AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

EGEMEN TÜRKGENCİ

ÖMER FARUK ÖZGÜL

AIM OF THE ARTICLE

- Offer a less expensive solution for Image Recognition
- Utilize Transformers used in NLP
- Compare popular CNN solutions to ViT

INTRODUCTION

- Self-Attention based structures such as Transformers work well in unprecented data
- It relies on a pre-train of large data and fine tune
- Without a preprocessing phase, unusable due to the high computation

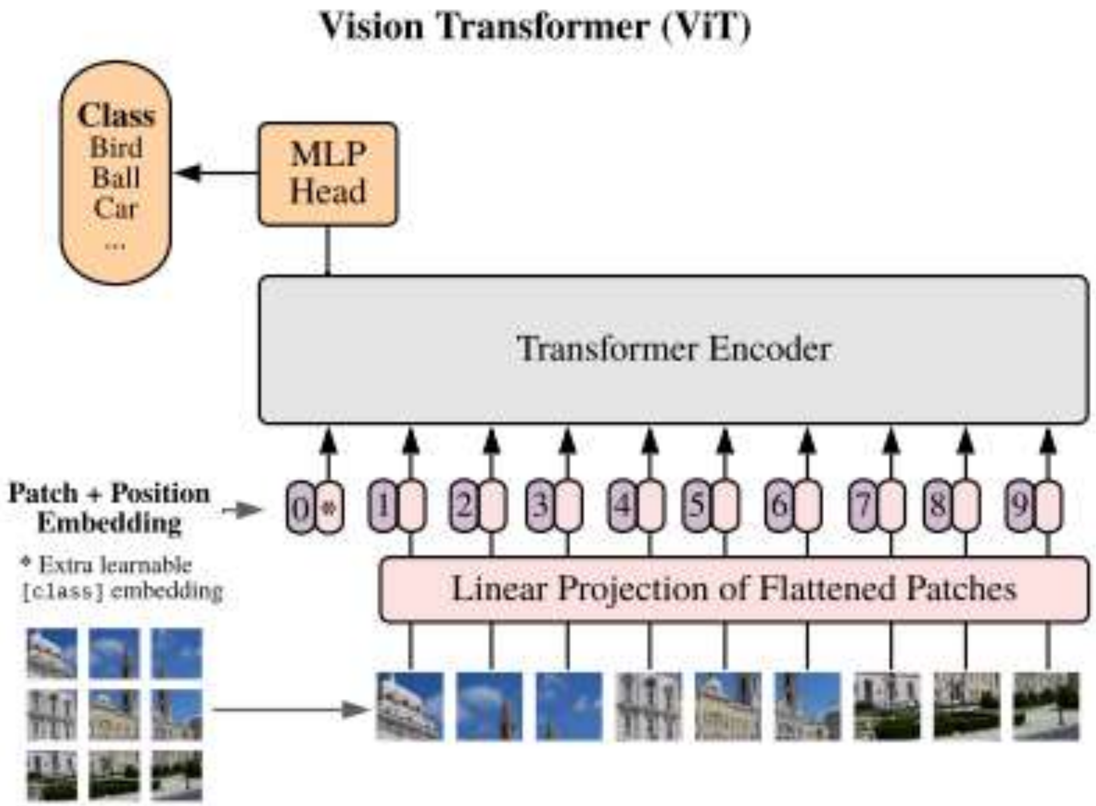
INTRODUCTION

- The Image is split into patches and processed before fed to the transformer
- Compared to ResNet, when fed huge amount of data, produces better results.

RELATED WORK

- There are previous attempts in utilizing Transformers for image recognition
- To illustrate: local calculations instead of global calculations
- 2x2 patches are used instead of individual pixels but can only perform on small images.
- Reducing image resolution and colour space(iGPT)

METHOD



Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP
Head
Transformer Encoder
Patch + Position
Embedding
* Extra learnable
[class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

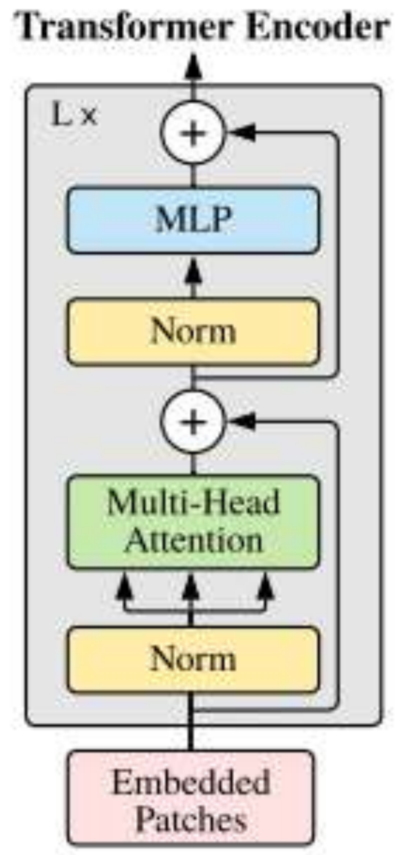
METHOD

- Transformer receives an input of 1D but our patches are 2D
- 2D Patches are first flattened into vectors.(16x16 → 256x1)
- Transformers use a latent vector size of D
- The flattened patches are mapped onto the vector of dimension D
- Utilize a learnable linear projection to map

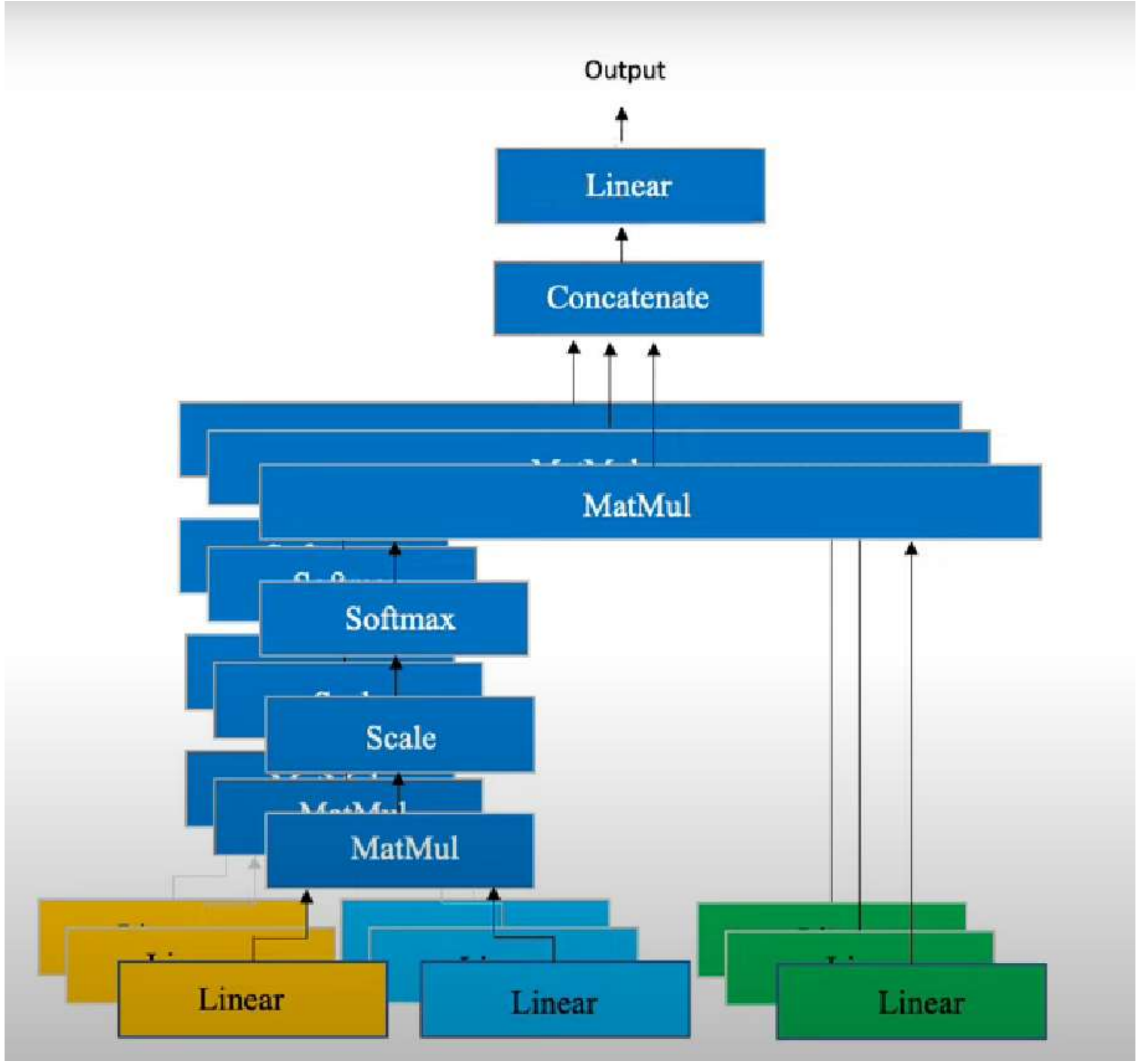
METHOD

- Later a positional embedding is added on top of patch embeddings
- A class token is hardcoded (Extra learnable class embedding)
- The result is then fed to the transformer

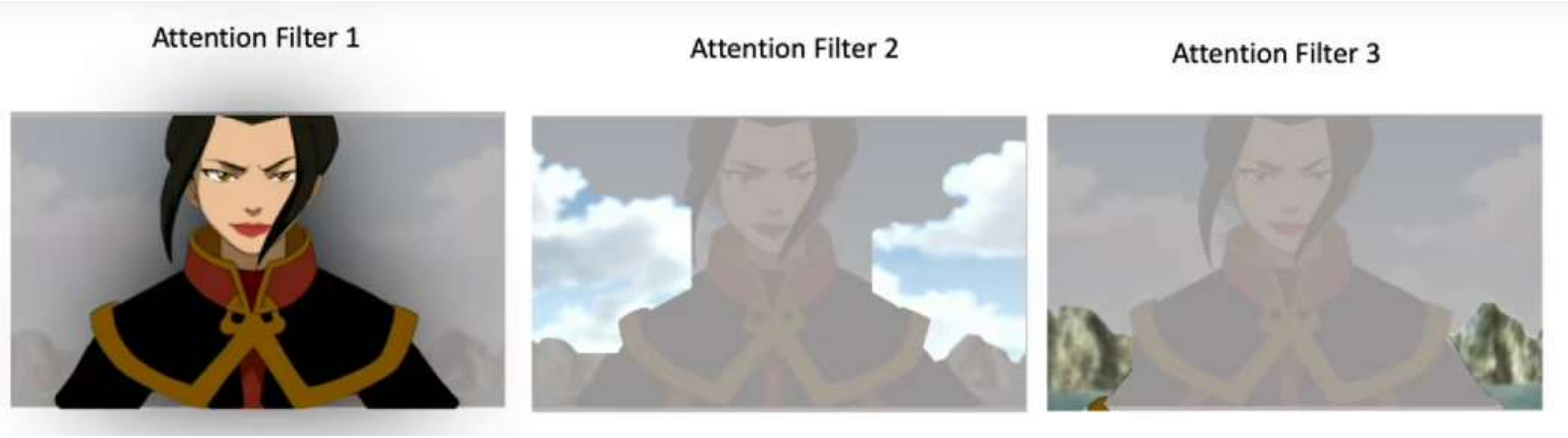
TRANSFORMER STRUCTURE



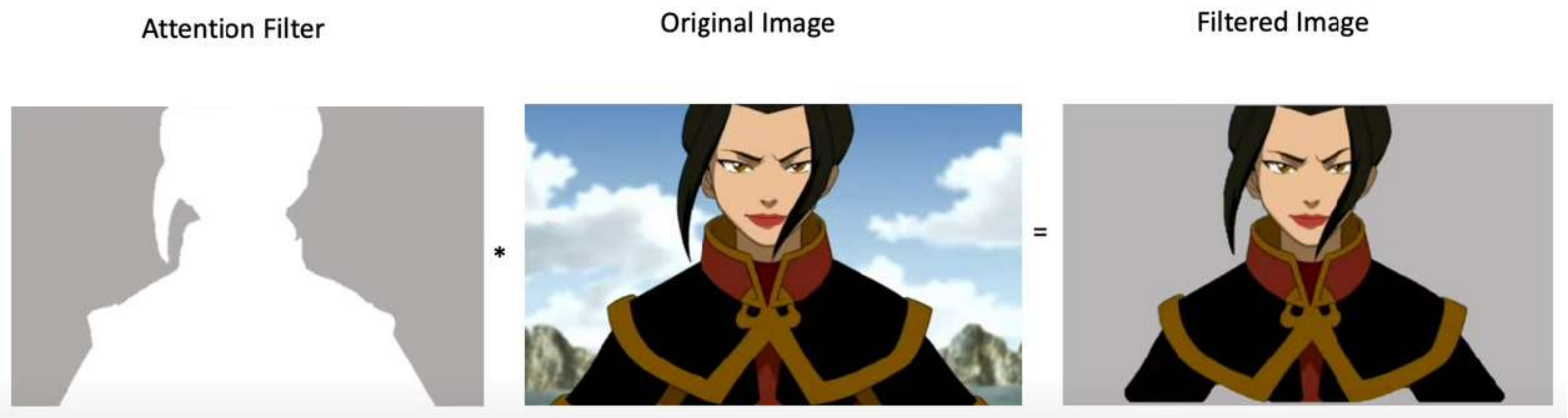
MULTI-HEAD ATTENTION



MULTI-HEAD ATTENTION



ATTENTION FILTER



FINE TUNING AND HIGHER RESOLUTION

- When fine tuning, we remove the pre-trained prediction head(MLP)
- We instead place a zero-initialized D x K dimensioned feed forward layer, where K is the number of downstream classes.

FINE TUNING AND HIGHER RESOLUTION

- In order to obtain better results, one should use High resolution data for fine tuning
- But this creates problems
- Since patch size remain constant, higher resolution images will produce more patches
- This will effect the positional embedding
- This is solved by using 2D interpolation

MODEL VARIANTS

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

COMPARISON TABLE

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

CONCLUSION

- There are still challenges
  - Applying ViT to other computer vision tasks, such as detection and segmentation
  - Continue to exploring self-supervised pre-training methods