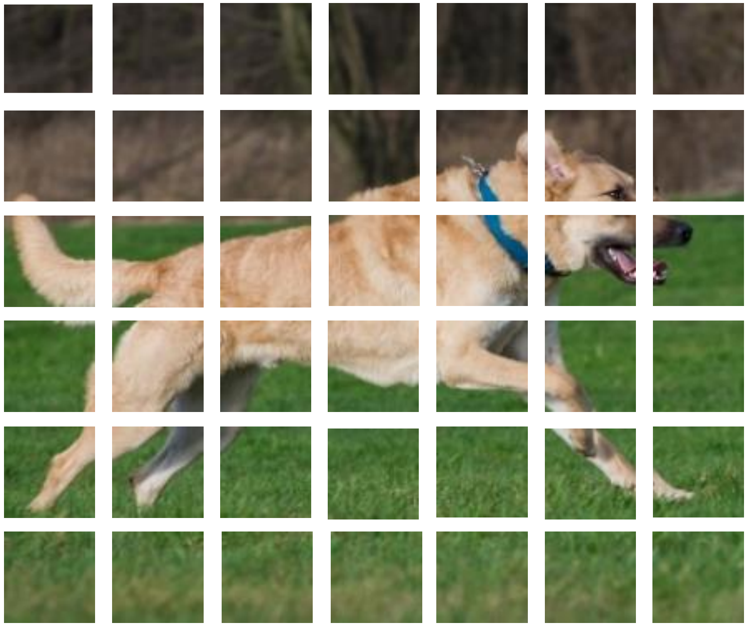




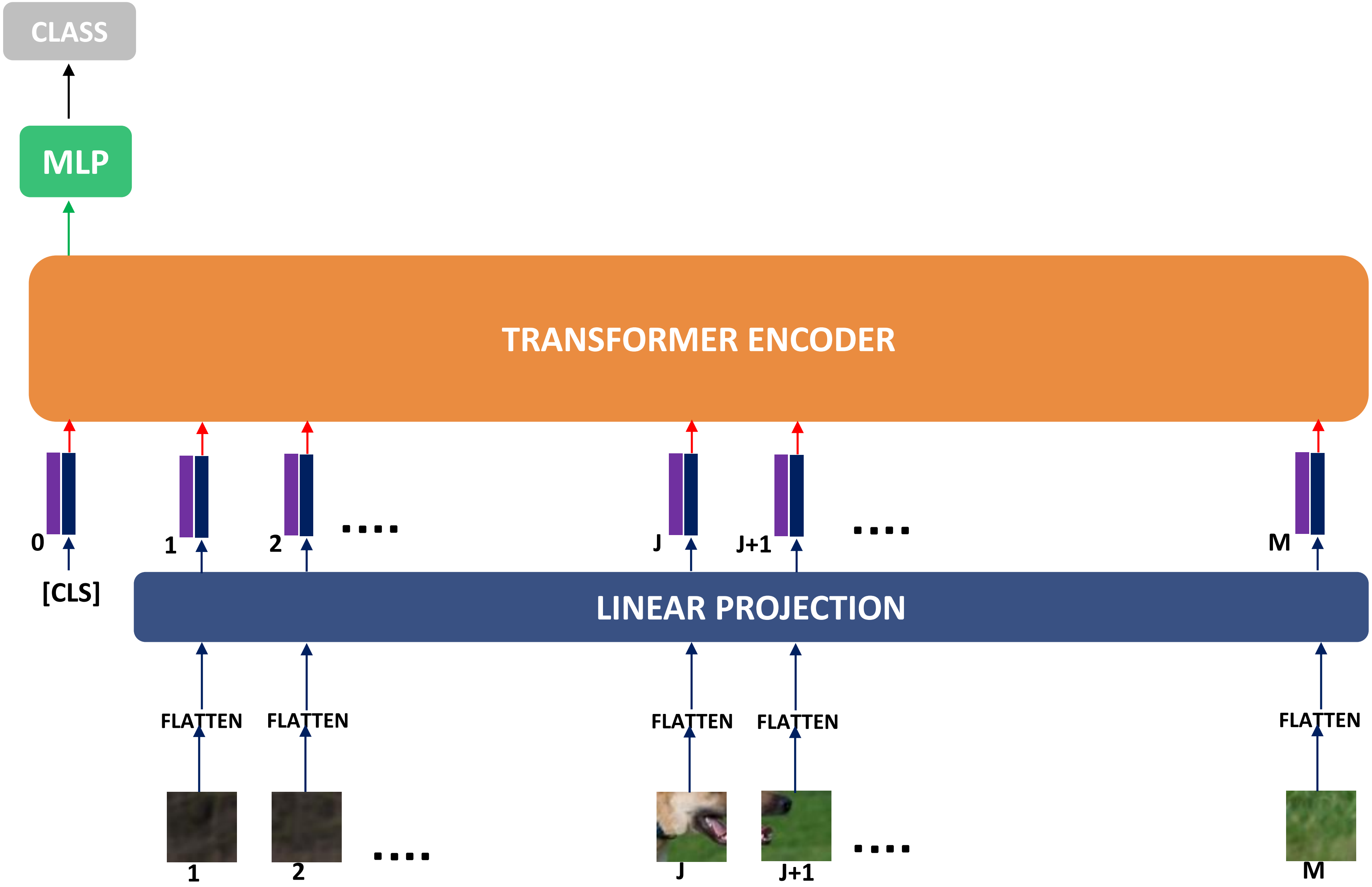
- ResNets beat Transformers of similar size when trained on ’mid-sized’ datasets (e.g. ImageNet)
- Transformer are devoid of some of the inductive biases inherent in CNN models
- However ViTs outperform CNNs when the pretraining dataset is sufficiently large ($\sim$ 100 million)

Introduction

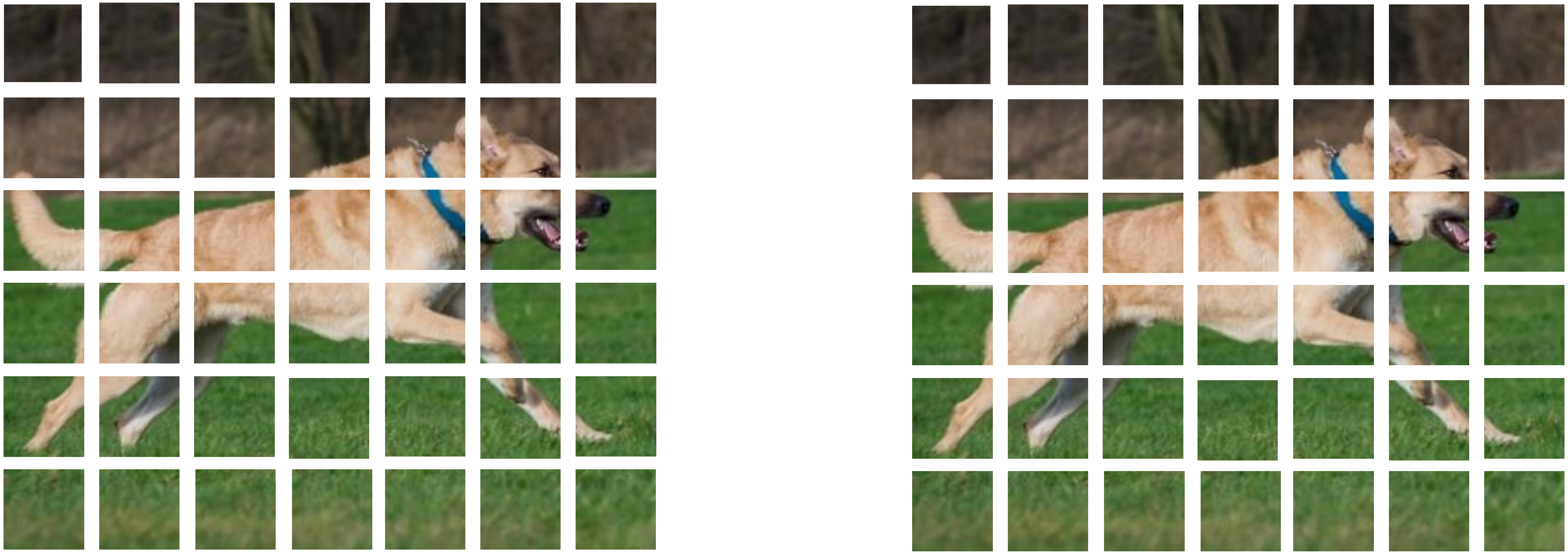
- An input image is represented as a sequence of image patches



Vision Transfomer

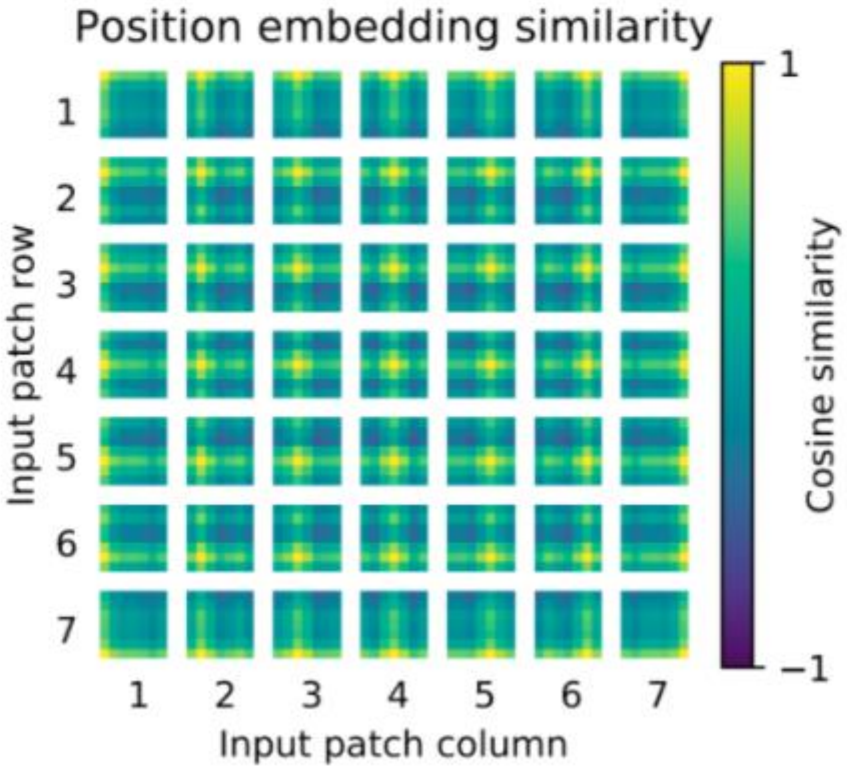


Effect of Positional Encoding



- The position of some of the patches in the original image has been altered
- A Vision Transformer (without PE) does not know about location of patches in the image
  - It is also not aware of the 2D structure of the image

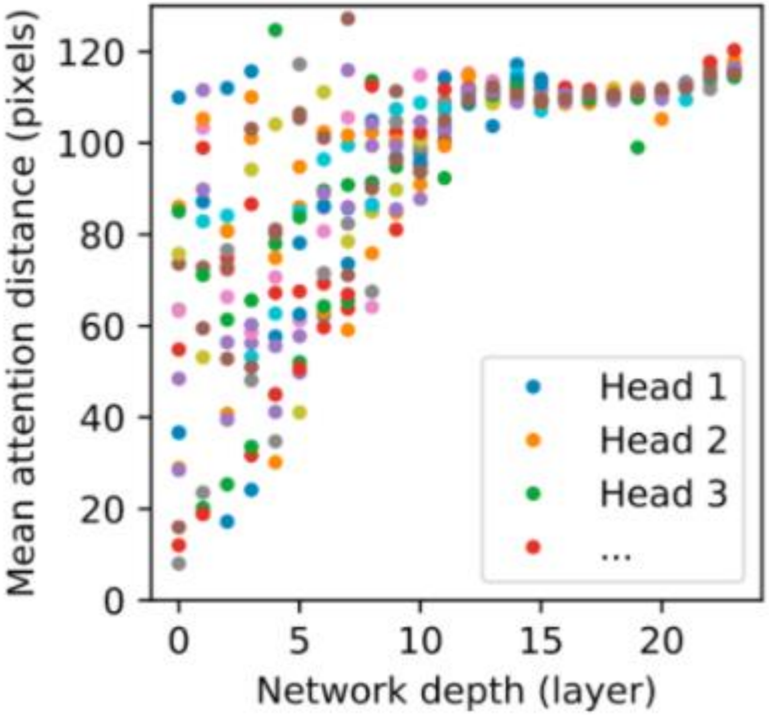
Similarity



- Position embedding: Parameters of the model that encode the relative position of patches
- Position embeddings are most similar to those in the same row and column
  – This demonstrates the models ability to recover the grid structure of image datasets

Fig. & Ref.: https://blog.research.google/2020/12/transformers-for-image-recognition-at.html

Attention distance vs Network depth



- Larger spatial attention distance indicates an element attending to another element located far from it in the same transformer block
  – This implies that the model is capturing global features
- Deeper layers use only global features
- Lower layers capture both local and global features

Fig. & Ref.: https://blog.research.google/2020/12/transformers-for-image-recognition-at.html