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

Prabal Ghosh
Côte d'Azur University

AIQDSC34 - ECUE More on Deep Learning Algorithms
28/03/2024

https://arxiv.org/abs/2010.11929

Outline

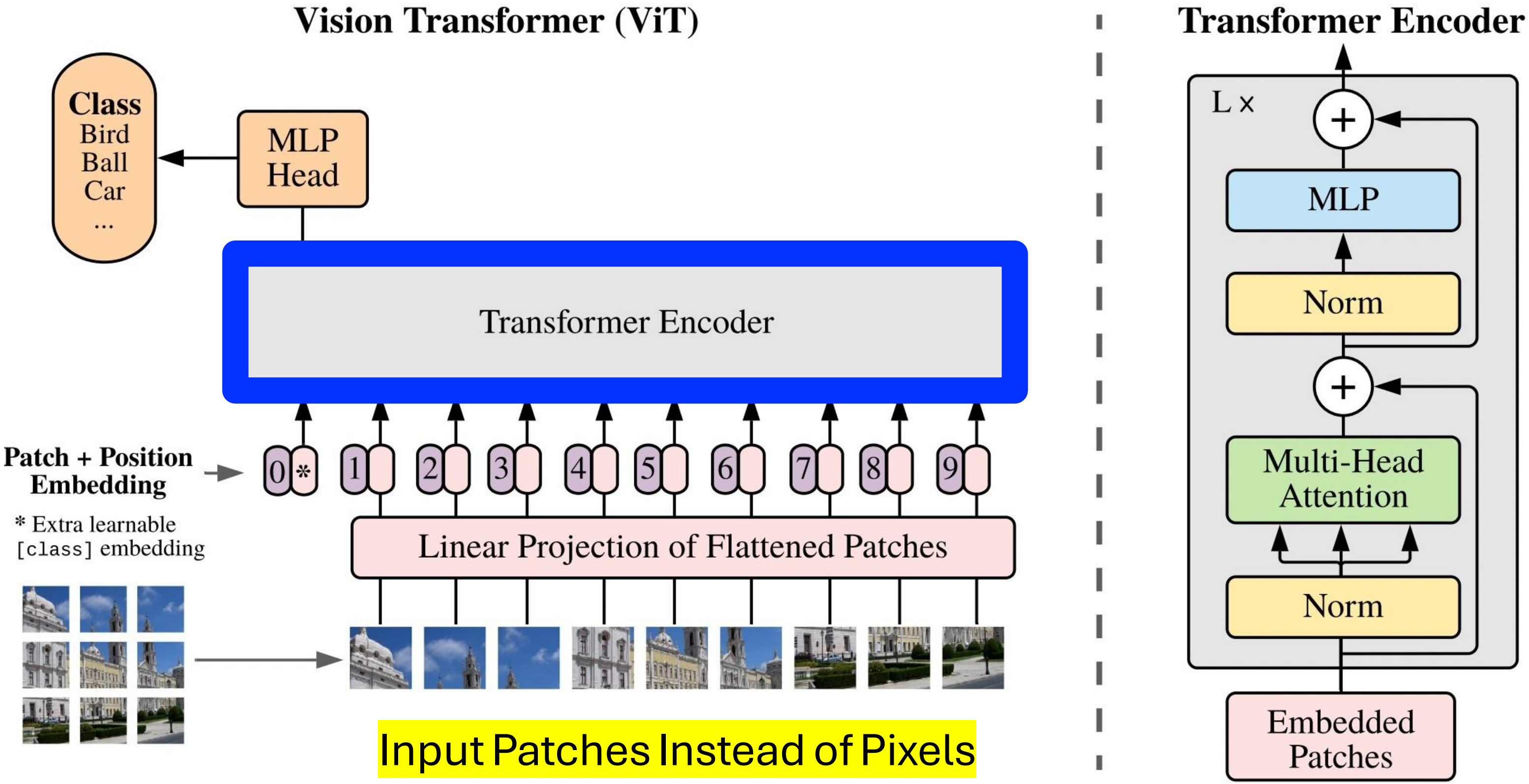
- Motivation for Vision Transformer
- Vision Transformer architecture
- Training Vision Transformer
- Result Discussion
- Conclusion

Motivation and idea :

- The encoder-decoder architecture of Transformer achieved state of the art performance on machine translation tasks , by allowing significantly more parallelization (i.e less time to train).
- **RestNet** (CNN) was the **best** solution **for image classification**.

- Can we apply Transformers to images and get state-of-the-art results ?
- Use the **Transformer Encoder architecture** with fewest possible modifications and apply on **image classification tasks**.

Vision Transformers Architecture: Uses Popular BERT Architecture



Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

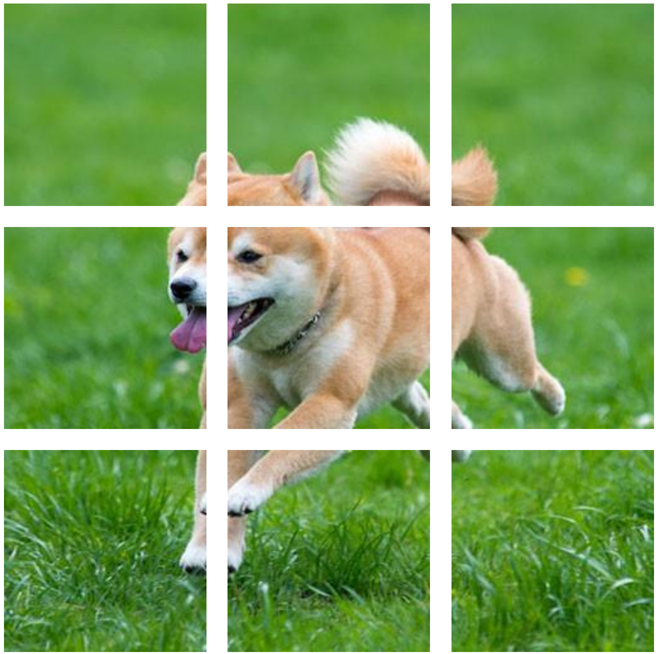
Standard Transformer on Patches

Dog image is free for use

Dosovitskiy et al, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR 2021

Image Ref. http://wangshusen.github.io/

Standard Transformer on Patches



Split Image into Patches

- Here, the patches do not overlap.
- User specifies:
  - patch size, e.g., 16×16;

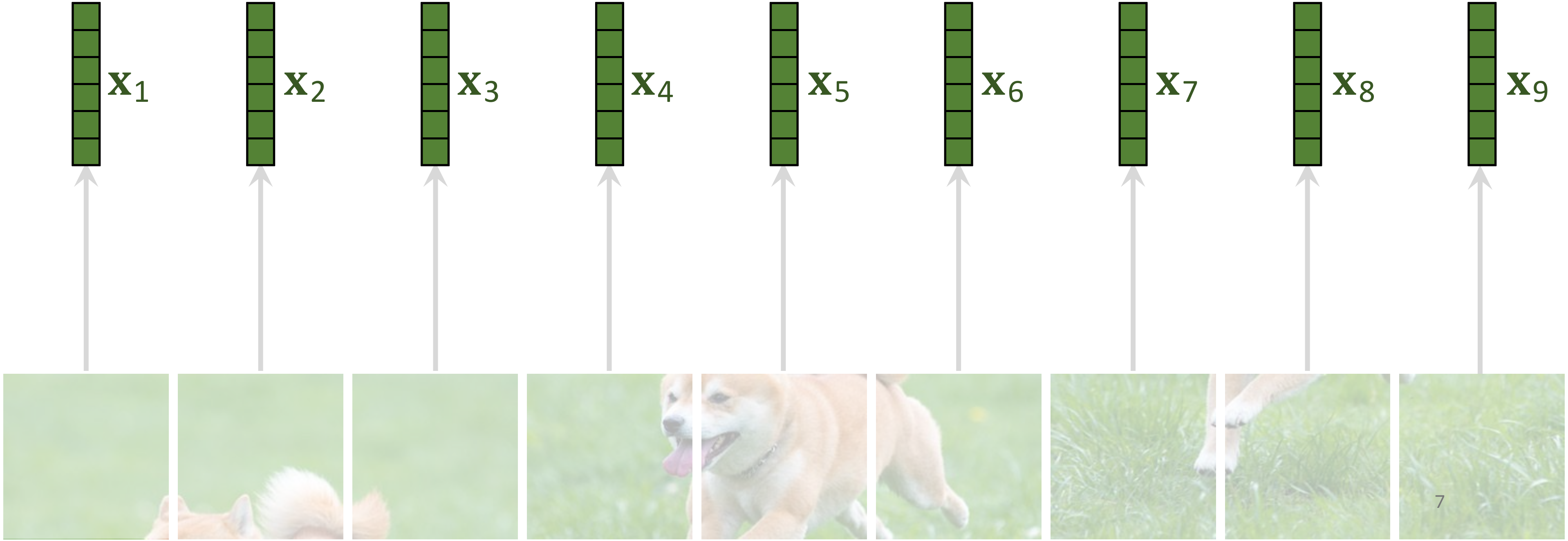
9 input patches, each of shape 3x16x16

Dosovitskiy et al, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, ICLR 2021

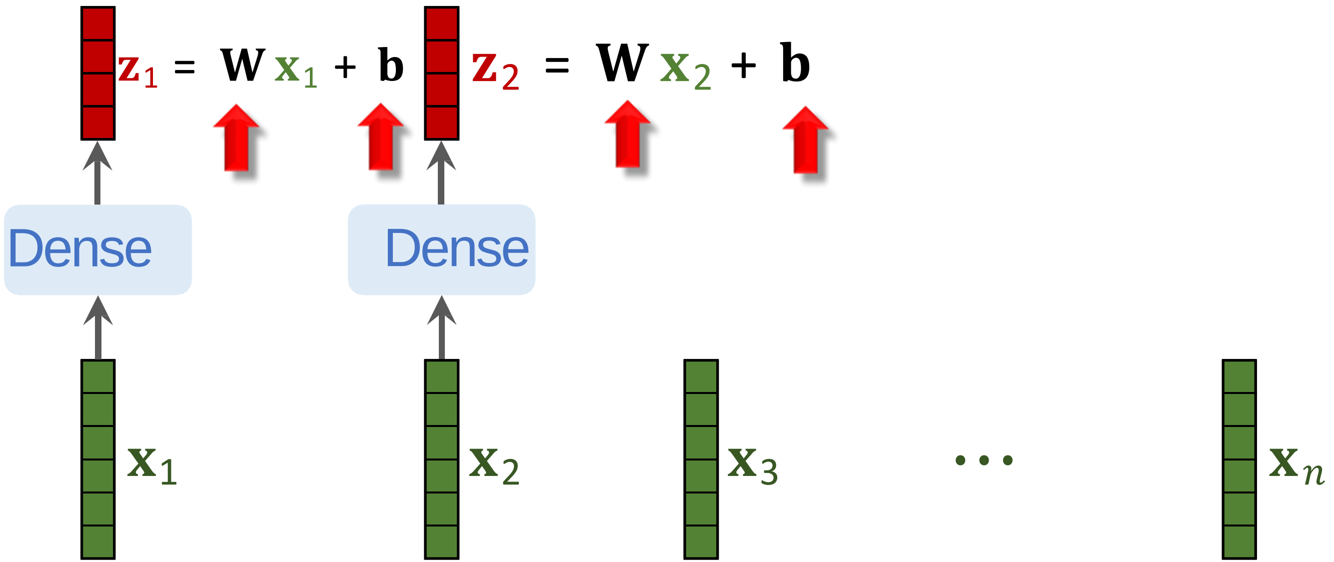
Image Ref. http://wangshusen.github.io/

Vectorization

If the patches are $H \times W \times C$ tensors, then the vectors are $HWC \times 1$.



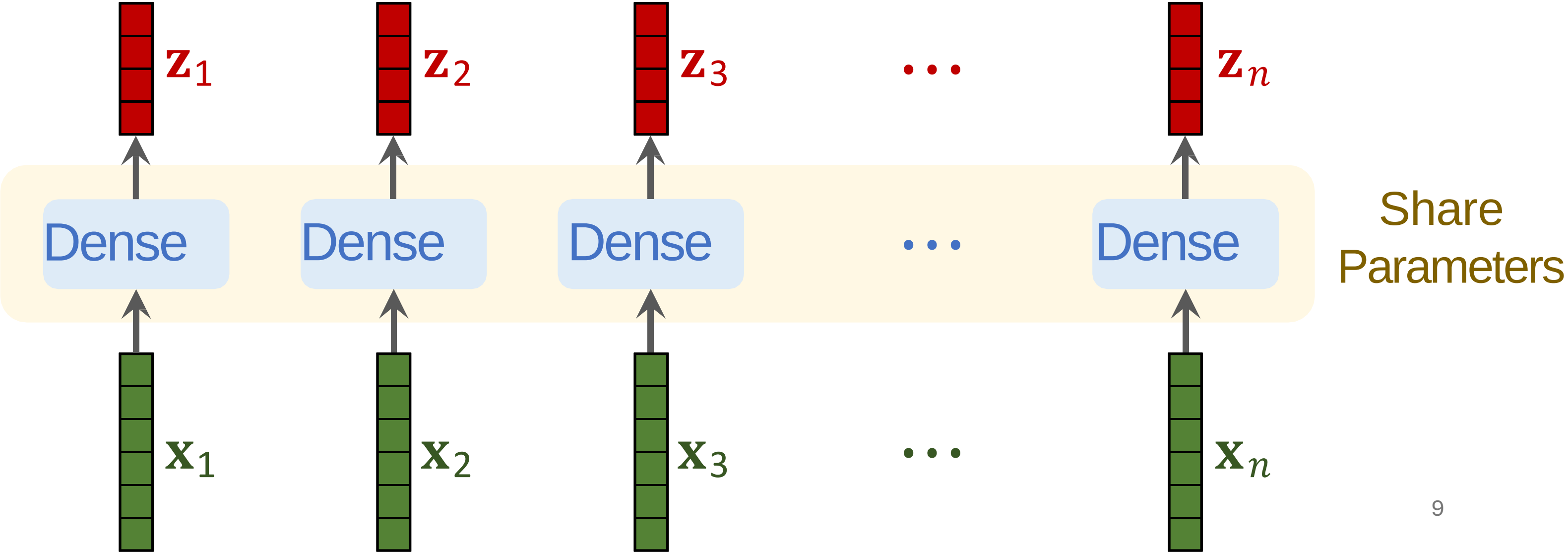
Vision Transformer Architecture Explained:



Vision Transformer Architecture Explained:

Linear projection to D-dimensional vector

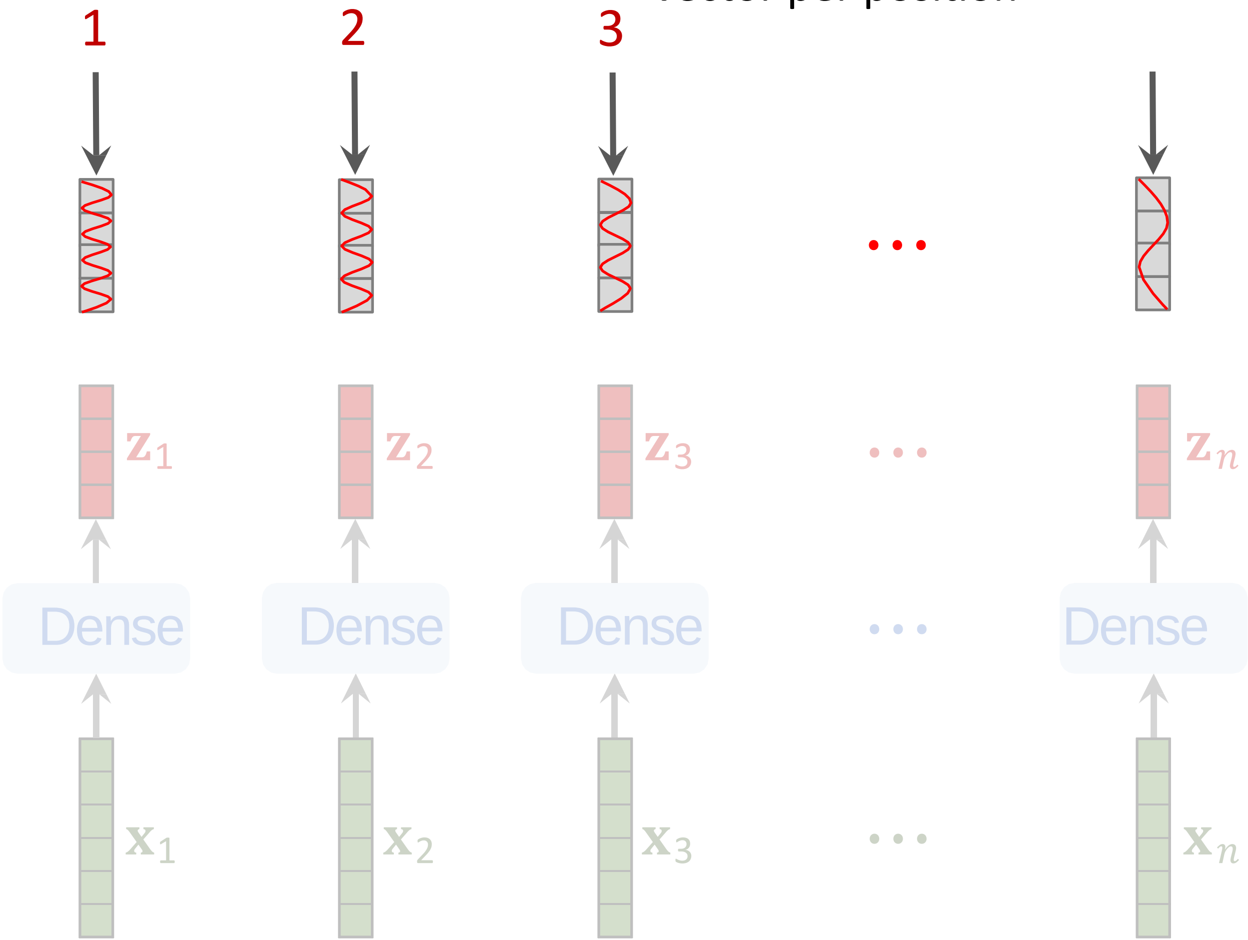
N input patches, each of shape Cx16x16



Vision Transformer Architecture Explained:

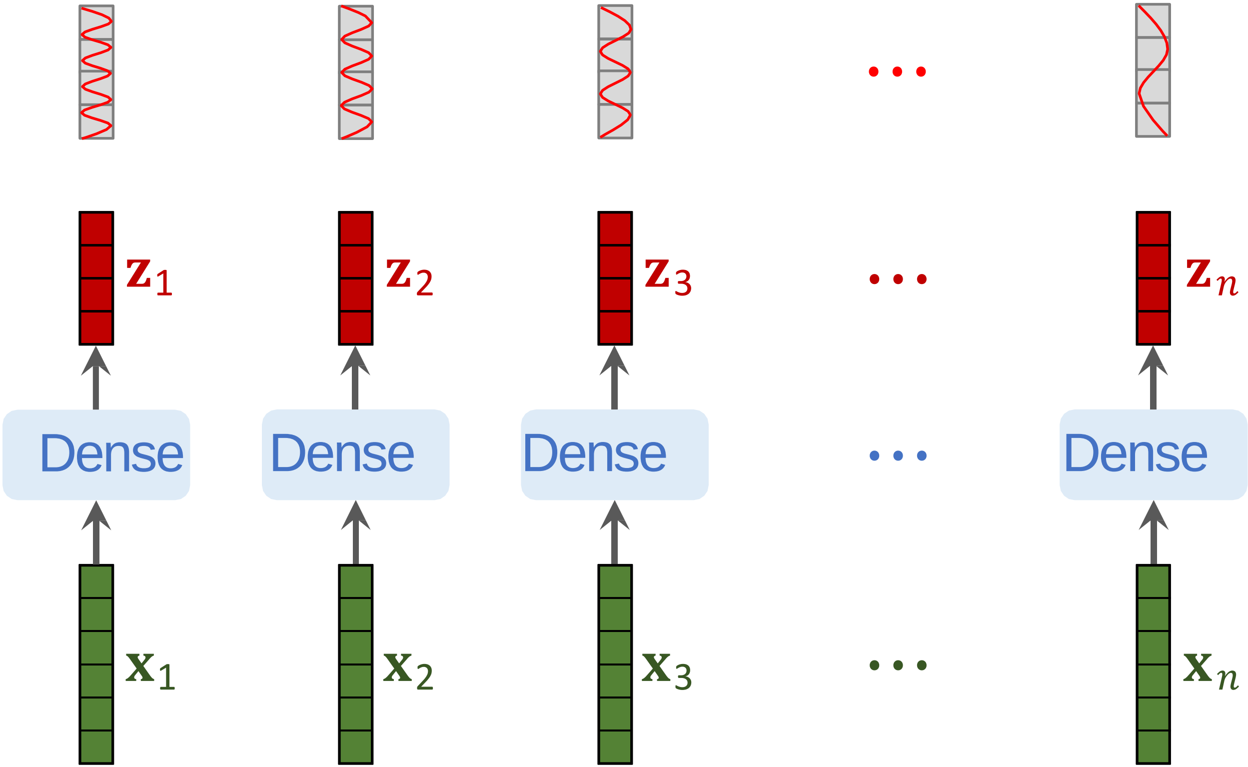
Add positional embedding: learned D- dim vector per position

Positional Embedding:



Vision Transformer Architecture Explained:

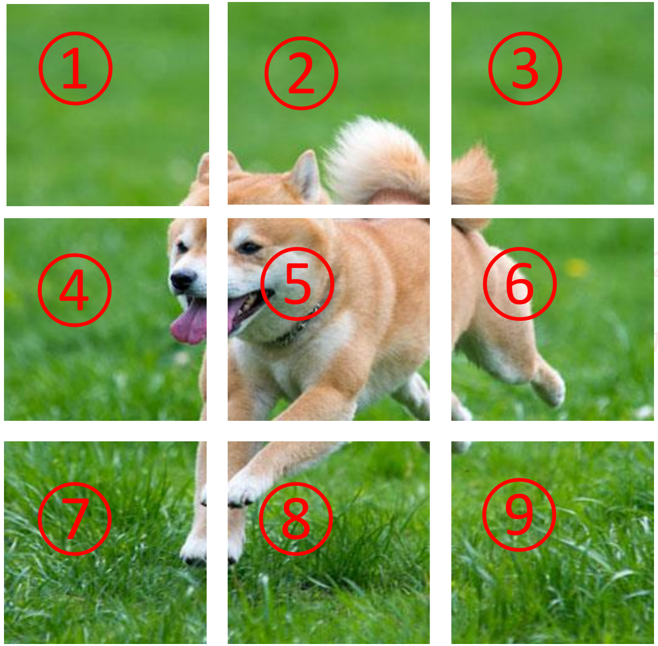
Add positional embedding vectors to $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n$.



Vision Transformer Architecture Explained:

Add positional encoding vectors to $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n$. (Why?)

- 3% drop in accuracy is observed, if we do not apply positional.
- Positional embedding can be 1D or 2D, but no significant improvement is found in using 2D embedding, hence the paper uses 1D position embedding.



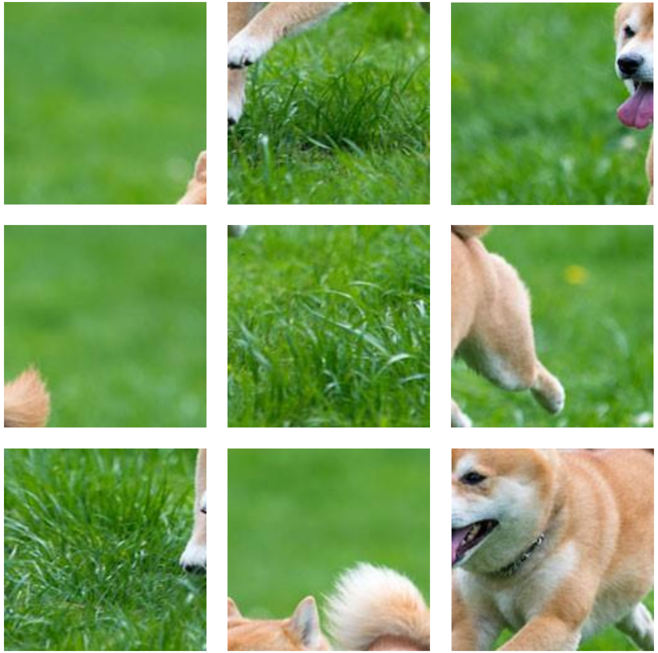
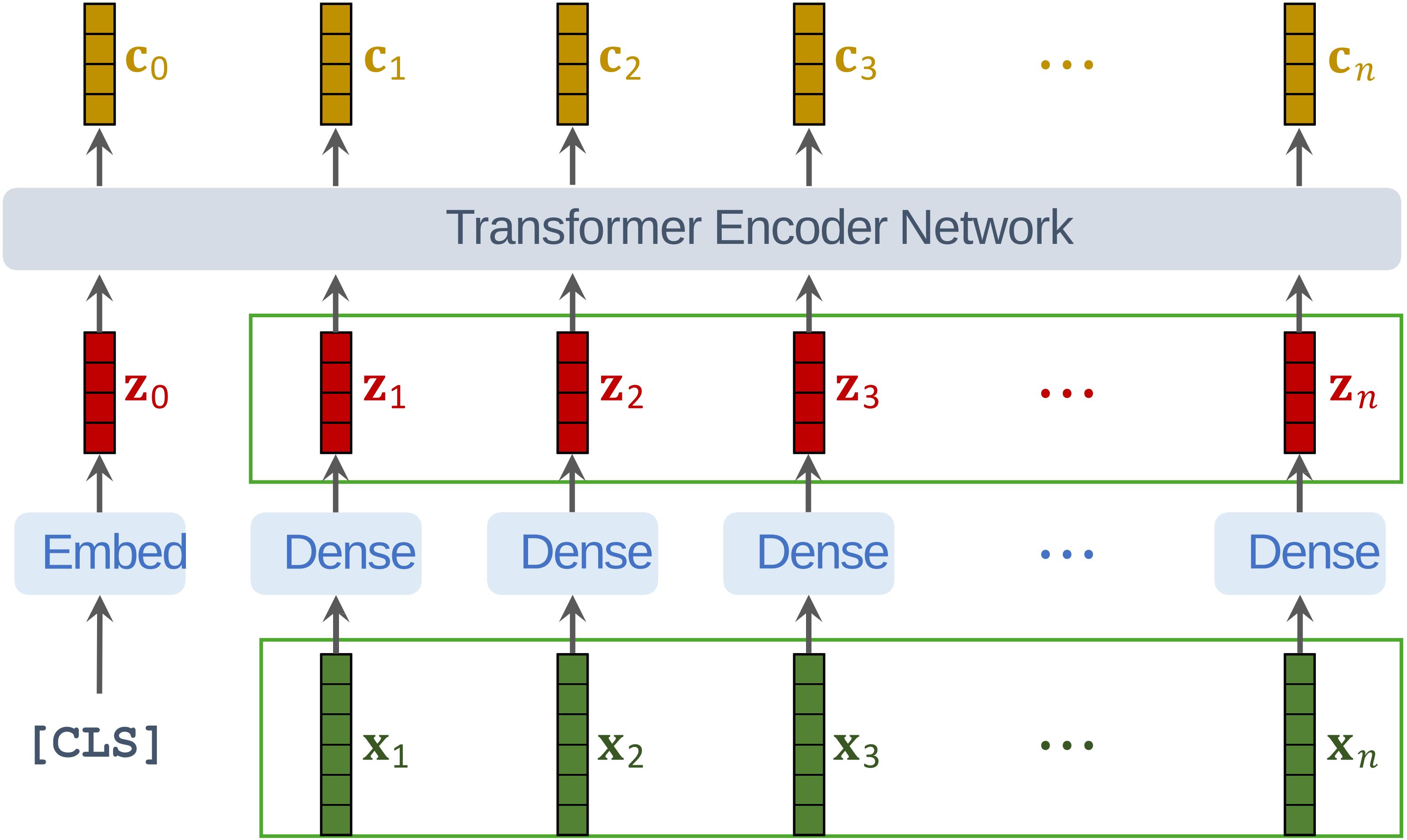
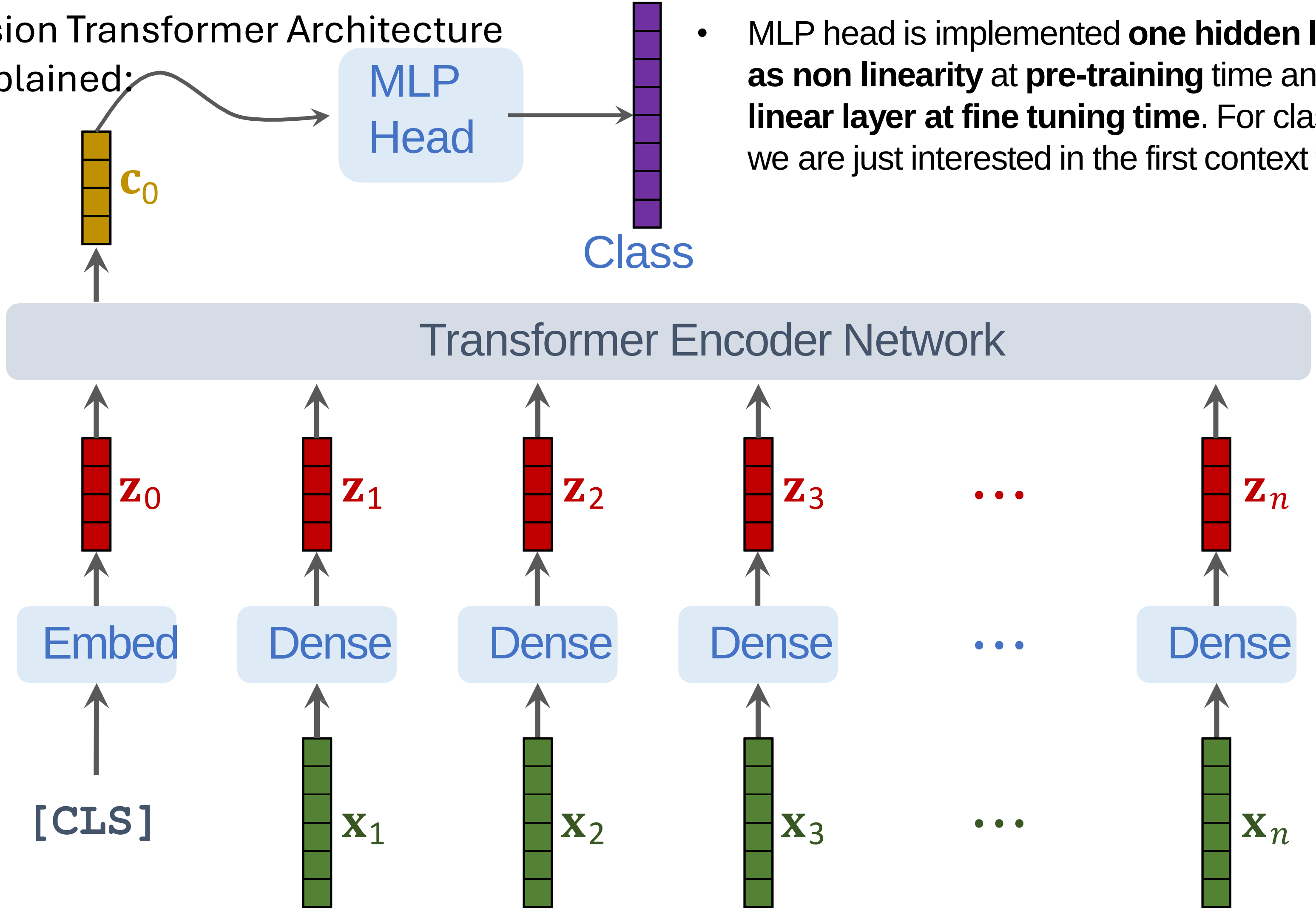


Image Ref. http://wangshusen.github.io/

Vision Transformer Architecture Explained:



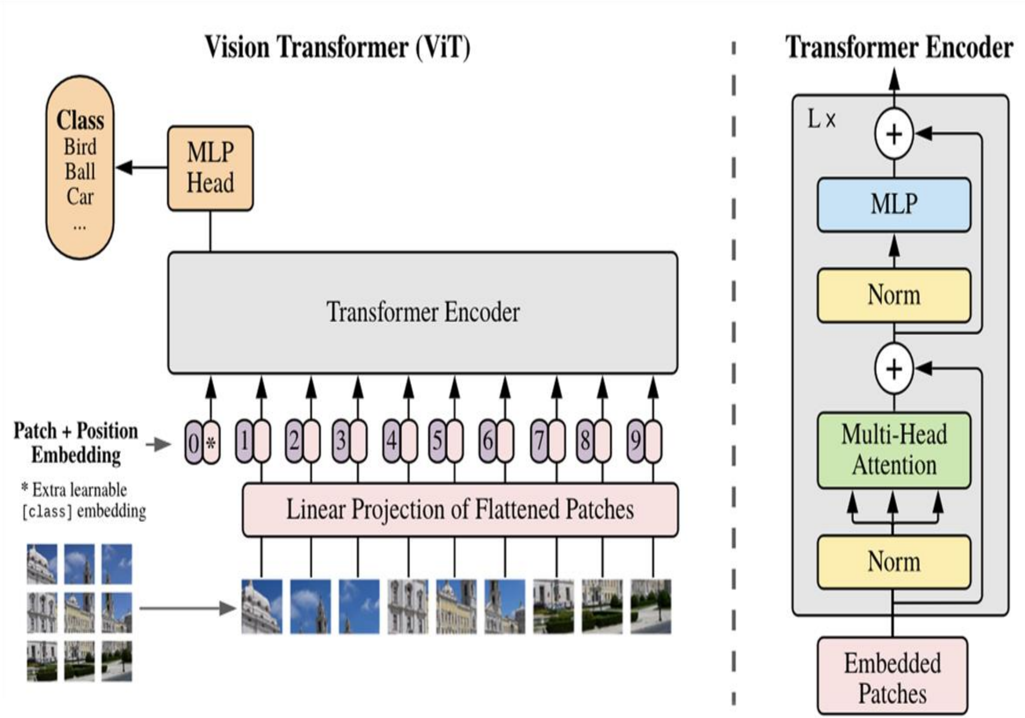
Vision Transformer Architecture Explained:



- MLP head is implemented **one hidden layer and tanh as non linearity** at **pre-training** time and by **a single linear layer at fine tuning time**. For classification task we are just interested in the first context token(C0).

Vision Transformer Architecture

Image Patches $\equiv$ Tokens (Words) in NLP
$x \in \mathbb{R}^{H \times W \times C} \rightarrow$ image
$x_p \in \mathbb{R}^{N \times (P^2 C)} \rightarrow$ sequence of flattened 2D patches (reshape x)
$P \times P \rightarrow$ resolution of each image patch
$N = \dfrac{HW}{P^2} \rightarrow$ resulting number of patches
$D \rightarrow$ latent vector size

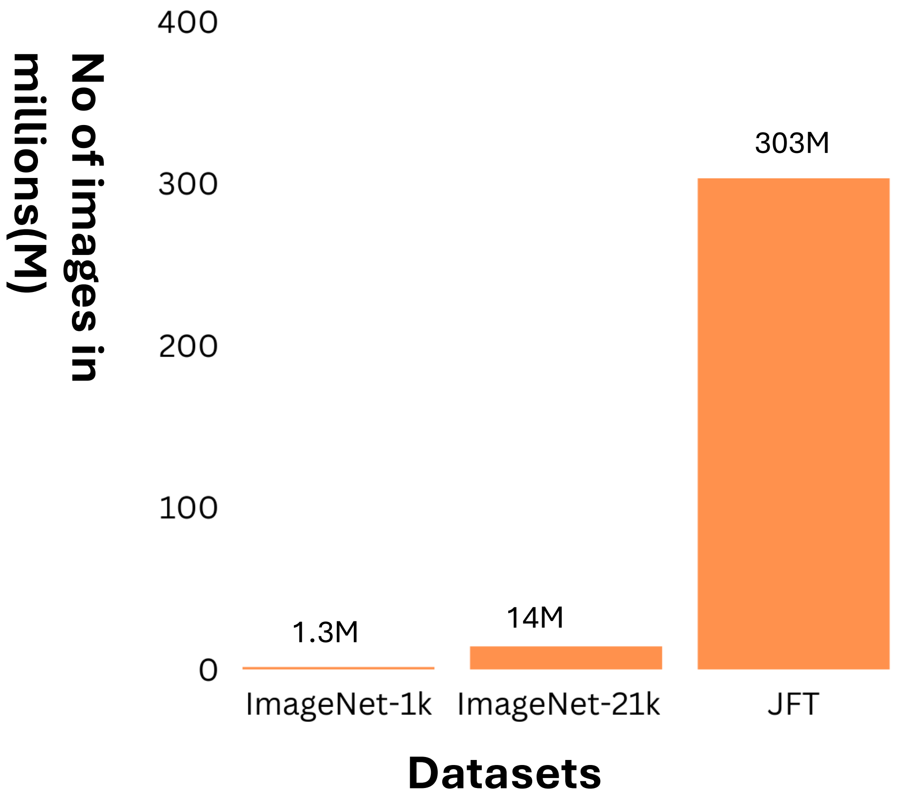


The MLP contains two layers with a GELU non-linearity.

$$
\begin{aligned}
\mathbf{z}_0 &= [\mathbf{x}_{\text{class}};\, \mathbf{x}_p^1\mathbf{E};\, \mathbf{x}_p^2\mathbf{E};\cdots;\, \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, && \mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D} && (1)\\
\mathbf{z'}_\ell &= \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, && \ell = 1 \ldots L && (2)\\
\mathbf{z}_\ell &= \mathrm{MLP}(\mathrm{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell, && \ell = 1 \ldots L && (3)\\
\mathbf{y} &= \mathrm{LN}(\mathbf{z}_L^0) && && (4)
\end{aligned}
$$

Training Dataset Vision Transformer

Vision Transformer(ViT) was pre-trained on 3 datasets of varying size and scale.



Pre-training datasets

- ImageNet-1K: 1.3M images, 1K classes
- ImageNet-21k: 14M images, 21K classes
- JFT: 303M images, 18K classes

ImageNet-1k and ImageNet-21K are also used for fine-tuning!

Pre-Training Vision Transformer

Vision Transformer(ViT) was pre-trained with same configurations of as BERT.

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

Table 1: Details of Vision Transformer model variants.

All the models are pre-trained using ADAM optimizer with batch size of 4096

Fine-Tuning Vision Transformer

- All the models are fine tuned using SGD optimizer with monentum and batch size of 512 for ViT-L/16 and 518 for ViT-H/14.

	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

Image Classification Accuracies

- Pretrained on ImageNet (small), ViT is slightly worse than ResNet.
- Pretrained on ImageNet-21K (medium), ViT is comparable to ResNet.
- Pretrained on JFT (large), ViT is slightly better than ResNet.

Self-supervision:

Most of Transformers success in NLP is the result of large-scale self-supervised pre-training where Transformer is trained on massive unlabelled data from the web.

Using masked work prediction technique that were used in BERT(randomly masking words in input sentence), ViT designers also tried the same technique where they masked 50% of patches(masked patch prediction) but achieved less performance than supervised pre-training(79.9% ACC on ImageNet while supervised pre-training is ~85%).

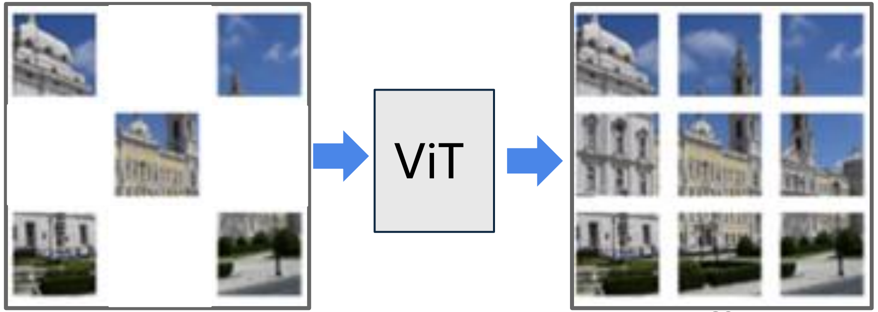
Masked-word prediction in BERT

Transformer is an efficient deep learning architecture

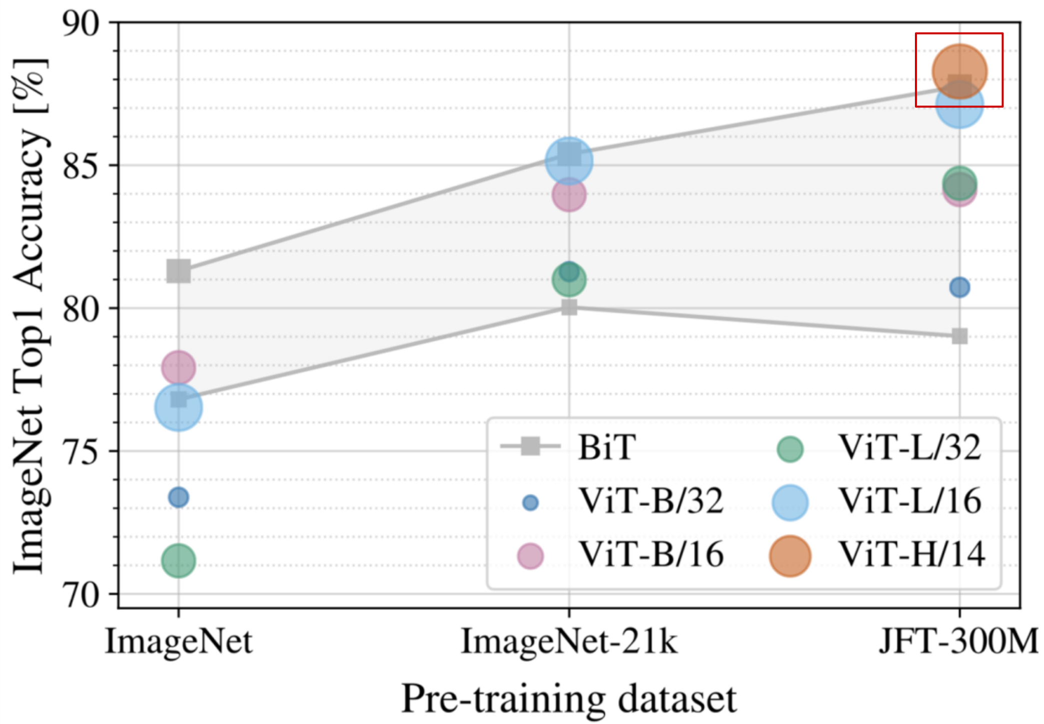
Transformer Encoder

Transformer is an efficient deep learning architecture

Masked patch prediction ViT



Vision Transformer vs SOTA CNN(ResNet)

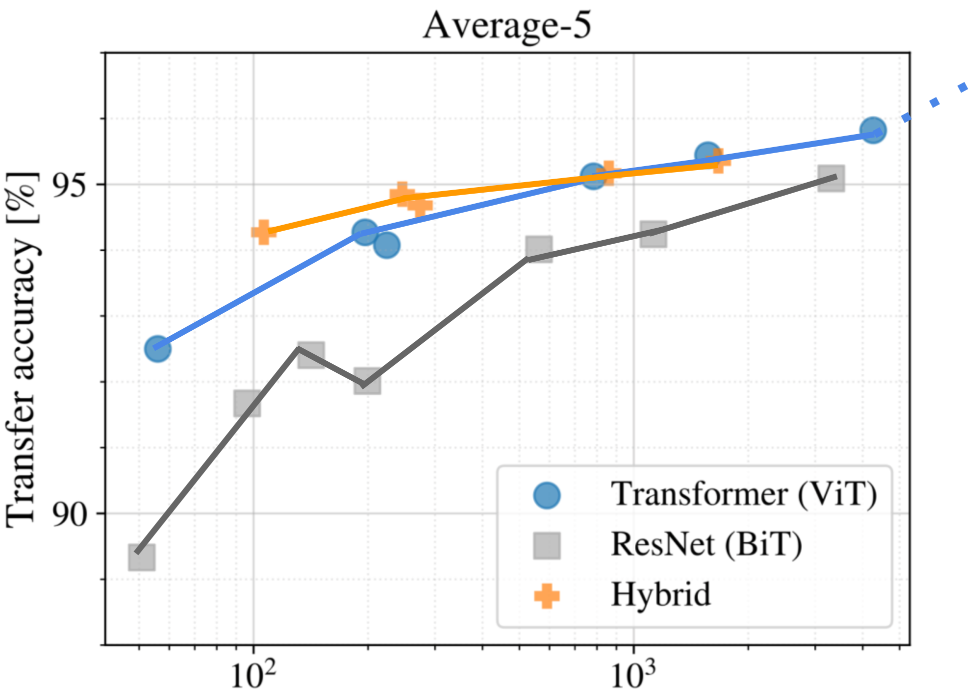


B: Base, L: Large, H: Huge
14, 16, 32: Patch size(the smaller patch size, the more the patches, and the bigger the model) >> **N = HW/PP**
Ex: ViT-B/16: Base ViT with **16x16** patch size

Results

- On small pre-training dataset(ImageNet-1k, 1.3M images), ResNet performs better than ViT due to CNN spatial inductive biases that compensate for small dataset.
- On medium pre-training dataset(Imagenet-21k, 14M images), ViTs and ResNet performance are almost similar although ViTs perform slightly better.
- On large pre-training dataset(JFT, 303M images), large ViT outperforms ResNet and show no sign of plateau.

Vision Transformer vs SOTA CNN(ResNet)



Pre-training compute(in ExaFLOPs)

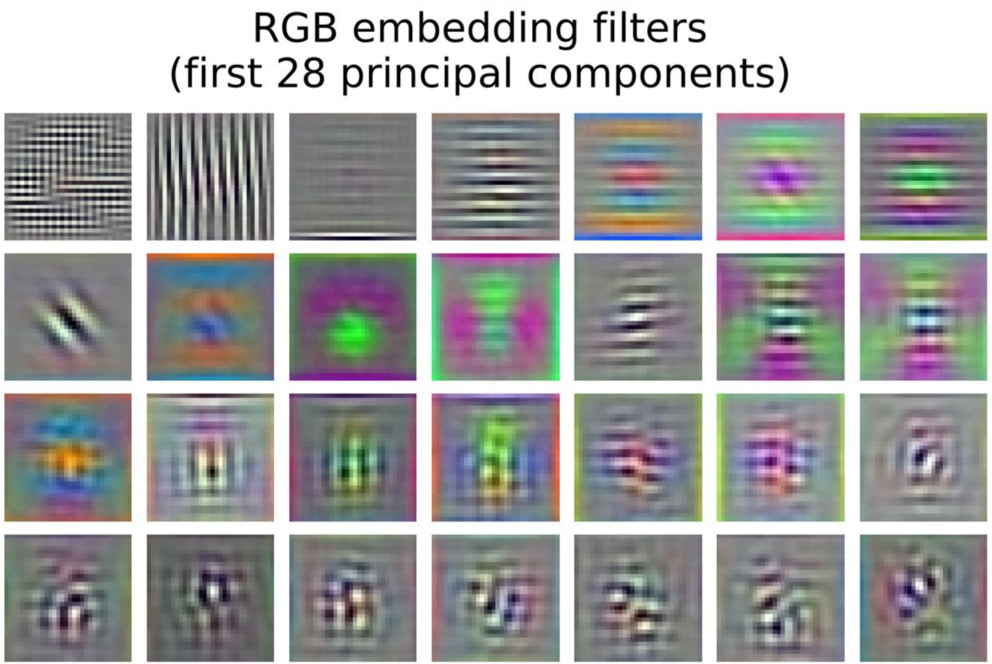
1 FLOP = 1 multiply-add(wx+b) per second
FLOPs: floating point operations per second

Pre-training compute

- ViT clearly outperforms ResNet on performance/compute trade-off.
- ViT uses approximately 2-4x less compute to achieve the same transfer accuracy(average of all downstream datasets).
- Hybrid(CNN+ViT) slightly outperforms ViT on relatively small compute, but vanishes on large compute budget.
- ViT shows extreme scaling behavior. Its performance doesn't seem to saturate for increased compute.

Inspecting ViT Representation

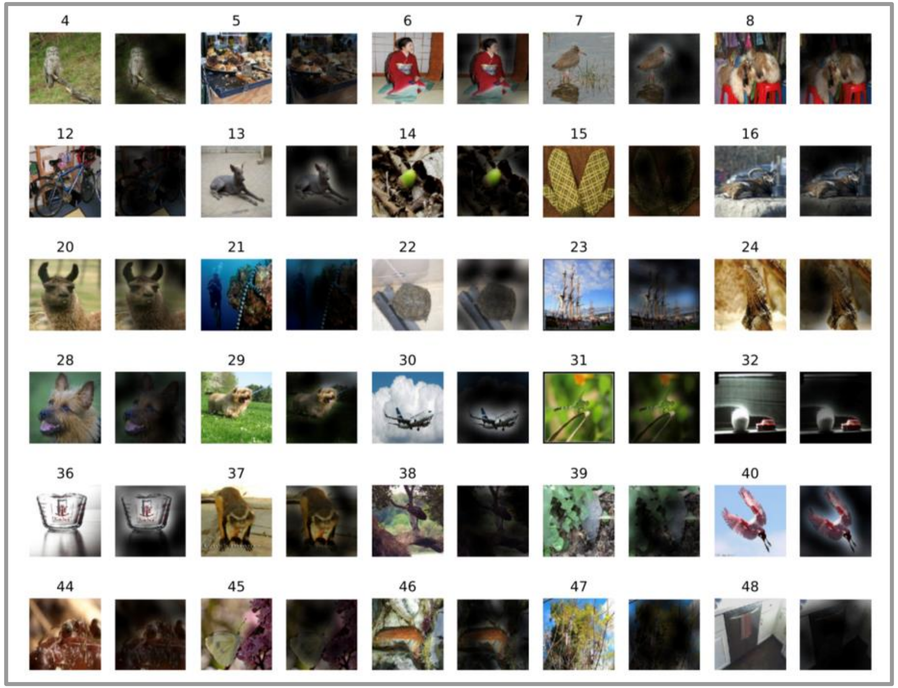
Vision Transformer shows remarkable performance when trained on massive datasets.



The visualized linear embedding of flattened patches shows that the first layer of ViT(linear projection) learns the low level features(such as edges, blobs) of the input image much like ConvNets do!

Inspecting ViT Representation

Vision Transformer shows remarkable performance when trained on massive datasets. How does it processes images internally?



On global level, ViT attends to the meaningful part of the image and ignore the rest.

Conclusion

Vison Transformer matches or exceeds the state of art on many image classification datasets, while being relatively cheap to pre-train.

While initial results are encouraging, we need to analyse the performane of ViT on other computer vison tasks, such as detection and segmentation.

Image: Google AI blog

Thank You!