2025-04-21/23

VISION TRANSFORMERS

CMSC 475/675 NEURAL NETWORKS

(TEJAS GOKHALE)

Relevant paper: "An image is worth 16×16 words"
Dosovitskiy et al. ICLR 2021

- so far we've seen transformers that work ~~really well~~ on NLP applications
  - ↳ input & output are "language" (words, sentences, paragraphs, poems, ...)
- before this, we saw many image-based applications (CNN, ResNet etc. for image classification, object detection etc.)

* CAN WE USE TRANSFORMERS for IMAGES?

→ what are some technical issues with this?

① input format? how do we get vectorization?

(actually this is easier for images as images are already vectors

... for language we had to first do vectorization through Word2Vec etc.)

② considering each pixel as a token

(possible, but a LOT of tokens)

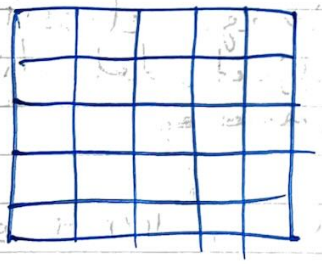
↳ even a tiny image 100 x 100 → 10^4 tokens

↳ a "standard" HD image: 1280 x 720 = 921,600

↳ a "full" HD imge: 1920 x 1080 ≈ 2 Million tokens

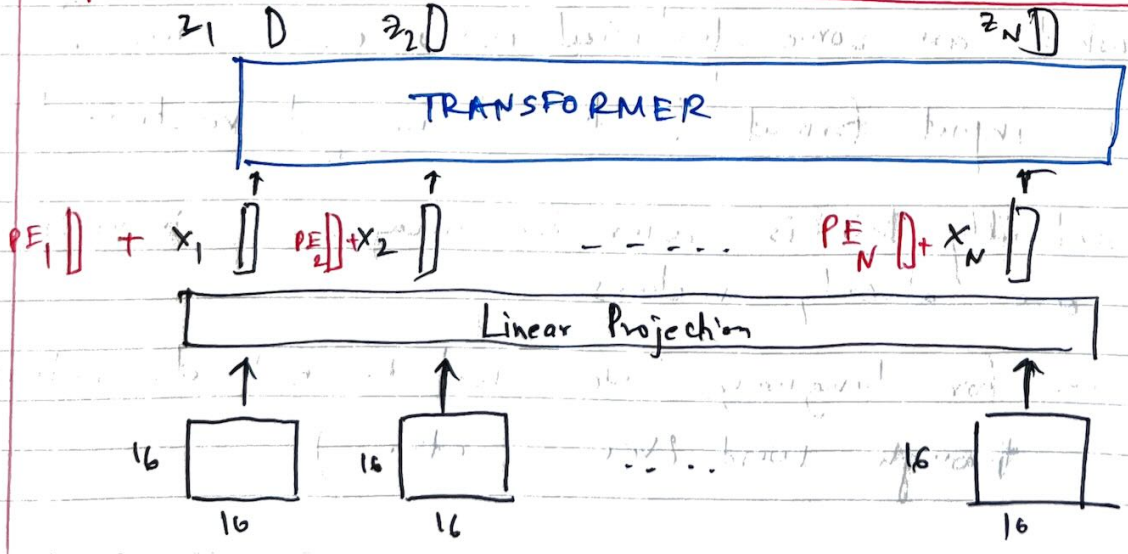
HOW DO WE DEAL WITH THIS?

USE PATCHES!



→ convert each patch into a token

(in the paper: 16x16 patches)



* How to do image classification with this ?

(outputs of the Transformer are also tokens)

One idea: add a classifier on top.

(Better) trick

add a dummy token ([CLS] token)

↳ the embedding for [CLS] is learnable