Project Report

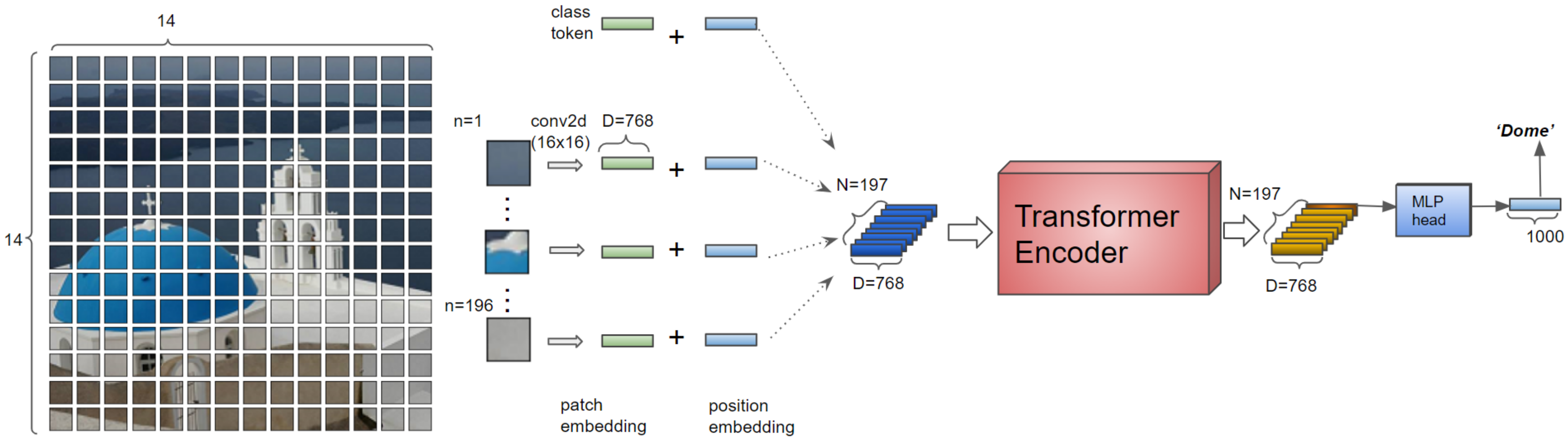
Machine Learning for Computer Vision

Hanying Zhang

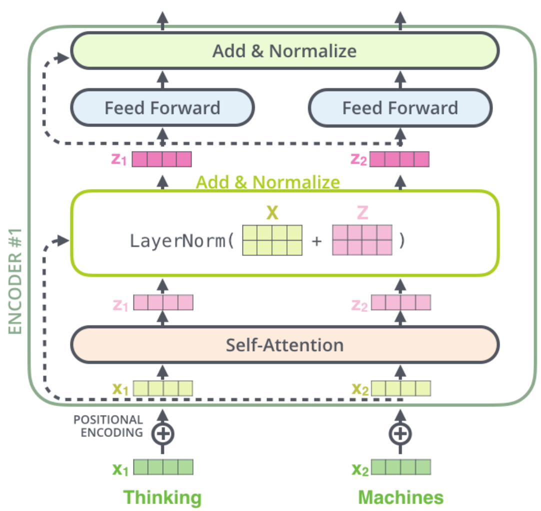
Executive Summary

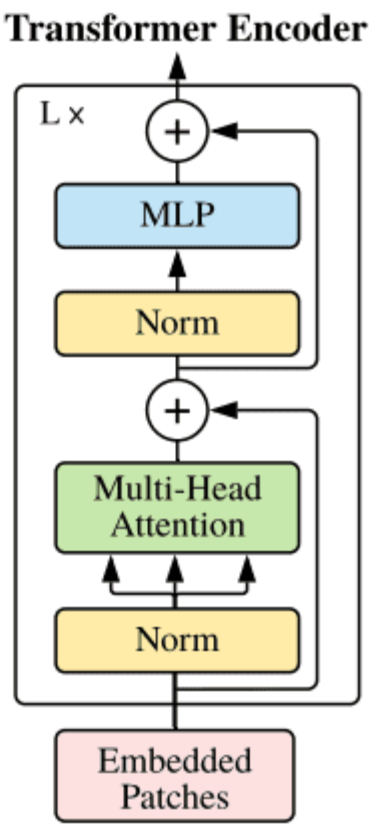
- In this project the Vision Transformer(ViT) paper is studied and implemented.
- A classification task on a new dataset is carried out to test the performance of ViT as backbone.
- The positional embedding and attention matrix are also studied.

ViT – Vision Transormer



Transformer in ViT





Classification – Dataset

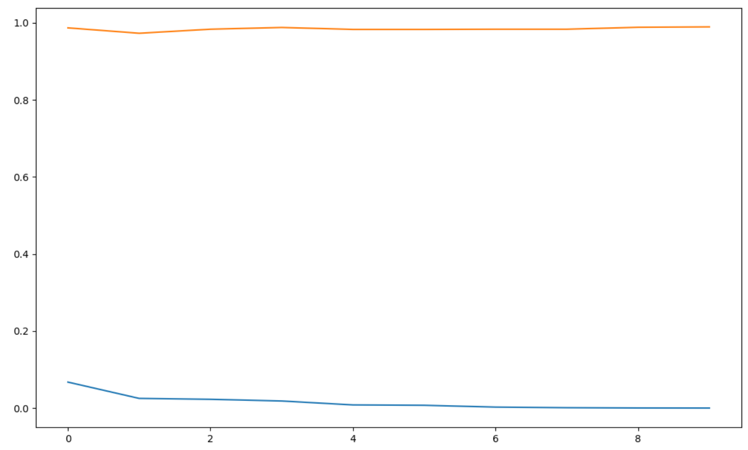
- A dogs vs. cats competition that still accepts test submissions

- The dataset is slightly different from the competition on Kaggle

- Train set: 20k images
- Validation set: 2k images
- Test set: 2k images

Classification – Fine Tune

- Pre-trained model: jx_vit_base_p16_224-80ecf9dd.pth
- Replace the head with one FC layer, the weight is [2,768] where 768 is the dim of the output of Transformer
- First experiment: All weights in the model except for the ones in the head is fixed
- Fine Tune: All weights in the model are trainable
- Settings
  - 5 epochs
  - Batch size: 16
  - Optimizer: Adam
  - Learning rate: 0.0001
  - CosineAnnealingLR decay



Classification – Results

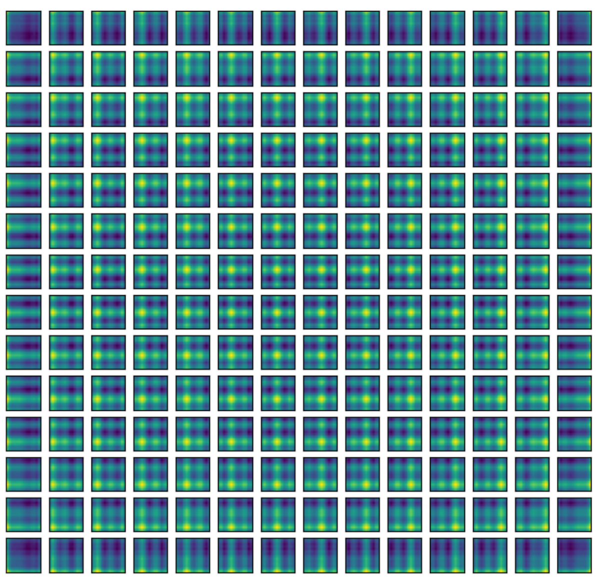
- Test Accuracy:
  - 96.5% before Fine Tune
  - 98.65% after Fine Tune

- Obvious that the accuracy becomes better after fine-tune

- Training with more epochs and bigger learning rates are also explored, but these parameters failed to give better results

16	小白U1635036757	98.75	7	2021-10-24 14:58:16	2021-10-24 14:58:16	暂无
17	小白U1632542321	98.65	3	2021-09-26 20:06:04	2021-09-26 20:06:04	暂无
18	剑飞	98.65	11	2022-06-10 17:41:57	2022-06-10 18:39:39	上传方案
19	独苍	98.6	11	2020-11-03 18:36:12	2020-11-03 18:36:12	暂无
20	花前月下意	98.55	3	2020-07-24 12:39:33	2020-07-24 12:39:33	暂无

Visualization – Positional Embedding



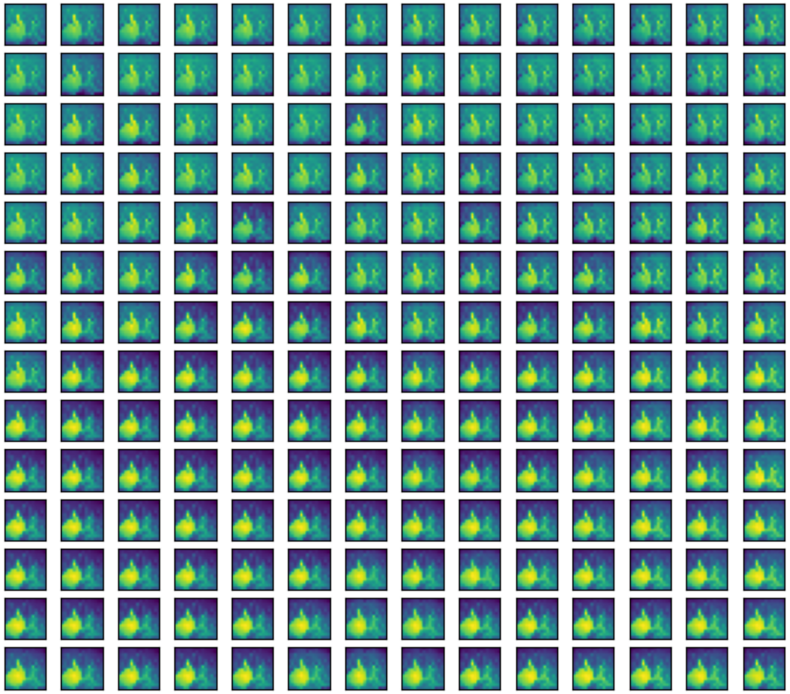
- Several variations of position embedding methods were tried in the original paper
  - No positional embedding
  - 1D positional embedding
  - 2D positional embedding
  - Relative positional embedding
- The latter 3 ones give similar performance
- Visualize the 1D positional embedding for explanation - computing the cosine similarity between the i-th embedding and all the embeddings(except for the class_token one)
- Each patch of the image has the embedding correctly represents its position in the image. Thus we can say that 1D positional embedding is enough for this model

Visualization – Attention Matrix

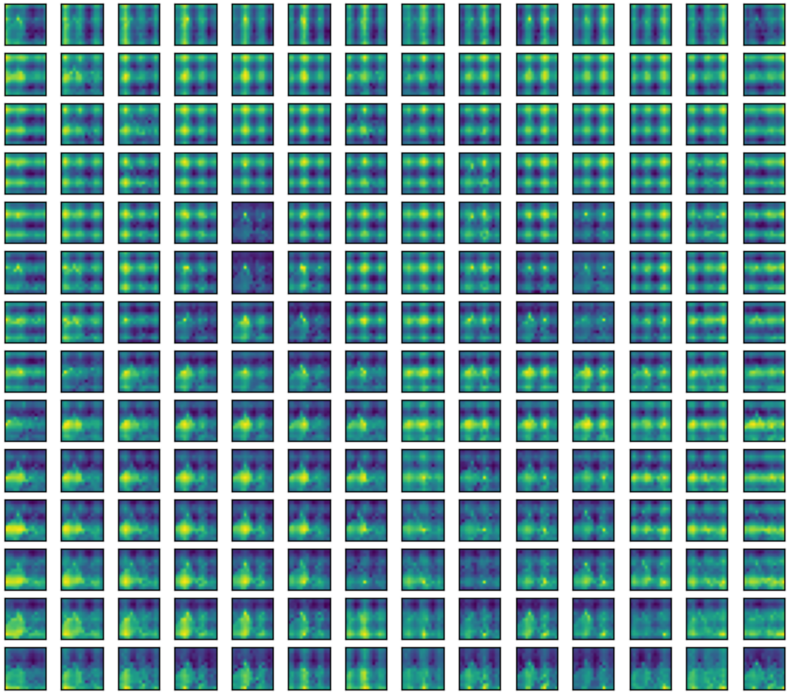
- The attention matrix, matmul(Q, K^T) / (sqrt(D_K))
- To see how the attention mechanism works
- One image from the test set which contains two objects is selected to show the attentions

Visualization – Attention Matrix

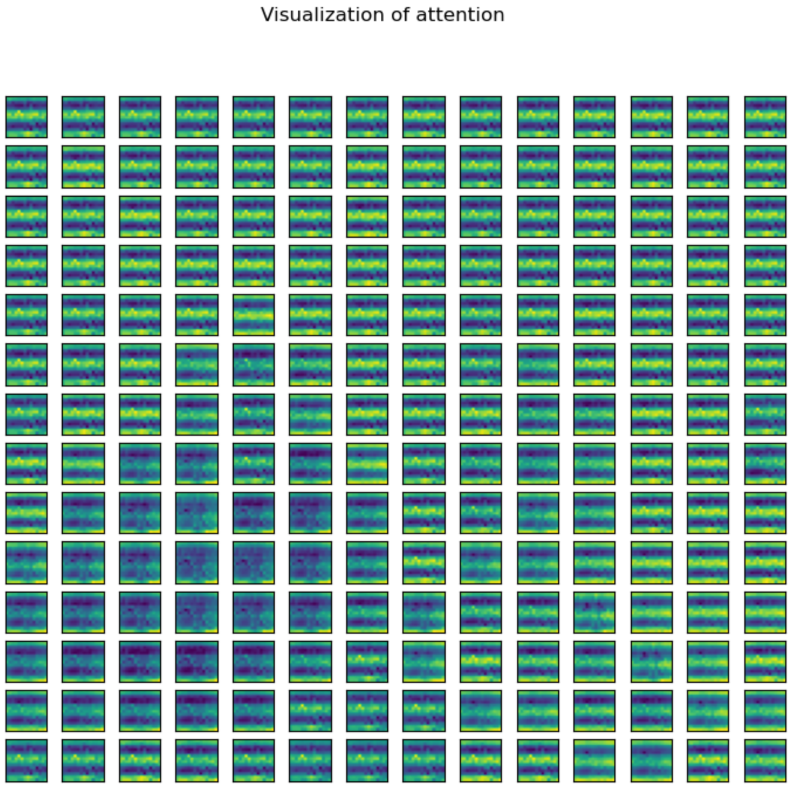
Visualization of attention



Visualization of attention



Visualization – Attention Matrix



Thank you for your ATTENTION.