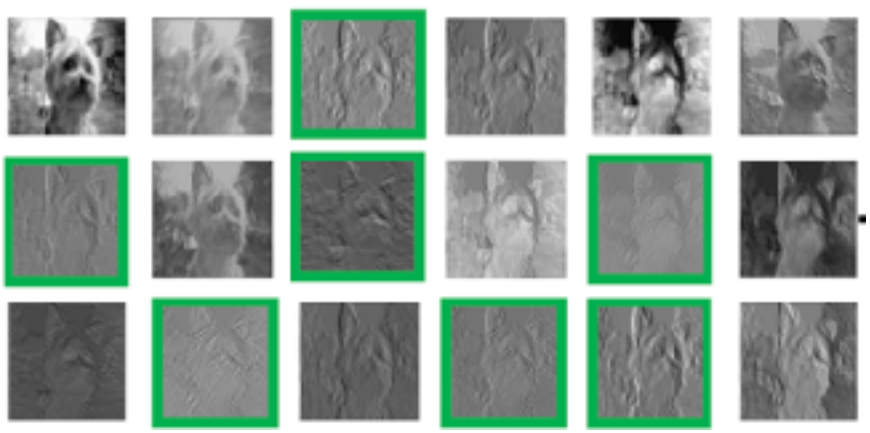
Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

ICCV 2021

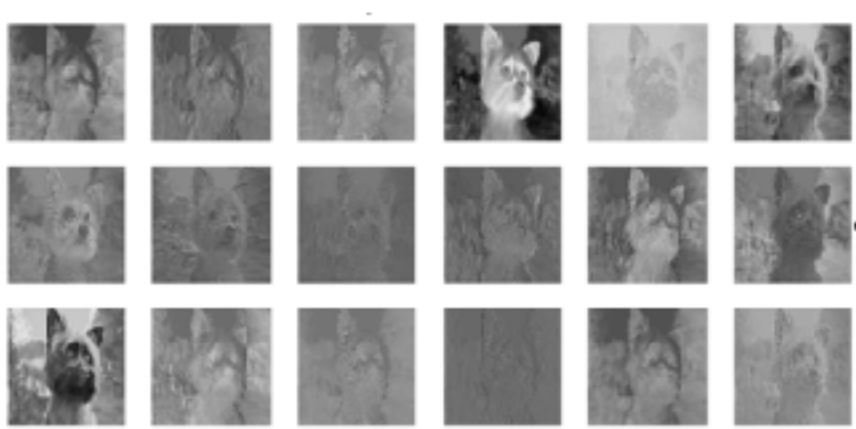
Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi,
Zihang Jiang, Francis EH Tay, Jiashi Feng, Shuicheng Yan

Limitations of ViTs

- ViT tokenization of images (i.e., into patches) makes it harder for the model to learn local structures such as edges, lines, etc.



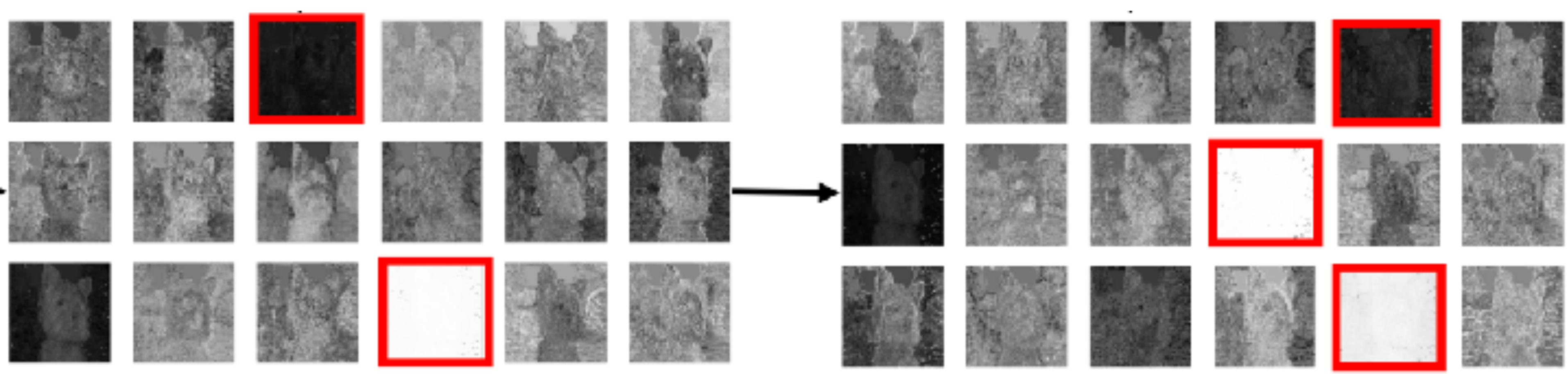
ResNet50 conv1 activations



ViT-L/16 block1 activations

Limitations of ViTs

- ViT backbone is over-parameterized for midsize datasets like ImageNet-1K.
- This leads to redundant features.

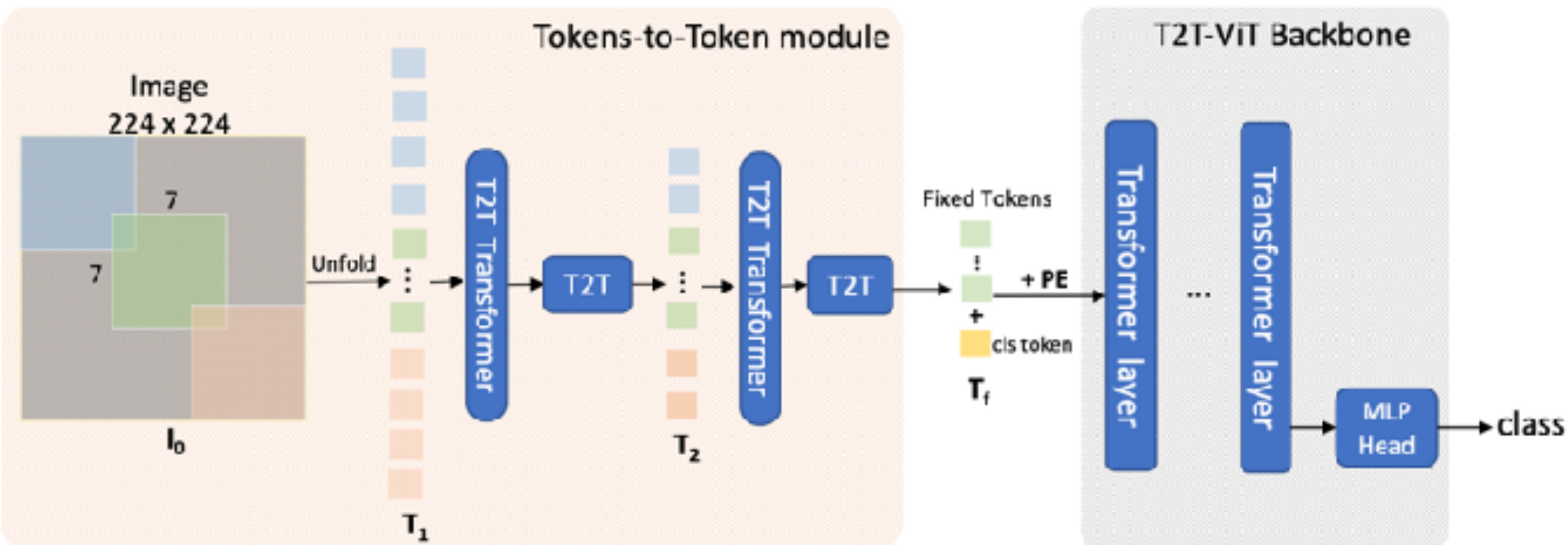


ViT-L/16 block12 activations

ViT-L/16 block24 activations

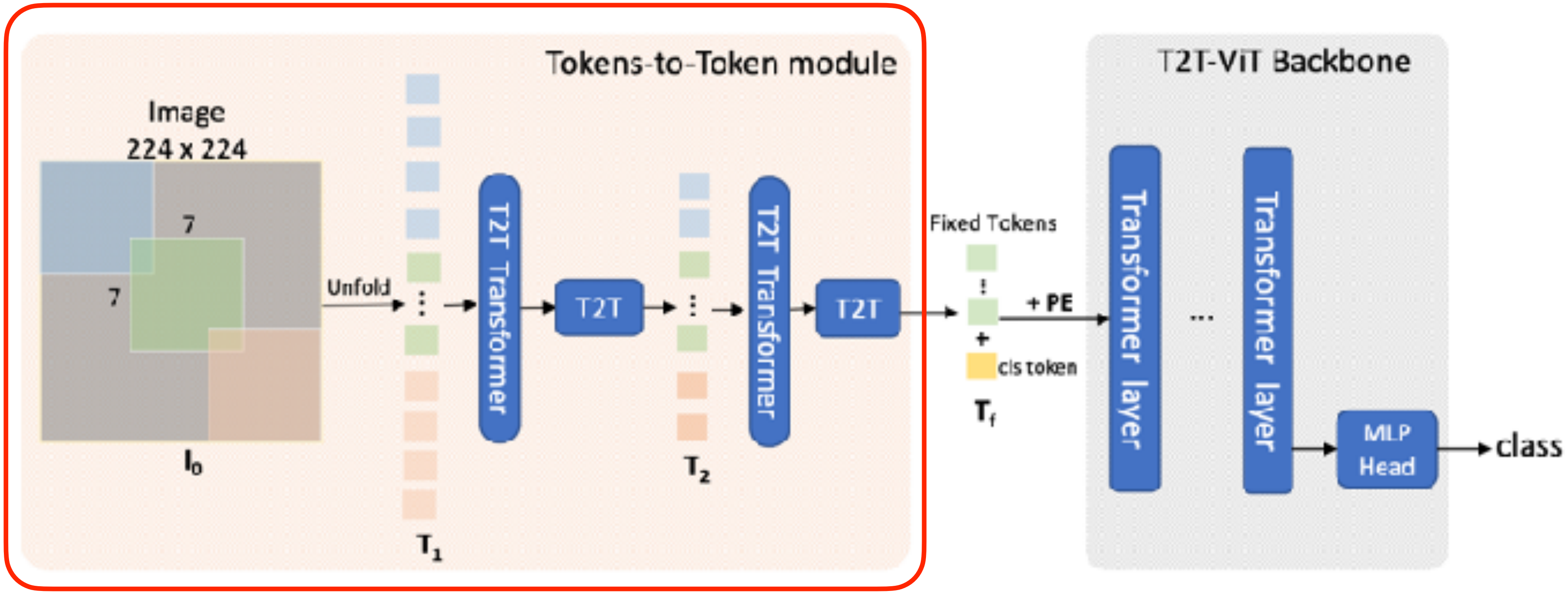
Tokens-to-Token ViT

- The authors propose a layer-wise “Tokens-to-Token module” (T2T) to model local structures in the image.
- Instead of a parameter-heavy backbone, an efficient “T2T-ViT backbone” is used to process the resulting T2T tokens.



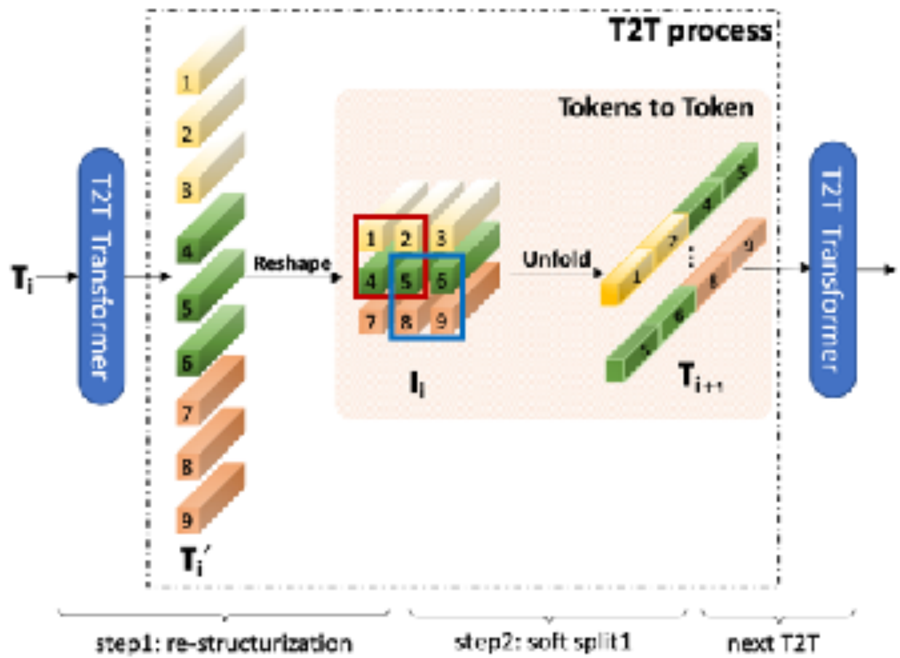
Tokens-to-Token ViT

- The authors propose a layer-wise “Tokens-to-Token module” (T2T) to model local structures in the image.
- Instead of a parameter-heavy backbone, an efficient “T2T-ViT backbone” is used to process the resulting T2T tokens.



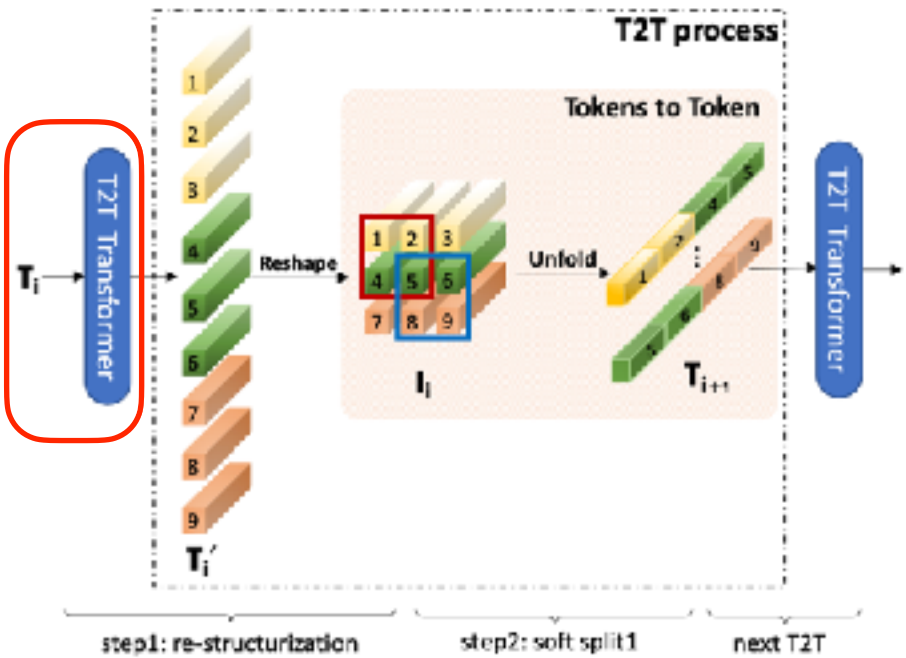
Tokens-to-Token Module

- First, patch-level tokens are processed using self-attention.
- Then, the patch-level tokens are reshaped into a 2D grid.
- Lastly, the authors apply a sliding window operation to aggregate information within local neighborhoods of tokens.



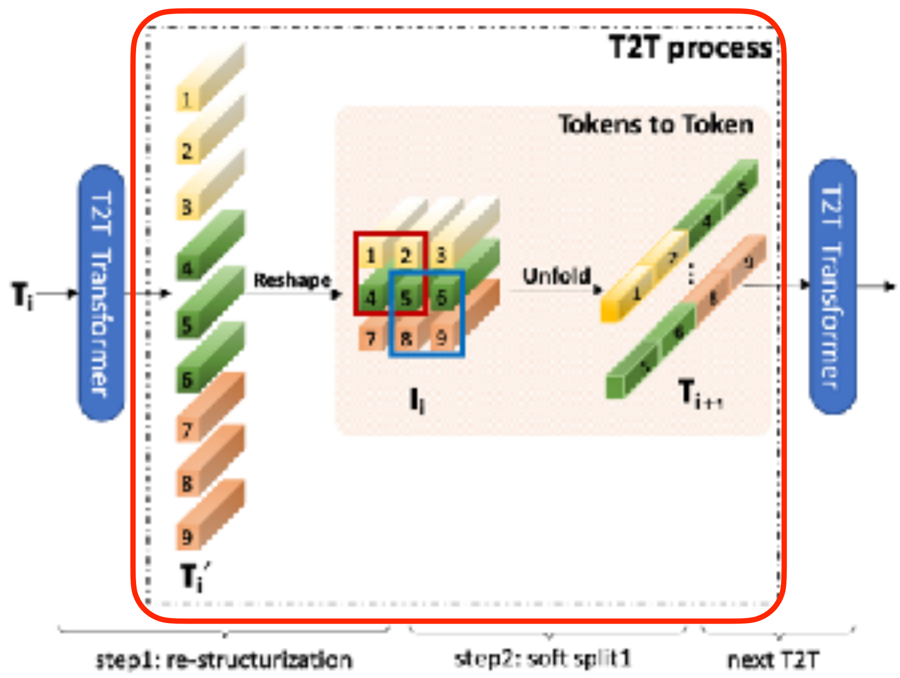
Tokens-to-Token Module

- First, patch-level tokens are processed using self-attention.
- Then, the patch-level tokens are reshaped into a 2D grid.
- Lastly, the authors apply a sliding window operation to aggregate information within local neighborhoods of tokens.



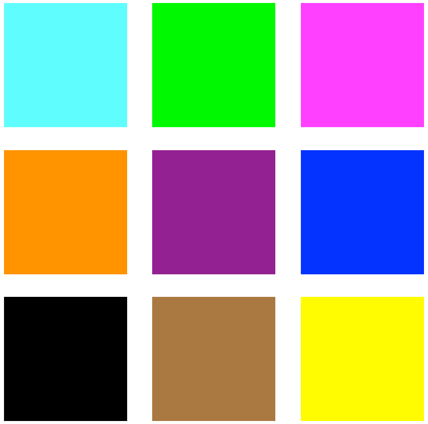
Tokens-to-Token Module

- First, patch-level tokens are processed using self-attention.
- Then, the patch-level tokens are reshaped into a 2D grid.
- Lastly, the authors apply a sliding window operation to aggregate information within local neighborhoods of tokens.



Local Patch Aggregation

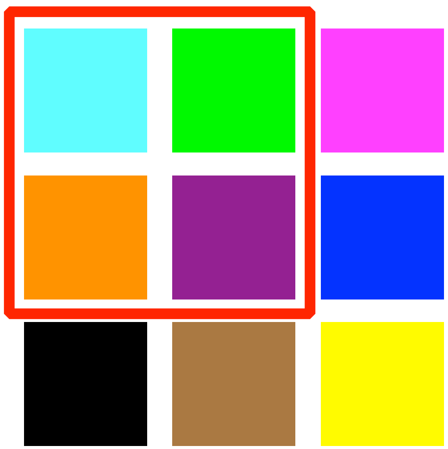
- The patch-level tokens are reshaped back into a 2D grid.
- The sliding window operation is applied on the resulting 2D grid, and the neighboring patches/tokens are concatenated.



a) A 2D grid of d dimensional tokens

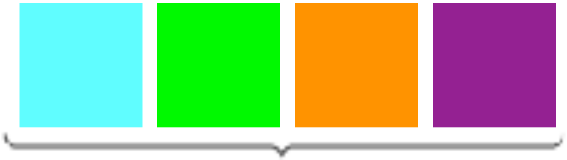
Local Patch Aggregation

- The patch-level tokens are reshaped back into a 2D grid.
- The sliding window operation is applied on the resulting 2D grid, and the neighboring patches/tokens are concatenated.



a) A 2D grid of d dimensional tokens

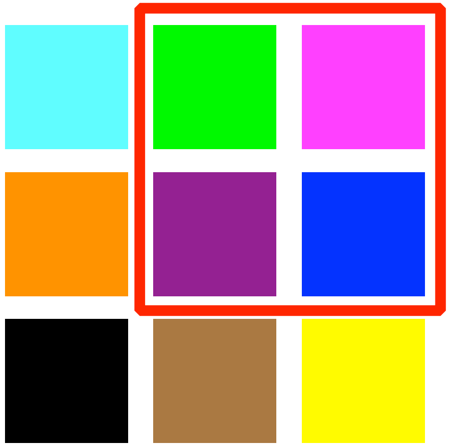
New Token 1



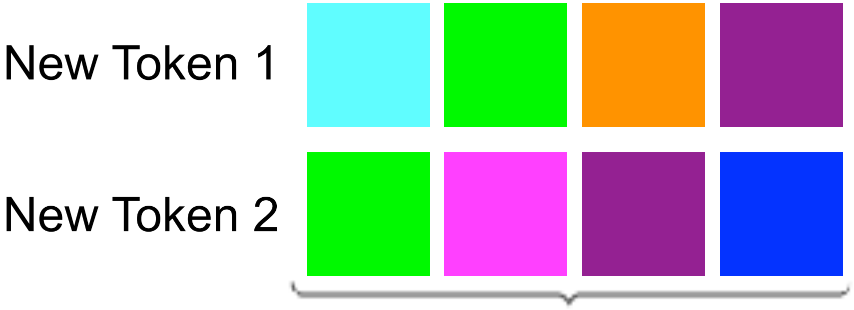
b) Concatenation of the Neighboring Tokens

Local Patch Aggregation

- The patch-level tokens are reshaped back into a 2D grid.
- The sliding window operation is applied on the resulting 2D grid, and the neighboring patches/tokens are concatenated.



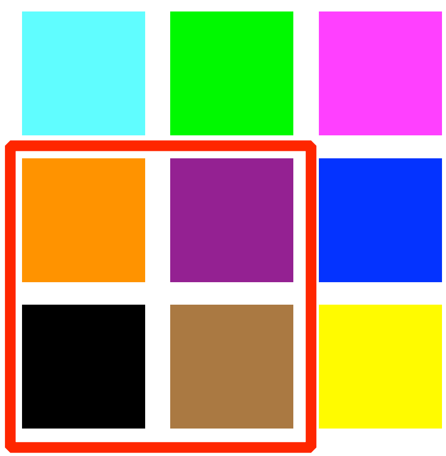
a) A 2D grid of d dimensional tokens



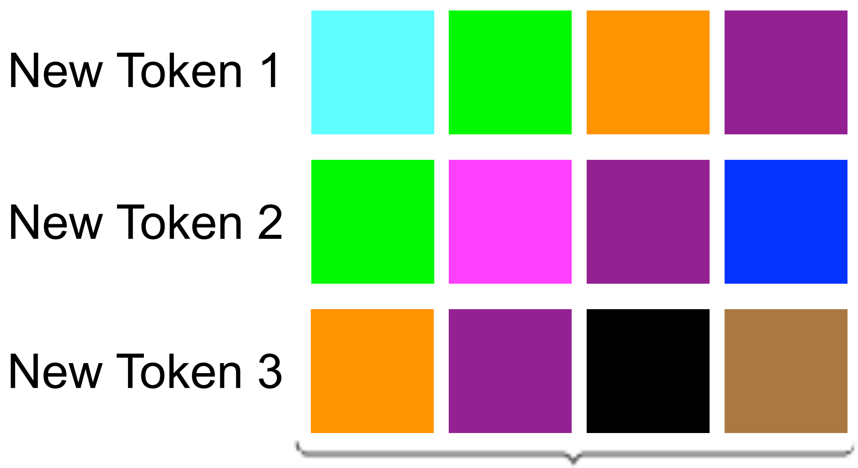
b) Concatenation of the Neighboring Tokens

Local Patch Aggregation

- The patch-level tokens are reshaped back into a 2D grid.
- The sliding window operation is applied on the resulting 2D grid, and the neighboring patches/tokens are concatenated.



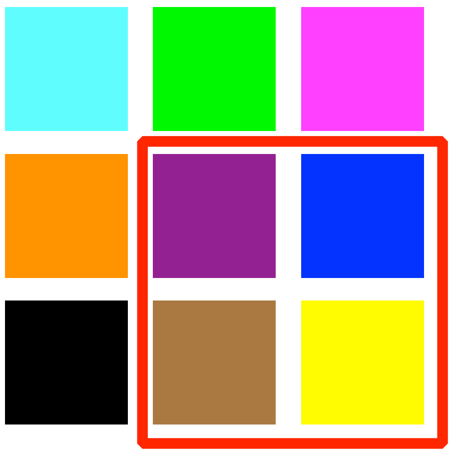
a) A 2D grid of d dimensional tokens



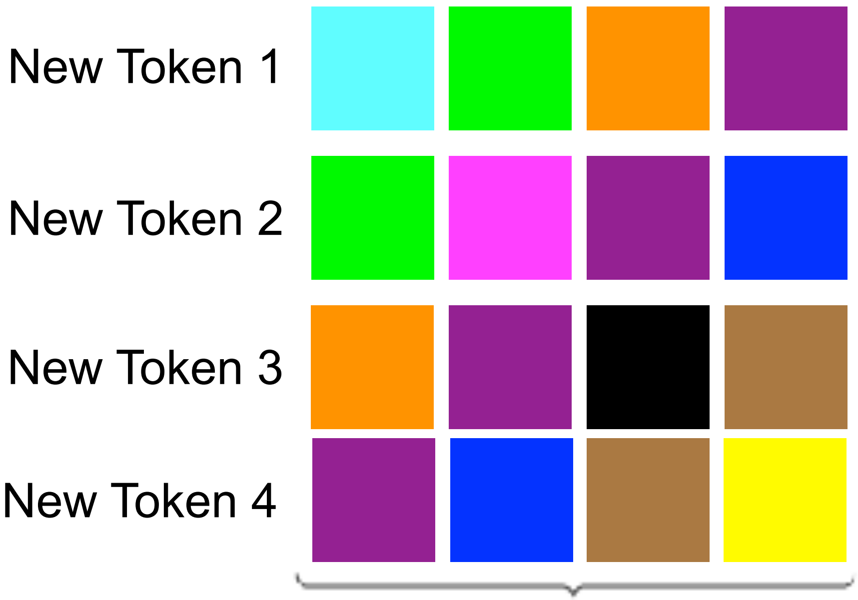
b) Concatenation of the Neighboring Tokens

Local Patch Aggregation

- The patch-level tokens are reshaped back into a 2D grid.
- The sliding window operation is applied on the resulting 2D grid, and the neighboring patches/tokens are concatenated.



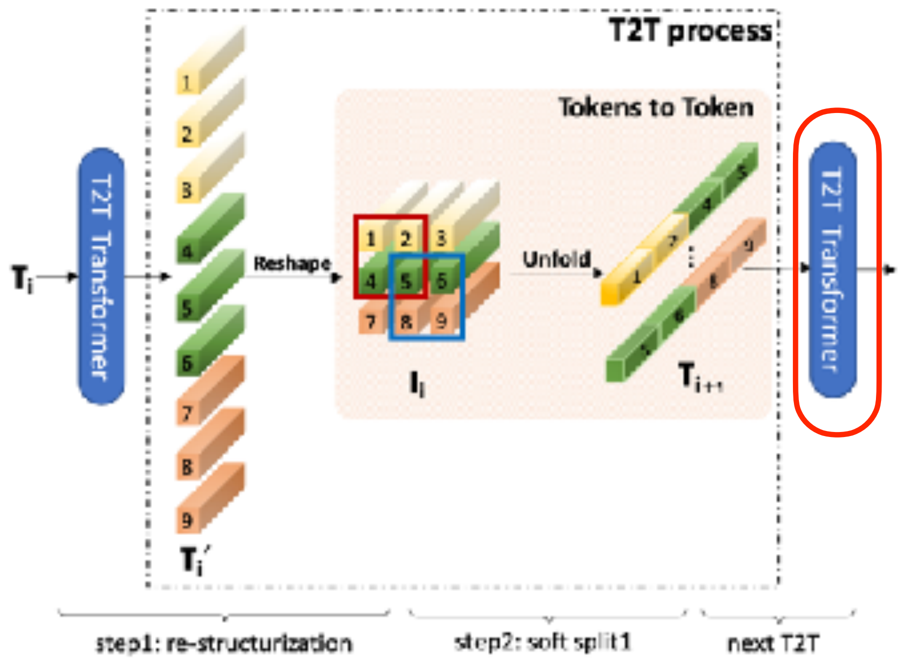
a) A 2D grid of d dimensional tokens



b) Concatenation of the Neighboring Tokens

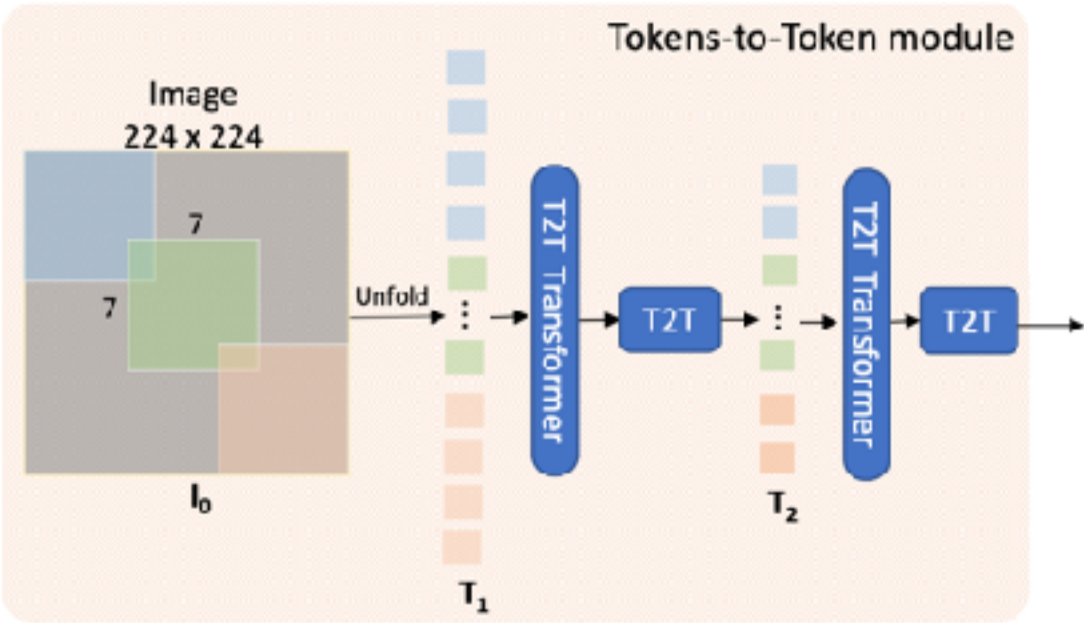
Tokens-to-Token Module

- First, patch-level tokens are processed using self-attention.
- Then, the patch-level tokens are reshaped into a 2D grid.
- Lastly, the authors apply a sliding window operation to aggregate information within local neighborhoods of tokens.



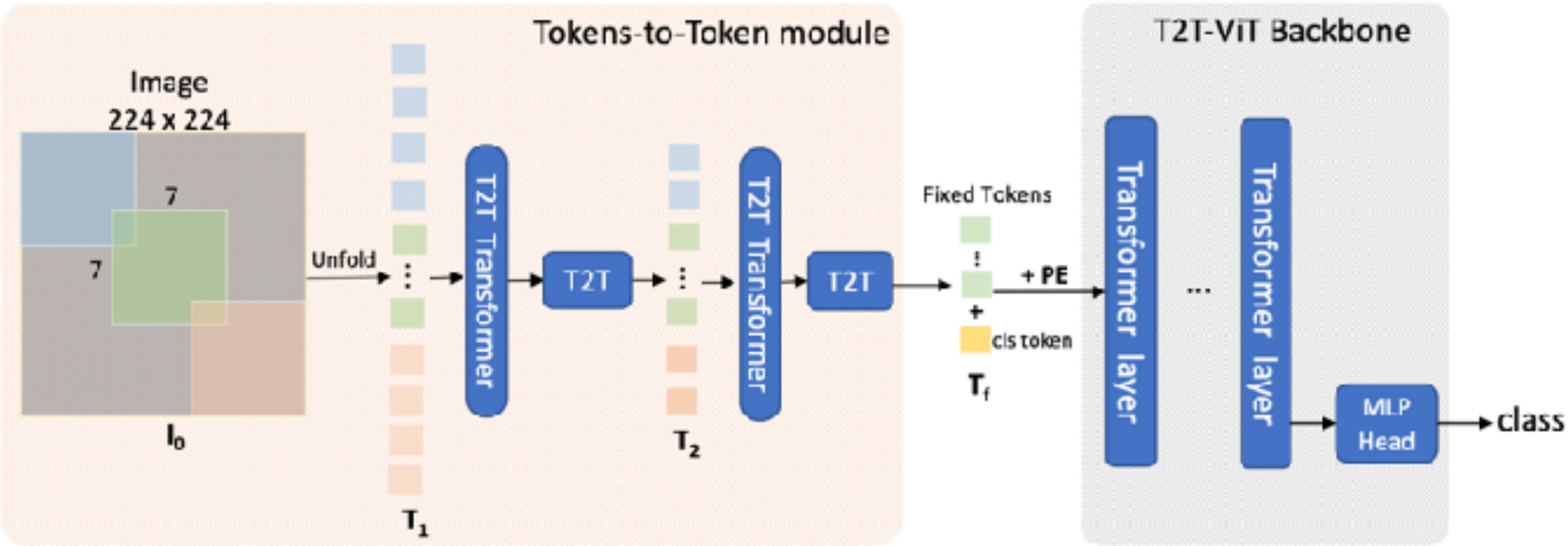
Tokens-to-Token Module

- These two steps can be conducted iteratively as many times as needed.



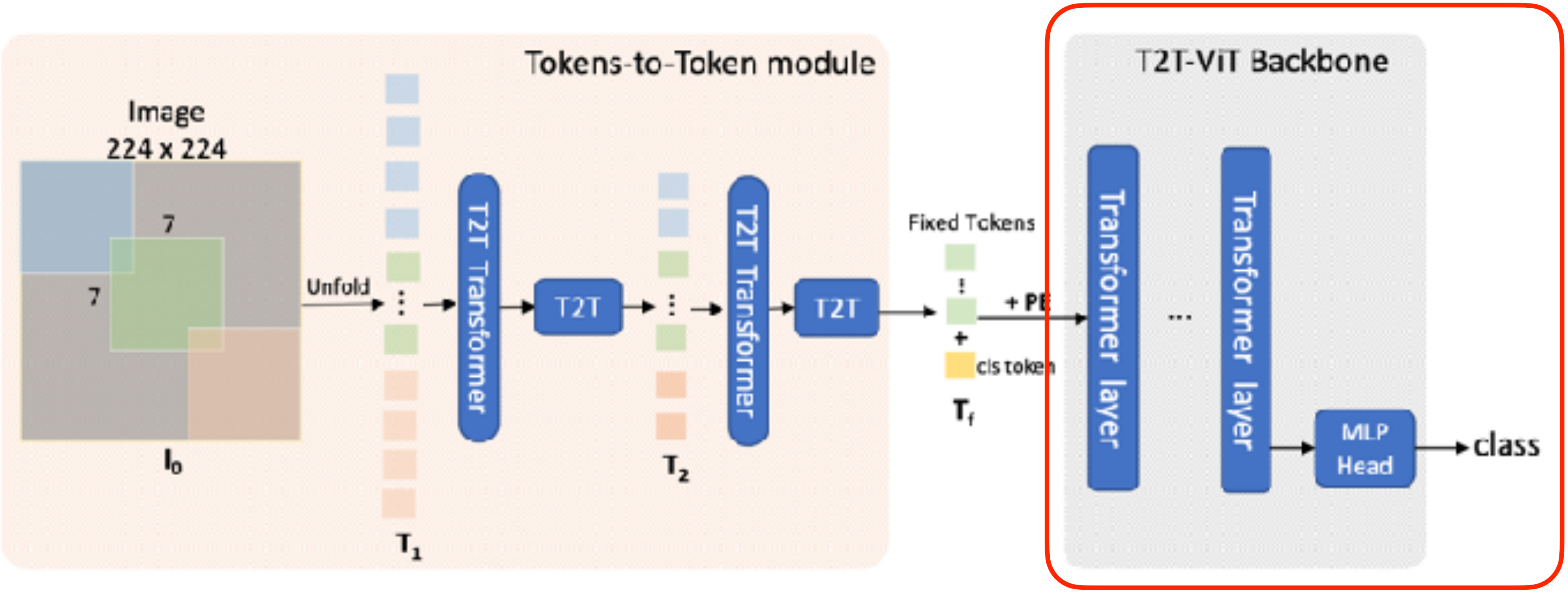
Tokens-to-Token ViT

- The authors propose a layer-wise “Tokens-to-Token module” (T2T) to model local structures in the image.
- Instead of a parameter-heavy backbone, an efficient “T2T-ViT backbone” is used to process the resulting T2T tokens.



Tokens-to-Token ViT

- The authors propose a layer-wise "Tokens-to-Token module" (T2T) to model local structures in the image.
- Instead of a parameter-heavy backbone, an efficient "T2T-ViT backbone" is used to process the resulting T2T tokens.



T2T-ViT vs ViT Architecture

- Unlike ViT, T2T-ViT employs a deep-narrow architecture.

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 vs ViT Architecture

- Unlike ViT, T2T-ViT employs a deep-narrow architecture.

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 vs ViT Architecture

- Unlike ViT, T2T-ViT employs a deep-narrow architecture.

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

Ablation Studies

- Ablation study validating the effectiveness of (1) the T2T module, and (2) Deep-Narrow ViT architecture

Ablation type	Models	Top1-Acc (%)	Params (M)	MACs (G)
T2T module	T2T-ViT-14$_{wo\ T2T}$	79.5	21.1	4.2
	T2T-ViT-14	81.5 (+2.0)	21.5	4.8
DN Structure	T2T-ViT-14	81.5	21.5	4.8
	T2T-ViT-d768-4	78.8 (-2.7)	25.0	5.4

Ablation Studies

- Ablation study validating the effectiveness of (1) the T2T module, and (2) Deep-Narrow ViT architecture

Ablation type	Models	Top1-Acc (%)	Params (M)	MACs (G)
T2T module	T2T-ViT-14$_{wo\ T2T}$	79.5	21.1	4.2
	T2T-ViT-14	81.5 (+2.0)	21.5	4.8
DN Structure	T2T-ViT-14	81.5	21.5	4.8
	T2T-ViT-d768-4	78.8 (-2.7)	25.0	5.4

Ablation Studies

- Ablation study validating the effectiveness of (1) the T2T module, and (2) Deep-Narrow ViT architecture

Ablation type	Models	Top1-Acc (%)	Params (M)	MACs (G)
T2T module	T2T-ViT-14$_{wo\ T2T}$	79.5	21.1	4.2
	T2T-ViT-14	81.5 (+2.0)	21.5	4.8
DN Structure	T2T-ViT-14	81.5	21.5	4.8
	T2T-ViT-d768-4	78.8 (-2.7)	25.0	5.4

Qualitative Results

- The proposed approach addresses the two previously discussed limitations of ViTs.