UCF Center for Research in Computer Vision

CAP6412
Advanced Computer Vision

Mubarak Shah

shah@crcv.ucf.edu

Lecture-4_VIT_DeIT_SWIN

UCF Center for Research in Computer Vision

Contents

- VIT (Vision Transformer)
- DeIT (Data Efficient Image Transformer)
- SWIN (Shifted Window Transformer)

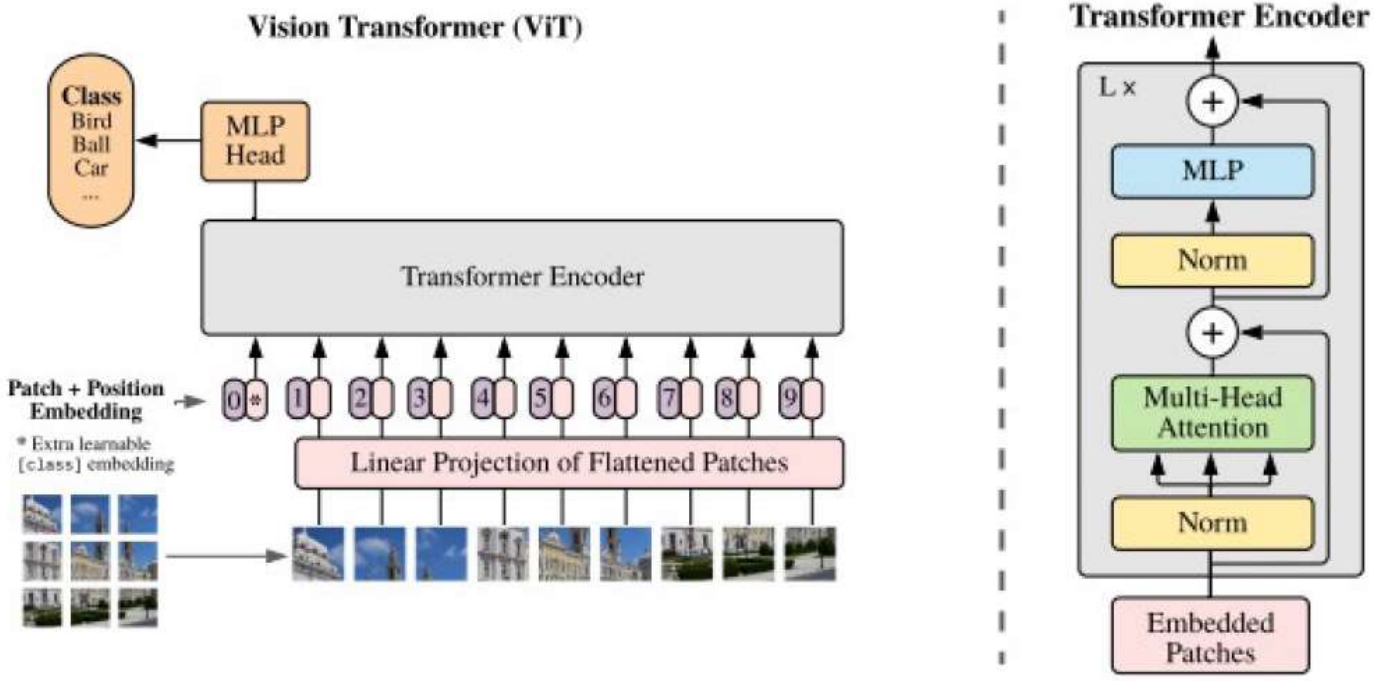
Vision Transformer (VIT)

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020. (ICLR 2021; 2,196 citations!)

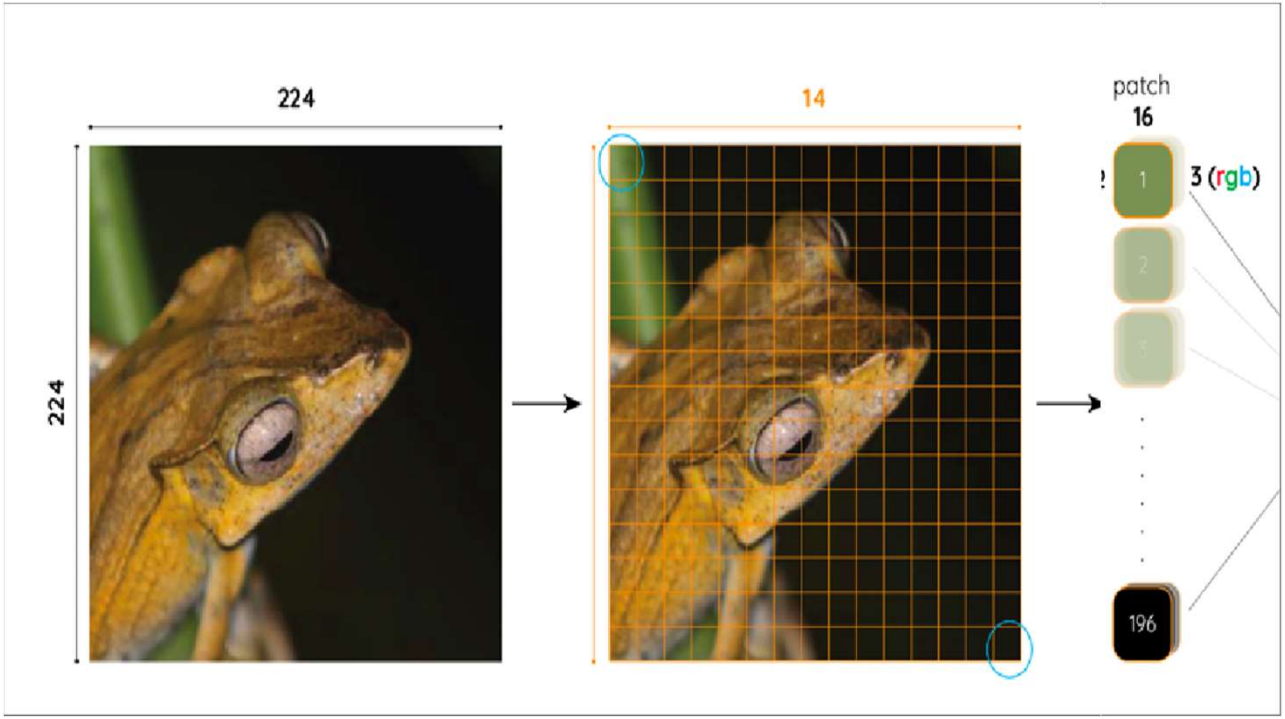
Vision Transformer (VIT)

- Naive application of self-attention to images requires high computation
- Divide an image into 16x16 patches (tokens)
- Transformers need to be trained on large datasets
- Vision Transformer (ViT) attains excellent results when pre-trained on JFT-300M
  - 88:55% on ImageNet,
  - 90:72% on ImageNet-ReaL,
  - 94:55% on CIFAR-100,
  - 77:63% on the VTAB suite of 19 tasks

Vision Transformer (VIT)

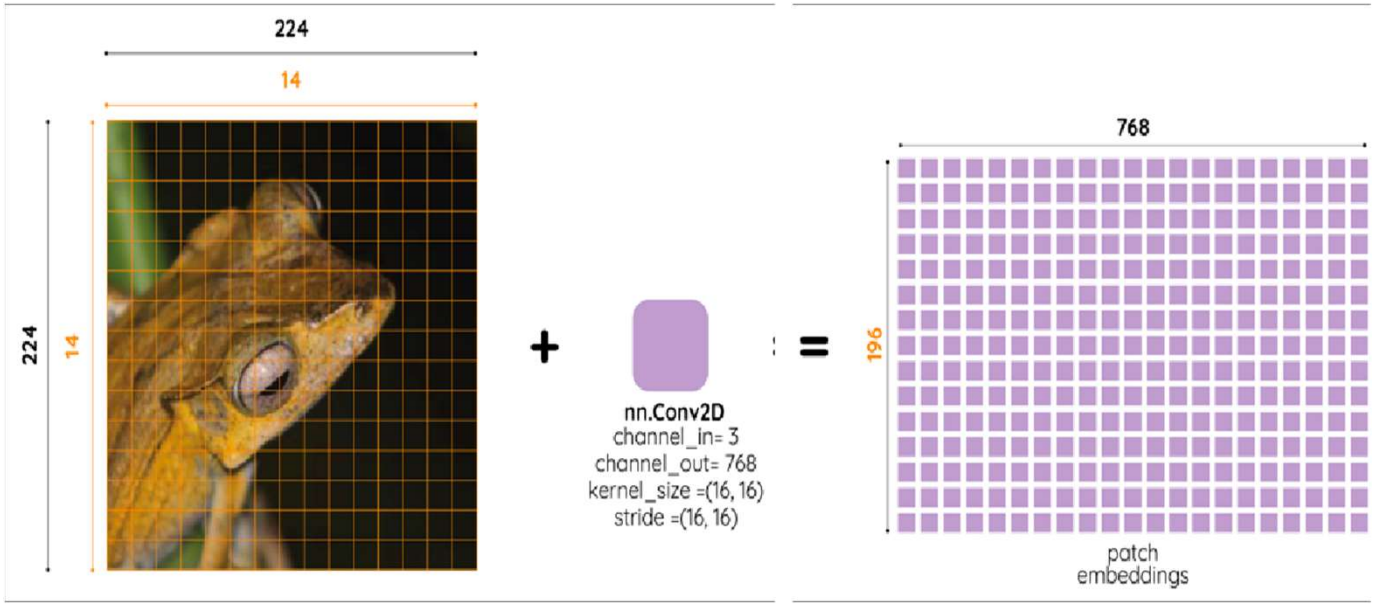


Divide image into 16x16 patches



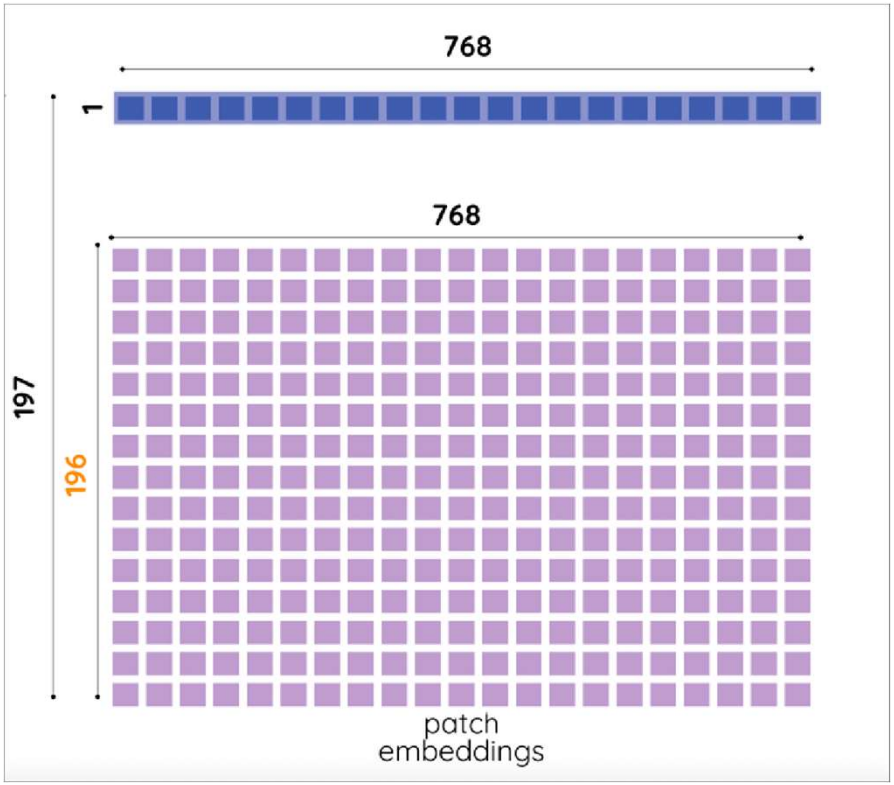
Slide credit: Piotr Mazurek

Generate embedding for each patch



Slide credit: Piotr Mazurek

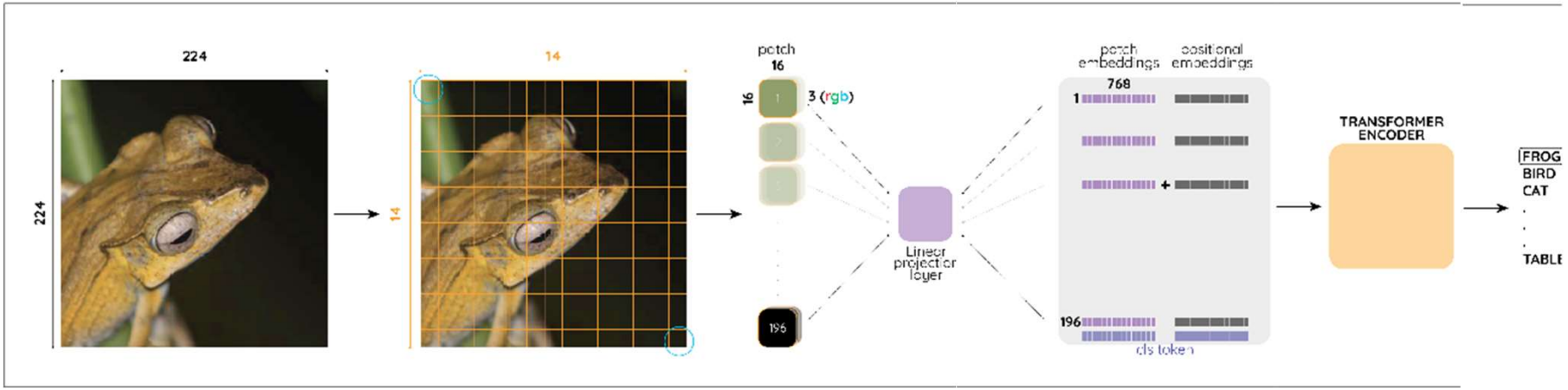
CLS Token



Slide credit: Piotr Mazurek

UCF Center for Research in Computer Vision

Complete VIT



Slide credit: Piotr Mazurek

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}};\ \mathbf{x}_p^1\mathbf{E};\ \mathbf{x}_p^2\mathbf{E};\ \cdots\ ;\ \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos},$$

$$\mathbf{z'}_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell, \qquad \ell = 1 \ldots L$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0)$$

VIT Model Variants

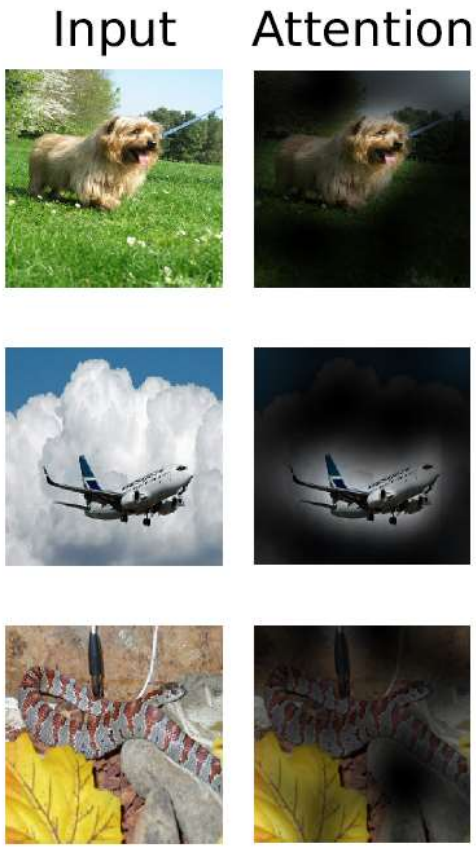
Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

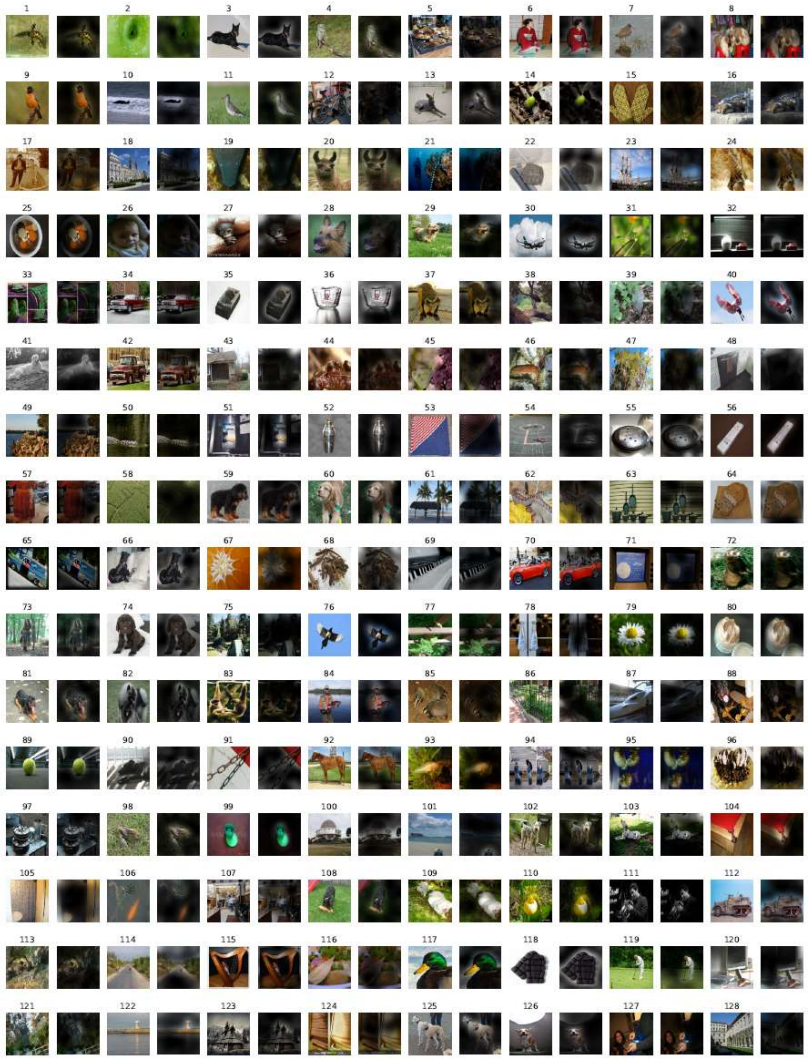
Table 1: Details of Vision Transformer model variants.

Results

	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

Attention





UCF Center for Research in Computer Vision

Vision Transformer (DeIT)

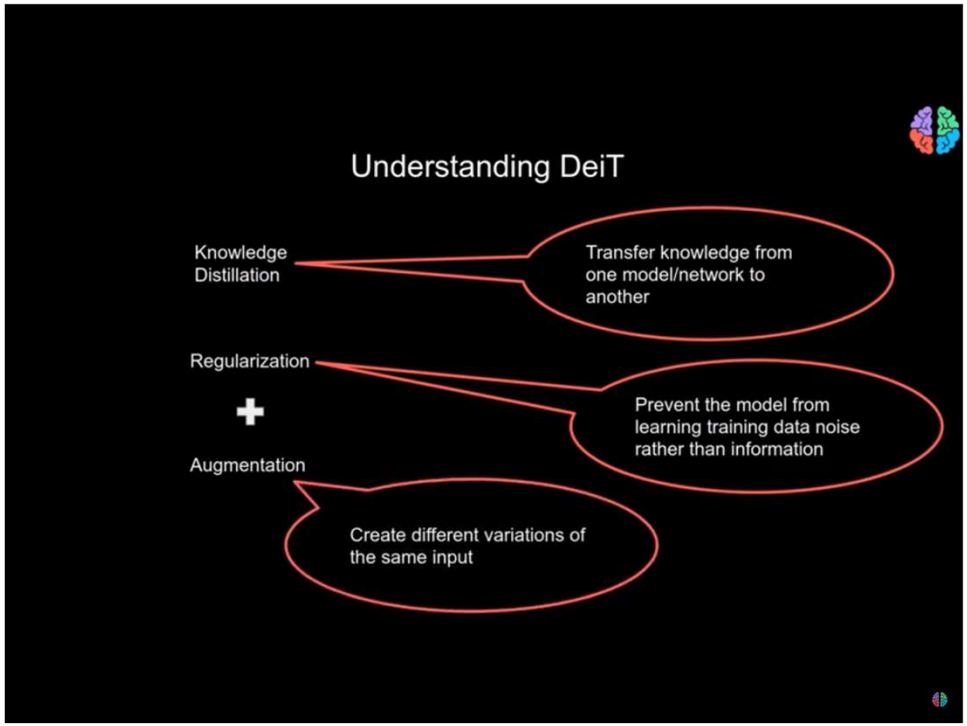
Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou, ” Training data-efficient image transformers & distillation through attention,” arXiv preprint arXiv:2012.12877v2 , 2021, (PMLR, 601 citations)

DeIT

- Only difference between VIT and DeIT is the training strategies, and the distillation token.
- Transformer as a Student while CNN as a Teacher
- Distillation tokens to produce estimated labels from teacher
- Attention between class and distillation tokens
- Competitive convolution-free transformers by training on Imagenet only
- Trained on a single computer in less than 3 days
- DeIT (86M parameters) achieves top-1 accuracy of 83.1% (single-crop) on ImageNet (1 million) with no external data
- VIT was trained on (JFT-300M , 300 millions images)

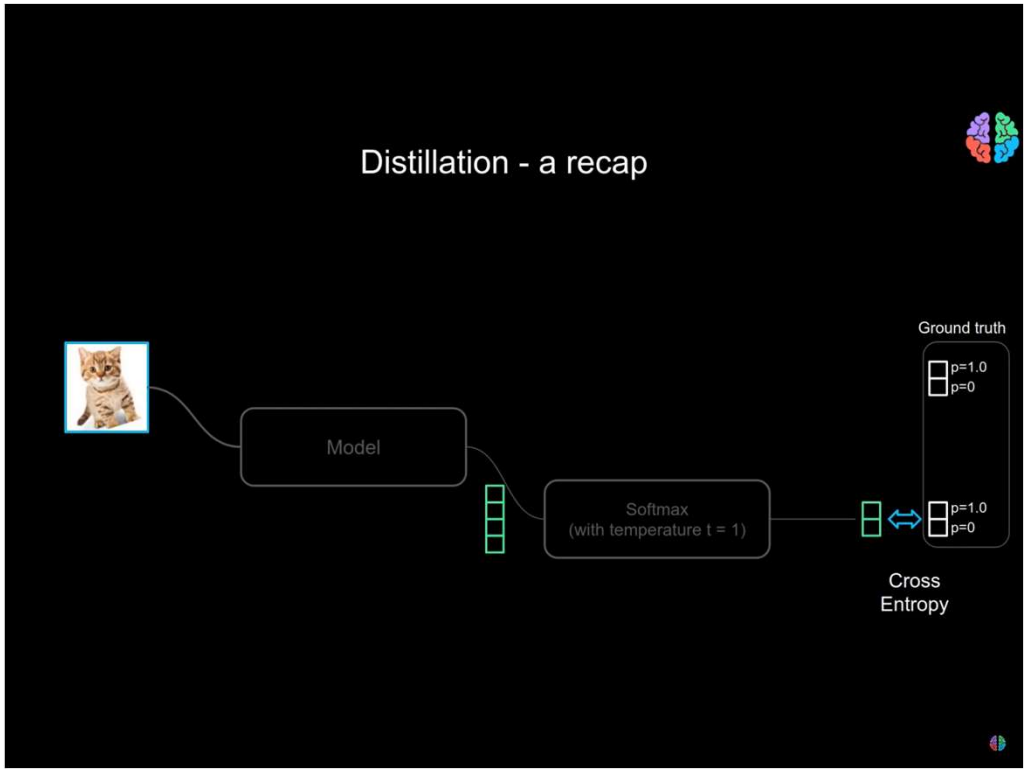
DeIT

Slide courtesy of AI Bites, Youtube Channel: https://www.youtube.com/c/AIBites



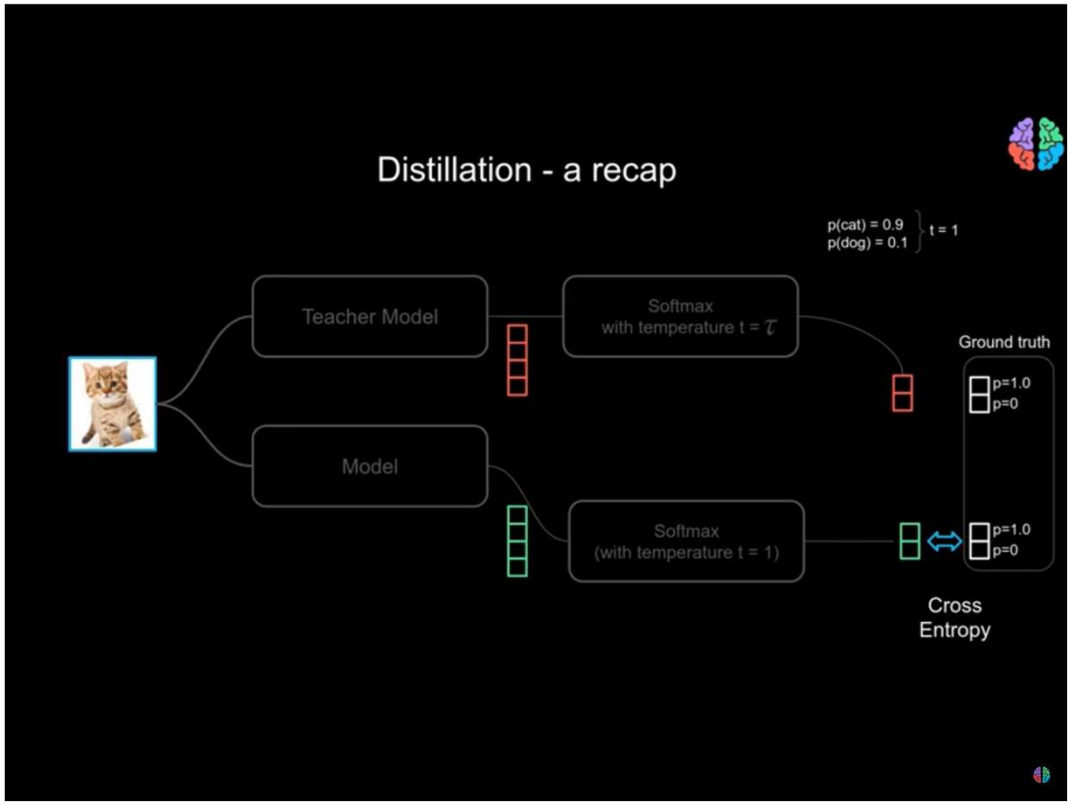
Knowledge Distillation

Slide courtesy of AI Bites, Youtube Channel: https://www.youtube.com/c/AIBites

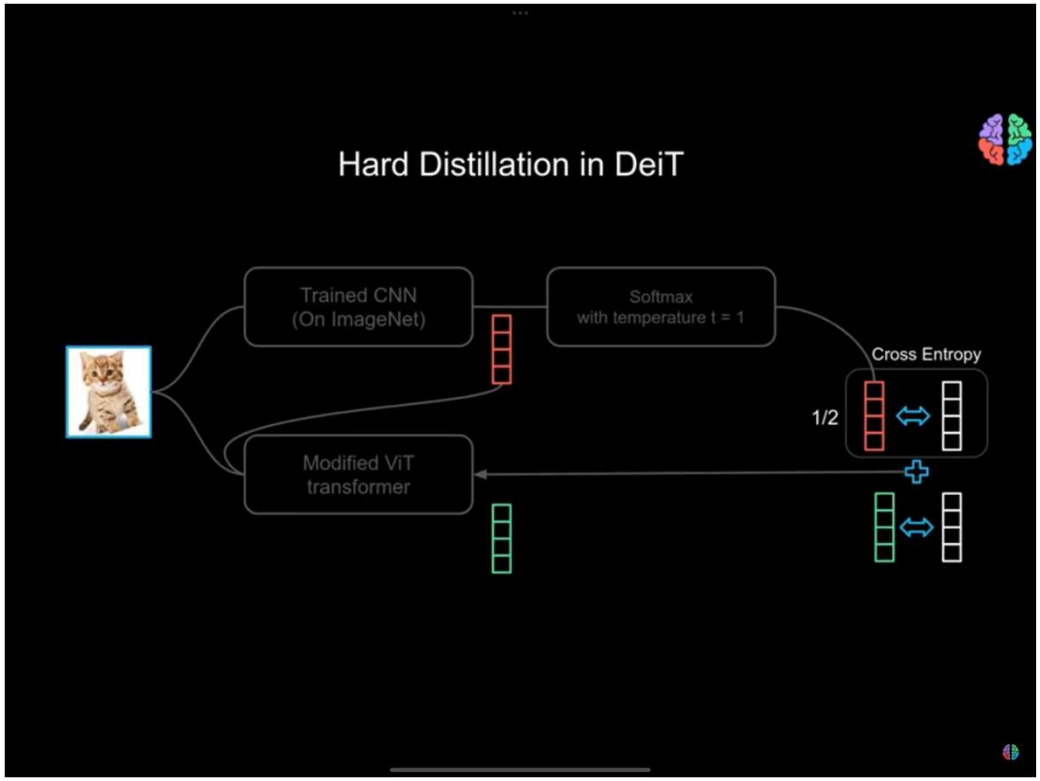


Knowledge Distillation

Slide courtesy of AI Bites, Youtube Channel: https://www.youtube.com/c/AIBites

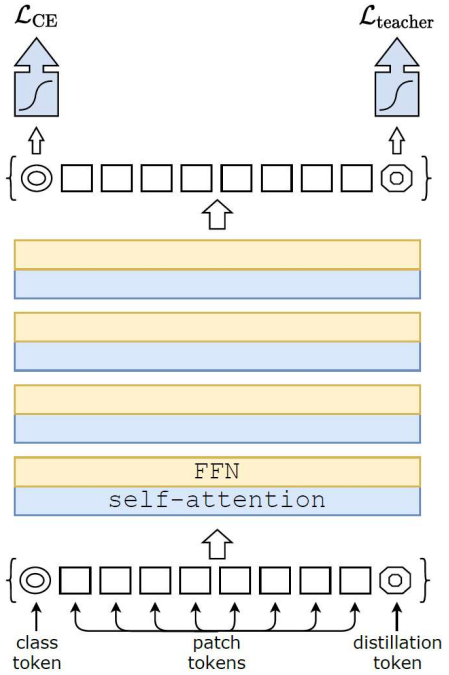


Slide courtesy of AI Bites, Youtube Channel: https://www.youtube.com/c/AIBites



UCF Center for Research in Computer Vision

DeIT



Different DeIT Configurations

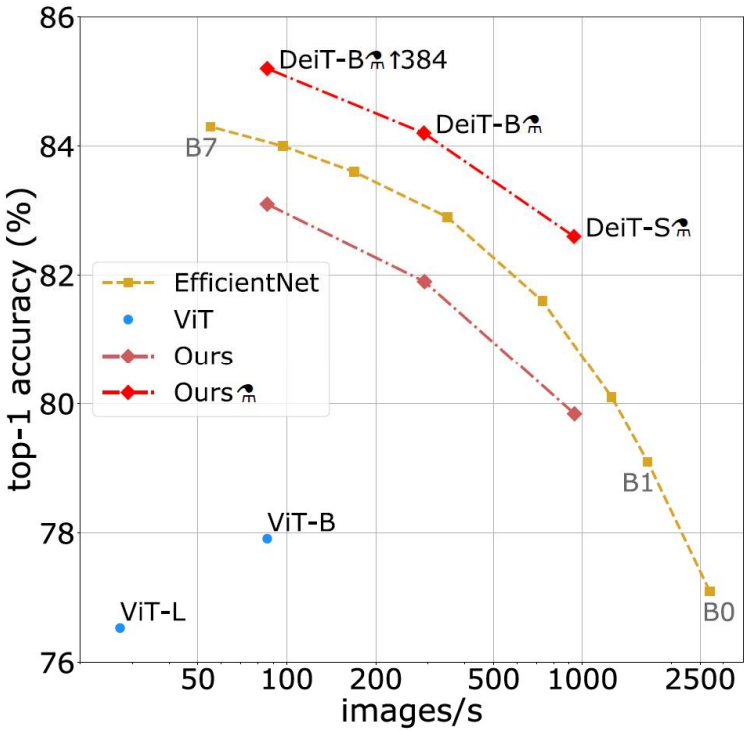
Model	ViT model	embedding dimension	#heads	#layers	#params	training resolution	throughput (im/sec)
DeiT-Ti	N/A	192	3	12	5M	224	2536
DeiT-S	N/A	384	6	12	22M	224	940
DeiT-B	ViT-B	768	12	12	86M	224	292

Comparison Between VIT and DeIT

Methods	ViT-B [15]	DeiT-B
Epochs	300	300
Batch size	4096	1024
Optimizer	AdamW	AdamW
learning rate	0.003	$0.0005 \times \frac{\text{batchsize}}{512}$
Learning rate decay	cosine	cosine
Weight decay	0.3	0.05
Warmup epochs	3.4	5
Label smoothing ε	✗	0.1
Dropout	0.1	✗
Stoch. Depth	✗	0.1
Repeated Aug	✗	✓
Gradient Clip.	✓	✗
Rand Augment	✗	9/0.5
Mixup prob.	✗	0.8
Cutmix prob.	✗	1.0
Erasing prob.	✗	0.25

Table 9: Ingredients and hyper-parameters for our method and Vit-B.

Performance



Results

Table 6: Datasets used for our different tasks.

Dataset	Train size	Test size	#classes
ImageNet [42]	1,281,167	50,000	1000
iNaturalist 2018 [26]	437,513	24,426	8,142
iNaturalist 2019 [27]	265,240	3,003	1,010
Flowers-102 [38]	2,040	6,149	102
Stanford Cars [30]	8,144	8,041	196
CIFAR-100 [31]	50,000	10,000	100
CIFAR-10 [31]	50,000	10,000	10

Model	ImageNet	CIFAR-10	CIFAR-100	Flowers	Cars	iNat-18	iNat-19	im/sec
Grafit ResNet-50 [49]	79.6	_	_	98.2	92.5	69.8	75.9	1226.1
Grafit RegNetY-8GF [49]	_	_	_	99.0	94.0	76.8	80.0	591.6
ResNet-152 [10]	_	_	_	_	_	69.1	_	526.3
EfficientNet-B7 [48]	84.3	98.9	91.7	98.8	94.7	_	_	55.1
ViT-B/32 [15]	73.4	97.8	86.3	85.4	_	_	_	394.5
ViT-B/16 [15]	77.9	98.1	87.1	89.5	_	_	_	85.9
ViT-L/32 [15]	71.2	97.9	87.1	86.4	_	_	_	124.1
ViT-L/16 [15]	76.5	97.9	86.4	89.7	_	_	_	27.3
DeiT-B	81.8	99.1	90.8	98.4	92.1	73.2	77.7	292.3
DeiT-B↑384	83.1	99.1	90.8	98.5	93.3	79.5	81.4	85.9
DeiT-B⚗	83.4	99.1	91.3	98.8	92.9	73.7	78.4	290.9
DeiT-B⚗ ↑384	84.4	99.2	91.4	98.9	93.9	80.1	83.0	85.9

Ablation on ↓	Pre-training	Fine-tuning	Rand-Augment	AutoAug	Mixup	CutMix	Erasing	Stoch. Depth	Repeated Aug.	Dropout	Exp. Moving Avg.	top-1 accuracy: pre-trained 224^2	top-1 accuracy: fine-tuned 384^2
none: DeiT-B	adamw	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8 ±0.2	83.1 ±0.1
optimizer	SGD	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✗	74.5	77.3
	adamw	SGD	✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8	83.1
data augmentation	adamw	adamw	✗	✗	✓	✓	✓	✓	✓	✗	✗	79.6	80.4
	adamw	adamw	✗	✓	✓	✓	✓	✓	✓	✗	✗	81.2	81.9
	adamw	adamw	✓	✗	✗	✓	✓	✓	✓	✗	✗	78.7	79.8
	adamw	adamw	✓	✗	✓	✗	✓	✓	✓	✗	✗	80.0	80.6
	adamw	adamw	✓	✗	✗	✗	✓	✓	✓	✗	✗	75.8	76.7
regularization	adamw	adamw	✓	✗	✓	✓	✗	✓	✓	✗	✗	4.3*	0.1
	adamw	adamw	✓	✗	✓	✓	✓	✗	✓	✗	✗	3.4*	0.1
	adamw	adamw	✓	✗	✓	✓	✓	✓	✗	✗	✗	76.5	77.4
	adamw	adamw	✓	✗	✓	✓	✓	✓	✓	✓	✗	81.3	83.1
	adamw	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✓	81.9	83.1

SWIN

- Z Liu, Y Lin, Y Cao, H Hu, Y Wei, Z Zhang, S Lin, “Swin transformer: Hierarchical vision transformer using shifted windows”, ICCV-2021. (Marr Prize) 628 Citations

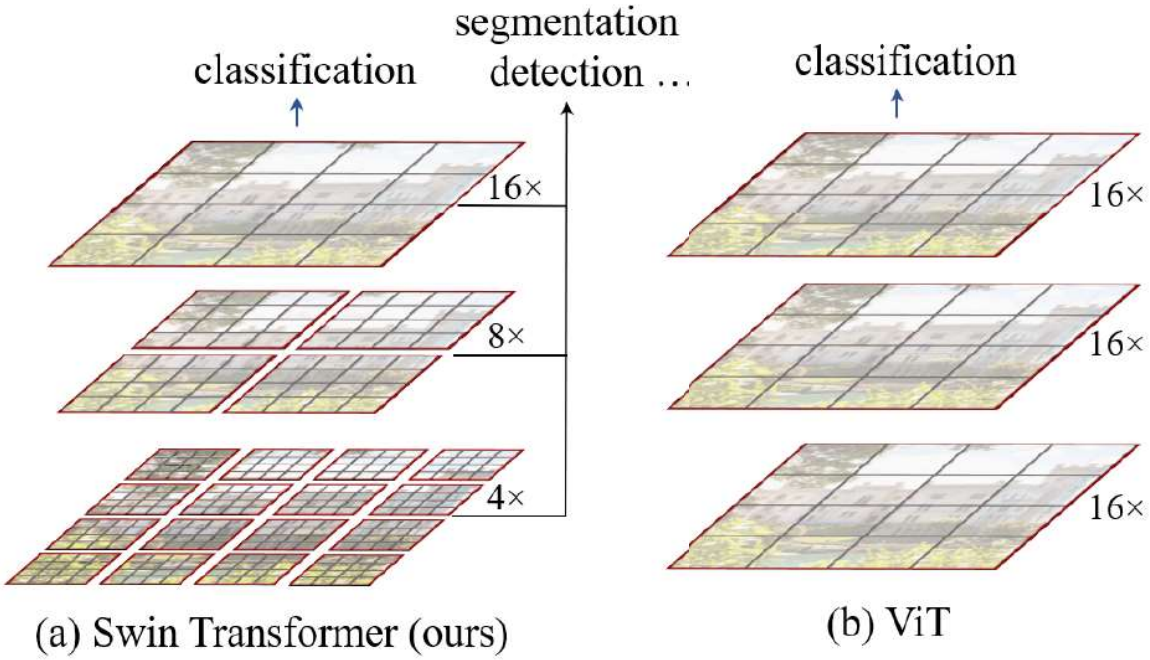
SWIN

- Adapting Transformer from language to vision is challenging
- Differences between language and vision Domains
  - Large variations in the scale of visual entities
  - High resolution of pixels in images compared to words in text
- To address these differences, SWIN proposes
  - A hierarchical Transformer whose representation is computed with Shifted windows
  - Shifted Windowing limit attention
    - To local windows and
    - Allowing cross window connections

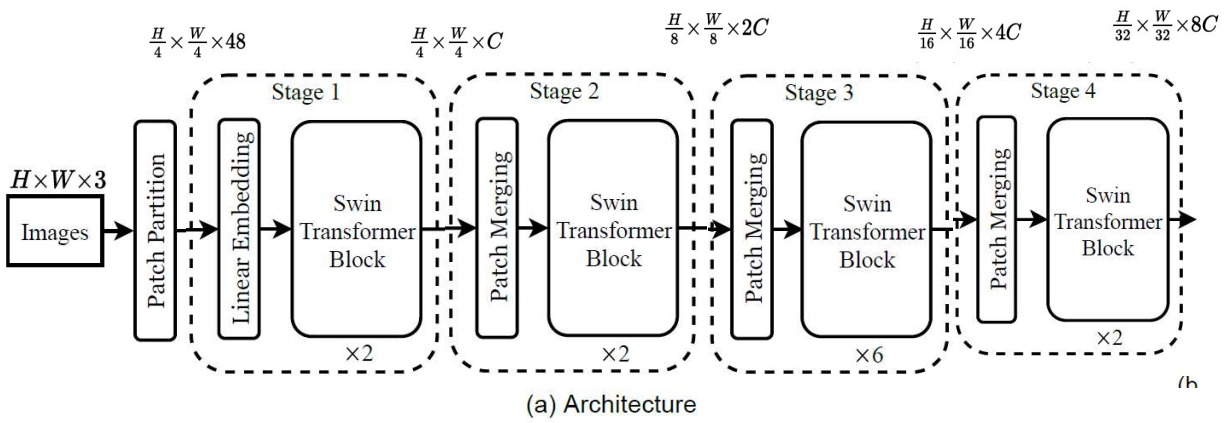
SWIN

- 87.3 top-1 accuracy on ImageNet-1K
- Dense prediction tasks such as
  - Object detection (58.7 box AP and 51.1 mask AP on COCO)
  - Semantic segmentation (53.5 mIoU on ADE20K)
- Performance surpasses the previous state-of-the art by
  - +2.7 box AP and +2.6 mask AP on COCO, and
  - +3.2 mIoU on ADE20K,

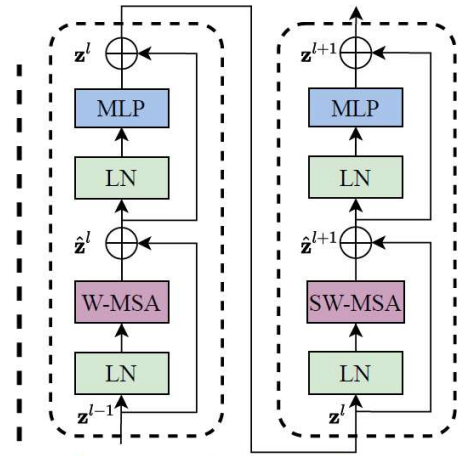
Hierarchical Feature Maps and Local Attention



SWIN Architecture

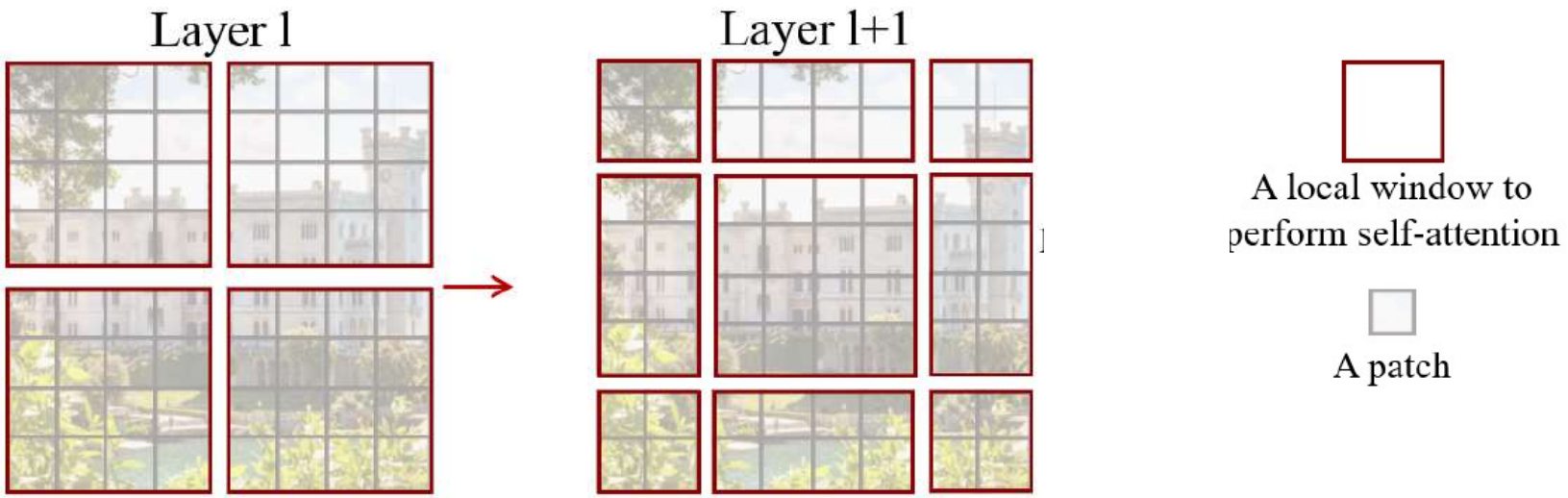


(a) Architecture



) Two Successive Swin Transformer Blocks

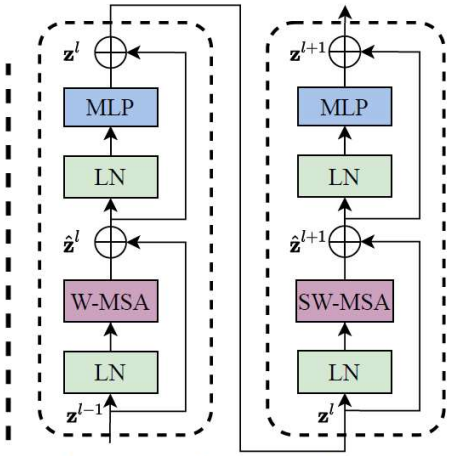
Self-Attention within each window and shifted windows



Transformer Blocks

$$\hat{\mathbf{z}}^{l} = \text{W-MSA}\left(\text{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1},$$
$$\mathbf{z}^{l} = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l}\right)\right) + \hat{\mathbf{z}}^{l},$$

$$\hat{\mathbf{z}}^{l+1} = \text{SW-MSA}\left(\text{LN}\left(\mathbf{z}^{l}\right)\right) + \mathbf{z}^{l},$$
$$\mathbf{z}^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1},$$



) Two Successive Swin Transformer Blocks

Different Configurations

- Swin-T: $C = 96$, layer numbers = $\{2, 2, 6, 2\}$
- Swin-S: $C = 96$, layer numbers =$\{2, 2, 18, 2\}$
- Swin-B: $C = 128$, layer numbers =$\{2, 2, 18, 2\}$
- Swin-L: $C = 192$, layer numbers =$\{2, 2, 18, 2\}$

Results

(a) Regular ImageNet-1K trained models					
method	image size	#param.	FLOPs	throughput (image / s)	ImageNet top-1 acc.
RegNetY-4G [44]	224^2	21M	4.0G	1156.7	80.0
RegNetY-8G [44]	224^2	39M	8.0G	591.6	81.7
RegNetY-16G [44]	224^2	84M	16.0G	334.7	82.9
ViT-B/16 [19]	384^2	86M	55.4G	85.9	77.9
ViT-L/16 [19]	384^2	307M	190.7G	27.3	76.5
DeiT-S [57]	224^2	22M	4.6G	940.4	79.8
DeiT-B [57]	224^2	86M	17.5G	292.3	81.8
DeiT-B [57]	384^2	86M	55.4G	85.9	83.1
Swin-T	224^2	29M	4.5G	755.2	81.3
Swin-S	224^2	50M	8.7G	436.9	83.0
Swin-B	224^2	88M	15.4G	278.1	83.5
Swin-B	384^2	88M	47.0G	84.7	84.5

Results

(b) ImageNet-22K pre-trained models

method	image size	#param.	FLOPs	throughput (image / s)	ImageNet top-1 acc.
R-101x3 [34]	384^2	388M	204.6G	-	84.4
R-152x4 [34]	480^2	937M	840.5G	-	85.4
ViT-B/16 [19]	384^2	86M	55.4G	85.9	84.0
ViT-L/16 [19]	384^2	307M	190.7G	27.3	85.2
Swin-B	224^2	88M	15.4G	278.1	85.2
Swin-B	384^2	88M	47.0G	84.7	86.4
Swin-L	384^2	197M	103.9G	42.1	87.3

Results

(b) Various backbones w. Cascade Mask R-CNN

	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	param	FLOPs	FPS
DeiT-S†	48.0	67.2	51.7	41.4	64.2	44.3	80M	889G	10.4
R50	46.3	64.3	50.5	40.1	61.7	43.4	82M	739G	18.0
Swin-T	**50.5**	**69.3**	**54.9**	**43.7**	**66.6**	**47.1**	86M	745G	15.3
X101-32	48.1	66.5	52.4	41.6	63.9	45.2	101M	819G	12.8
Swin-S	**51.8**	**70.4**	**56.3**	**44.7**	**67.9**	**48.5**	107M	838G	12.0
X101-64	48.3	66.4	52.3	41.7	64.0	45.1	140M	972G	10.4
Swin-B	**51.9**	**70.9**	**56.5**	**45.0**	**68.4**	**48.7**	145M	982G	11.6

ADE20K Method	Backbone	val mIoU	test score	#param.	FLOPs	FPS
DLab.v3+ [11]	ResNet-101	44.1	-	63M	1021G	16.0
DNL [65]	ResNet-101	46.0	56.2	69M	1249G	14.8
OCRNet [67]	ResNet-101	45.3	56.0	56M	923G	19.3
UperNet [63]	ResNet-101	44.9	-	86M	1029G	20.1
OCRNet [67]	HRNet-w48	45.7	-	71M	664G	12.5
DLab.v3+ [11]	ResNeSt-101	46.9	55.1	66M	1051G	11.9
DLab.v3+ [11]	ResNeSt-200	48.4	-	88M	1381G	8.1
SETR [73]	T-Large‡	50.3	61.7	308M	-	-
UperNet	DeiT-S†	44.0	-	52M	1099G	16.2
UperNet	Swin-T	46.1	-	60M	945G	18.5
UperNet	Swin-S	49.3	-	81M	1038G	15.2
UperNet	Swin-B‡	51.6	-	121M	1841G	8.7
UperNet	Swin-L‡	**53.5**	**62.8**	234M	3230G	6.2

Table 3. Results of semantic segmentation on the ADE20K val

Summary

- VIT is first Vision Transformer, but trained on huge dataset of 300M
- DeIT employs knowledge distillation, augmentation and efficient training strategy, but trained on ImageNet of 1 million
- SWIN employs window attention, in addition performs well on other tasks: object detection, semantic segmentation