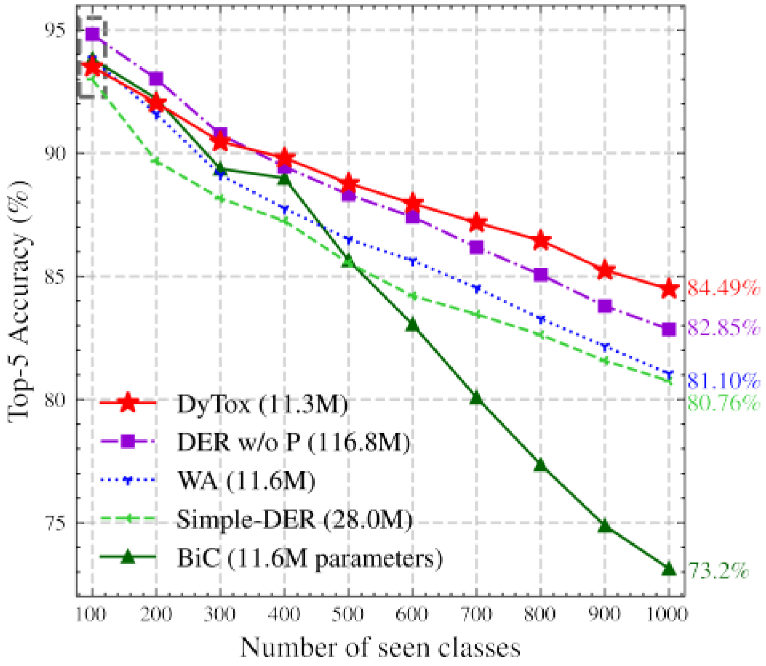
DyTox: Transformers for Continual Learning with DYnamic TOken eXpansion (CVPR 2022)

- Young Jo Choi
- Department of Digital Analytics

Abstract

- Transformer를 Incremental learning 분야에 처음으로 적용
- ViT와 다르게 Encoder / Decoder를 다 씀



Continual Learning

Transformers (ViT)

Transformers (ViT)

Token
Embedding
수식

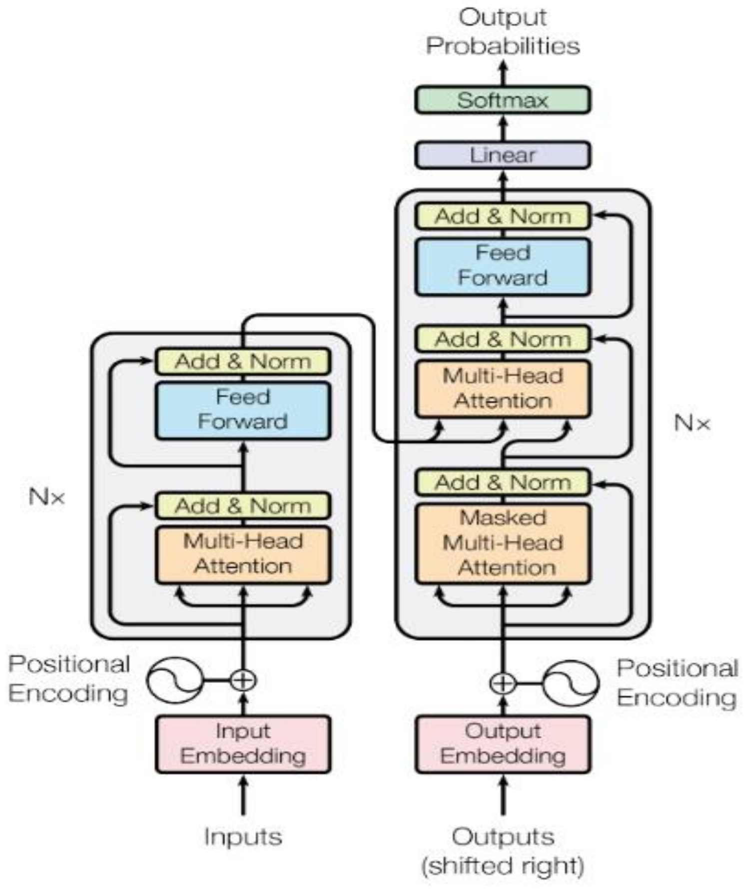
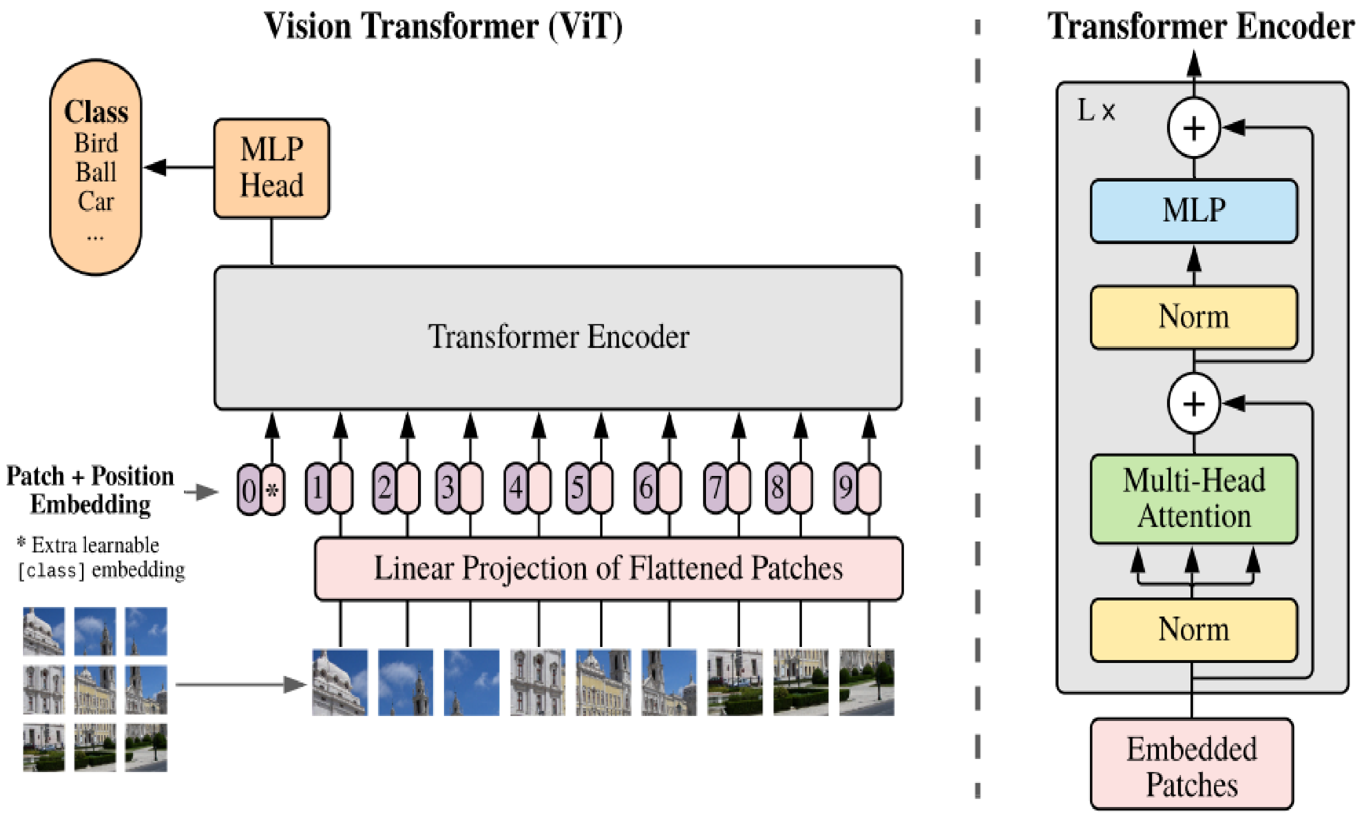


Figure 1: The Transformer - model architecture.



Structure

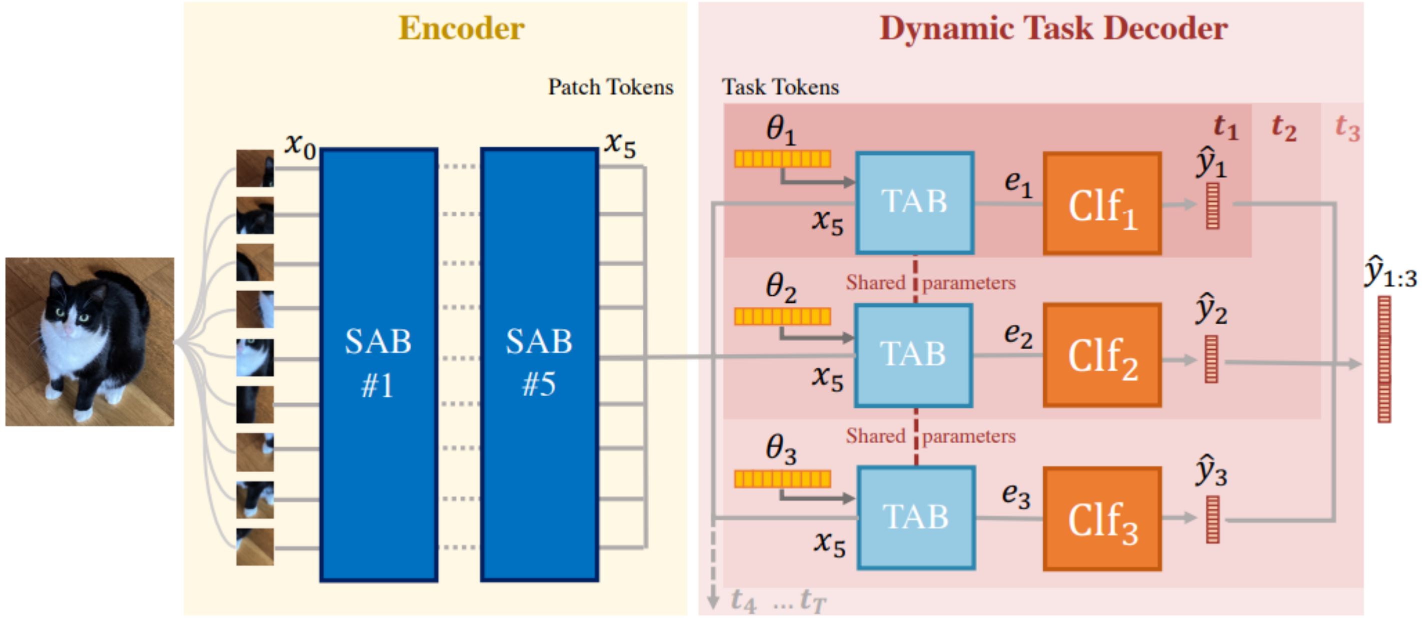


Figure 2: **DyTox transformer model**. An image is first split into multiple patches, embedded with a linear projection. The resulting patch tokens are processed by 5 successive Self-Attention Blocks (SAB) (Sec. 3.1). For each task ($t = 1 \dots T$), the processed patch tokens are then given to the Task-Attention Block (TAB) (Sec. 3.2): each forward through the TAB is modified by a different task-specialized token $\boldsymbol{\theta}_t$ for $t \in \{1 \dots T\}$ (Sec. 3.3). The T final embeddings are finally given separately to independent classifiers Clf_t each predicting their task's classes C^t. All $|C^{1:T}|$ logits are activated with a sigmoid. For example, at task $t = 3$, one forward is done through the SABs and three task-specific forwards through the unique TAB.

Structure

- Patch tokenizer
- Self-Attention (SA) based encoder
- Task-Attention Block

Symbol	Meaning
(x_i^t, y_i^t)	Input sample & its label from the t^{th} task
C^t	Label set of the t^{th} task
$C^{1:t}$	All labels from all seen tasks
$\boldsymbol{\theta}_t$	Task token of the t^{th} task
Clf_t	Independent classifier of the t^{th} task
SAB_l	l^{th} Self-Attention Block
TAB	Task-Attention Block

Structure

- Dynamic task token expansion (task specific)

Algorithm 1 DyTox's forward pass at step t

Input: x_0 (initial patch tokens), y (ground-truth labels)
Output: $\hat{y}_{1:t}$ (predictions for all classes of $\mathcal{C}^{1:t}$)

$x_L \leftarrow \mathrm{SAB}_{l=L} \circ \ldots \mathrm{SAB}_{l=1}(x_0)$ ▷ Sec. 3.1
for $i \leftarrow 1$; $i \leq t$; i++ **do**
  $e_i \leftarrow \mathrm{TAB}([\boldsymbol{\theta}_i, x_L])$ ▷ Sec. 3.2
  $\hat{y}_i \leftarrow \mathrm{Clf}_i(e_i)$ ▷ Sec. 3.3
end for
$\hat{y}_{1:t} \leftarrow [\hat{y}_1, \ldots, \hat{y}_t]$

Structure

- Context
- Losses

Structure

Experiments

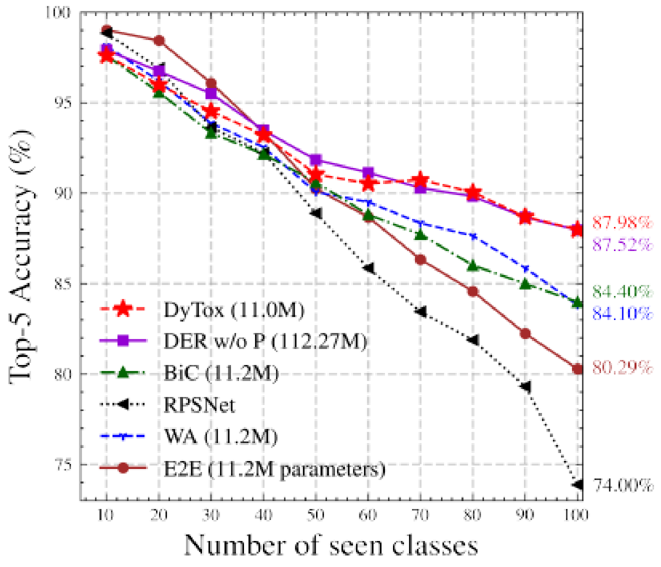
- details

Hyperparameter	CIFAR	ImageNet
# SAB	5	
# CAB	1	
# Attentions Heads	12	
Embed Dim	384	
Input Size	32	224
Patch Size	4	16

Table 1: **DyTox's architectures** for CIFAR and ImageNet. The only difference between the two architectures is the patch size, as the image sizes vary between datasets.

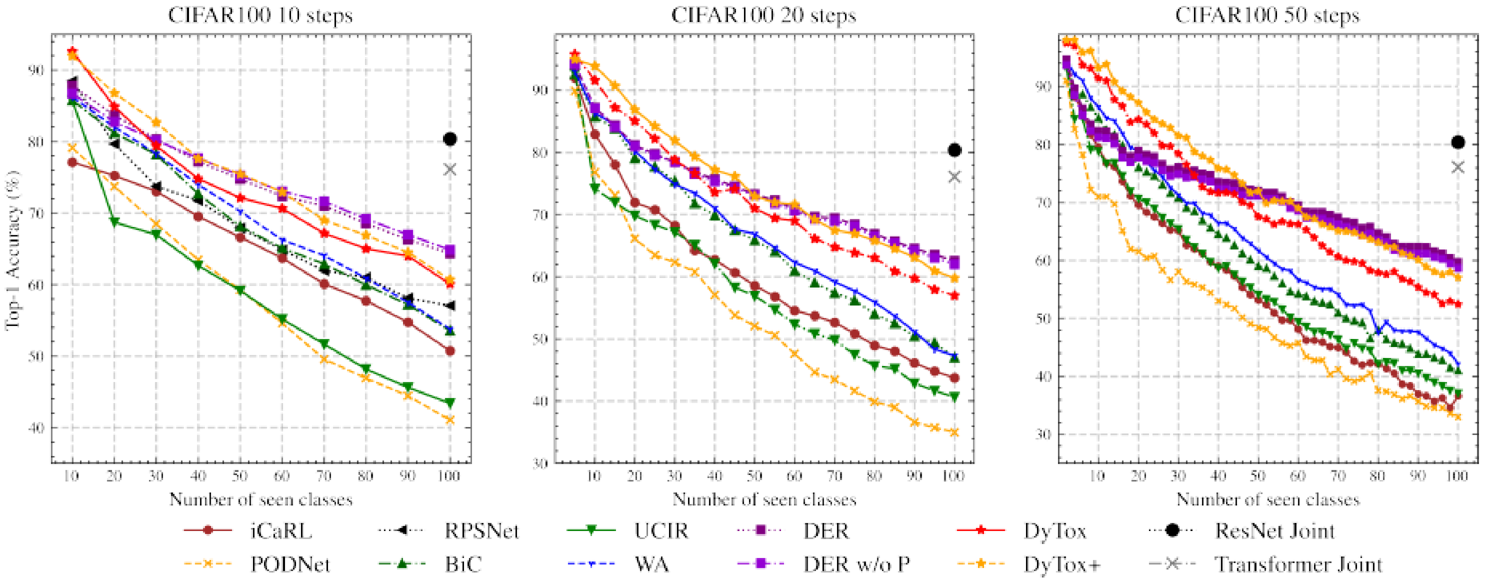
Experiments - Result

Methods	ImageNet100 10 steps #P	top-1 Avg	top-1 Last	top-5 Avg	top-5 Last	ImageNet1000 10 steps #P	top-1 Avg	top-1 Last	top-5 Avg	top-5 Last
ResNet18 joint	11.22	-	-	-	95.10	11.68	-	-	-	89.27
Transf. joint	11.00	-	79.12	-	93.48	11.35	-	73.58	-	90.60
E2E [5]	11.22	-	-	89.92	80.29	11.68	-	-	72.09	52.29
Simple-DER [48]	-	-	-	-	-	28.00	66.63	59.24	85.62	80.76
iCaRL [59]	11.22	-	-	83.60	63.80	11.68	38.40	22.70	63.70	44.00
BiC [32]	11.22	-	-	90.60	84.40	11.68	-	-	84.00	73.20
WA [81]	11.22	-	-	91.00	84.10	11.68	65.67	55.60	86.60	81.10
RPSNet [56]		-	-	87.90	74.00	-	-	-	-	-
DER w/o P [76]	112.27	**77.18**	66.70	**93.23**	87.52	116.89	68.84	60.16	88.17	82.86
DER† [76]	-	76.12	66.06	92.79	88.38	-	66.73	58.62	87.08	81.89
DyTox	11.01	77.15	69.10	92.04	87.98	11.36	71.29	63.34	88.59	84.49



Experiments - Result

Methods	10 steps #P	10 steps Avg	10 steps Last	20 steps #P	20 steps Avg	20 steps Last	50 steps #P	50 steps Avg	50 steps Last
ResNet18 Joint	11.22	-	80.41	11.22	-	81.49	11.22	-	81.74
Transf. Joint	10.72	-	76.12	10.72	-	76.12	10.72	-	76.12
iCaRL [59]	11.22	65.27 $\pm$ 1.02	50.74	11.22	61.20 $\pm$ 0.83	43.75	11.22	56.08 $\pm$ 0.83	36.62
UCIR [32]	11.22	58.66 $\pm$ 0.71	43.39	11.22					
BiC [75]	11.22	68.80 $\pm$ 1.20	53.54	11.22					
WA [81]	11.22	69.46 $\pm$ 0.29	53.78	11.22					
PODNet [19]	11.22	58.03 $\pm$ 1.27	41.05	11.22					
RPSNet [56]	56.5	68.60	57.05	-					
DER w/o P [76]	112.27	75.36 $\pm$ 0.36	65.22	224.5					
DER† [76]	-	74.64 $\pm$ 0.28	64.35	-					
DyTox	10.73	73.66 $\pm$ 0.02	60.67 $\pm$ 0.34	10.74					
DyTox+	10.73	75.54 $\pm$ 0.10	62.06 $\pm$ 0.25	10.74					



Improved training procedure

	1 step	50 steps	
Training	**Last** (↑)	**Last** (↑)	**Forgetting** (↓)
DyTox	76.12	52.34	33.15
DyTox+	77.51+1.39	57.09+4.75	31.50-1.65

Overhead

- Memory overhead
- Computational overhead
- Training procedure introspection

Ablation study

		Knowledge Distillation	Finetuning	Token Expansion	Divergence Classifier	Indendepent Classifiers	**Avg**	**Last**
DyTox	**Transformer**						60.69	38.87
		✓					61.62	39.35
		✓	✓				63.42	42.21
	Dynamic	✓	✓	✓			67.30	47.57
		✓	✓	✓	✓		68.28	49.45
		✓	✓	✓	✓	✓	70.20	52.34

Conclusion