AN IMAGE IS WORTH 16X16 WORDS: Vision Transformer (ViT)

Team:
Waasi A Jagirdar (waj2117)
Rushil Samir Patel (rp3268)
Sergey Sokolovskiy (ss7299)

EECS E6691 Advanced Deep Learning, 2025 Spring

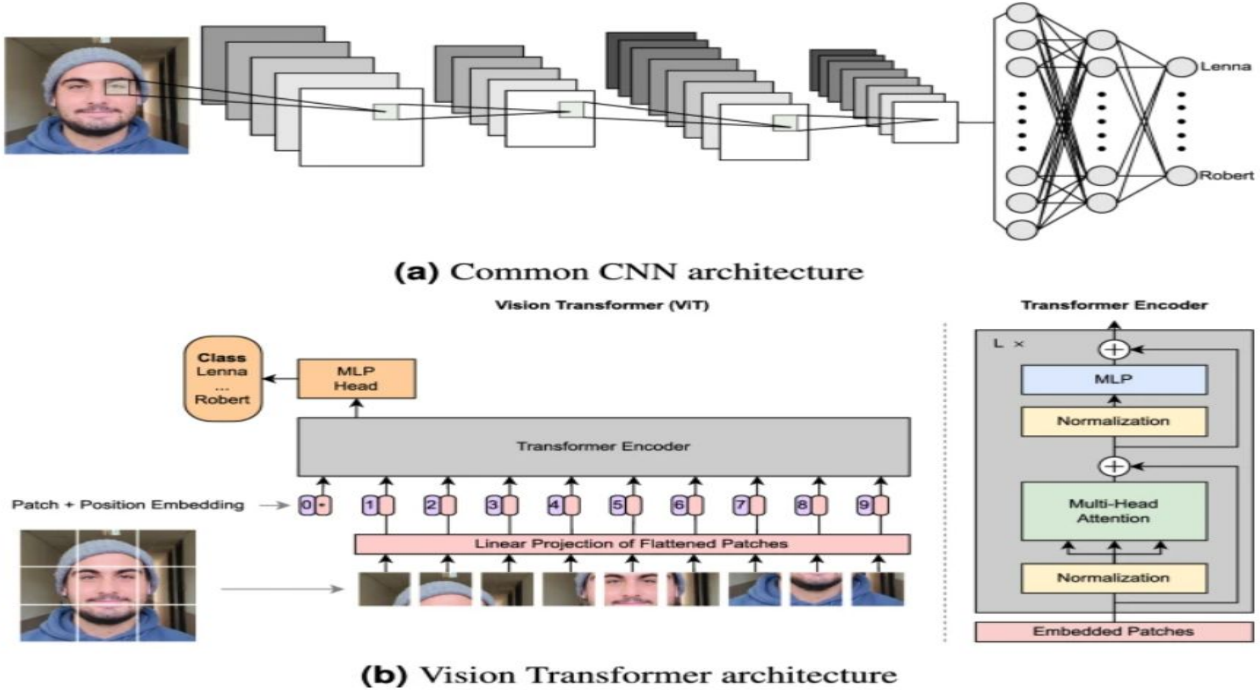
Outline of the Presentation

- Introduction
- Motivation for Vision Transformer (ViT)
- Key Ideas
- Data Preparation
- Architecture and Functionality
- Results
- Recent Advances in ViTs
- Future Works
- Demo (?)
- Conclusion
- References

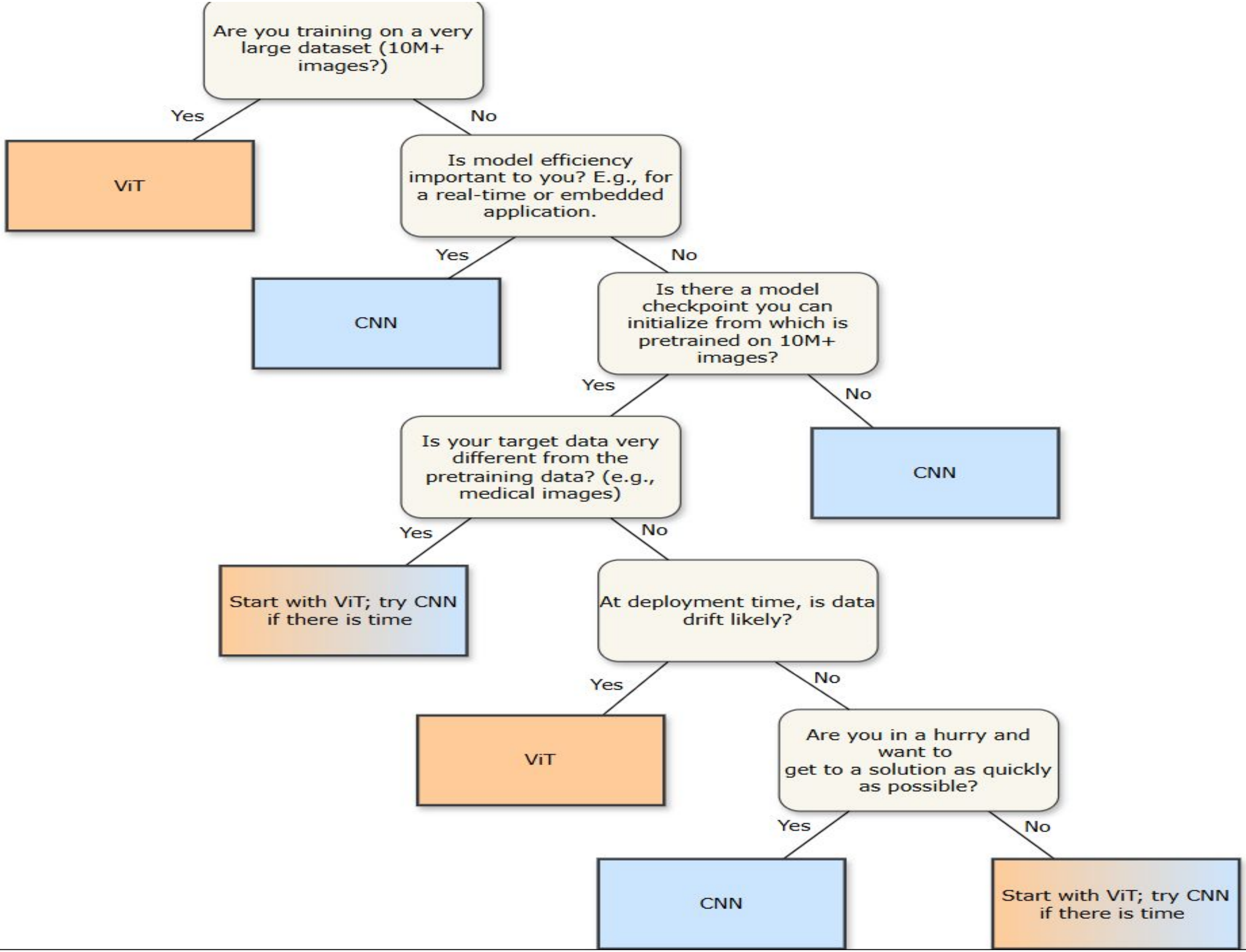
Introduction

- What is a Vision Transformer
- How does it differ from a Convolutional Neural Network(CNN)

Fig. 1



https://www.nature.com/articles/s41598-024-72254-w



- https://tobiasvanderwerff.com/2024/05/15/cnn-vs-vit.html

Motivation for the Vision Transformer (ViT)

- What are the core limitations of CNN's
- Transformers work well in NLP due to self-attention—can we apply them to images?

AlexNet (2012) → Resnet (2015) → EfficientNet (2019) → Vision Transformer (2021)

AlexNet (2012)

- ReLu activation over sigmoid/tanh
- Introduced Dropout
- GPU Usage

Resnet (2015)

- Skip Connections
- Batch Normalization
- Deeper Networks

EfficientNet (2019)

- Compound Scaling
- Better accuracy with fewer parameters
- Neural Architecture Search(NAS)

Vision Transformer (2021)

- Transformer based
- Self attention for global dependencies(long range)

Inductive Bias in CNN's

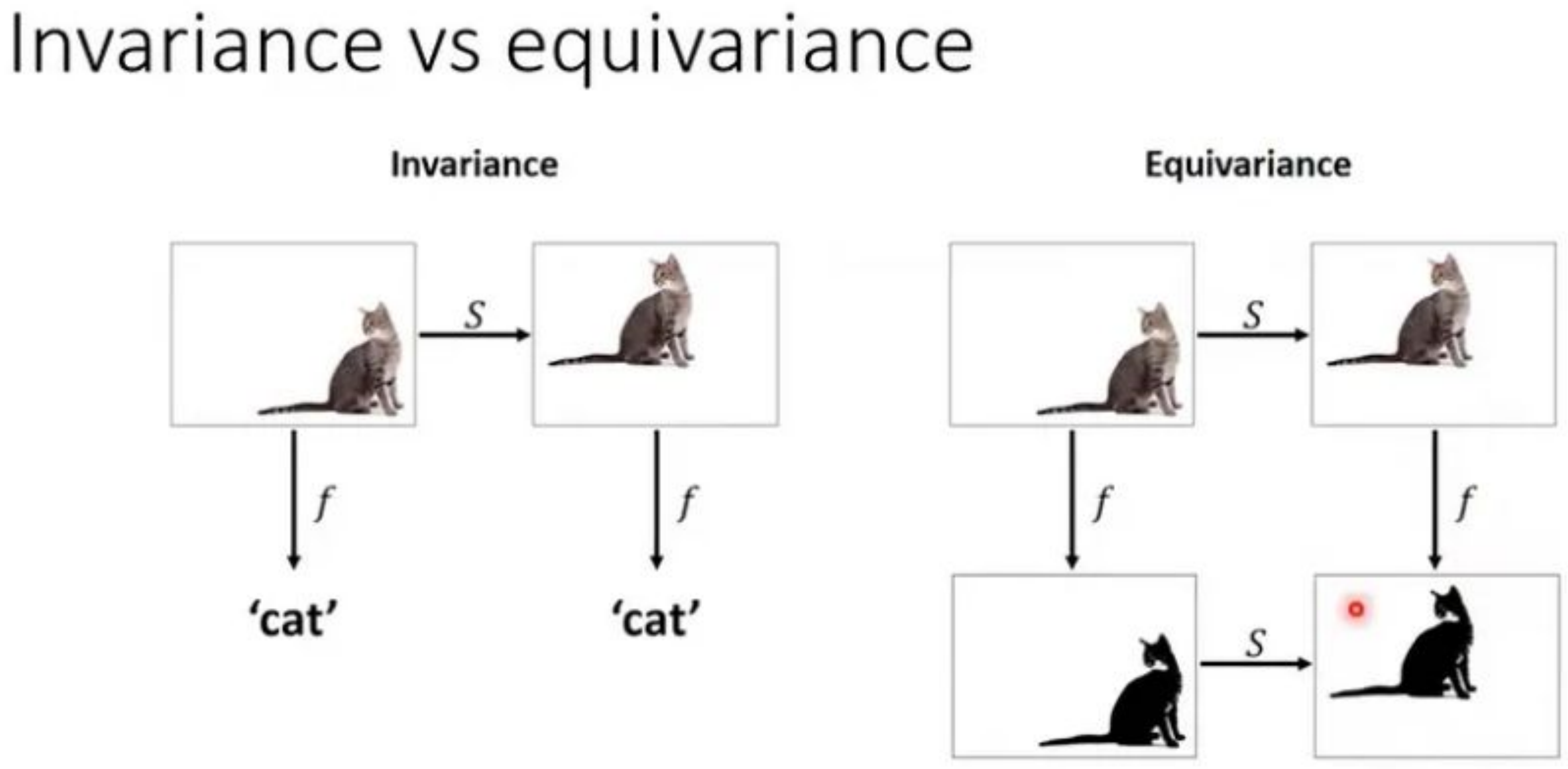
- Inductive Bias is the set of assumptions a model makes about the data to improve learning.

Local Feature Learning
↓
Capture local features before high level concepts (more robust model)

Weight Sharing
↓
Same convolutional filters across entire image(fewer params, less overfitting , high efficiency

Translational Equivariance
↓
Ability to detect patterns regardless of their position.

Translational Equivariance



- https://medium.com/@sue.sk.guo/all-you-should-know-about-translation-equivariance-invariance-in-cnn-cbf2a2ad33cd

Correlation with Vision Transformers

- Instead of translational equivariance , Positional Embeddings are added to maintain spatial relationships, but they are learned rather than built-in.
- Due to no weight sharing , ViTs need more training data because they lack this efficiency advantage.
- No local feature learning in ViTs, and hence they struggle with small datasets since they must learn spatial hierarchies from scratch.
- Vision Transformers (ViTs) do not have the inductive biases that CNNs naturally possess. This is a fundamental reason why ViTs require much larger datasets for training compared to CNNs.
- As a result , CNN's are superior for smaller datasets

Challenges & Early Attempts of applying vision to transformers

- Naive self-attention would require each pixel to attend to every other pixel.
- Pre - ViT phase (2017-2021):

Local Attention (2018)	Sparse Transformers(2019)	Block Attention(2019)	ViT(2021)

Key Breakthroughs of ViT:

- Scaled Transformers for vision by using patch-based tokenization instead of pixel-level attention.
- Large-scale pre-training (JFT-300M, ImageNet-21k) was key to making ViTs competitive with CNNs.

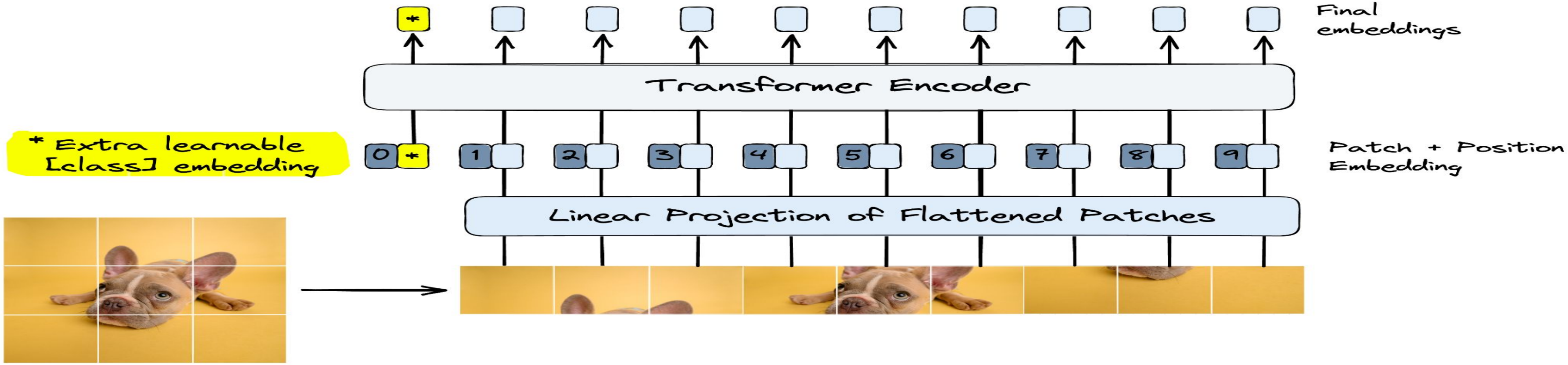
Data Preparation

Conv2D Layer

Manual segmentation

Patch Embeddings

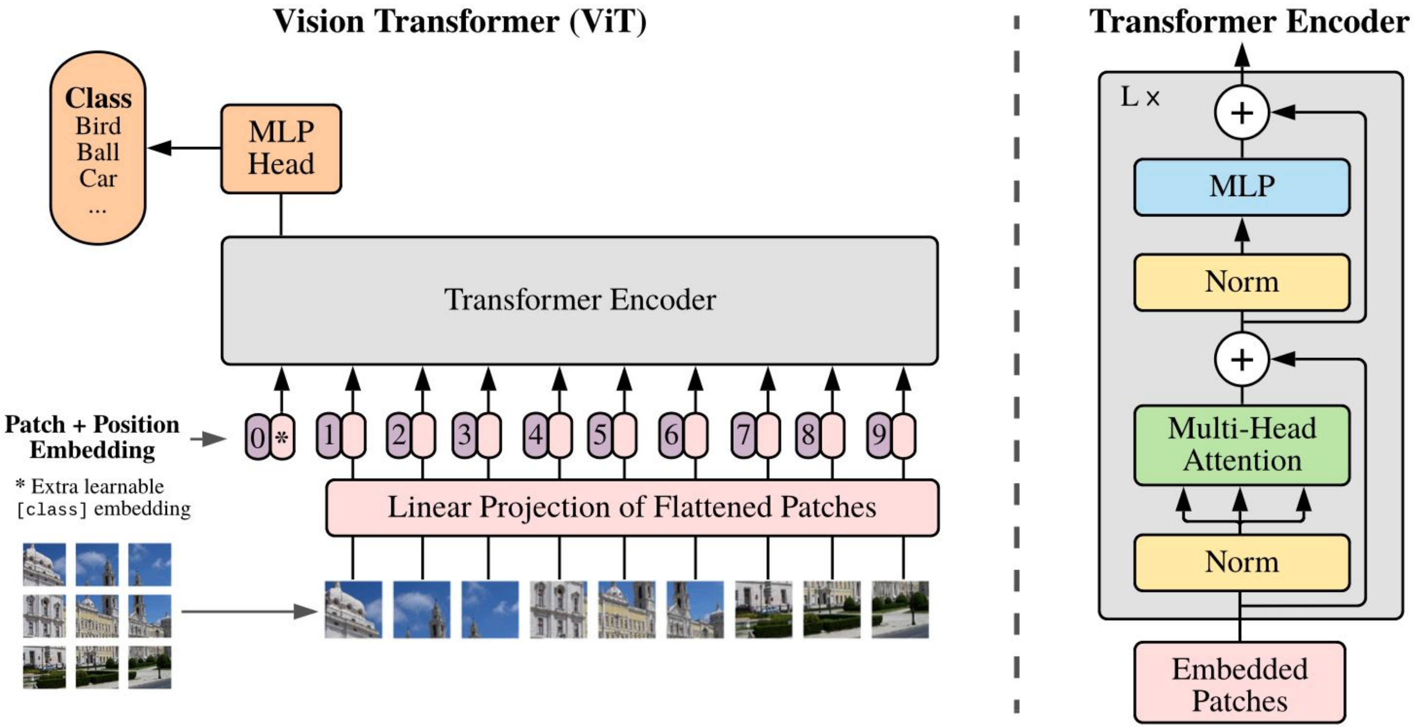
- ViTs split images into fixed-size patches (eg 16x16)
- Instead of word tokens , ViT consumes image patches
- Each patch is flattened and passed through a **linear projection layer (MLP)** to create **patch embeddings**.



- https://www.pinecone.io/learn/series/image-search/vision-transformers/

Architecture

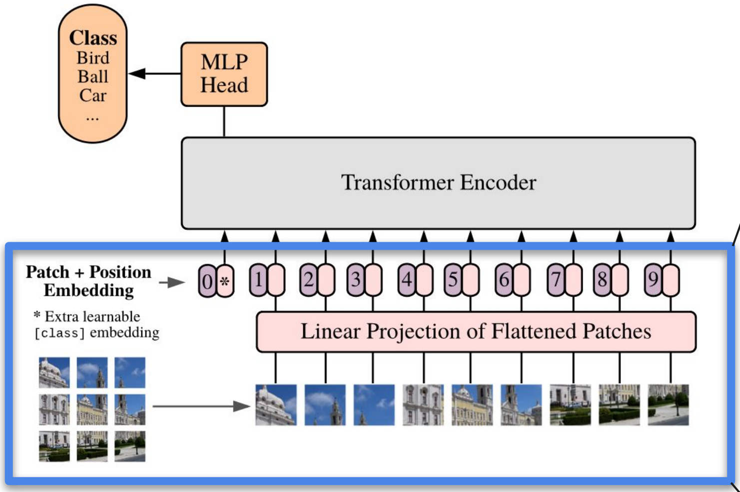
High Level



- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. https://arxiv.org/abs/2010.11929
- https://towardsdatascience.com/vision-transformers-explained-a9d07147e4c8/

Architecture

Input

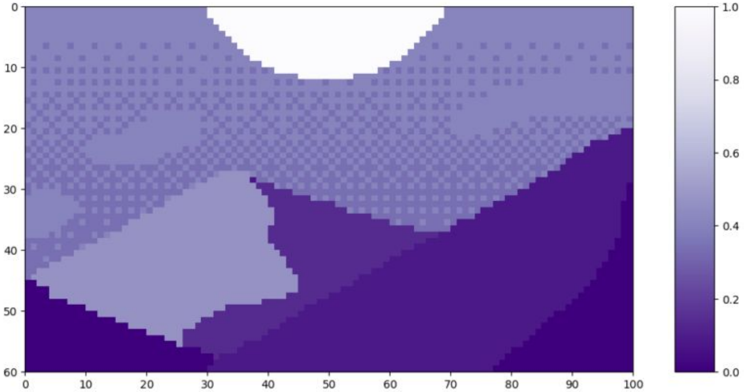


*Some papers may directly use CNN features.

- https://towardsdatascience.com/vision-transformers-explained-a9d07147e4c8/

Architecture

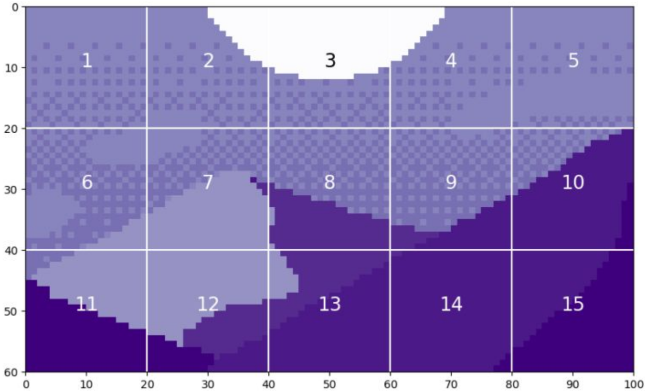
Creating patches / tokenization



Mountain at Dusk is H = 60 and W = 100 pixels.

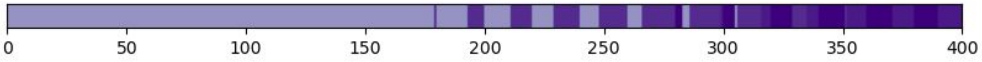
$$N = \frac{HW}{P^2}$$

*Padding
-zero
-reflection



There will be 15 patches, each 20 by 20.

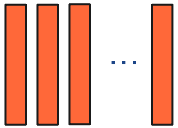
Flatten Patches



Each patch will make a token of length 400. Length = P^2.

Linear Layer

x Activation



N Patches -

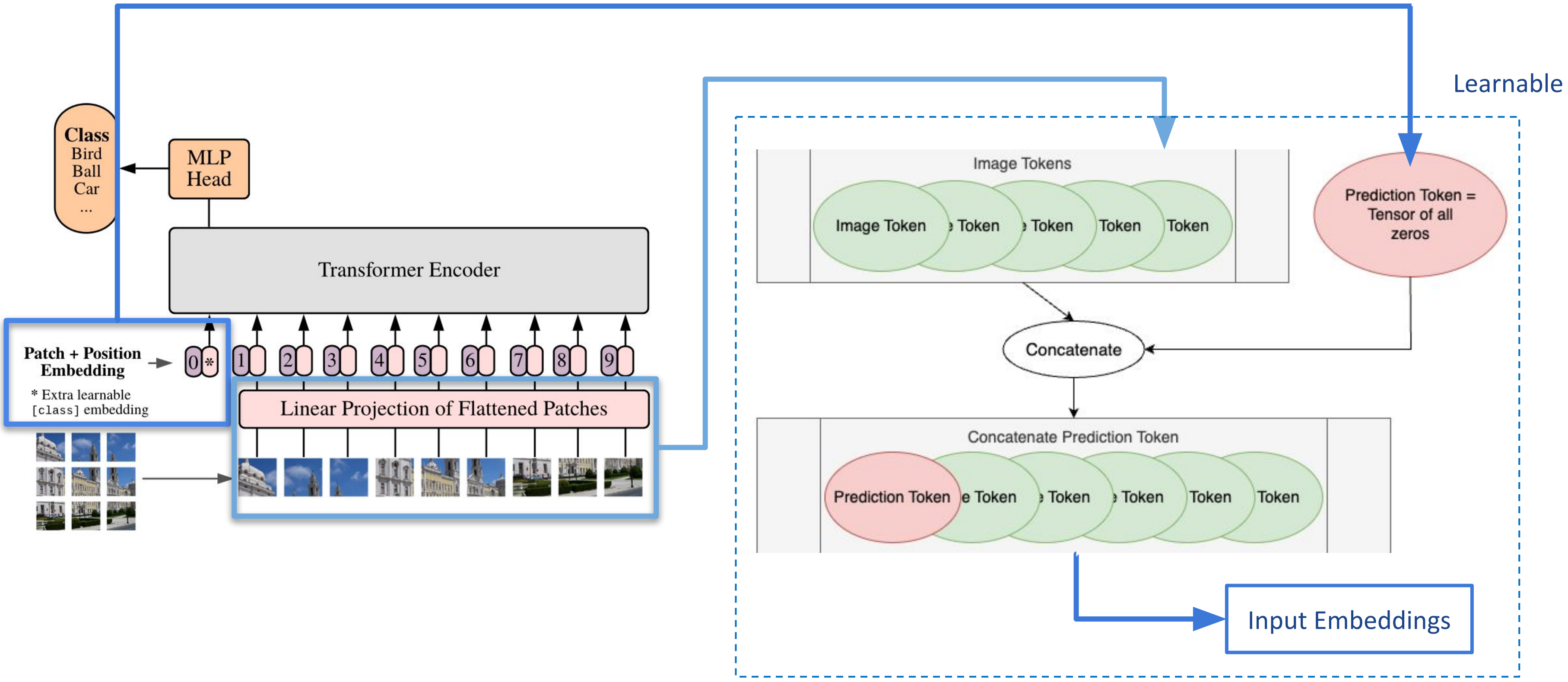
N Tokens

TokenLen x N

- https://towardsdatascience.com/vision-transformers-explained-a9d07147e4c8/

Architecture

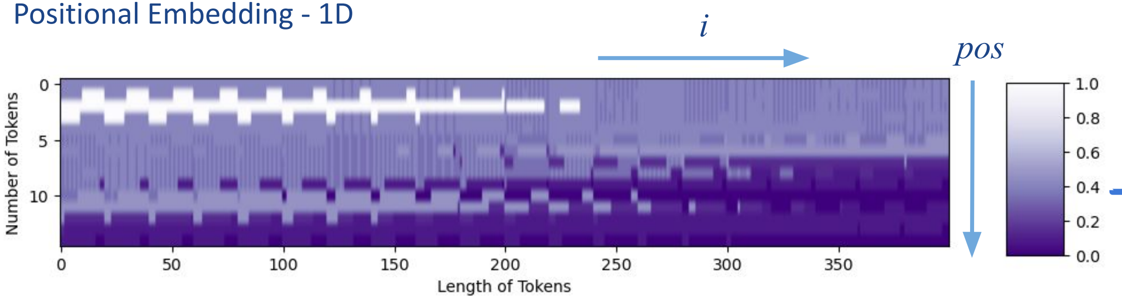
Creating patches / tokenization

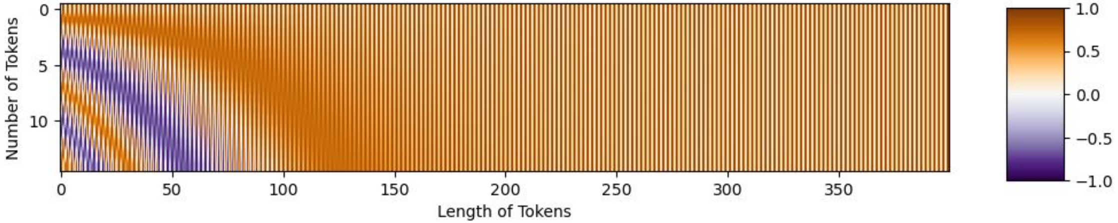


- https://towardsdatascience.com/vision-transformers-explained-a9d07147e4c8/

Architecture

Positional Embedding - 1D





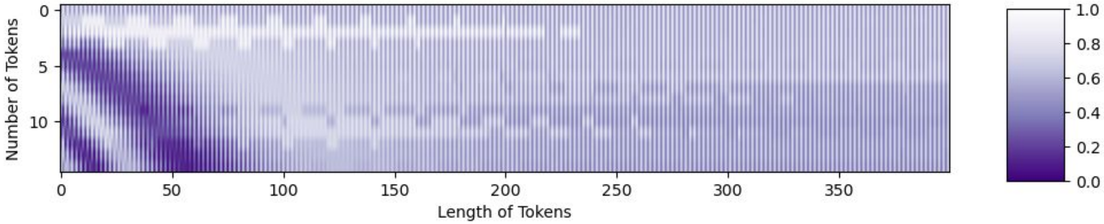
Can also be learnable

$$PE(pos, 2i) = \sin \frac{pos}{10000^{\frac{2i}{d_{model}}}}$$

$$PE(pos, 2i+1) = \cos \frac{pos}{10000^{\frac{2i}{d_{model}}}}$$

where d_{model} = TokenLen

Element-wise Addition

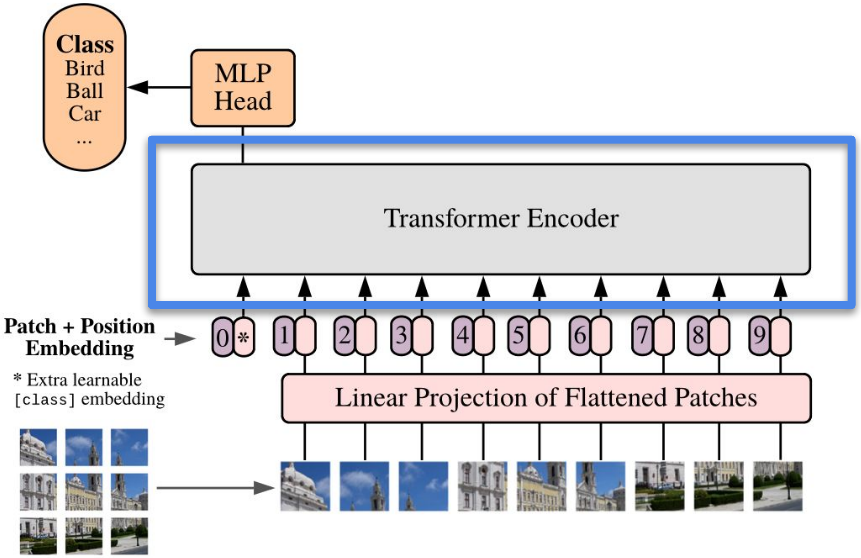


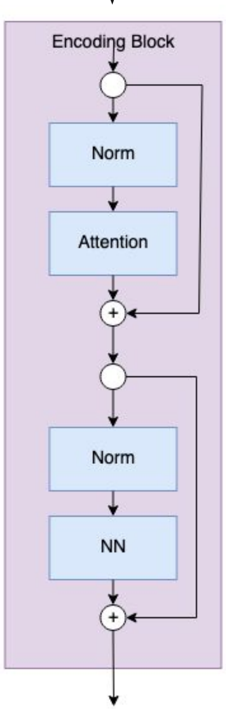
- https://towardsdatascience.com/vision-transformers-explained-a9d07147e4c8/
- https://github.com/hkproj/transformer-from-scratch-notes

Architecture

Encoder

Input Embeddings + Positional Embeddings

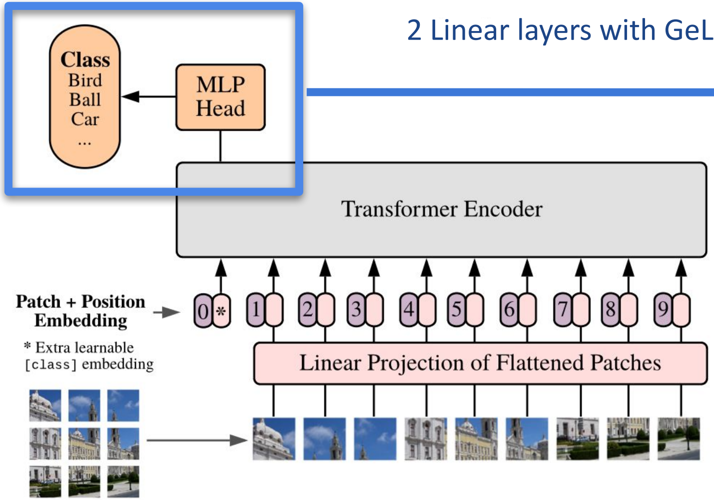




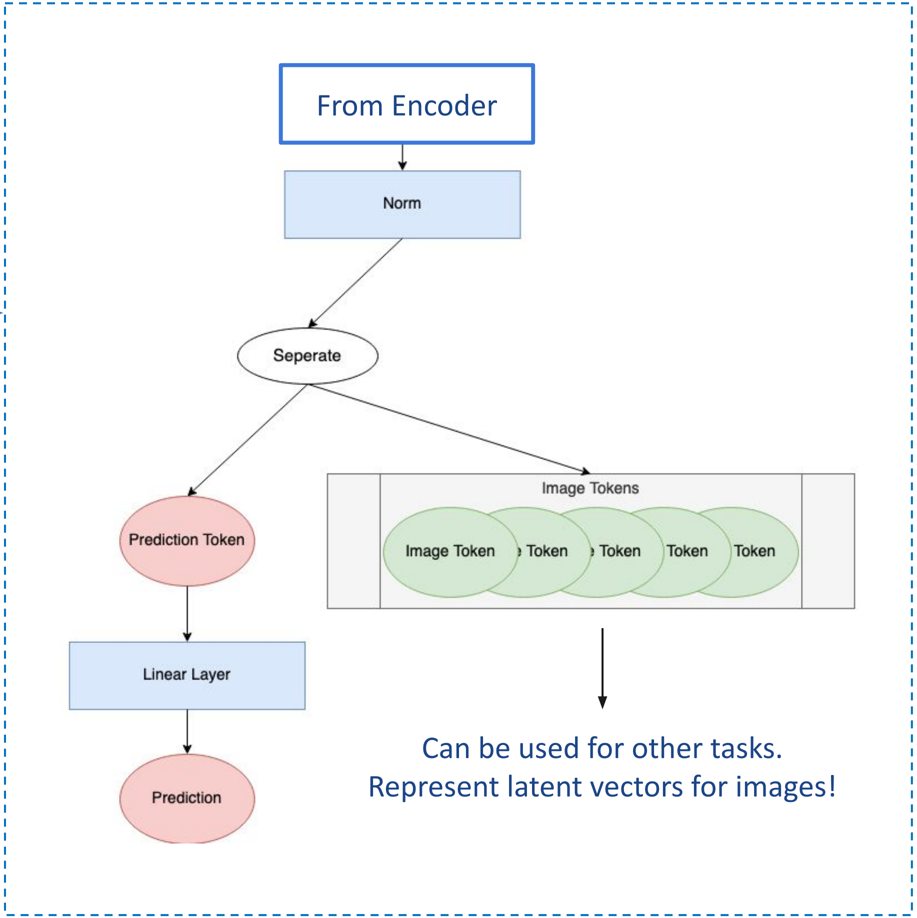
https://towardsdatascience.com/vision-transformers-explained-a9d07147e4c8/

Architecture

Classification



2 Linear layers with GeLU

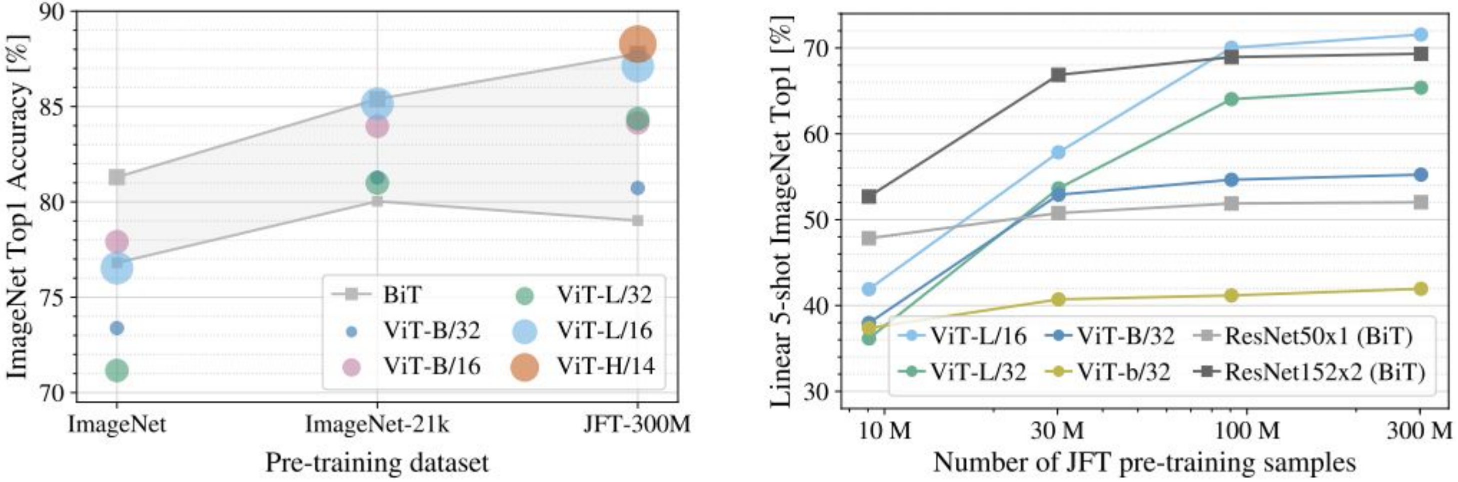


- https://towardsdatascience.com/vision-transformers-explained-a9d07147e4c8/

Results of the Original Paper: General Results

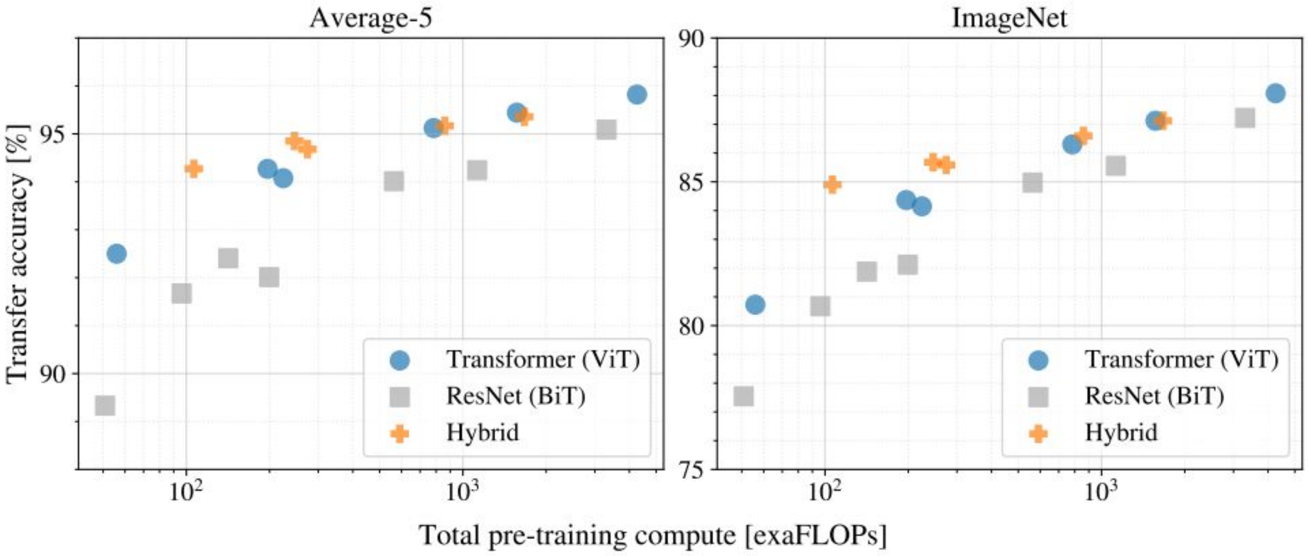
Dataset	ViT-H/14 (JFT-300)	ViT-L/16 (JFT-300)	BiT-L (ResNet152x4)
ImageNet	**88.55**	87.76	87.54
CIFAR-100	**94.55**	93.90	93.51
Oxford-IIIT Pets	**97.56**	97.32	96.62
Oxford Flowers-102	99.68	**99.74**	99.63
VTAB	**77.63**	76.28	76.29
Training Time (TPU-core-days)	2,500	680	9,900

Results of the Original Paper: Transfer Learning



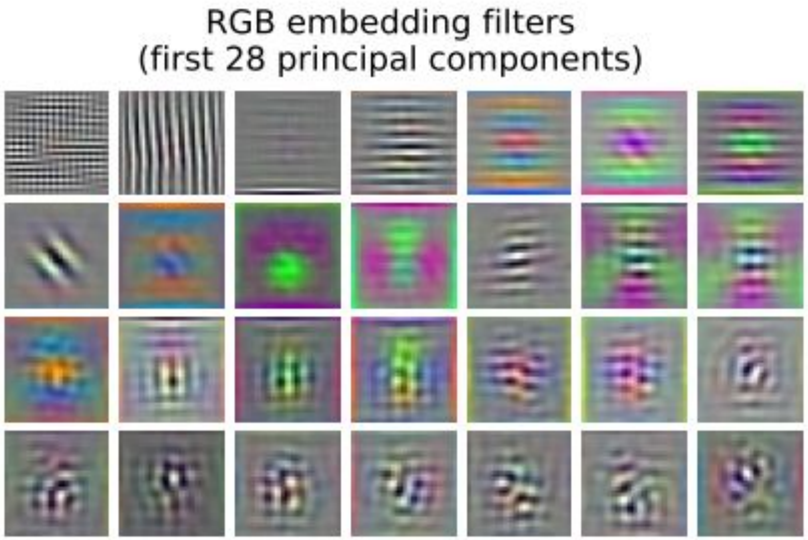
Transfer Learning Results for Fine-Tuning (left) and Linear Few-Shot (right) approaches. For larger pre-training datasets, ViTs tend to outperform ResNets.

Results of the Original Paper: Scaling Study

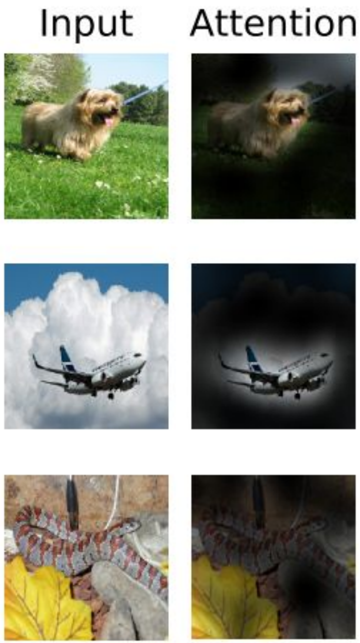


Performance versus pre-training compute for different architectures.

Results of the Original Paper: Behind the Scenes of ViT



Filter masks of the first linear layer that provides embedding for the flattened patches

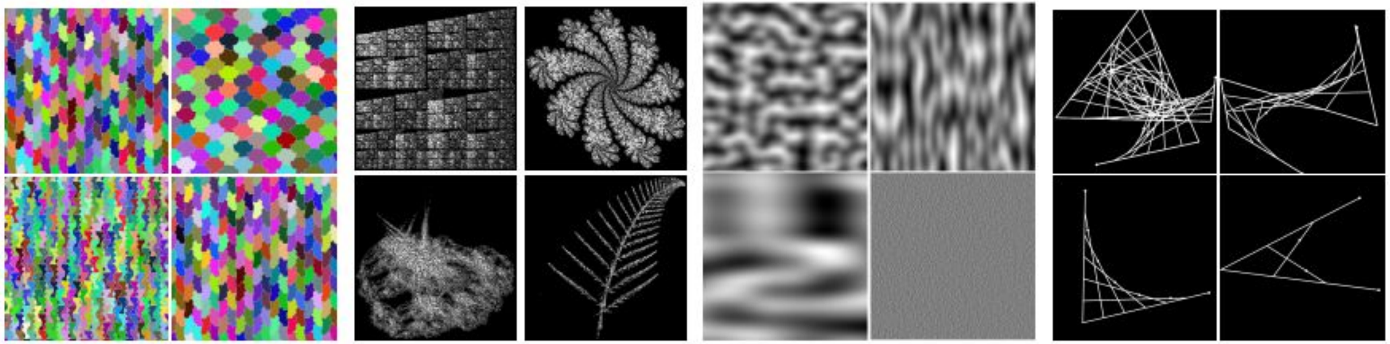


ViT Attention Examples

Challenges and Future Work

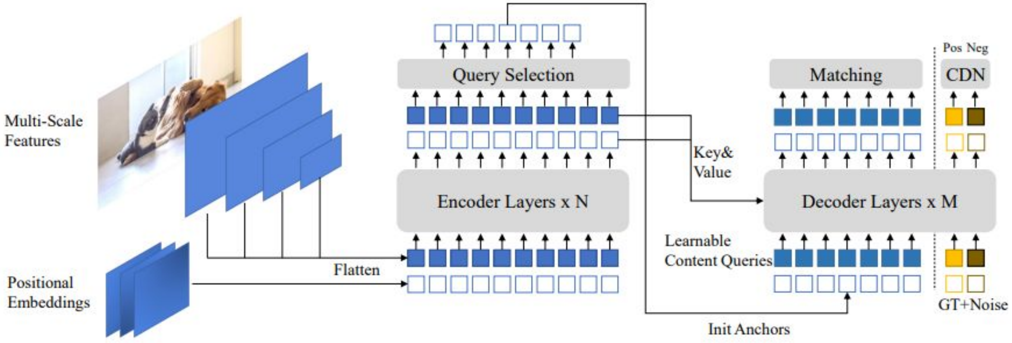
- Deeper exploration of the pre-training methods
  - Several research directions have been explored in order to study alternative pre-training approaches
- Applying ViT to other computer vision tasks such as object detection and segmentation:
  - Has been addressed in DETR (Carion et al.), and DINO (Zhang et al.) but these architectures still rely on CNNs
  - Other models like WB-DETR (Liu et al.) and Swin Transformer (Lui et al.) have been developed that do not explicitly rely on CNN backbone
- Scaling of ViT to further improve performance of these models:
  - Google DeepMind introduced a 22B-parameter ViT in 2023

Recent Advances in ViT: Pre-Training



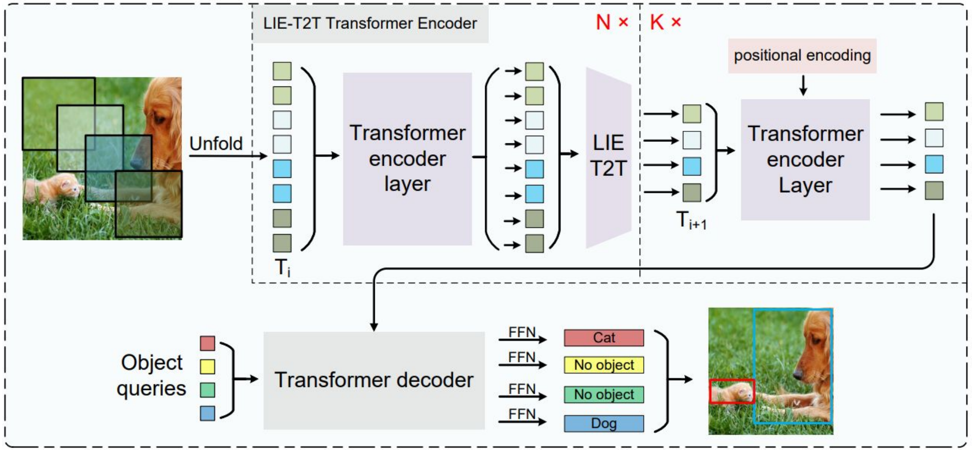
Synthetic Kernels used in Formula-Driven Supervised Learning (FDSL).

Recent Advancements in ViTs: Object Detection

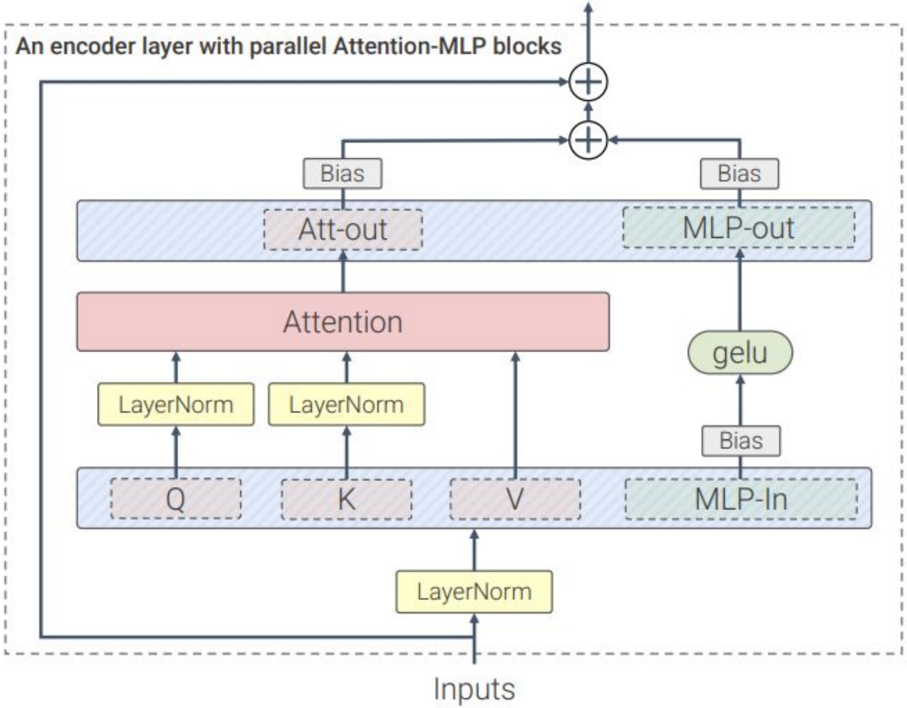


DETR with improved de-noising anchor boxes (DINO) architecture.

Transformer-Based Detector without Backbone (WB-DETR) architecture.

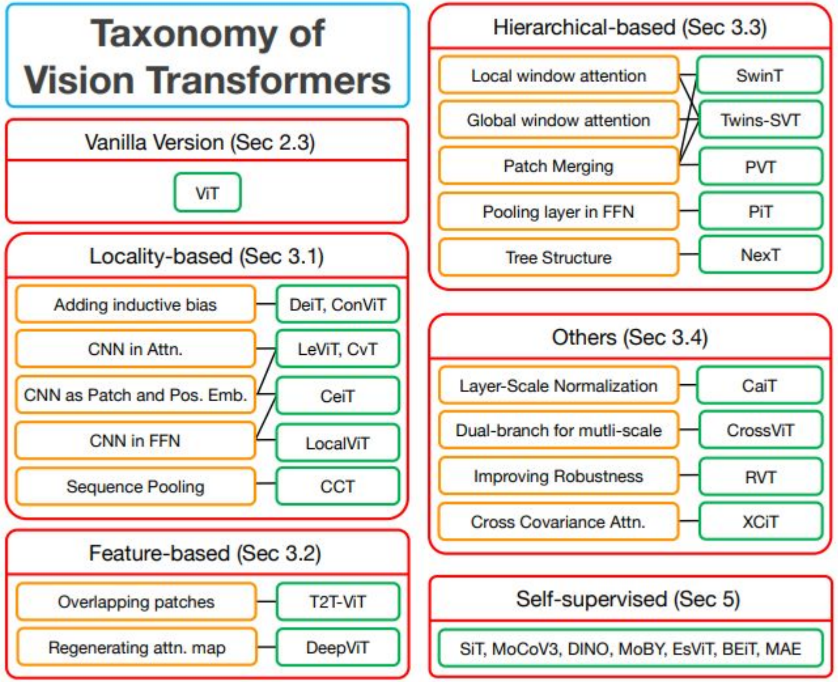


Recent Advancements in ViTs: Model Scaling



- ViT have recently been scaled to the models that incorporate up to 22B parameters
- Structure is the same as the one of original ViT with the following modifications:
  - Parallel Layers
  - QK Normalization
  - Omitting Biases in QKV Projections
- Achieves the best results on many benchmark datasets and is computationally efficient

Vision Transformers: Current State of the Field



Conclusion

References

- Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229). Cham: Springer International Publishing.
- Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., ... & Shum, H. Y. (2022). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).
- Liu, F., Wei, H., Zhao, W., Li, G., Peng, J., & Li, Z. (2021). WB-DETR: Transformer-based detector without backbone. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2979-2987).
- Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z. H., ... & Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 558-567).
- Nakamura, R., Kataoka, H., Takashima, S., Noriega, E. J. M., Yokota, R., & Inoue, N. (2023). Pre-training Vision Transformers with Very Limited Synthesized Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20360-20369).
- Kataoka, H., Matsumoto, A., Yamada, R., Satoh, Y., Yamagata, E., & Inoue, N. (2021). Formula-driven supervised learning with recursive tiling patterns. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4098-4105).
- Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., ... & Houlsby, N. (2023, July). Scaling vision transformers to 22 billion parameters. In International Conference on Machine Learning (pp. 7480-7512). PMLR.
- Ruan, B. K., Shuai, H. H., & Cheng, W. H. (2022). Vision transformers: state of the art and research challenges. arXiv preprint arXiv:2207.03041.