An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Koorye

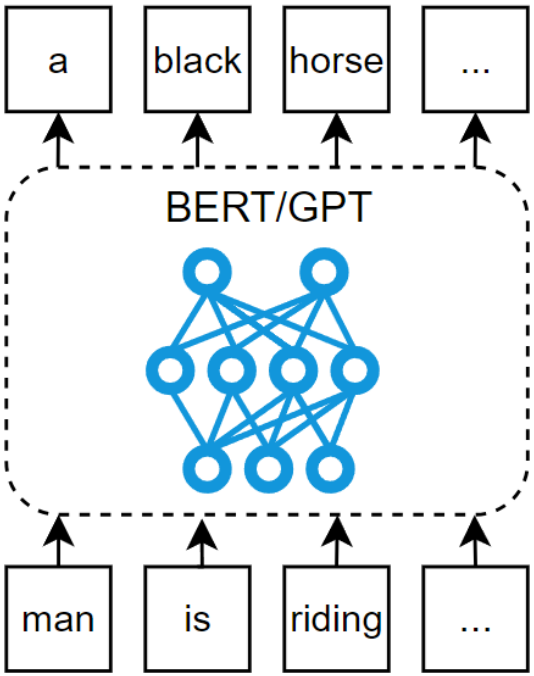
University of Electronic and Science Technology of China

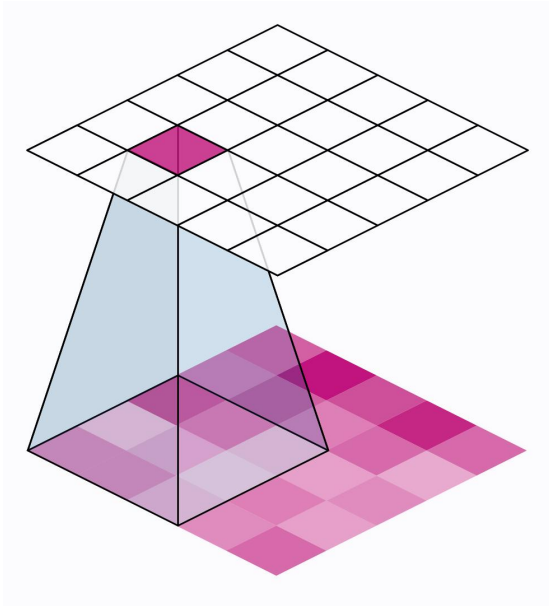
13th May 2024

Outline

Research Aim

The GAP between Language Models and Vision Models:

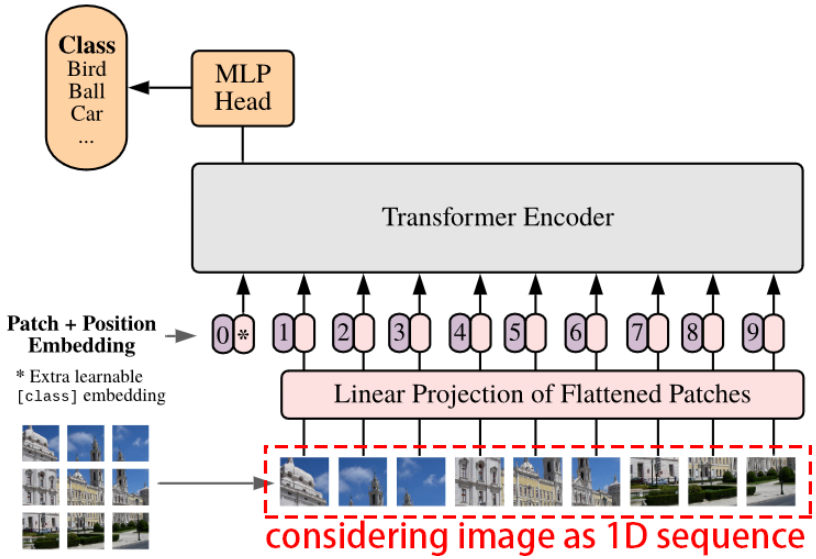




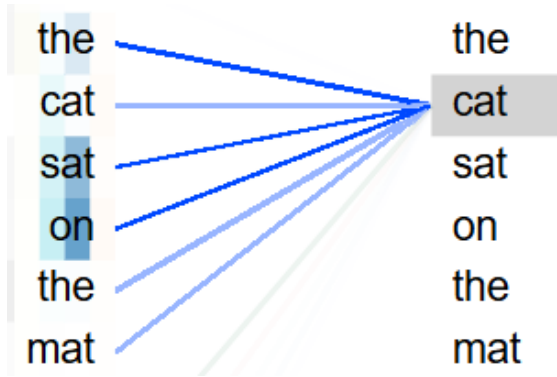
(a) Language Model receives 1D sequence (b) Vision Model receives 2D image pixels

Could the approach used in Language Models inspire a new era for visual tasks?

Method



(c) Vision Transformer Model

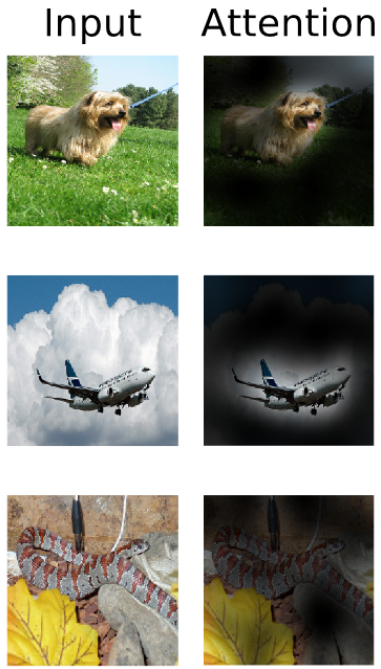


(d) Attention Module

- Considering images as **1D sequences** of patches, like text sentences.
- Model the **relationship** between patches, like the process of human understanding.

Results

(e) Model Accuracy Comparison



(f) Attention Viusalization

- The **accuracy** of model reaches a new level with smaller model **size**.
- Attention modules capture **visually rich areas** of information.

Conclusion & Discussion

Conclusion:

- Inspired by Language Models, processing images into sequences similar to text.
- Surpassing traditional models in accuracy and being more compact.

Discussion:

- Need more training data and larger models to maximize its potential.
- Unknown performances in more visual tasks.