Vision Transformers

Naeemullah Khan
naeemullah.khan@kaust.edu.sa

KAUST Academy
King Abdullah University of Science and Technology

June 11, 2025

Table of Contents

أكاديمية كاوست
KAUST ACADEMY

Why move beyond CNNs?

- **Limitations of CNNs:** While CNNs excel at capturing local patterns, their strong inductive biases can restrict modeling of global context.
- **Vision Transformers (ViTs):** ViTs leverage self-attention mechanisms to effectively capture long-range dependencies across an image.
- **Performance and Flexibility:** On large-scale vision tasks, ViTs often match or surpass CNNs in accuracy, and offer greater architectural flexibility.
- **Broader Applicability:** The transformer framework enables unified modeling across different modalities, bridging vision and language tasks.

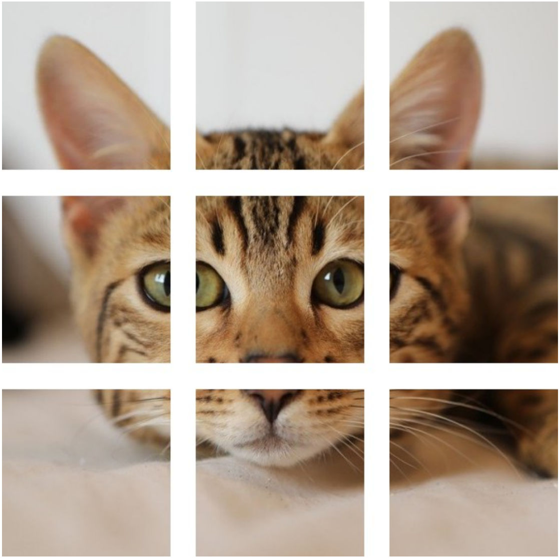
Learning Outcomes

By the end of this lecture, you will be able to:

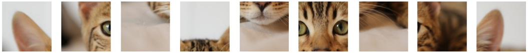
- Explain the ViT architecture and patch embedding process.
- Discuss optimization challenges and solutions (e.g., distillation).
- Contrast CNNs vs ViTs in terms of inductive bias, data efficiency, and performance.
- Describe hierarchical models (e.g., Swin) and detection frameworks (DETR).
- Understand ConvNeXt as a modern CNN influenced by transformer insights.
- Recognize current limitations and deployment considerations.

Tokenizing Images:

- **Input:** An image of size $H \times W \times C$ is divided into non-overlapping patches of size $P \times P$.
- Each patch is flattened into a vector and projected linearly to form patch embeddings.
- Positional encodings are added to retain spatial information.



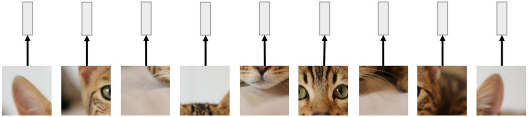
N input patches, each of shape 3x16x16



Vision Transformers (ViTs) Architecture (cont.)

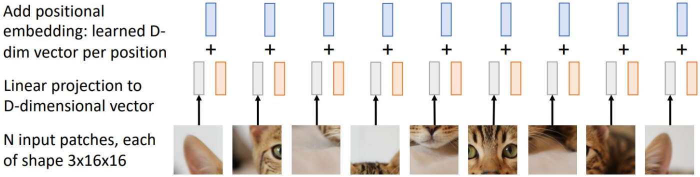
Linear projection to
D-dimensional vector

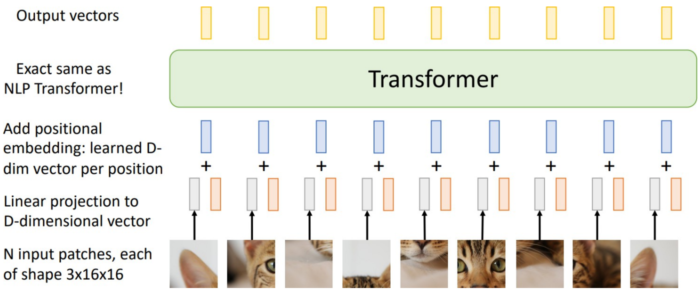
N input patches, each
of shape 3x16x16



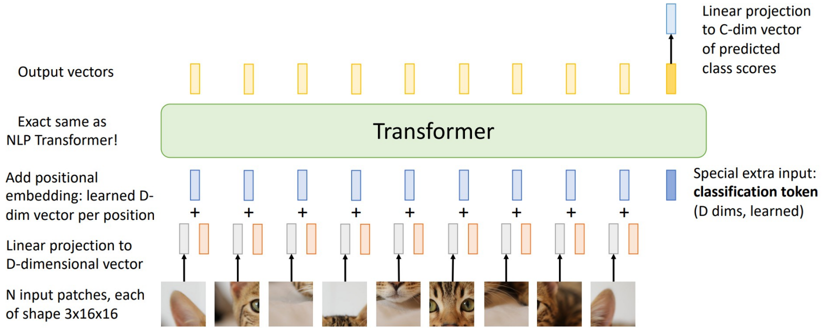
Transformer Encoder:

- The sequence of patch embeddings, along with a special [CLS] token, is processed by standard Transformer encoder blocks.
- Each block consists of multi-head self-attention and feed-forward layers.


Add positional embedding: learned D-dim vector per position
Linear projection to D-dimensional vector
N input patches, each of shape 3x16x16


Output vectors
Exact same as
NLP Transformer!
Transformer
Add positional
embedding: learned D-
dim vector per position
Linear projection to
D-dimensional vector
N input patches, each
of shape 3x16x16

Vision Transformers (ViTs) Architecture (cont.)



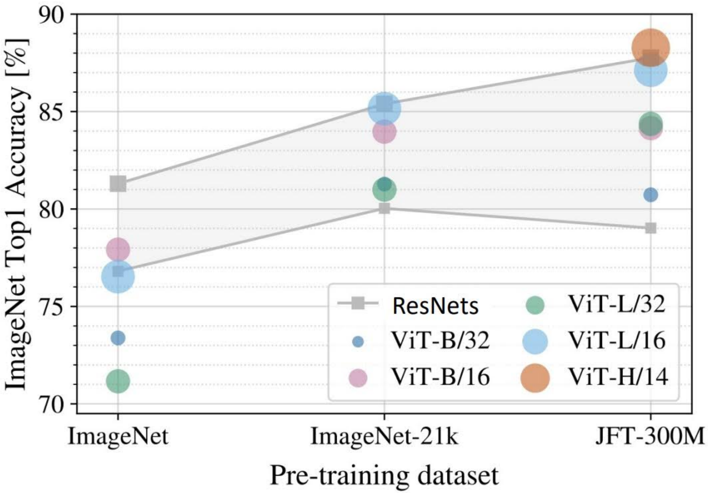
Classification:

- The output corresponding to the [CLS] token is used for image classification.

Results:

- ViT achieves competitive, often state-of-the-art, performance (e.g., 88–89% top-1 accuracy on ImageNet) when trained on large datasets.

Vision Transformers (ViTs) Architecture (cont.)



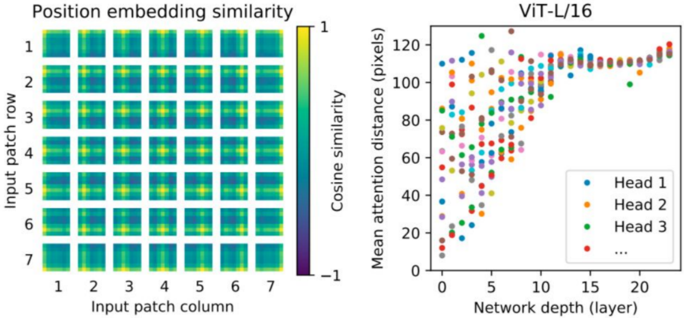
[1]

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

Patch Embedding

- **Patch Embedding:** Achieved via a linear projection of flattened patches ($P = 16$ typical)
- **Differences from CNN stems:**
  - Large stride causes optimization instability
  - Mitigated via convolutional preprocessing (e.g., small conv stem)

ViT: Learning Patterns



- ViT learns the grid-like structure of the image patches via its position embeddings.
- The lower layers contain both global and local features, while the higher layers contain only global features.

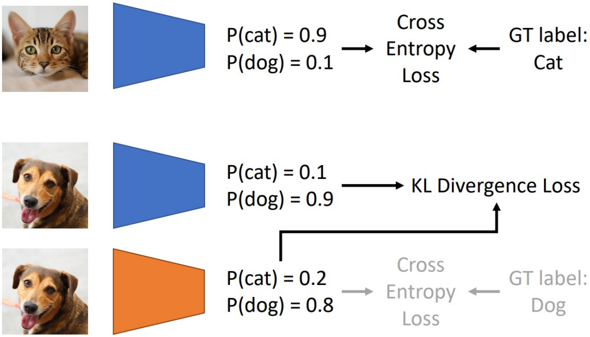
Improving ViT: Distillation

- Data-efficient ViT & DeiT approach uses *knowledge distillation* from CNN teacher models to improve data efficiency and convergence.
- Benefits: better performance with fewer data and better optimization stability.

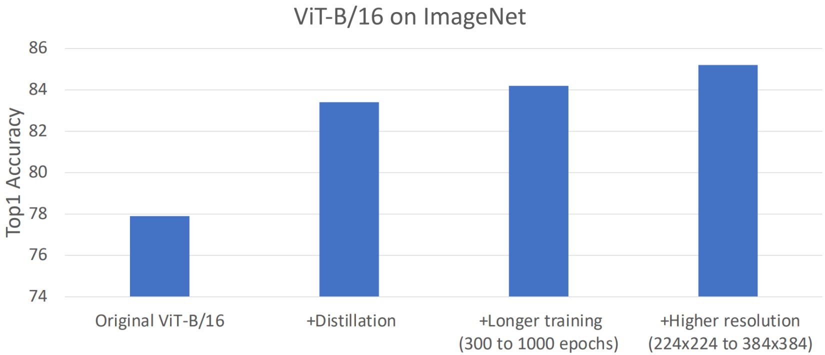
Improving ViT: Distillation (cont.)

Step 1: Train a teacher CNN on ImageNet

Step 2: Train a student ViT to match ImageNet predictions from the teacher CNN (and match GT labels)



Improving ViT: Distillation (cont.)



[2]

[2] Touvrom et al, "Training data-efficient image transformers distillation through attention", ICML 2021

CNN Vs ViT

CNN		ViT
✓	1. Maintain 2D structure logic	✓
✓	2. Shift equivariant	✗
✓	3. Consider only local correlations	✗
✓	4. Hierarchically growing field of view	✗
✓	5. Hierarchically progressing complexity	✓
✓	6. Reasonable amount of params	✓
✗	7. Global representation	✓

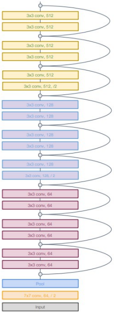
CNN Vs ViT (cont.)

Stage 3:
256 x 14 x 14

Stage 2:
128 x 28 x 28

Stage 1:
64 x 56 x 56

Input:
3 x 224 x 224

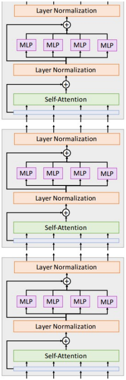


In most CNNs (including ResNets), **decrease** resolution and **increase** channels as you go deeper in the network (Hierarchical architecture)

Useful since objects in images can occur at various scales

In a ViT, all blocks have same resolution and number of channels (Isotropic architecture)

Can we build a **hierarchical** ViT model?



3rd block:
768 x 14 x 14

2nd block:
768 x 14 x 14

1st block:
768 x 14 x 14

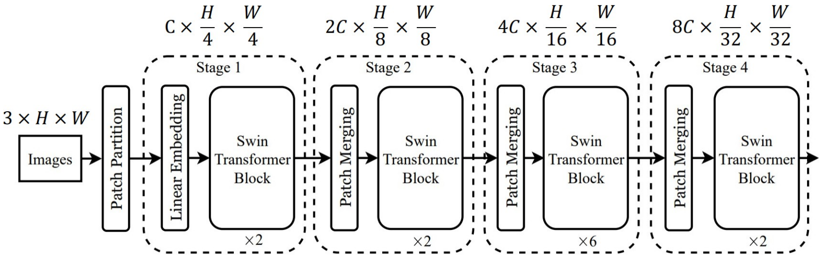
Input:
3 x 224 x 224

CNN Vs ViT (cont.)

Aspect	CNN	ViT
Inductive bias	Strong local focus, translation equivariance	Minimal bias—rely on data
Data efficiency	Good on small data	Typically needs large-scale
Global context	Local receptive fields	Global attention from day one
Optimization	Stable, flexible	Sensitive to hyperparams, regularization

Table 1: Comparison between CNN and ViT

Hierarchical ViT: Swin Transformer



Divide image into 4x4 patches and project to C dimensions

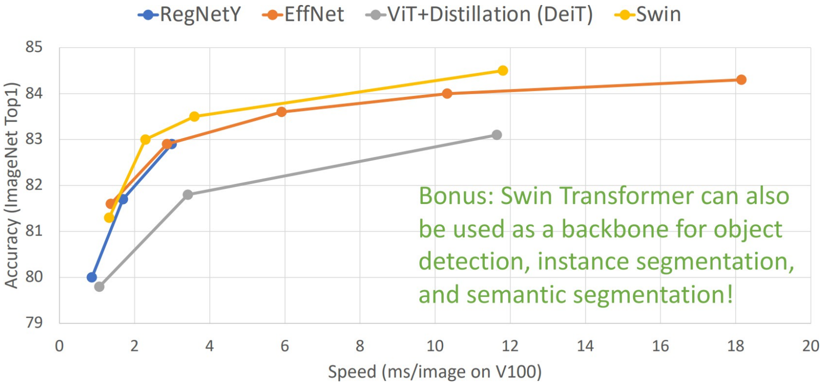
Merge 2x2 neighborhoods; now patches are (effectively) 8x8

Merge 2x2 neighborhoods; now patches are (effectively) 16x16

Merge 2x2 neighborhoods; now patches are (effectively) 32x32

Hierarchical ViT: Swin Transformer (cont.)

- Swin Transformer introduces windowed self-attention with shifting windows, merging tokens akin to pooling.[3]
- Builds a pyramid representation, enabling linear compute complexity and high performance on detection/segmentation (e.g. 58.7 COCO box AP).
- Hierarchical design allows for multi-scale feature extraction, similar to CNNs.



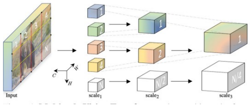
4

[3] https://arxiv.org/abs/2103.14030

[4] Liu et al, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", CVPR 2021

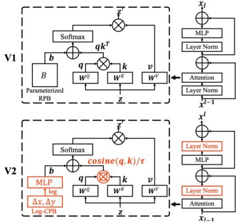
Other Hierarchical Vision Transformers

MViT



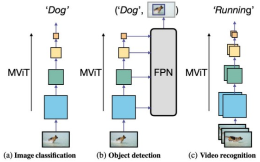
Fan et al, "Multiscale Vision Transformers", ICCV 2021

Swin-V2



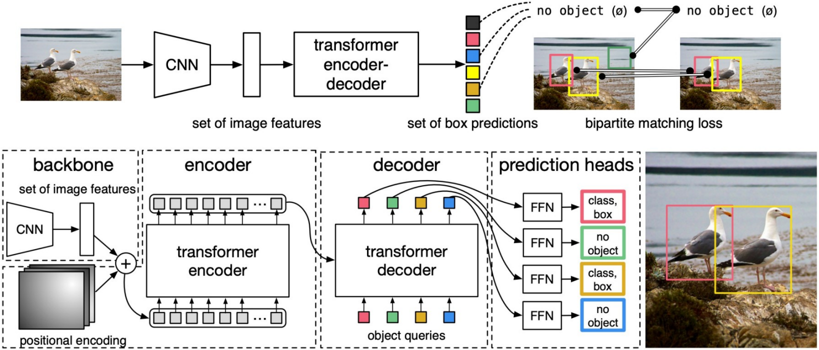
Liu et al, "Swin Transformer V2: Scaling up Capacity and Resolution", CVPR 2022

Improved MViT



Li et al, "Improved Multiscale Vision Transformers for Classification and Detection", arXiv 2021

- DETR (DEtection TRansformer, Carion et al., 2020) introduces a simple and unified object detection pipeline using Transformers.
- Transformers replace region proposal networks (RPNs) with attention-based object queries for end-to-end detection.
- DETR directly predicts a fixed set of bounding boxes and class labels, eliminating the need for anchors, hand-crafted box proposals, or non-maximum suppression (NMS).
- The approach simplifies the pipeline and enables global reasoning, though it may struggle with small objects and require higher compute.
- DETR uses bipartite matching (Hungarian algorithm) to uniquely match predicted boxes to ground truth boxes, and is trained end-to-end to regress box coordinates and classify objects.



5 6

[5] Carion et al., “End-to-End Object Detection with Transformers”, ECCV 2020 https://arxiv.org/abs/2005.12872

[6] https://docs.google.com/presentation/d/1X0BDhpJOa3IOf29DU8vAB4WJruyLugIX/edit?slide=id.p1#slide=id.p1

ConvNeXt: A Modern CNN

- ConvNeXt is a modern CNN architecture that incorporates insights from transformer models.
- It uses a hierarchical design similar to Swin Transformer, with a focus on simplicity and efficiency.
- Key features include:
  - Depthwise separable convolutions
  - Layer normalization
  - Global average pooling
- Achieves competitive performance on image classification tasks while maintaining the strengths of CNNs.
- ConvNeXt demonstrates that CNNs can still be effective with modern design principles, even in the era of transformers.

ConvNeXt: A Modern CNN (cont.)

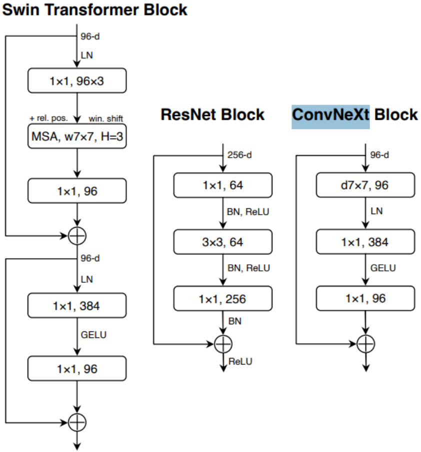


Figure 4. **Block designs** for a ResNet, a Swin Transformer, and a ConvNeXt. Swin Transformer's block is more sophisticated due to the presence of multiple specialized modules and two residual connections. For simplicity, we note the linear layers in Transformer MLP blocks also as "1×1 convs" since they are equivalent. [7]

[7] Liu et al., "A ConvNet for the 2020s", CVPR 2022 https://arxiv.org/abs/2201.03545

Limitations and Considerations

- **Data Requirements:**
  - ViTs require large-scale datasets or knowledge distillation to achieve strong performance.
  - They lack the strong inductive biases of CNNs.
- **Computational Cost:**
  - The self-attention mechanism has quadratic complexity with respect to input size, making ViTs computationally expensive.
  - Techniques such as windowed or masked attention help reduce this cost.
- **Inductive Bias vs. Flexibility:**
  - ViTs offer greater flexibility and generality.
  - They may be less sample-efficient compared to CNNs, which have built-in spatial priors.

Limitations and Considerations (cont.)

- **Interpretability:**
  - Attention maps can provide some interpretability.
  - They may also highlight irrelevant or noisy regions, limiting their usefulness.
- **Deployment Challenges:**
  - ViTs typically require more resources, making them less suitable for low-power or resource-constrained environments.
  - Research into lightweight variants, pruning, and quantization is ongoing to address this.

Limitations and Considerations (cont.)

- **Future Directions:**
  - Development of more efficient attention mechanisms to reduce computational demands.
  - Integration of CNN and transformer architectures to leverage the strengths of both.
  - Design of lightweight ViTs tailored for edge devices and real-time applications.

References

[1] Fei-Fei Li, Yunzhu Li, and Ruohan Gao. *Stanford CS231n: Deep Learning for Computer Vision*. http://cs231n.stanford.edu/index.html

[2] Assaf Shocher, Shai Bagon, Meirav Galun, and Tali Dekel. *WAIC DL4CV Deep Learning for Computer Vision: Fundamentals and Applications*. https://dl4cv.github.io/index.html

[3] Justin Johnson. *UMich EECS 498.008/598.008: Deep Learning for Computer Vision*. https://web.eecs.umich.edu/~justincj/teaching/eecs498/WI2022/

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, International Conference on Learning Representations (ICLR), 2021.

[5] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, *Training data-efficient image transformers & distillation through attention*, International Conference on Machine Learning (ICML), 2021.

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*, International Conference on Computer Vision (ICCV), 2021.

[7] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, S. Yan, *Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet*, International Conference on Computer Vision (ICCV), 2021.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, *End-to-End Object Detection with Transformers*, European Conference on Computer Vision (ECCV), 2020.

Credits

Dr. Prashant Aparajeya

Computer Vision Scientist — Director(AISimply Ltd)

p.aparajeya@aisimply.uk

This project benefited from external collaboration, and we acknowledge their contribution with gratitude.