# Transformers: Birds Eye View

D.A. Forsyth,

University of Illinois at Urbana Champaign

# An architecture

- Accepts
  - one or more streams of tokens
    - very often fixed length
- Produces
  - stream of tokens
  - OR single learned token

- Origins in NLP
- Apply to images by
  - tokenizing image
  - getting clever about decoding

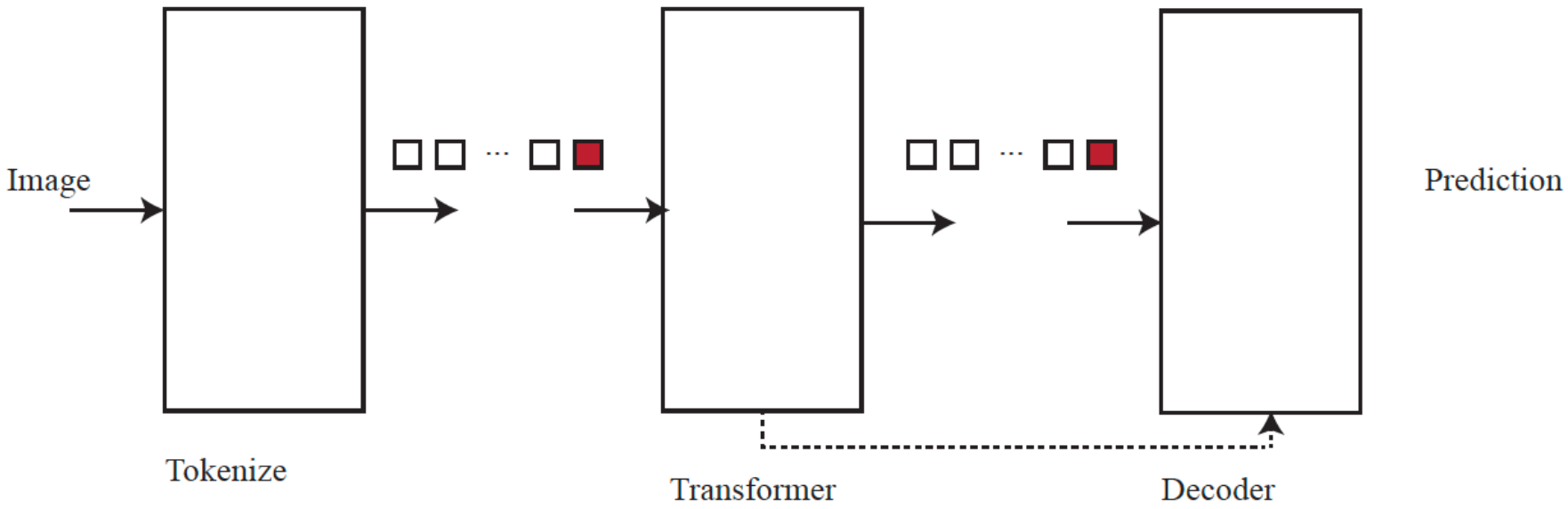
# Tokenizing an image

- Obtain feature representation of tokens
  - Convolutional encoder (waning)
  - Chop image into patches, project (more usual)

- One token per patch, one extra learnable token
  - the class token

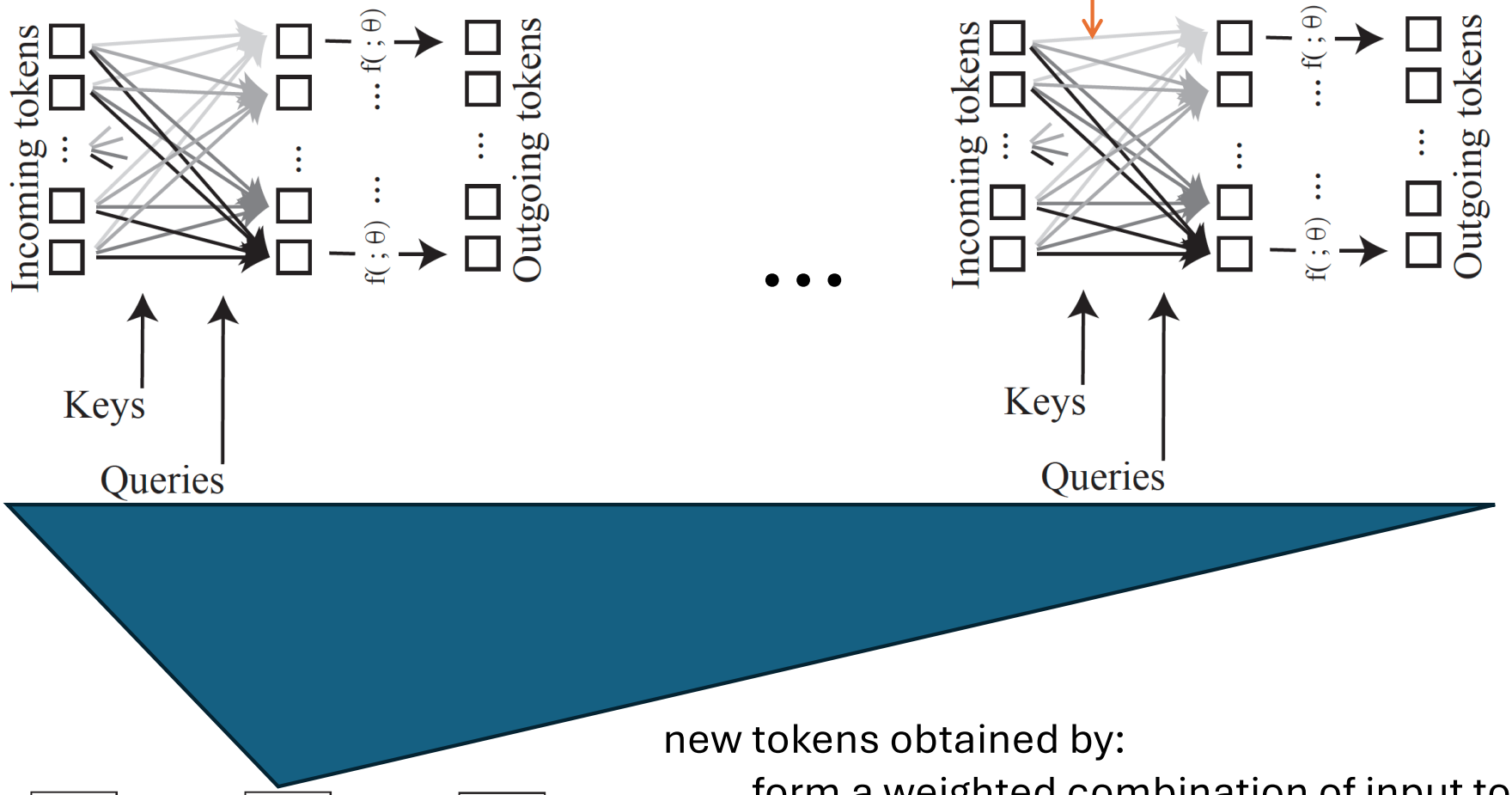
- Often (but not necessarily)
  - fixed size images into fixed numbers of tokens

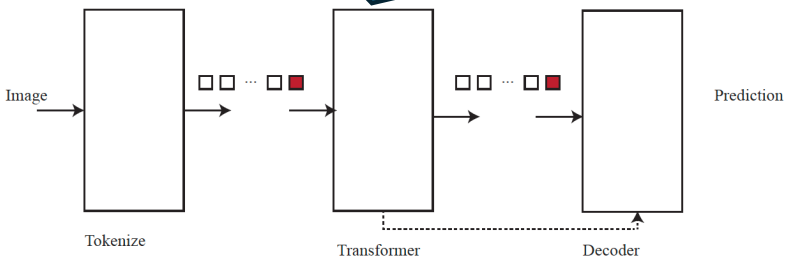
# General story



# General story, II

These weights are determined by attention – depend on keys and queries.




new tokens obtained by:

form a weighted combination of input tokens

weights determined by a similarity measure

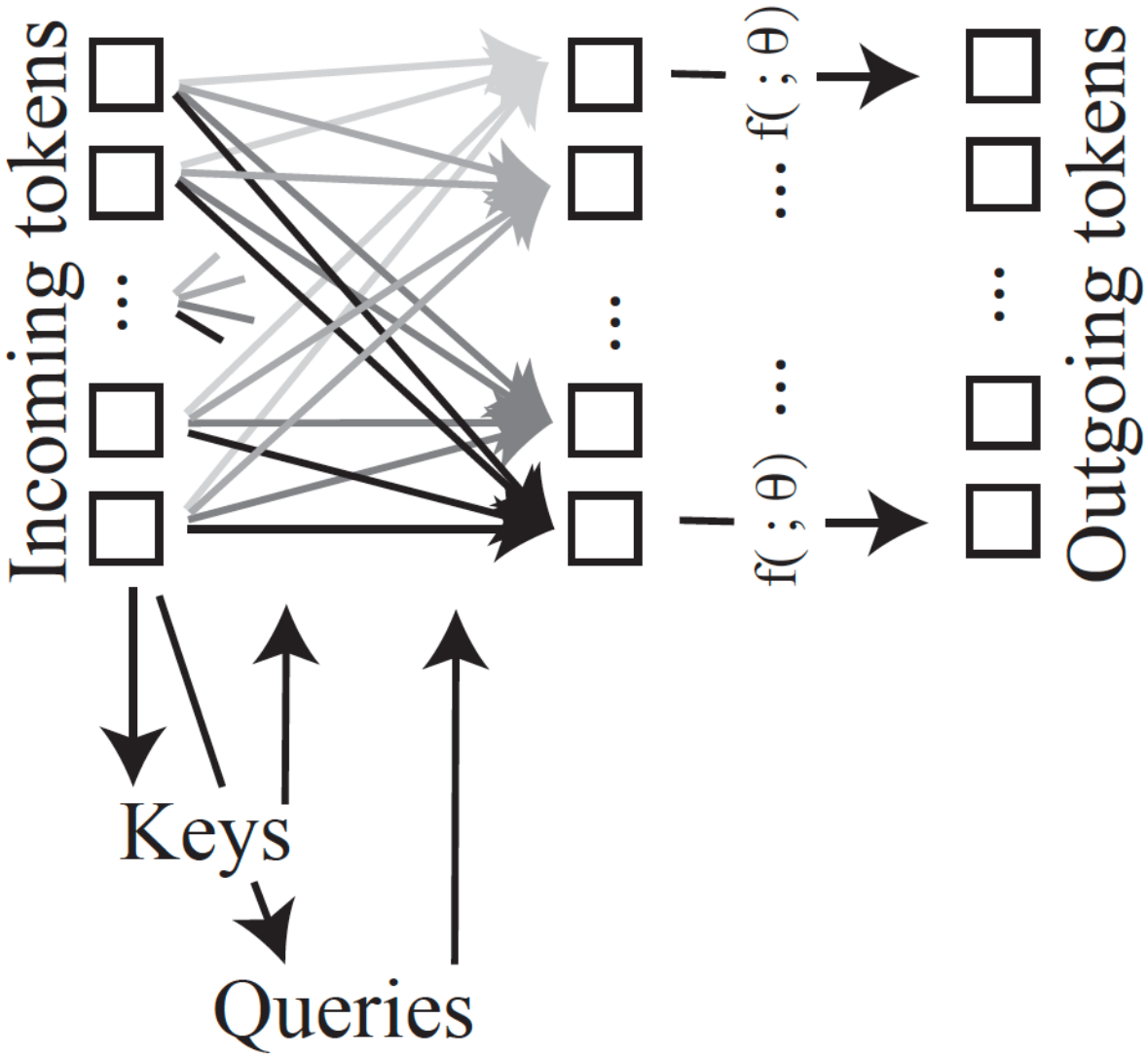
map each resulting token

report new tokens

# Self-attention

Keys and queries are a learnable function of tokens – typically, linear.

Important variants include:
Multi-headed self attention,
(more complex use of keys and queries)



# Properties of transformer w/ SA

- Can accept sequences of variable length
  - with some work
- Permutation invariant in the form described
  - obvious nuisance for images
    - procedure does not depend on where a patch is!
  - easily fixed
    - add to each token a positional encoding
      - learnable vector, different one for each location in stream
      - numerous interesting variations of what and where one provides

# Qualitative properties

- Every token communicates with every other
- Strong evidence of better encoders
- Inductive bias
  - Convolution, etc has inductive bias
    - mostly, pixels depend on nearby pixels
  - Transformers seem not to
    - or rather, it hasn't made a nuisance of itself
- Issue:
  - massive, expensive training demands
  - many/most vision transformers are fine-tuned from published weights

# Training effects

Benefit from large scale pretraining

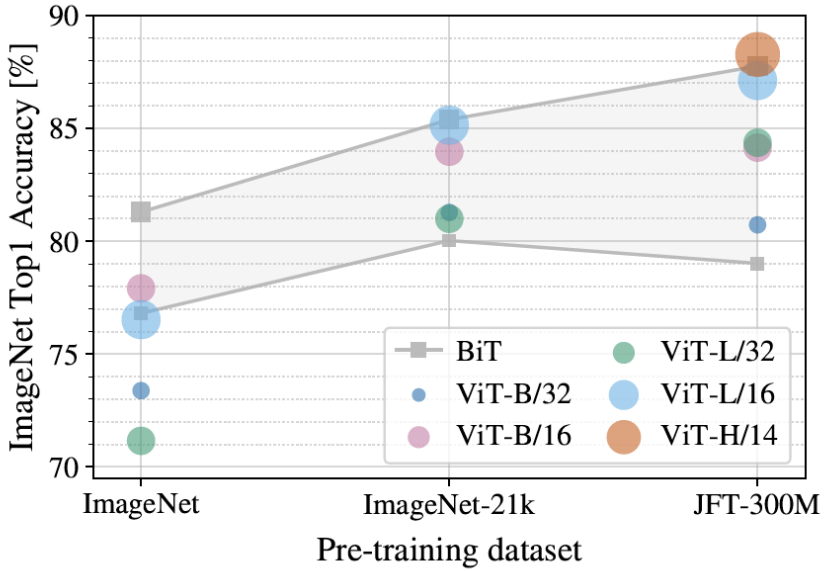



Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

Don't plateau at very large scale

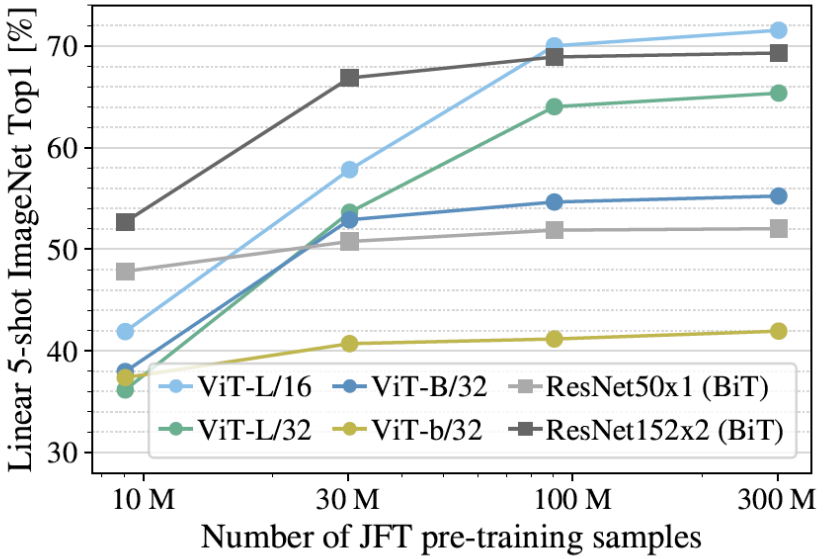



Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

*AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE, Dosovitsky et al 2021*

# Training effects

Training compute for given accuracy favors ViT's

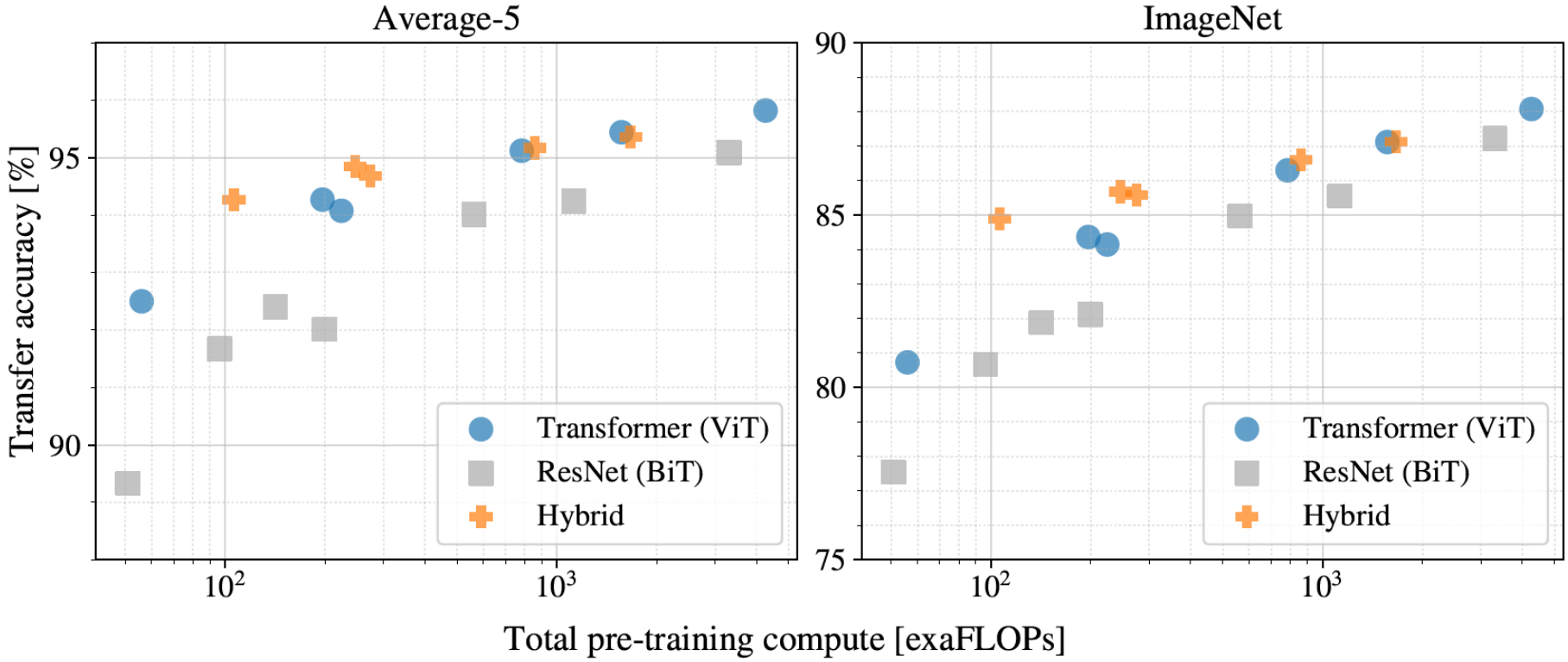


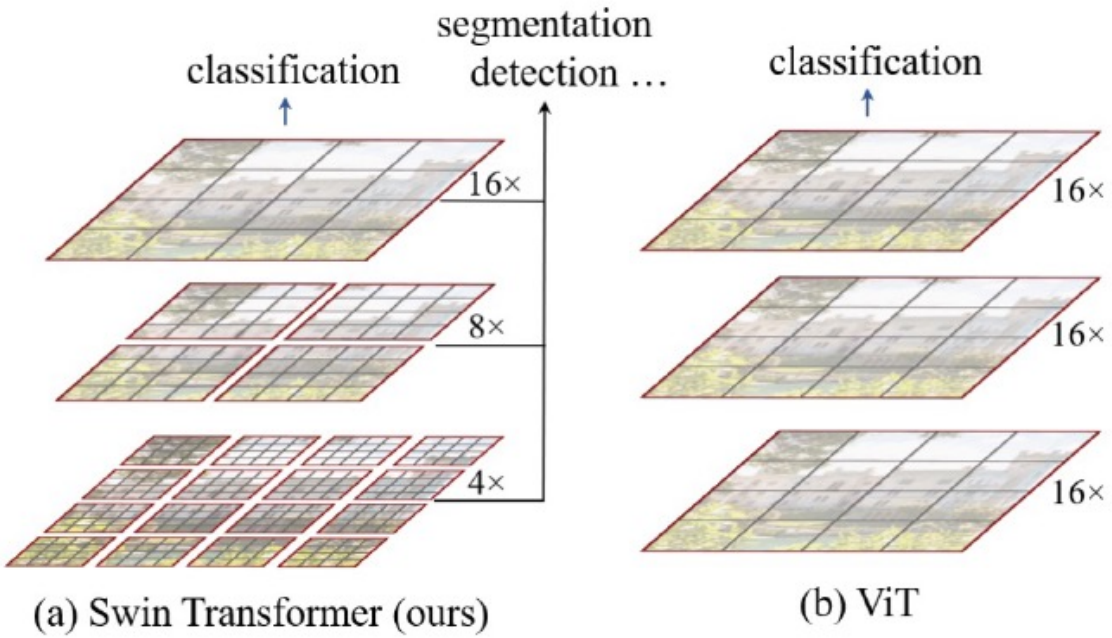

Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger models.

*AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE, Dosovitsky et al 2021*

# Variants

- How image is tokenized
  - very wide variation
- Where Q and K come from
- How to decode
  - very wide variation, following slides
- Training procedures
  - very wide variation, following slides

# Hierarchical transformer: Swin



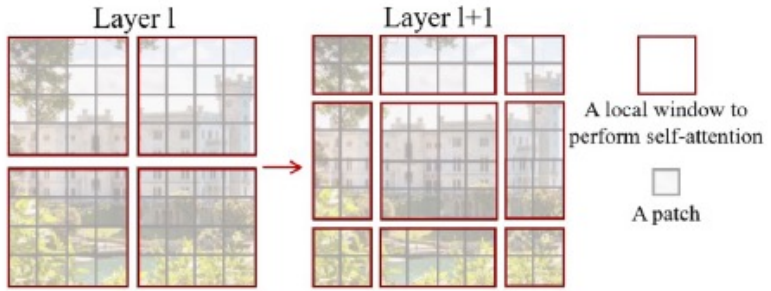
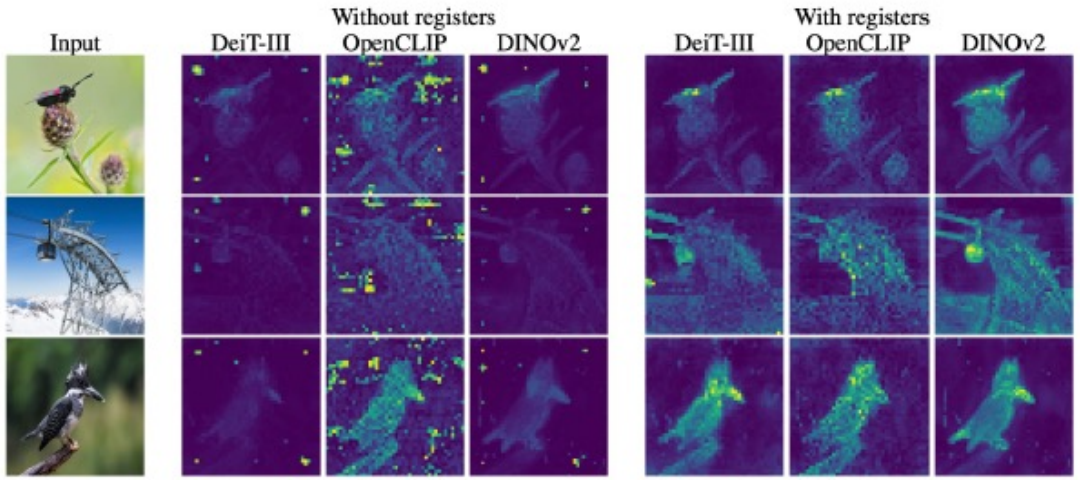



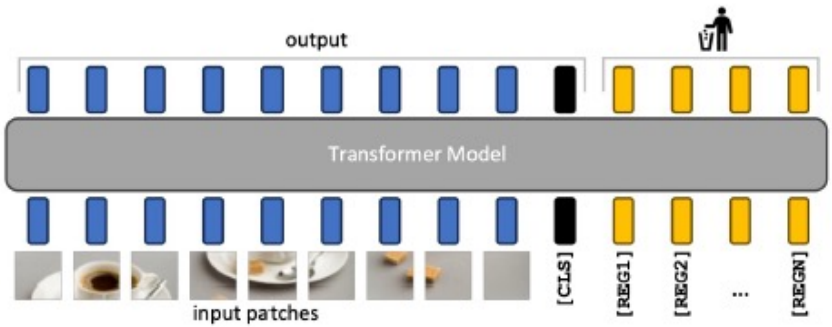

Figure 2. An illustration of the *shifted window* approach for computing self-attention in the proposed Swin Transformer architecture. In layer $l$ (left), a regular window partitioning scheme is adopted, and self-attention is computed within each window. In the next layer $l + 1$ (right), the window partitioning is shifted, resulting in new windows. The self-attention computation in the new windows crosses the boundaries of the previous windows in layer $l$, providing connections among them.

Z. Liu et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV 2021

# Registers

- "We make the following hypothesis: **large, sufficiently trained models learn to recognize redundant tokens, and to use them as places to store, process and retrieve global information**… While this behavior is not bad in itself, the fact that it happens inside the patch tokens is undesirable. Indeed, it leads the model to discard local patch information, possibly incurring decreased performance on dense prediction tasks. We therefore propose a simple fix to this issue: **we explicitly add new tokens to the sequence, that the model can learn to use as registers.**"




T. Darcet et al. Vision transformers need registers. ICLR 2024