

Image Classification

Image classification

Task

- Assign a single label to image

car



apple



Christmas tree



broccoli



Image classification

Task

- Input: Image
 - Fixed resolution
- Output: Class label
 - Cross entropy loss



Deep Network

class

Image Classification

Applications

- Visual search
- Testbed for network design
 - Basis of many vision tasks



Datasets

Datasets: CIFAR-10

- Low resolution: 32x32
- 10 classes
- 60,000 images
- Subset of MIT tiny images

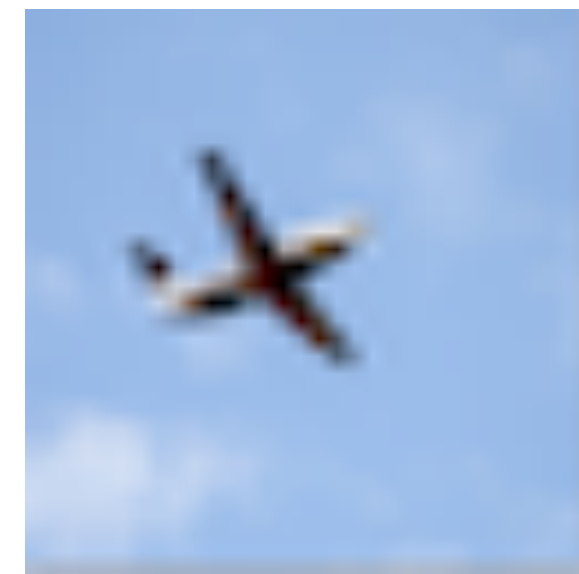
car



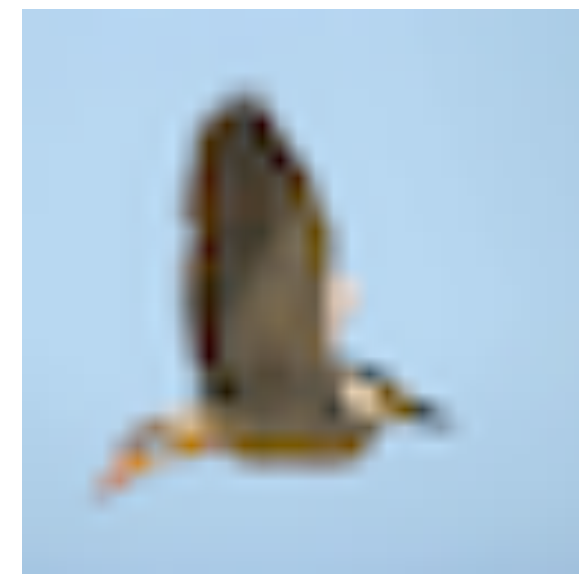
dog



airplane



bird



Datasets: ImageNet-1K

- 1000 classes
- 1.2M images
- Well balanced
- Well centered object



cheese



screwdriver



golf ball



harmonica



Datasets: ImageNet-21k

- 21,000 classes
- 14M images
- Fairly well balanced
- Well centered object



cheese



screwdriver



golf ball



harmonica



Datasets: Yahoo Flickr 100M

- >10K classes / tags
- 100M images
- noisy
- unbalanced



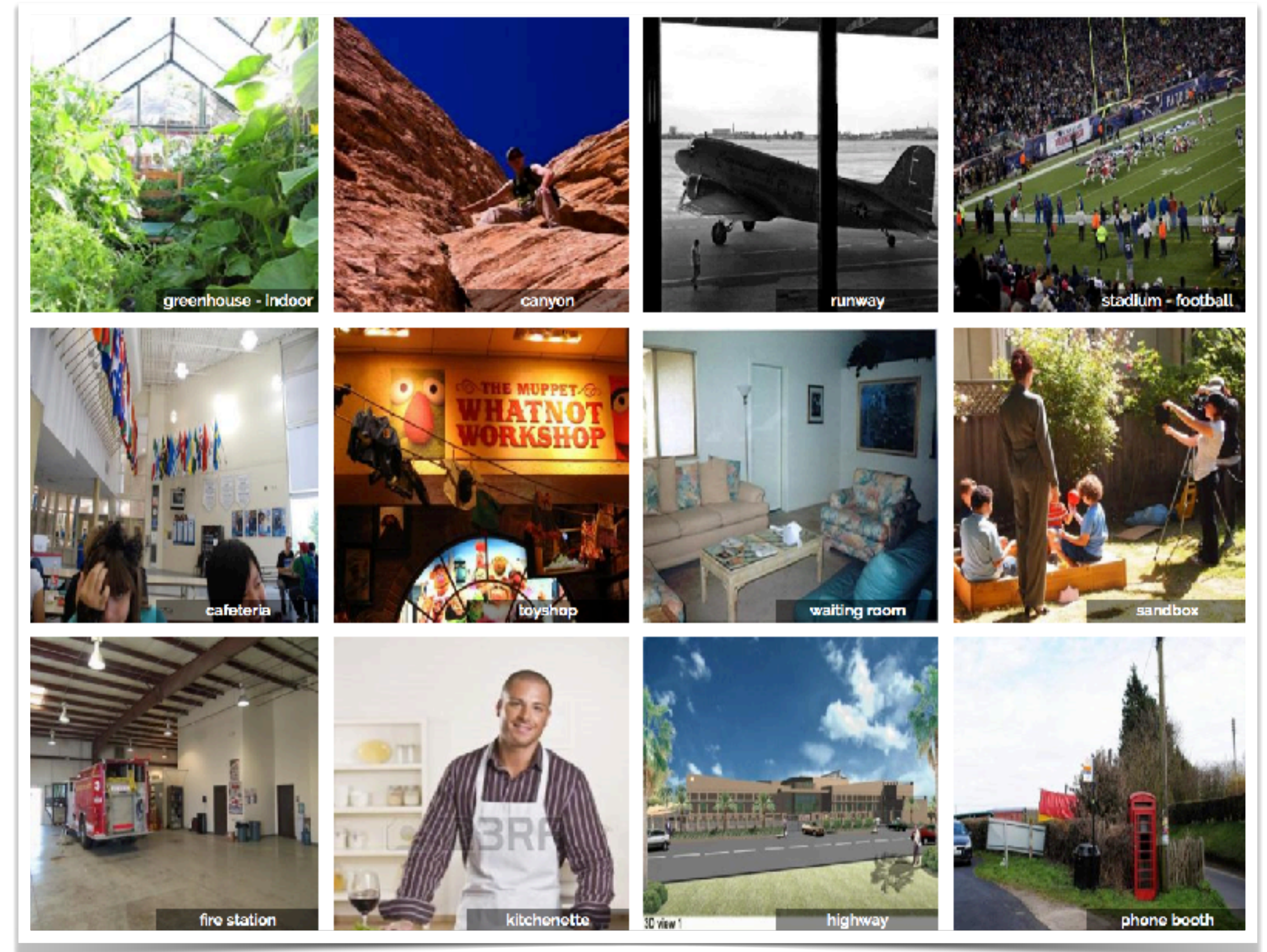
Datasets: Open Images

- 20,000 classes
- 9M images
- Image and object level annotations
- Creative commons



Datasets: MIT Places

- Scene categorization
- 400+ classes
- 10M images

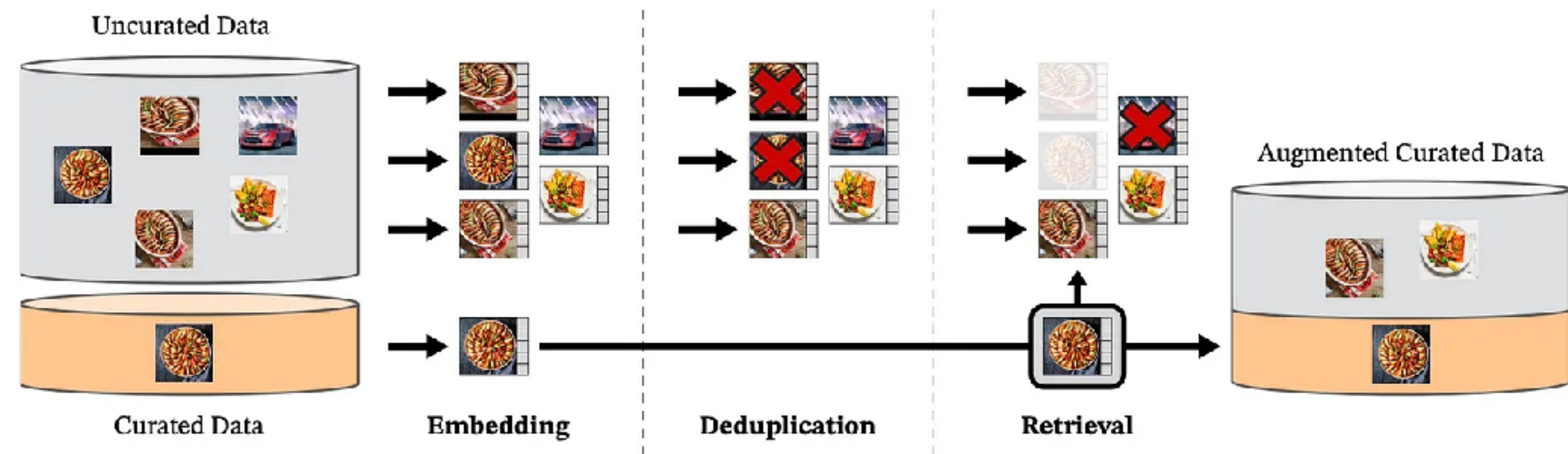


Datasets: JFT-3B

- Closed dataset (Google only)
- 30k classes (hierarchical)
- 3 billion images
 - Semi-automatically labeled



Datasets: LVD-142M

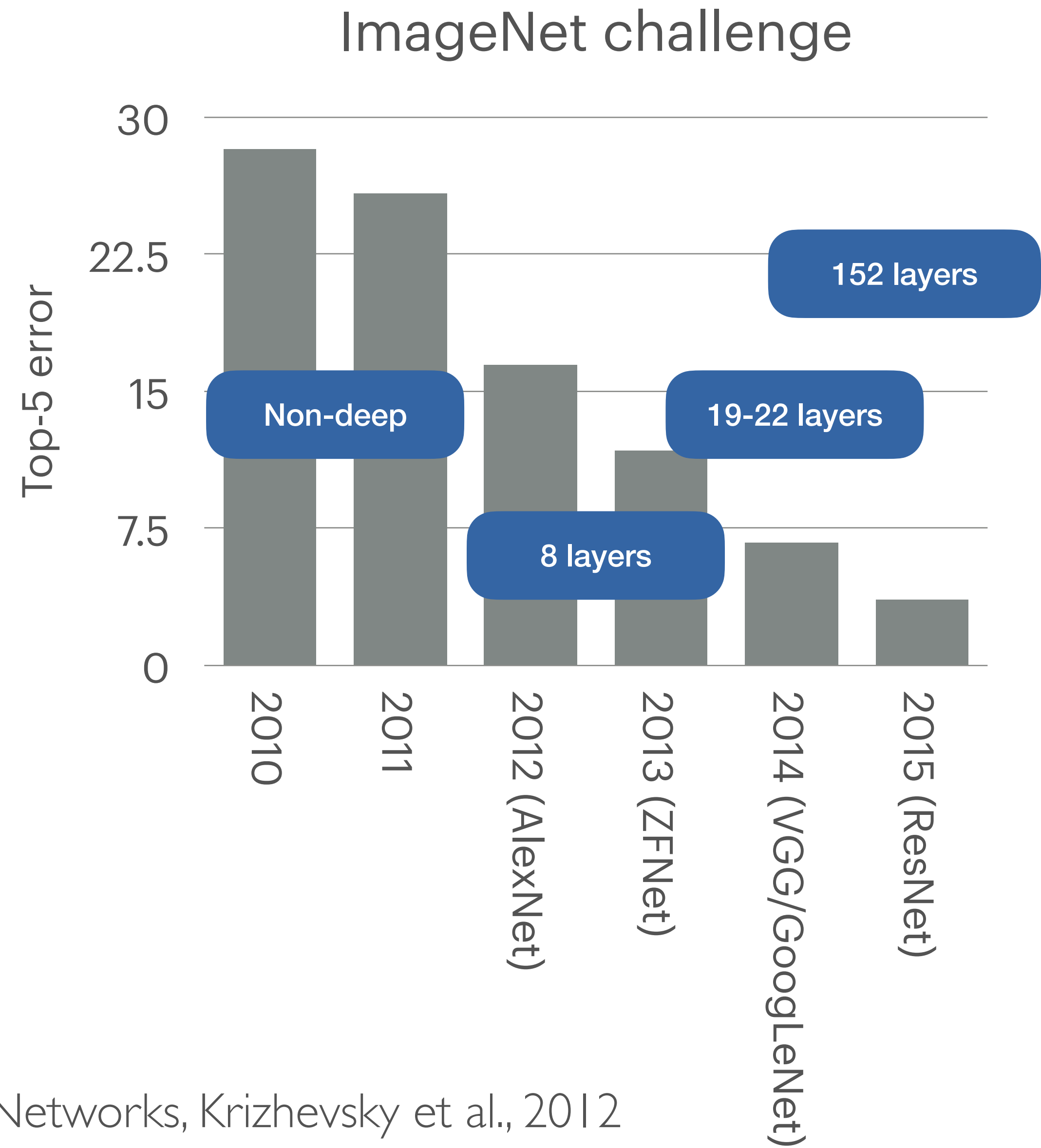


- Closed dataset (Meta only)
- Large uncurated data
 - Deduplicate
- ImageNet (or similar) dataset
 - Find $N=4$ similar uncurated images for each dataset image

Models

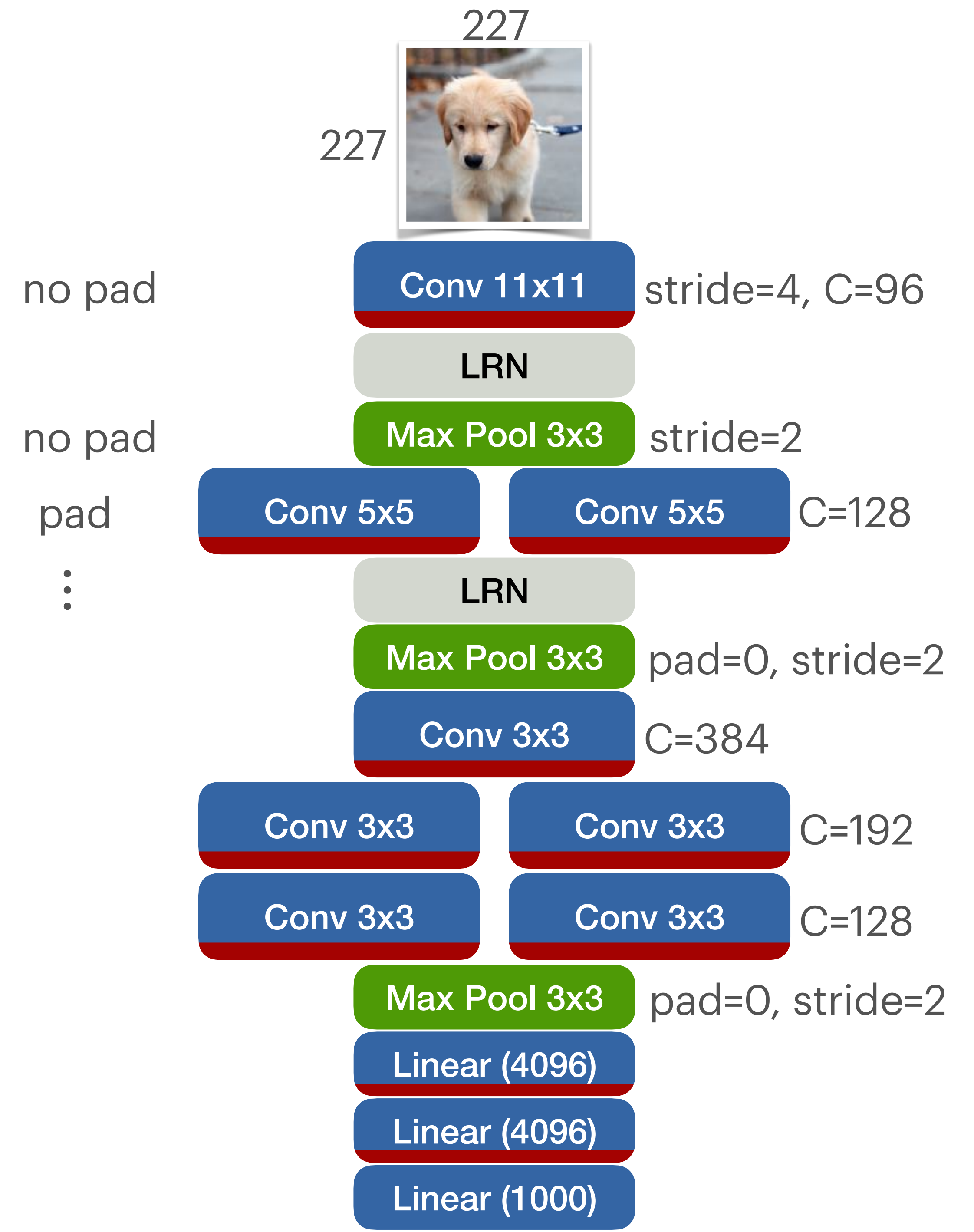
AlexNet

- Won the ImageNet competition 2012
- Start of deep learning revolution in computer vision



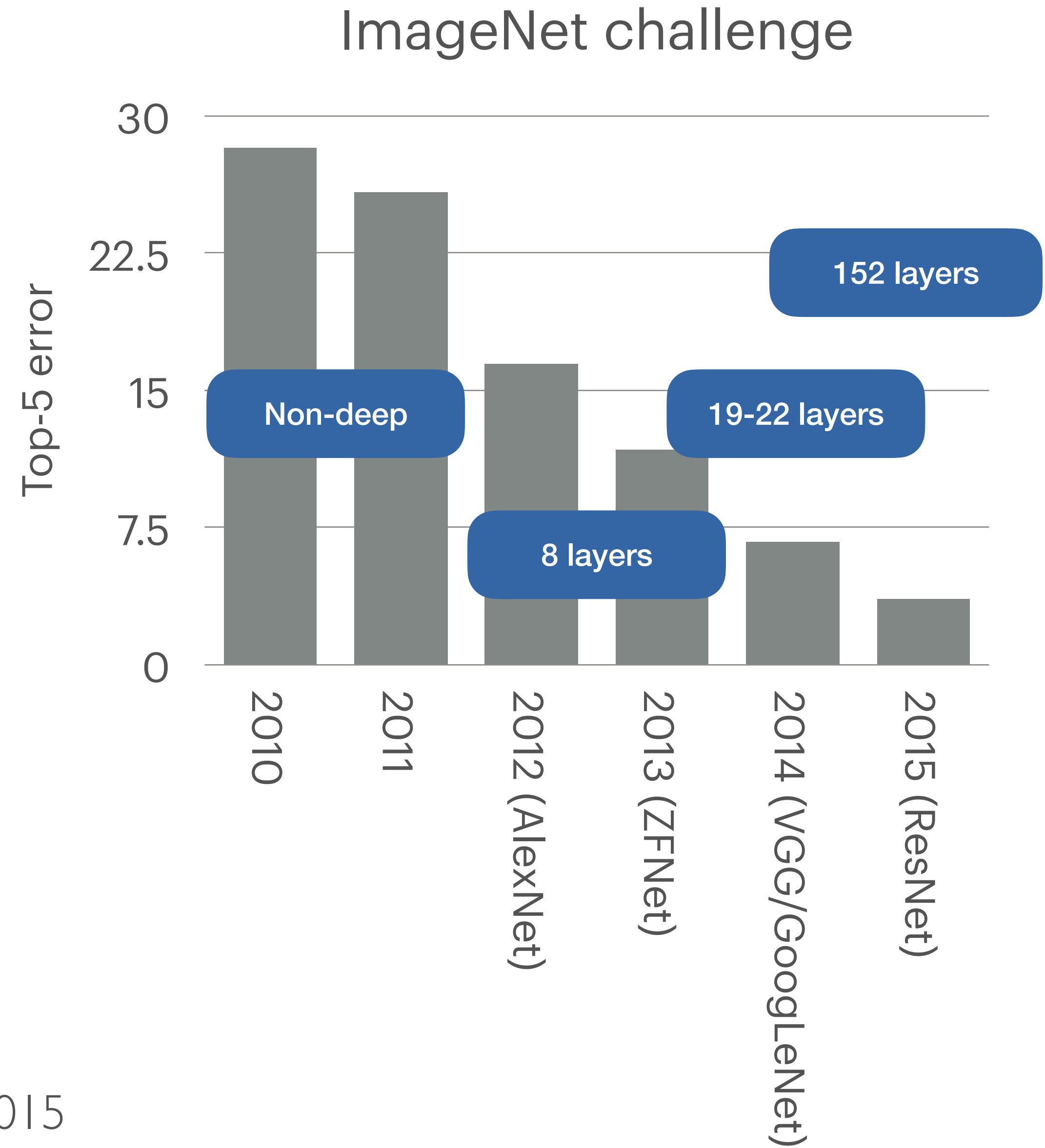
AlexNet

- First Deep Network to outperform non-deep vision systems
- Kicked off Deep Learning revolution
- Trained on two GPUs



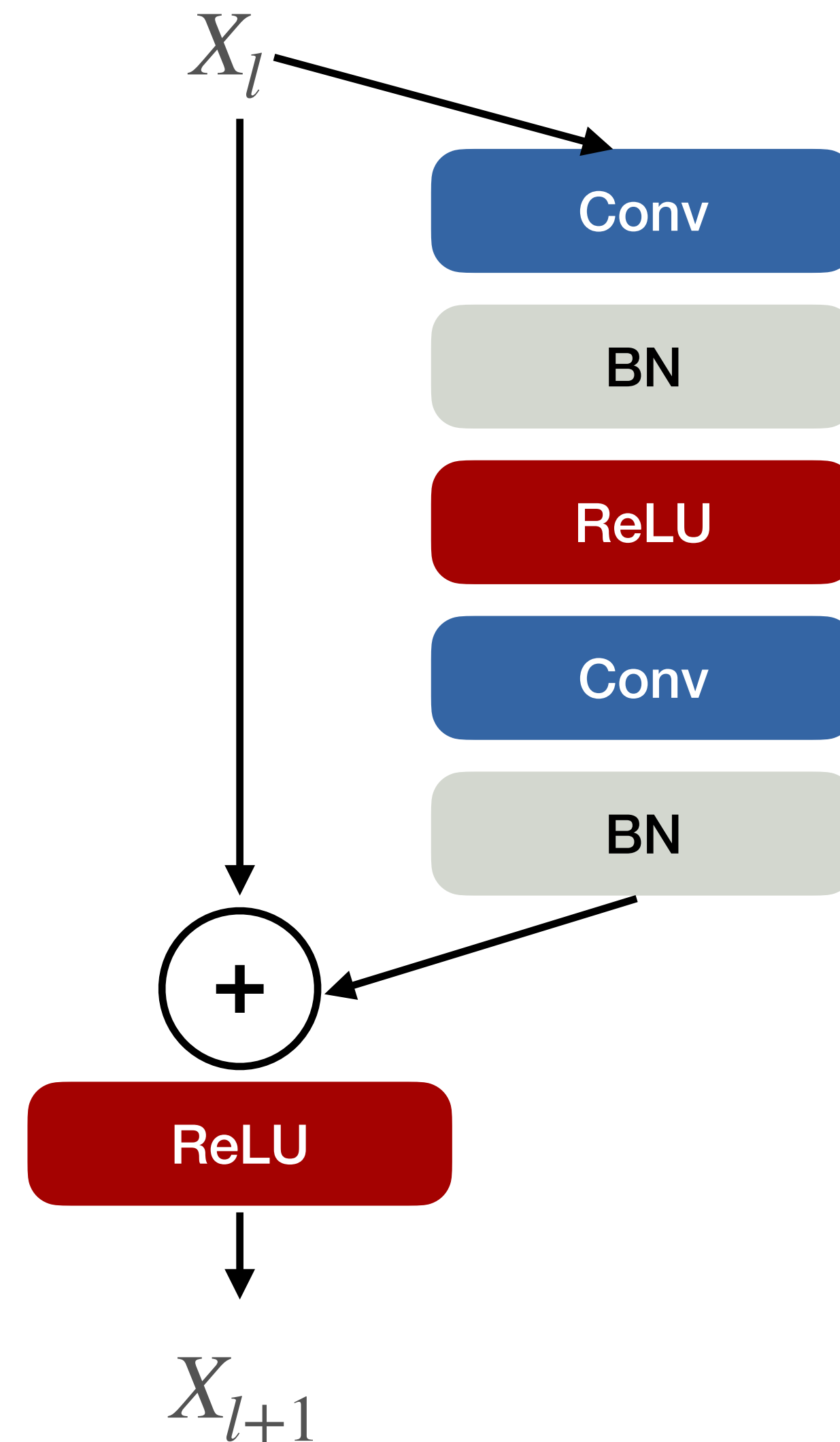
ResNet

- Won the ImageNet competition 2015
- Led to end of ImageNet competition



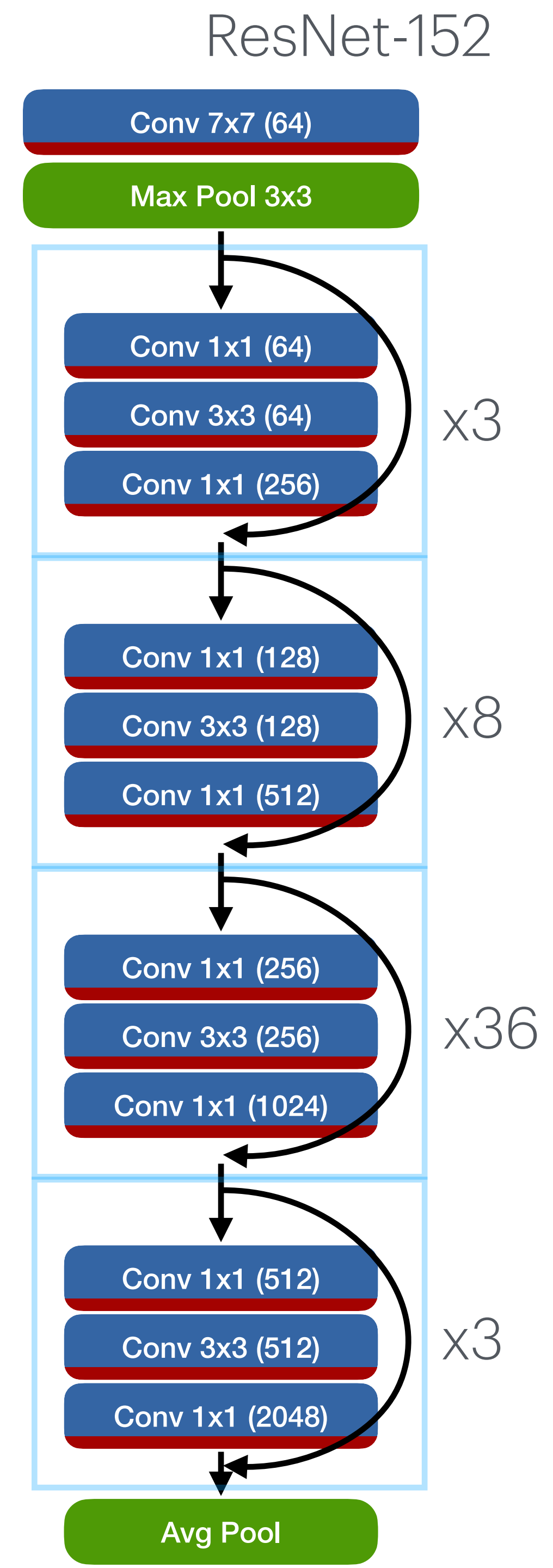
Residual blocks

- Add shortcut connections for gradients
 - Identity
 - Strided 1x1 convolution



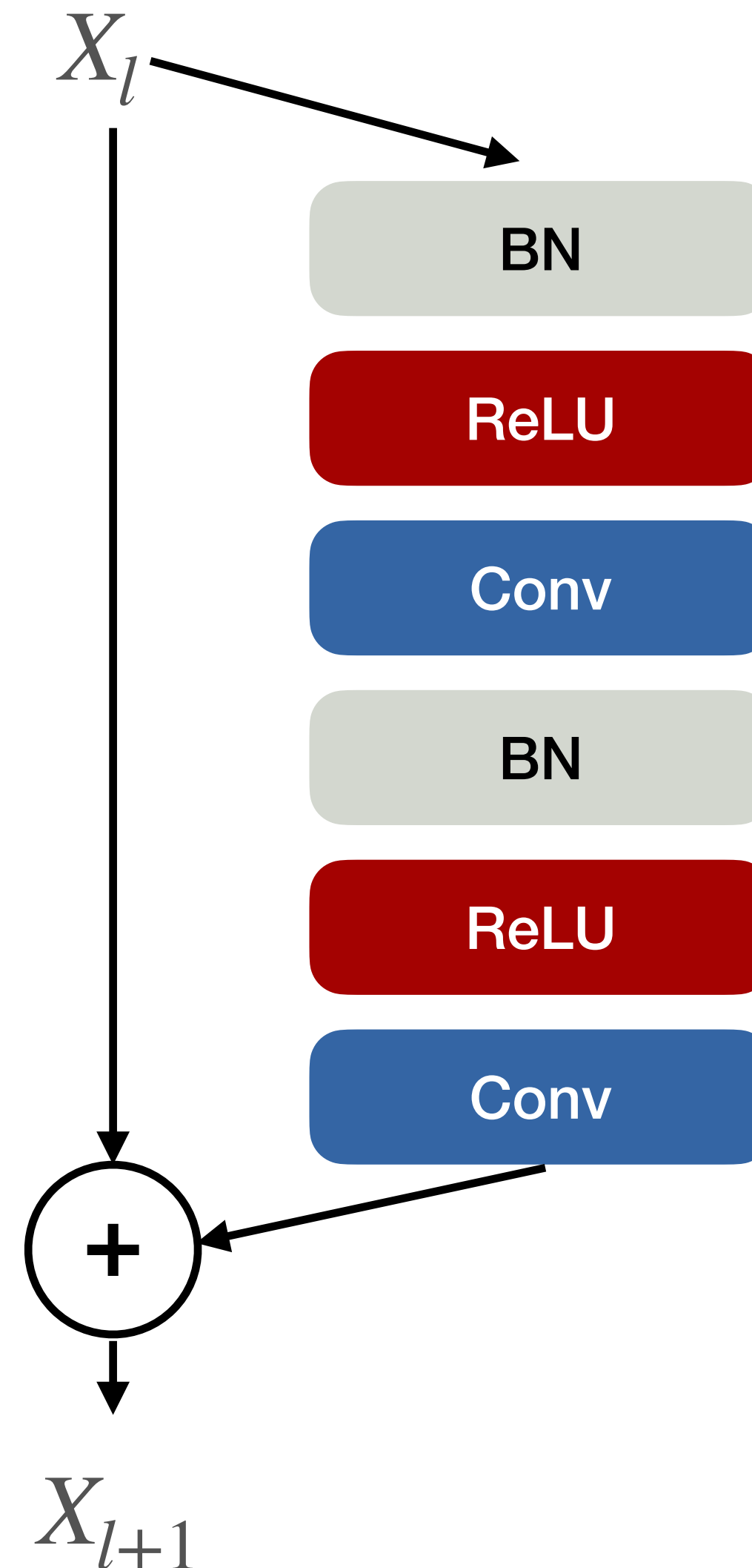
ResNet

- Multiple variants
 - ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152, ResNet-1001



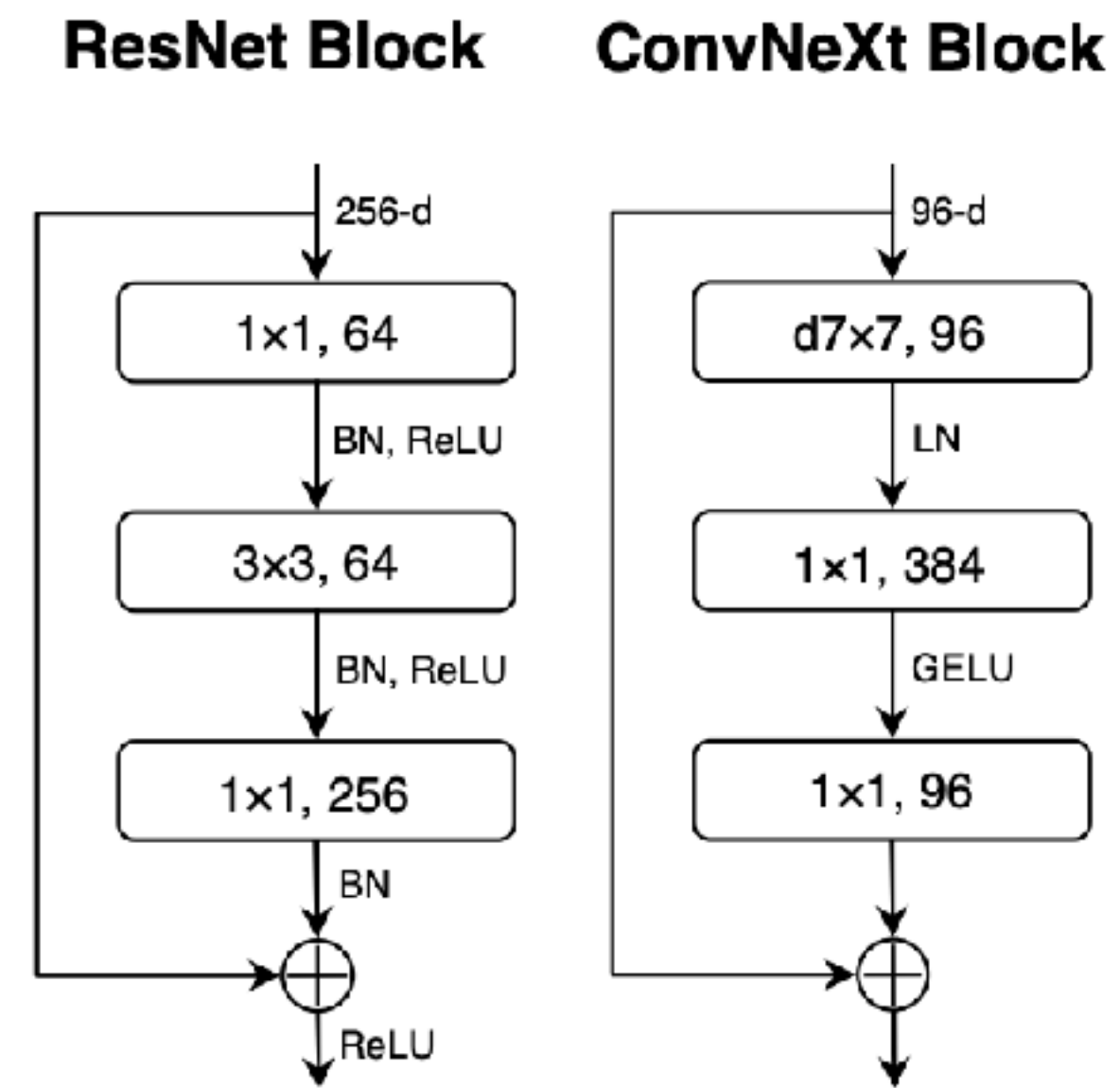
Pre-activation block

- Pure identity connections
 - Allows for deeper networks
 - Trains better



ConvNext

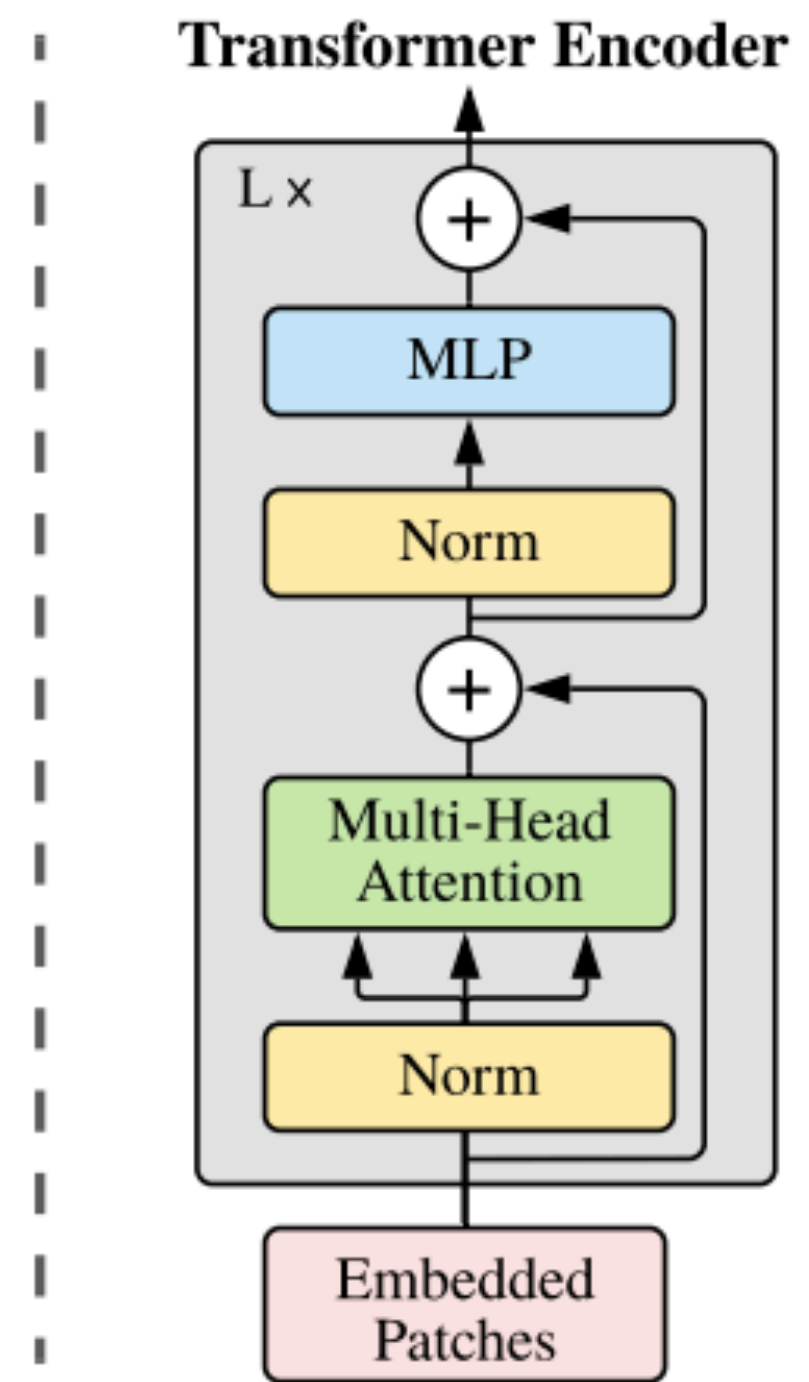
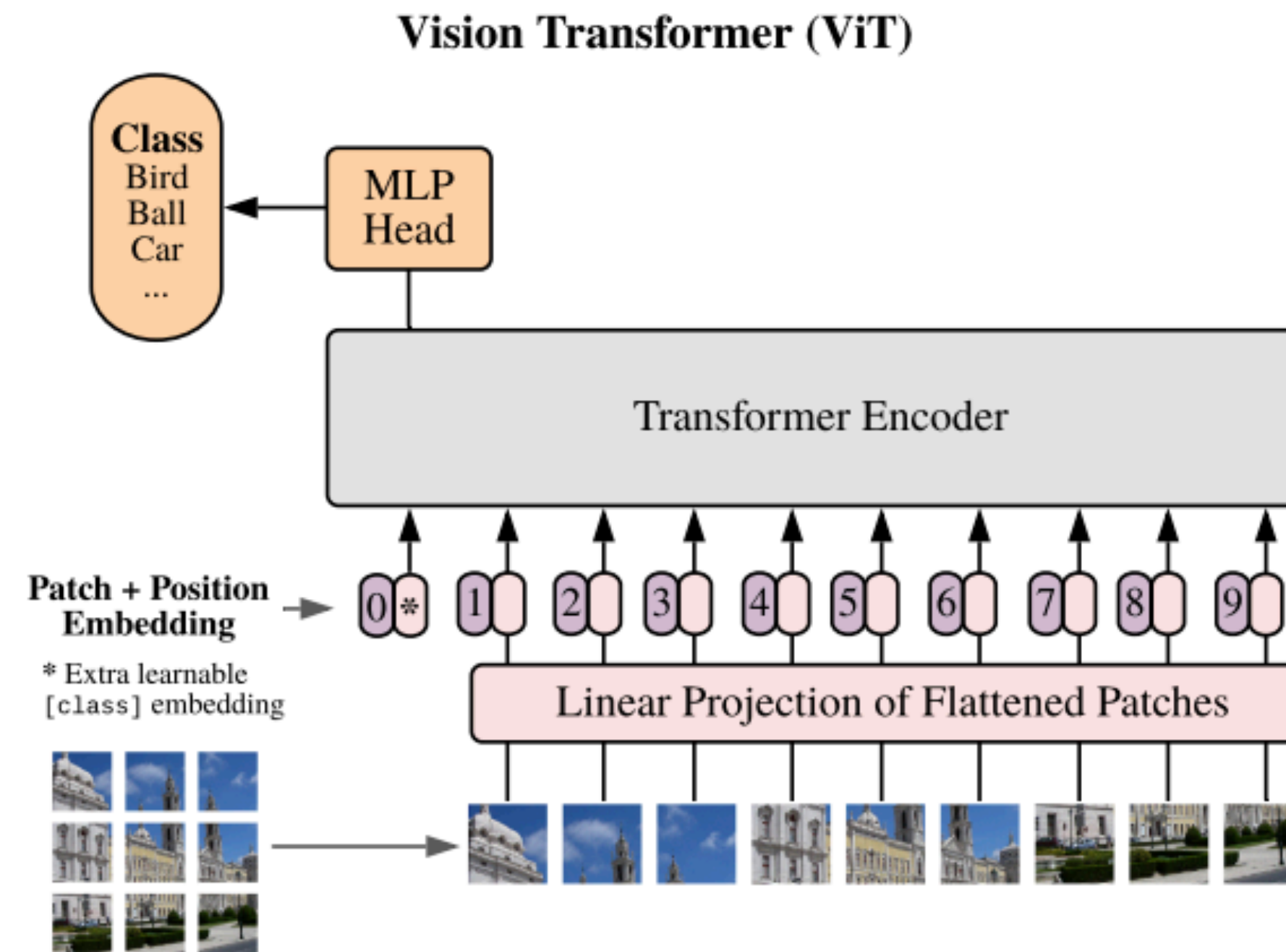
- Change the resnet block
 - Larger kernel size (depth-wise conv)
 - Followed by 1x1 conv
- Inspired by transformer



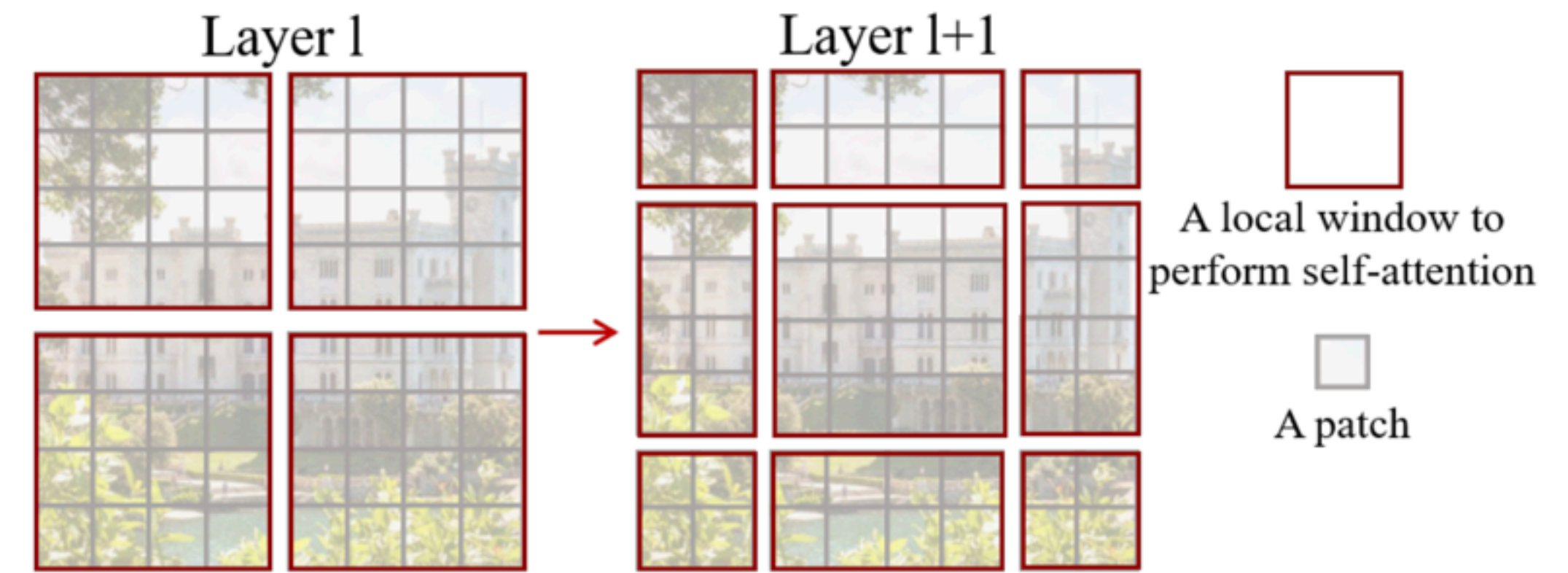
Vision Transformer

ViT

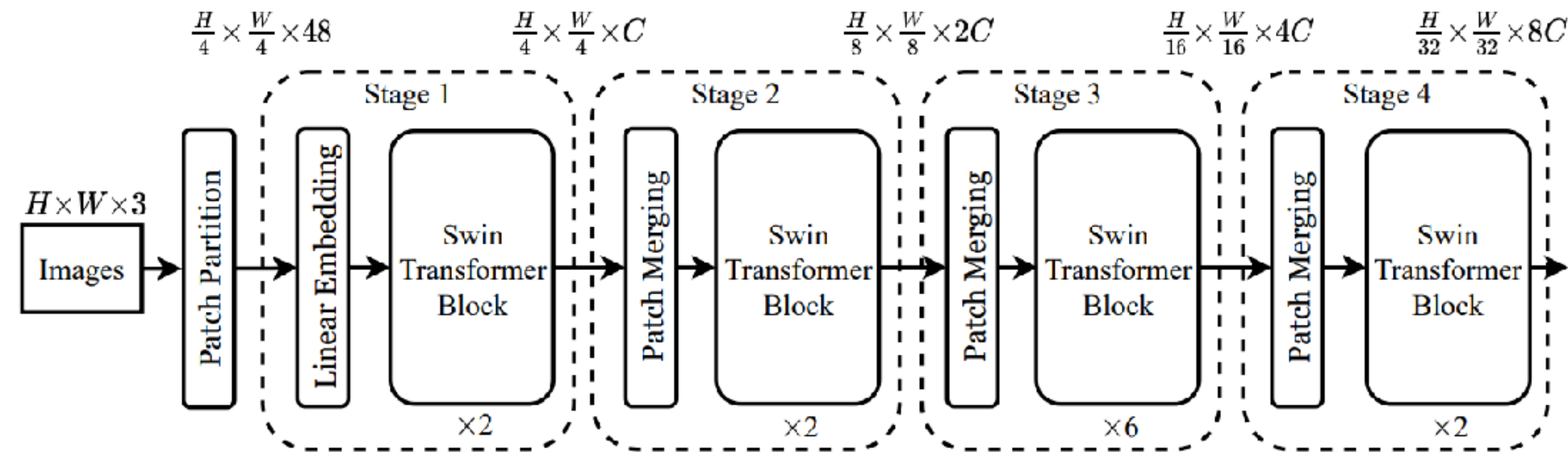
- Chop image into patches
- Feed them into transformer encoder
 - Bidirectional transformer
- Use classification token



SWIN

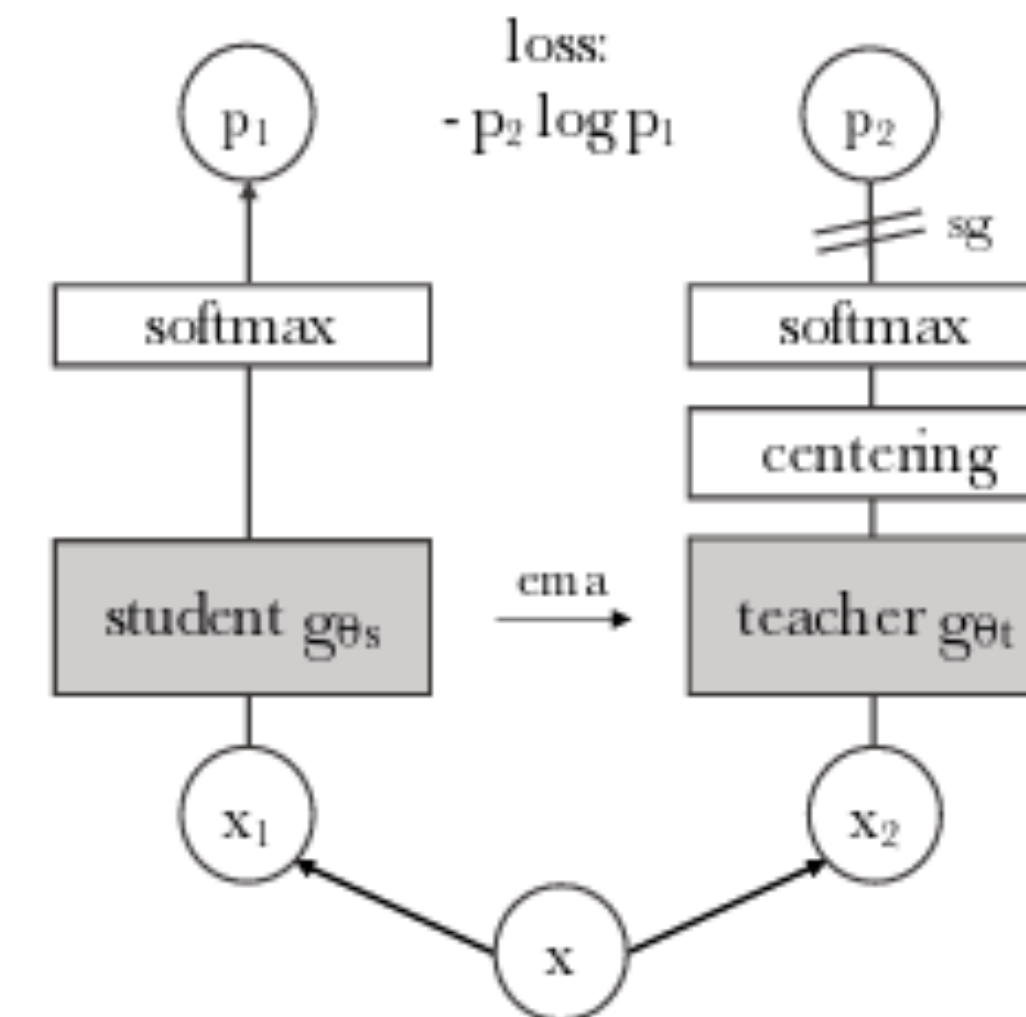
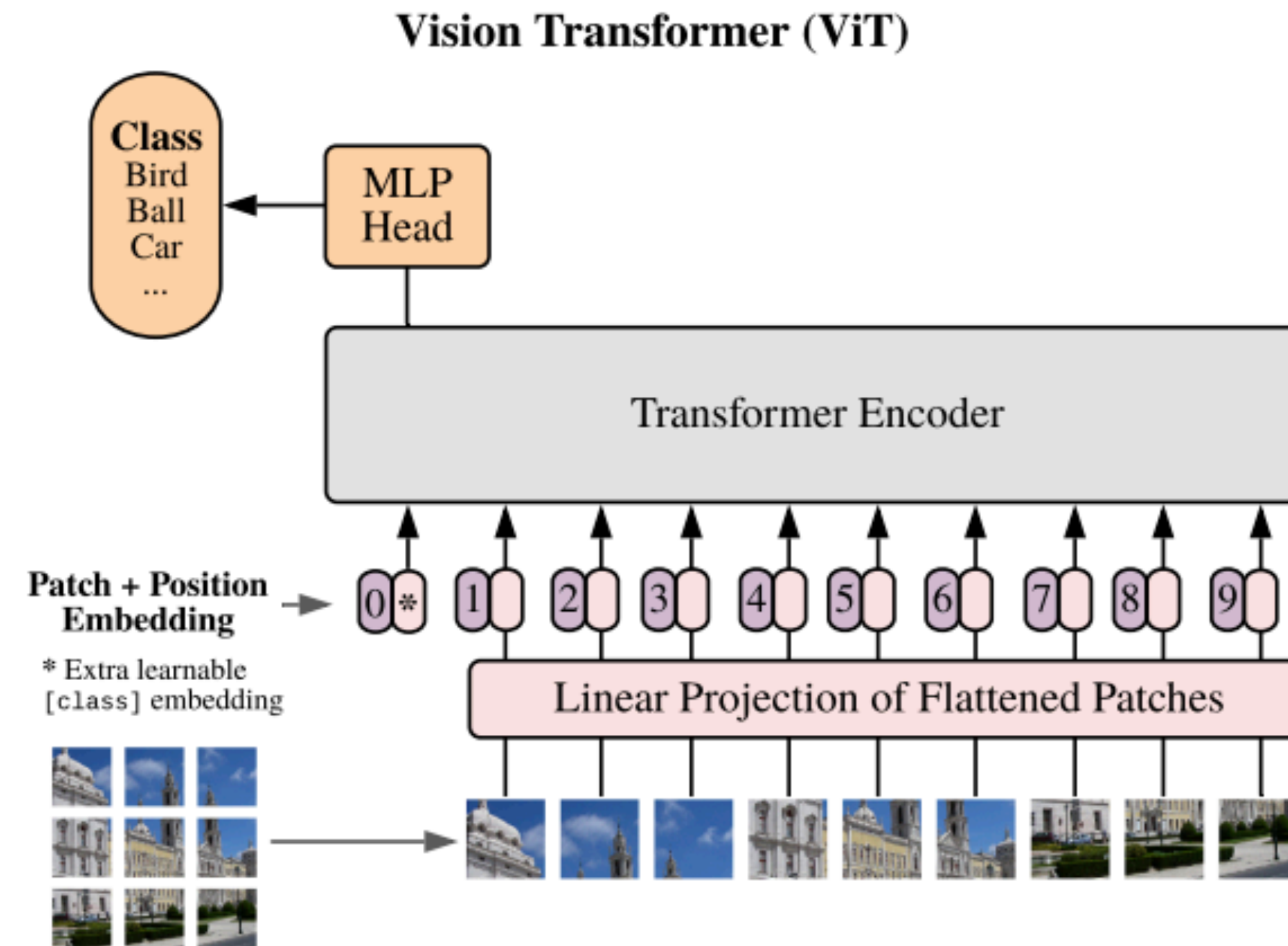


- ViT, but with windowed attention
- Divide image into non-overlapping windows
 - Attention only within window
- Shift window between layers



Dino v2

- ViT pre-trained on LVD-142M
- EMA-loss
 - Student: cls token
 - Teacher: Moving average of model
- iBOT loss
 - Student: Mask input tokens
 - Teacher: Unmasked model
- Higher-resolution end of training



Dino v2

- Result
 - Good general purpose network
 - Better than classification pretrained

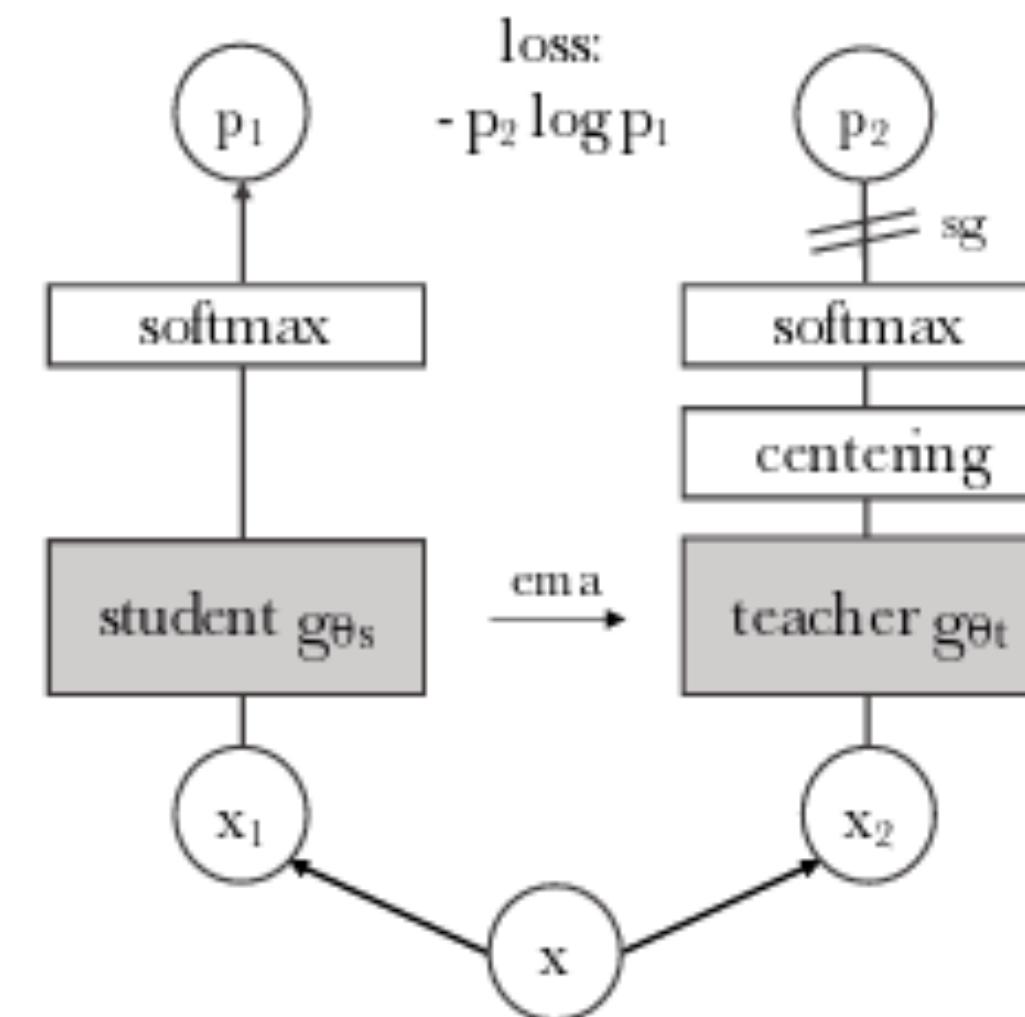
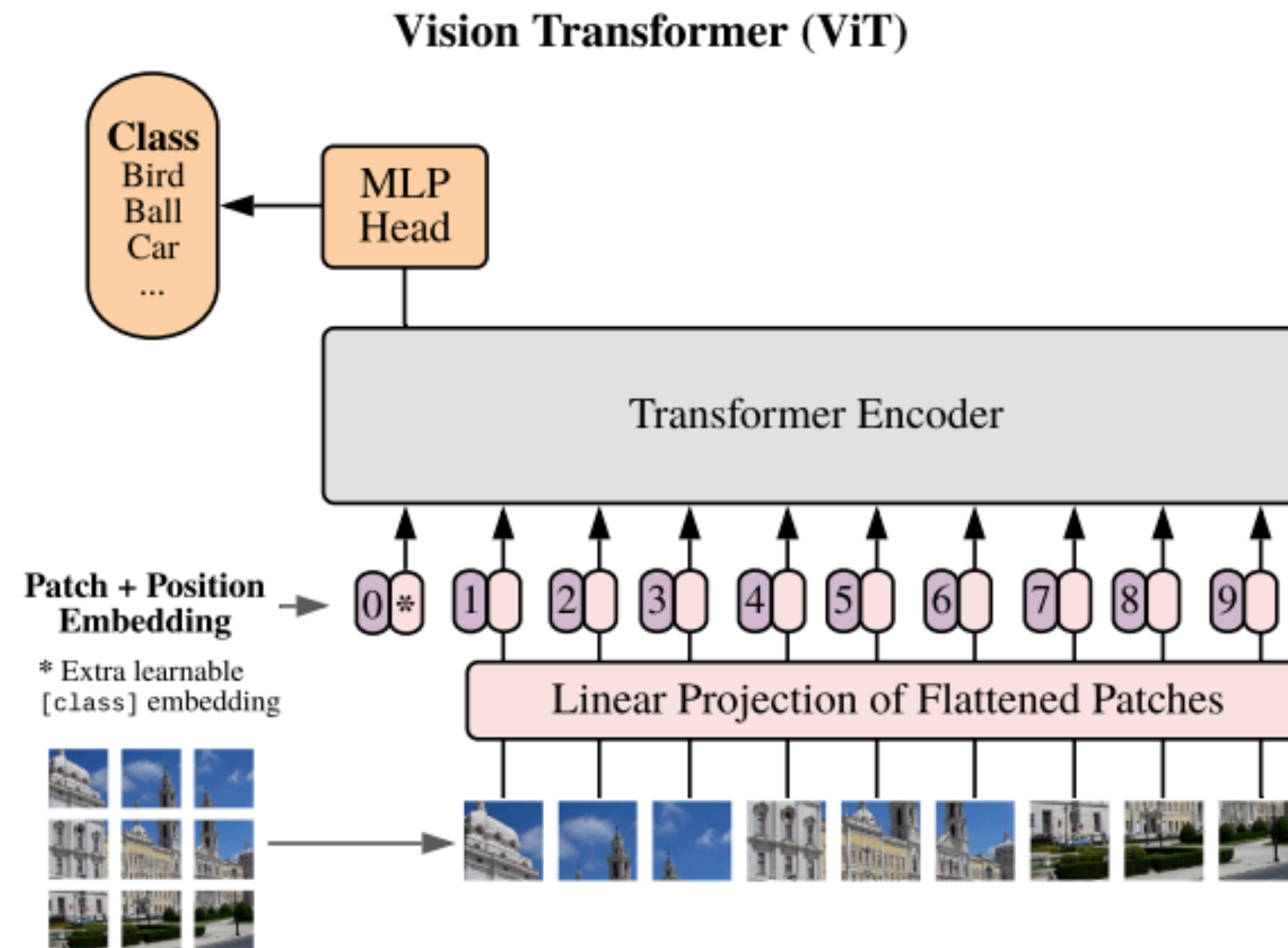


Image Classification

Trend 1

- **Architectures are very stable**
 - AlexNet lasted 1 year
 - ResNet lasted 4 years
 - ViTs lasted 5+ years
- Unclear we need new architectures for classification

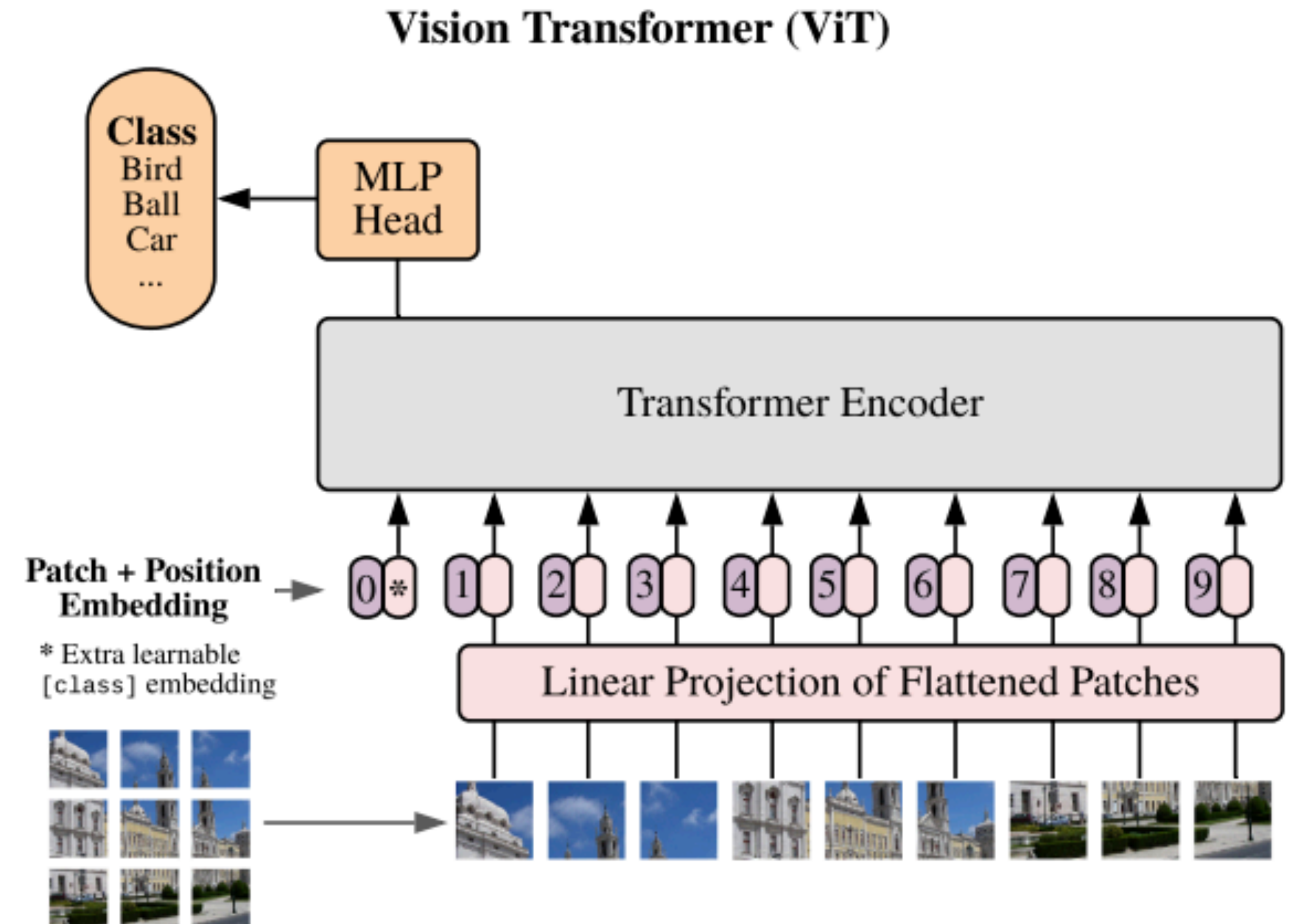


Image Classification

Trend 2

- **Classification label loosing value**
 - Dino v2: Unsupervised learning
 - Image Captioning data more popular nowadays
 - More data
 - Better supervision

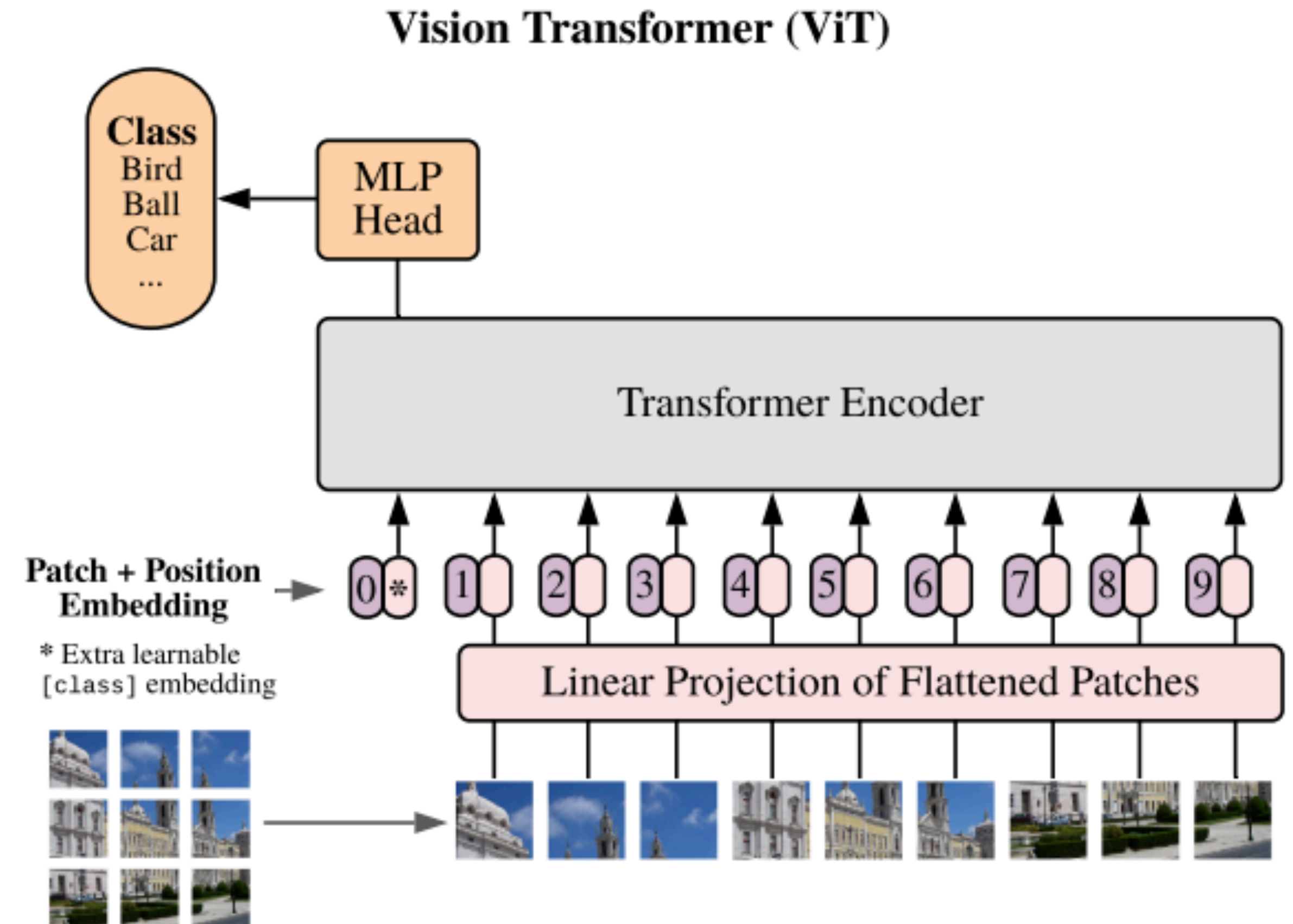
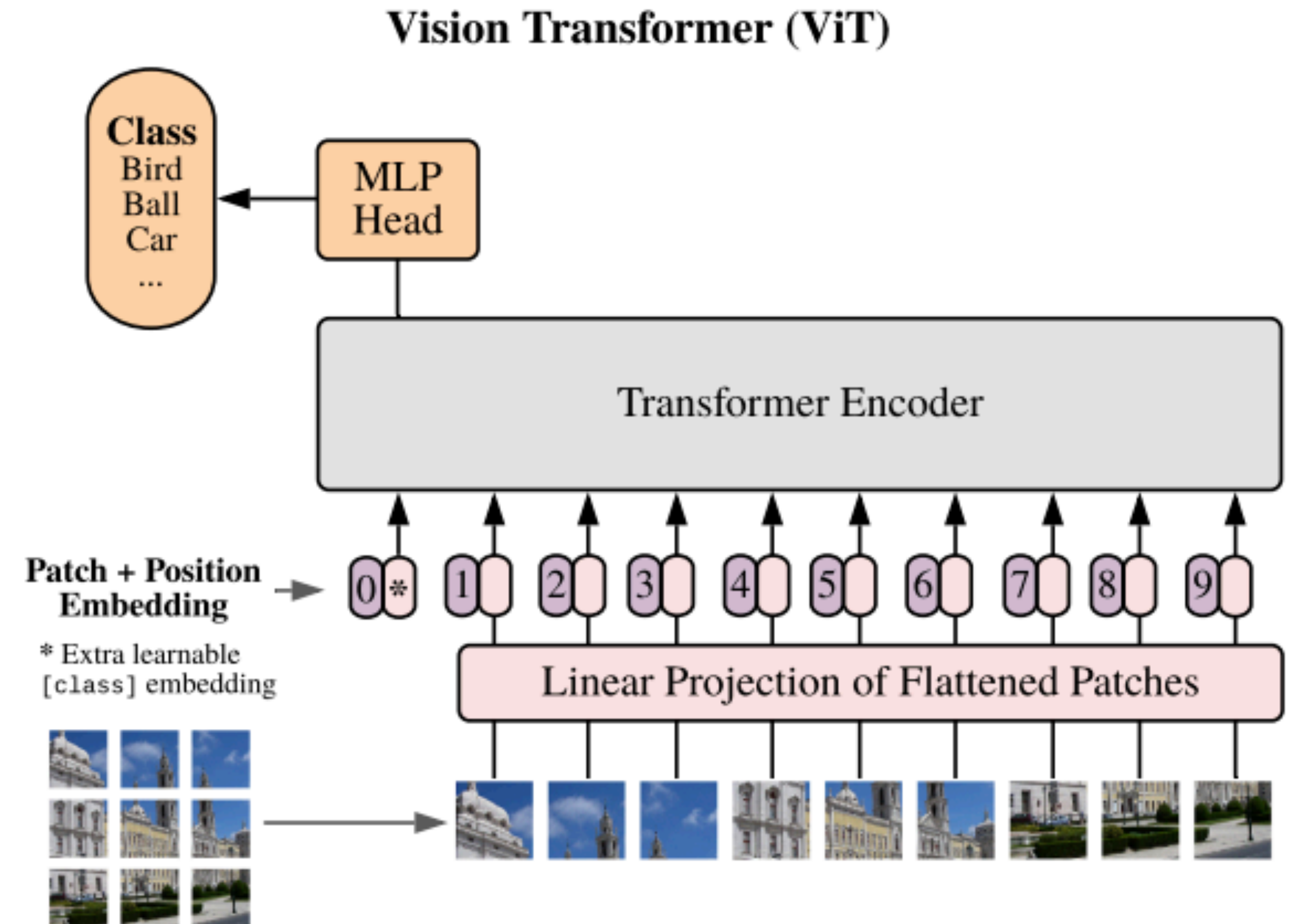


Image Classification

Trend 3

- **Zero-shot evaluations increase**
 - More and more models use classification just for evaluation
 - Zero-shot: without ever seeing the labels
 - Frozen-encoder: training a linear layer



References

- [1] Learning Multiple Layers of Features from Tiny Images, Krizhevsky & Hinton, Technical report, 2009.
- [2] 80 million tiny images: A large data set for nonparametric object and scene recognition, Torralba et al., PAMI 2008
- [3] ImageNet: A large-scale hierarchical image database, Deng et al., CVPR 2009
- [4] YFCC100M: The new data in multimedia research, Thomee et al., Communications of the ACM. 2016
- [5] The Open Images dataset v4: Unified image classification, object detection, and visual relationship detection at scale, Kuznetsova et al., IJCV 2020
- [6] Revisiting unreasonable effectiveness of data in deep learning era, Sun et al., ICCV 2017
- [7] Scaling vision transformers, Zhai et al., CVPR 2022
- [8] DINOv2: Learning Robust Visual Features without Supervision, Oquab et al., 2023
- [9] ImageNet Classification with Deep Convolutional Neural Networks, Krizhevsky et al., 2012
- [10] Deep Residual Learning for Image Recognition, He et al., 2015
- [11] Identity Mappings in Deep Residual Networks, He et al., 2016
- [12] A ConvNet for the 2020s, Liu et al., 2022
- [13] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, Dosovitskiy 2020
- [14] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, Liu 2021