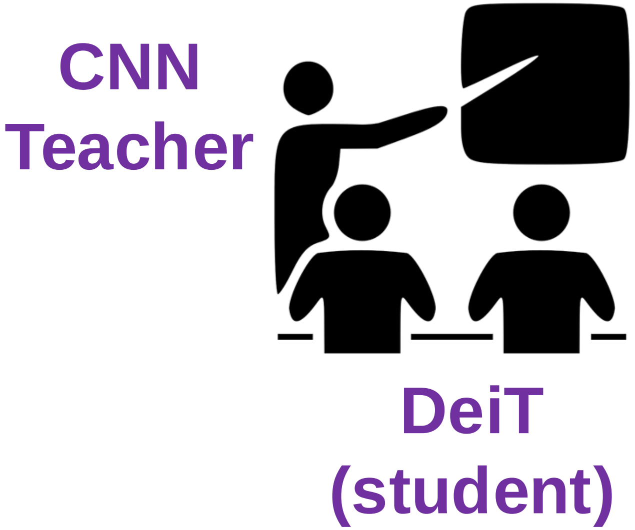
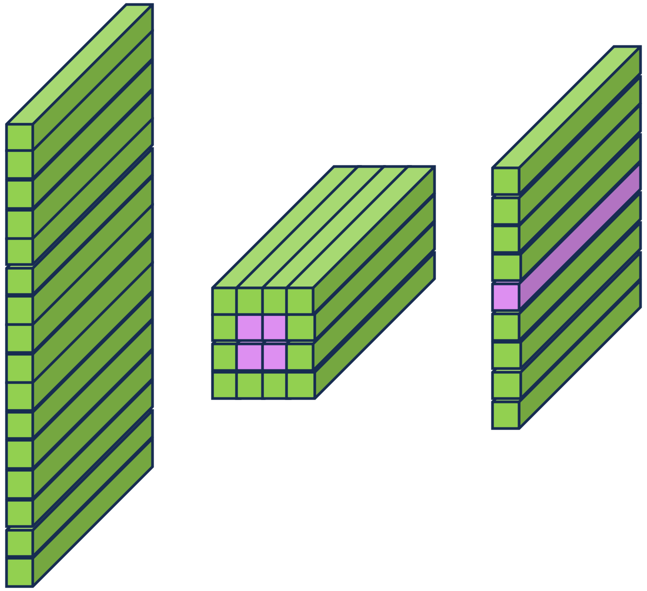
Variants of Vision Transformer

DeiT

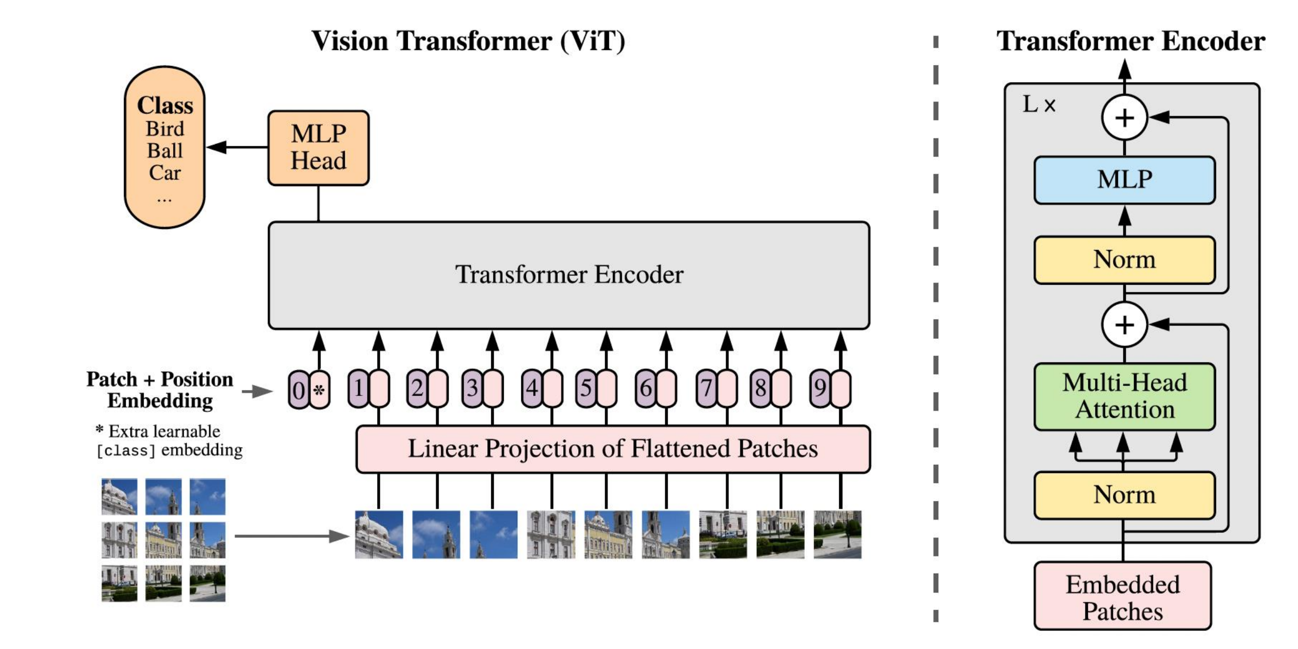


T2T-ViT



Vision Transformers Series

Recap of ViT



Generate a sequence of image tokens (patches) and apply standard transformer

Minimal inductive bias ➔ Learn everything from scratch

Pre-training on very large labeled dataset (JFT300M)

"An image is worth 16x16 words: Transformers for image recognition at scale.", Dosovitskiy et al., 2020 https://arxiv.org/pdf/2010.11929.pdf

ViT Variants

1. DeiT
2. T2T-ViT
3. BEiT
4. CaiT
5. SWIN
6. DINO
7. CLIP

...

This video

Motivation:
Vanilla ViT shows inferior performance to CNNs when trained from scratch on a mid-size image datasets

Part 1: DeiT

Data Efficient Image Transformers

Training data-efficient image transformers & distillation through attention, Touvron et al., 2021, https://arxiv.org/pdf/2012.12877.pdf

DeiT overview

- Original ViT requires hundred-million images for pre-training
  ➔ Inferior performance if trained on mid-size image datasets

- DeiT objective ➔ training on ImageNet1k (mid-size dataset)

Main idea:

➢ Teacher-Student Distillation

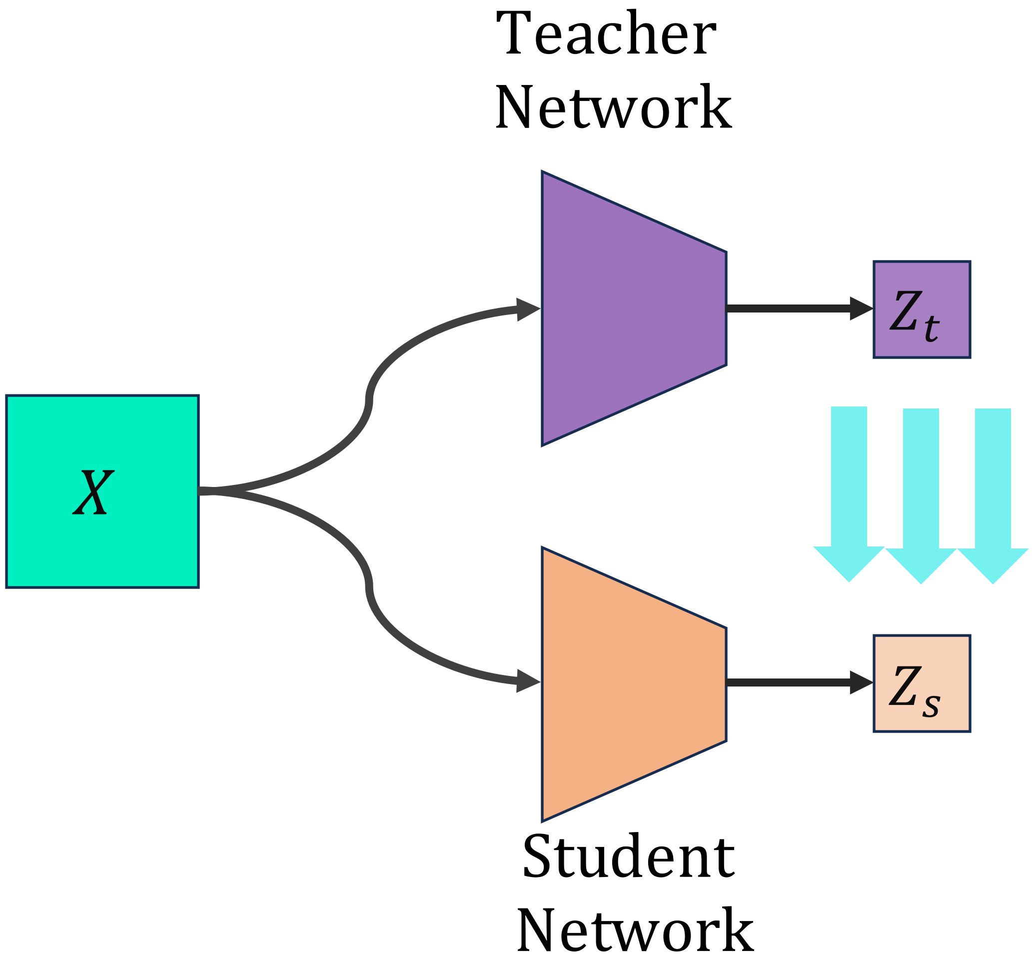
Knowledge Distillation

- Two networks:
  - Teacher model
  - Student model

- Teacher guides the training of student

Modes of training

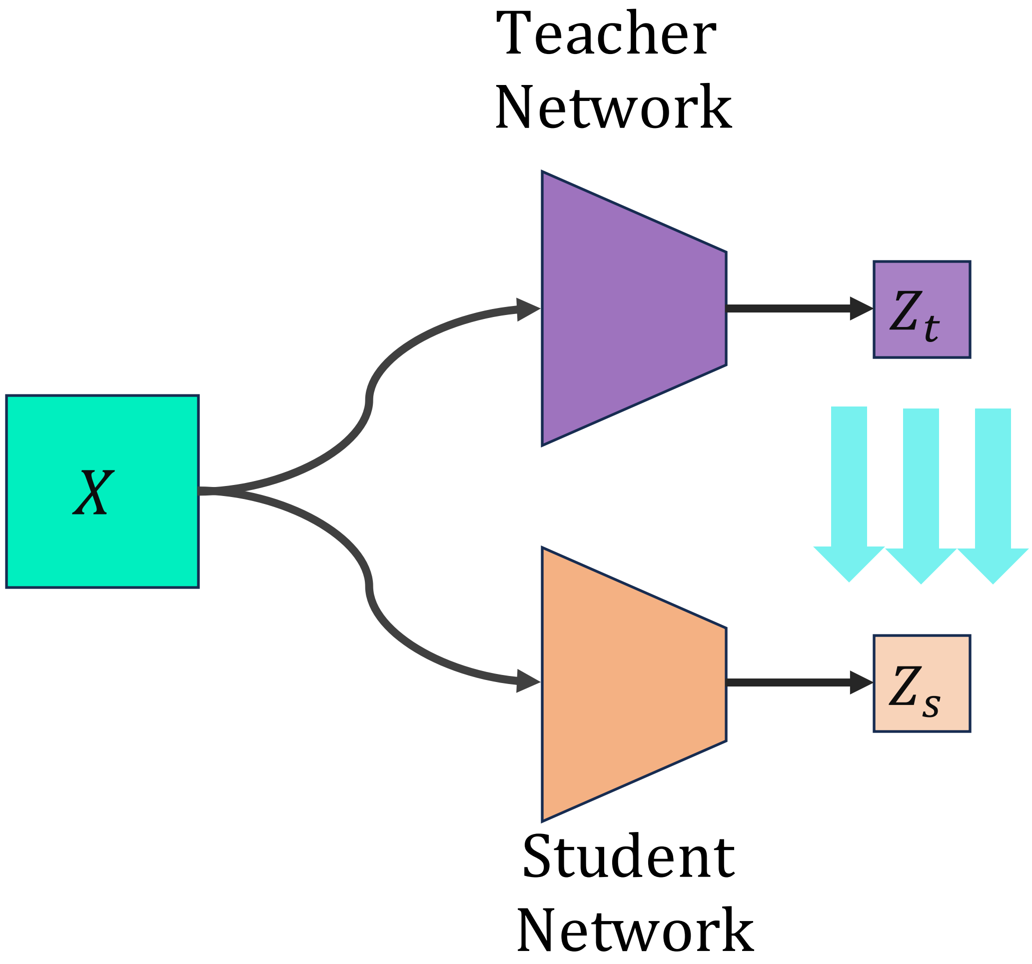
- Pre-trained teacher:
  Teacher is pre-trained and frozen during the training of the student model
- Simultaneous training:
  Training student and teacher models at the same time



Knowledge Distillation

Transferring targets from teacher to student:

1. Soft distillation
   Student's targets are derived from teacher's output probabilities
2. Hard distillation
   Student's targets are based on hard decision applied to teacher's output



Soft Distillation

Soft targets from teacher network

Minimizing KL-divergence between softmax output of teacher and softmax output of student network

$$\mathcal{L}_{\text{student}} = (1-\lambda)\mathcal{L}_{CE}(\sigma(Z_s), y) + \lambda\tau^2\,\mathrm{KL}\left(\sigma\left(\frac{Z_s}{\tau}\right), \sigma\left(\frac{Z_t}{\tau}\right)\right)$$

Z_t: Teacher logits

Z_s: Student logits

y: Ground truth label

λ: Balancing factor

τ: Temperature

Hard Distillation

Taking hard labels from teacher output as true label (y_t)

$$y_t = \mathrm{argmax_c}\, Z_t(c)$$

$$\mathcal{L}_{\text{student}} = \frac{1}{2}\mathcal{L}_{CE}(\sigma(Z_s), y) + \frac{1}{2}\mathcal{L}_{CE}(\sigma(Z_s), y_t)$$

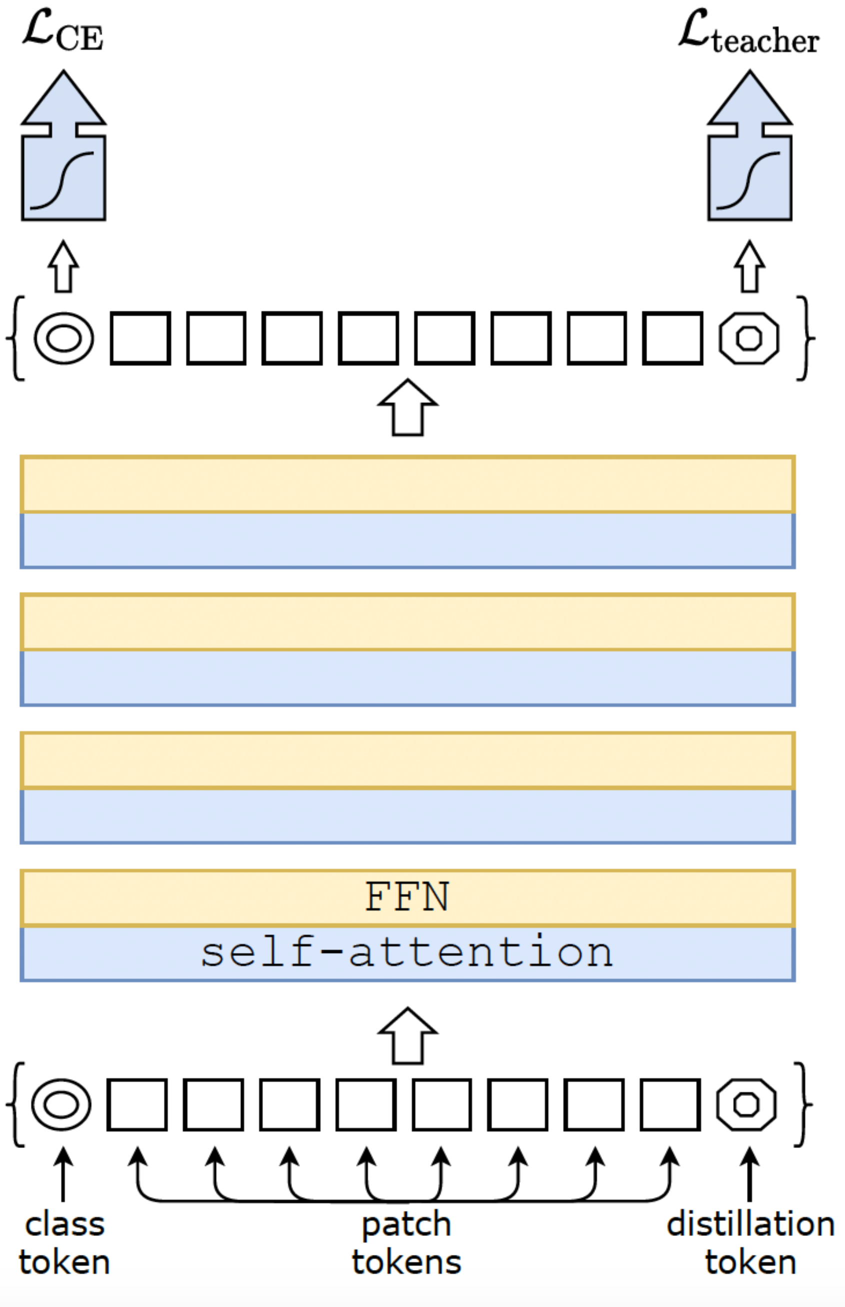
➔ Further enhance teacher hard labels with label-smoothing

- $(1-\epsilon)$: true label
- ϵ spread over others

DeiT Student Architecture

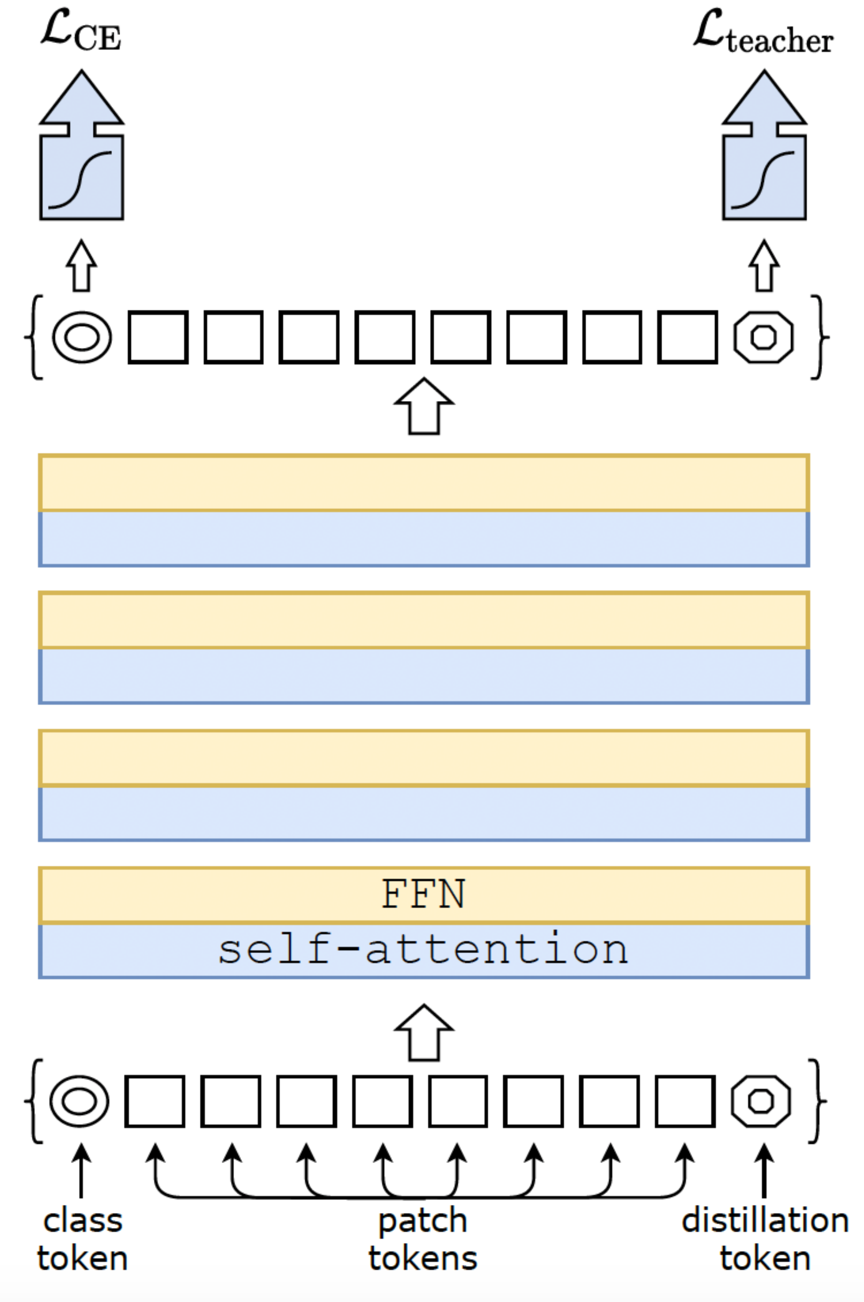
Distillation token: A special learnable token (like class token) added to the end of input sequence

➔ Allows the model to learn from the output of teacher network



https://arxiv.org/pdf/2012.12877.pdf

Test (inference)



Input to the classifier head:

1. Just the class embedding
2. Just the distillation embedding
3. Fusing both class and distillation embedding (**late fusion**)

https://arxiv.org/pdf/2012.12877.pdf

DeiT Model Variants

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

https://arxiv.org/pdf/2012.12877.pdf

Performance as a Function of the Teacher Network

Teacher Models	acc.	Student: DeiT-B ⚗ pretrain	↑384
DeiT-B	81.8	81.9	83.1
RegNetY-4GF	80.0	82.7	83.6
RegNetY-8GF	81.7	82.7	83.8
RegNetY-12GF	82.4	83.1	84.1
RegNetY-16GF	82.9	83.1	84.2

Student learns better from a **ConvNet** teacher than a transformer teacher

https://arxiv.org/pdf/2012.12877.pdf

Comparing distillation methods

method ↓	Supervision		ImageNet top-1 (%)			
	label	teacher	Ti 224	S 224	B 224	B↑384
DeiT– no distillation	✓	✗	72.2	79.8	81.8	83.1
DeiT– usual distillation	✗	soft	72.2	79.8	81.8	83.2
DeiT– hard distillation	✗	hard	74.3	80.9	83.0	84.0
DeiT⚗: class embedding	✓	hard	73.9	80.9	83.0	84.2
DeiT⚗: distil. embedding	✓	hard	74.6	81.1	83.1	84.4
DeiT⚗: class+distillation	✓	hard	74.5	81.2	83.4	84.5

Part 2: T2T – ViT

Tokens-to-Token

Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet,
Yuan et al., 2021, https://arxiv.org/pdf/2101.11986.pdf

T2T-ViT overview

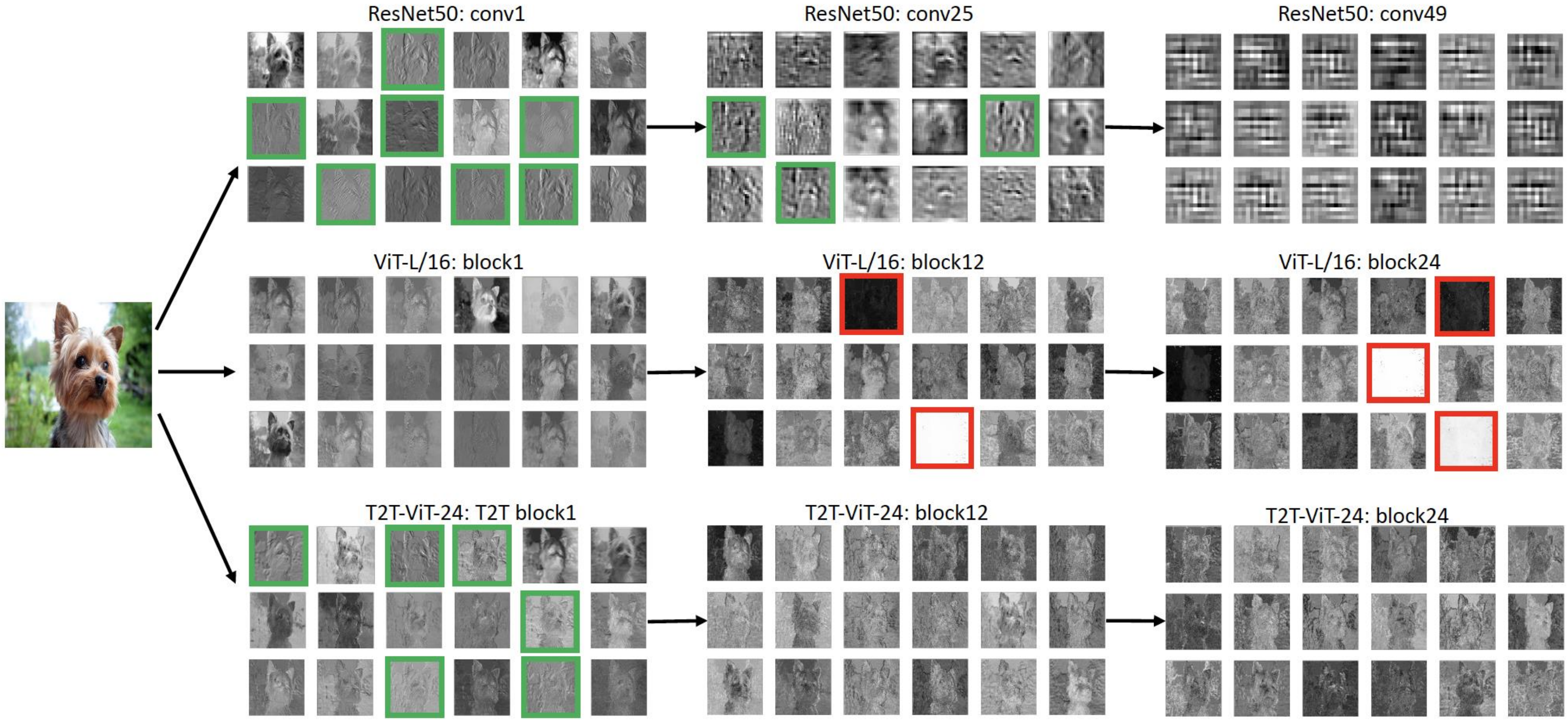
- Original ViT has inferior performance when trained from scratch on a mid-size dataset

- Hypothesized Reasons:
  - Tokenization process fails to model local structures inherence in images
  - Redundant attention backbone in ViT, which leads to limited feature richness

Proposed solution:

- ➢ Recursive (layer-wise) tokenization
- ➢ A deep-narrow backbone structure motivated by CNN architectures

CNN vs. ViT Features



https://arxiv.org/pdf/2101.11986.pdf

Tokens-to-Token

Steps:

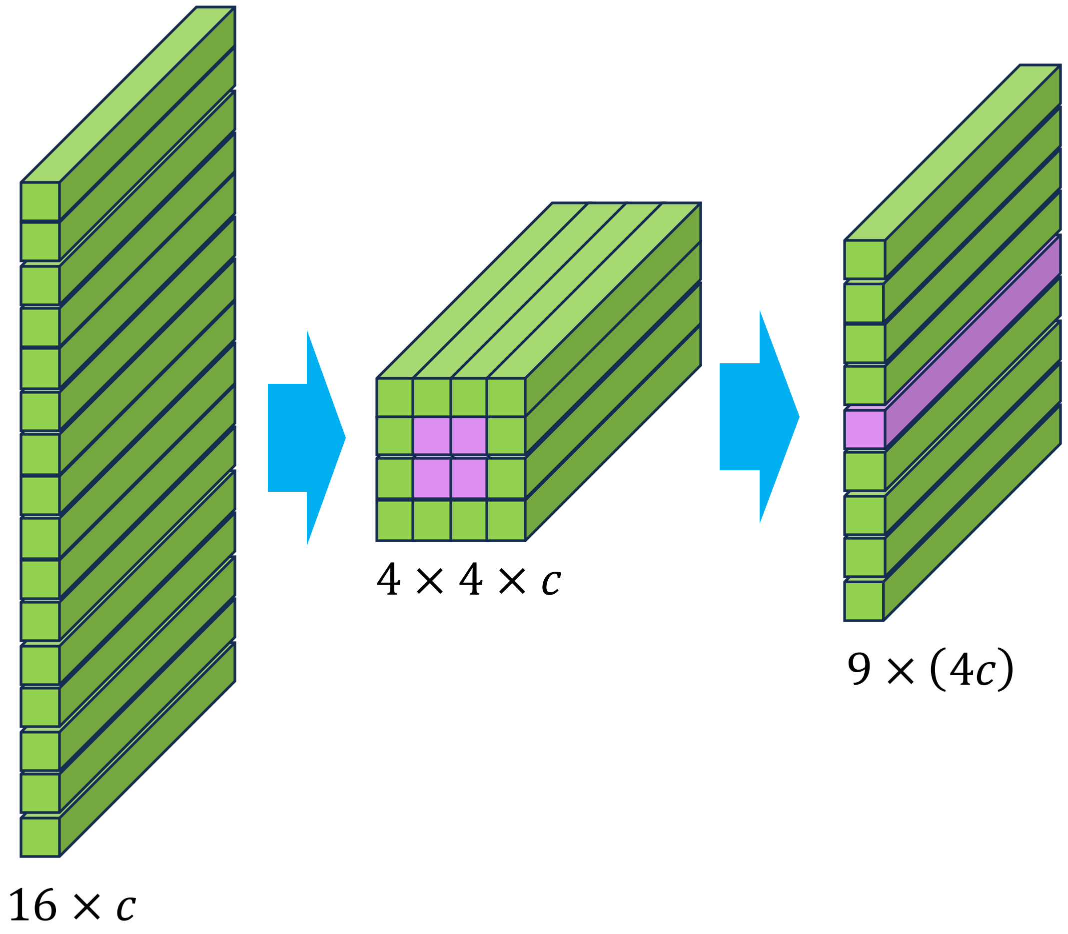
(1) Reconstruction

The output token embeddings from each transformer layer is restructured to an image format

(2) Soft split

Split the reconstructed image into tokens with overlaps

$$\left.\begin{array}{l} k\text{: patch size} \\ s\text{: overlap} \end{array}\right\} \Rightarrow \text{stride} = k - s$$



Tokens-to-Token

Steps:

(1) Restructurization (or reconstruction)

- The output token embeddings from each transformer layer is restructured to an image format

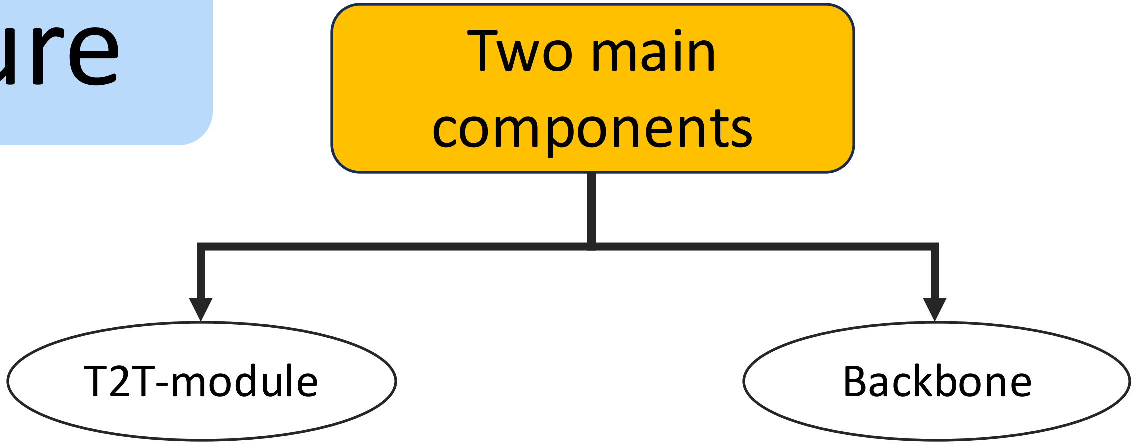
(2) Soft split

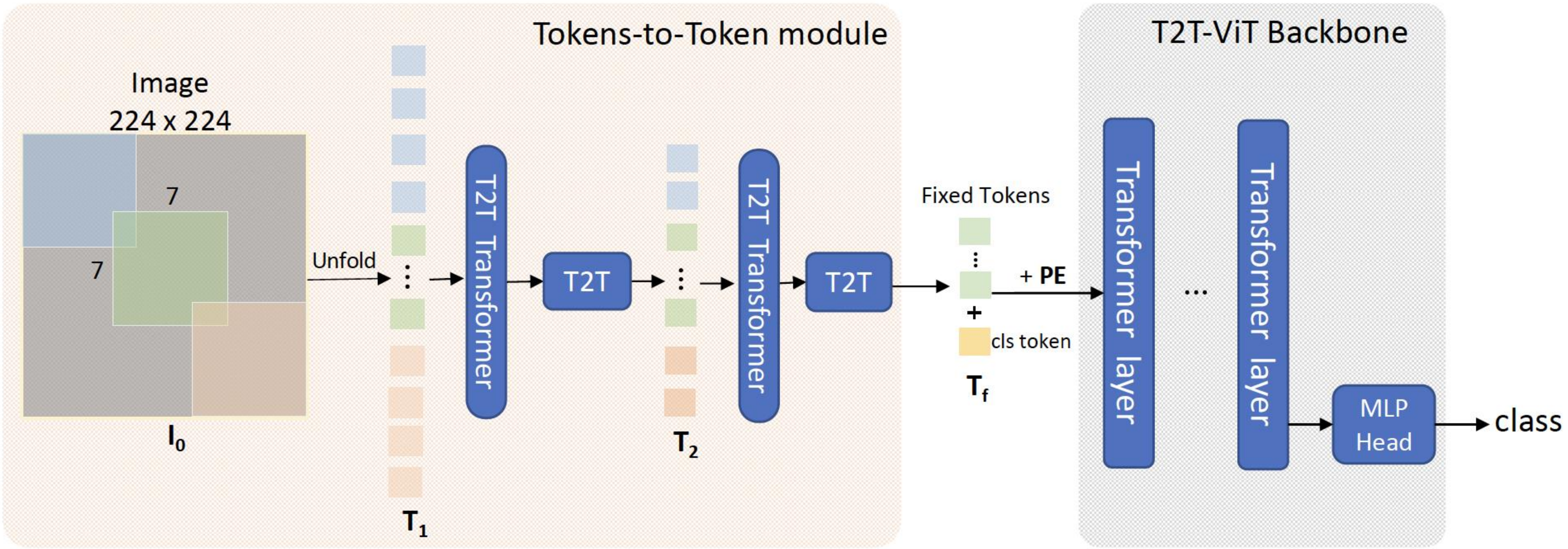
- Split the reconstructed image into tokens with overlaps
- Aggregate neighboring tokens into one token

Benefits of T2T

- Local structure is embedded into tokens
- Aggregation reduces the sequence lengths (# tokens)
- Enabling more efficient backbone architectures like CNNs

T2T Architecture





https://arxiv.org/pdf/2101.11986.pdf

T2T Module

Depth: $n = 2$

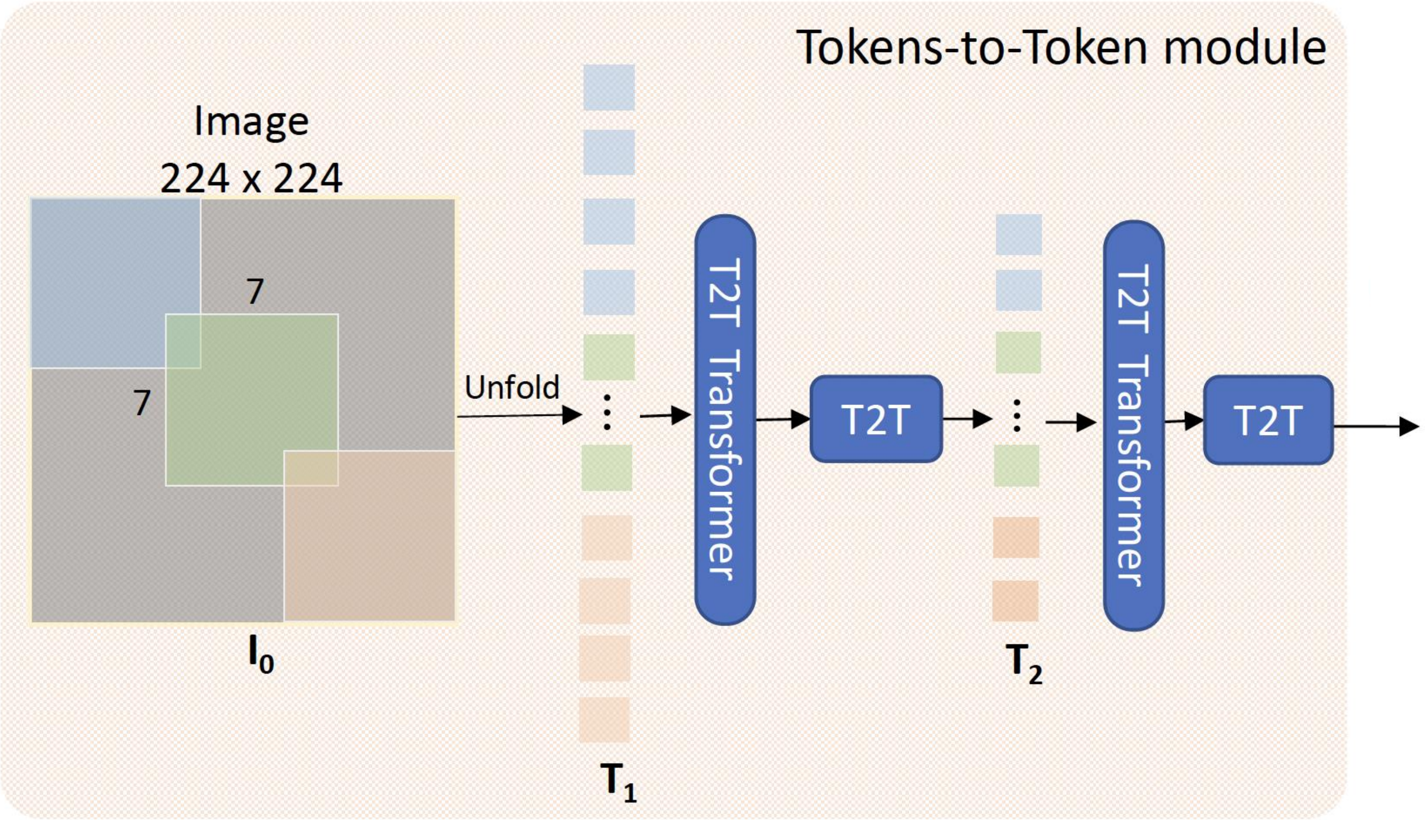
Composed of

- Two T2T Transformers
- Three soft-splits $(n + 1)$

- Patch sizes: $[7, 3, 3]$
- Overlap: $[3, 1, 1]$

➔

- Input: 224×224
- Output: 14×14



Architecture Details

Models	Tokens-to-Token module				T2T-ViT backbone			Model size	
	T2T transformer	Depth	Hidden dim	MLP size	Depth	Hidden dim	MLP size	Params (M)	MACs (G)
ViT-S/16 [12]	-	-	-	-	8	786	2358	48.6	10.1
ViT-B/16 [12]	-	-	-	-	12	786	3072	86.8	17.6
ViT-L/16 [12]	-	-	-	-	24	1024	4096	304.3	63.6
T2T-ViT-14	Performer	2	64	64	14	384	1152	21.5	4.8
T2T-ViT-19	Performer	2	64	64	19	448	1344	39.2	8.5
T2T-ViT-24	Performer	2	64	64	24	512	1536	64.1	13.8
T2T-ViT$_t$-14	Transformer	2	64	64	14	384	1152	21.5	6.1
T2T-ViT-7	Performer	2	64	64	8	256	512	4.2	1.1
T2T-ViT-12	Performer	2	64	64	12	256	512	6.8	1.8

https://arxiv.org/pdf/2101.11986.pdf

Results

Models	Top1-Acc (%)	Params (M)	MACs (G)
ViT-S/16 [12]	78.1	48.6	10.1
DeiT-small [36]	79.9	22.1	4.6
DeiT-small-Distilled [36]	81.2	22.1	4.7
T2T-ViT-14	**81.5**	21.5	4.8
T2T-ViT-14↑384	**83.3**	21.5	17.1
ViT-B/16 [12]	79.8	86.4	17.6
ViT-L/16 [12]	81.1	304.3	63.6
T2T-ViT-24	**82.3**	**64.1**	**13.8**

All models trained from scratch on ImageNet1k

- Vanilla ViT
- DeiT
- T2T-ViT

https://arxiv.org/pdf/2101.11986.pdf

Summary

- Vanilla ViT requires very large dataset for pre-training
- DeiT and T2T-ViT: Two ViT variants that can be pre-trained on a mid-size dataset and still outperforming CNNs

Thanks for watching