Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?

Bill Psomas[1,2], Ioannis Kakogeorgiou[1], Konstantinos Karantzalos[1], Yannis Avrithis[2]
[1]National Technical University of Athens; [2]Institute of Advanced Research in Artificial Intelligence (IARAI)
psomasbill@mail.ntua.gr

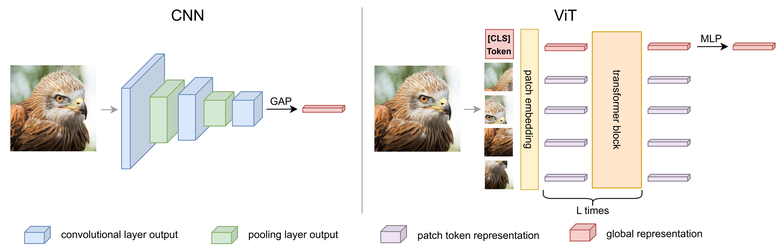
ICCV23 PARIS

IARAI institute of advanced research in artificial intelligence

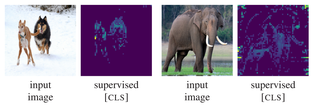
Introduction

Motivation

- Convolutional networks vs. vision transformers: different poolings?



- Vision transformers: [CLS] necessity?
- Supervised vision transformers: low-quality spatial attention?



Focus

- Pooling for both network types improving over default?
- Pooling for high-quality spatial attention?
- Validity in both supervised & self-supervised setting?

Approach

- Framework to inspect & compare pooling methods.
- Introduce SimPool: attention-based, universal pooling for robust vector representation.

Landscape of Poolings

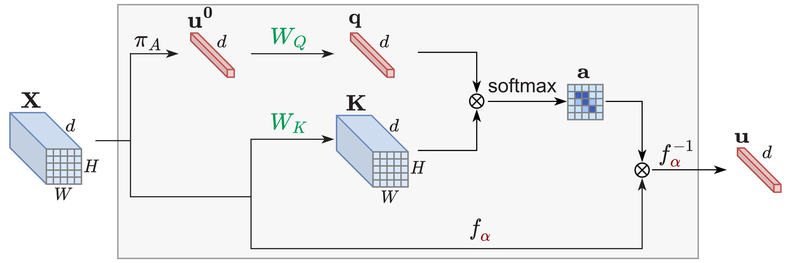
#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x},\mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)
1	GAP [16]	✓		1					$\mathbf{1}_p/p$	X	$f_{-1}(x)$		Z	55.0
	max [2]			1					$\mathbf{1}_p$	X	$f_{-\infty}(x)$		Z	53.9
	GeM [3]			1					$\mathbf{1}_p/p$	X	$f_\alpha(x)$		Z	55.9
	LSE [4]	✓		1					$\mathbf{1}_p/p$	X	e^{rx}		Z	55.3
	HOW [5]			1					$\mathrm{diag}(X^\top X)$	$\mathrm{FC}(\mathrm{avg}_3(X))$	$f_{-1}(x)$		Z	54.8
2	OTK [6]	✓		k	U	U	X	$-\|\mathbf{x}-\mathbf{y}\|^2$	$\mathrm{SINKHORN}(e^{S/\epsilon})$	$\psi(X)$	$f_{-1}(x)$		Z	55.9
	k-means		✓	k	random	U	X	$-\|\mathbf{x}-\mathbf{y}\|^2$	$\eta_2(\arg\max_1(S))$	X	$f_{-1}(x)$	X	Z	55.4
	Slot [7]*	✓	✓	k	U	$W_Q U$	$W_K X$	$\mathbf{x}^\top\mathbf{y}$	$\sigma_2(S/\sqrt{d})$	$W_V X$	$f_{-1}(x)$	X	$\mathrm{MLP}(\mathrm{GRU}(Z))$	56.7
3	SE [8]	✓		1	$\pi_A(X)$	$\sigma(\mathrm{MLP}(U))$				$\mathrm{diag}(\mathbf{q})X$		V		55.7
	CBAM [9]*	✓		1	$\pi_A(X)$	$\sigma(\mathrm{MLP}(U))$	X	$\mathbf{x}^\top\mathbf{y}$	$\sigma(\mathrm{conv}_7(S))$	$\mathrm{diag}(\mathbf{q})X$		$V\,\mathrm{diag}(\mathbf{a})$		55.6
4	ViT [10]*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\sigma_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	$\mathrm{MLP}(\mathrm{MSA}(X))$	$\mathrm{MLP}(g_m^{-1}(Z))$	56.0
	CaiT [11]*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\sigma_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	X	$\mathrm{MLP}(g_m^{-1}(Z))$	56.7
5	SimPool	✓		1	$\pi_A(X)$	$W_Q U$	$W_K X$	$\mathbf{x}^\top\mathbf{y}$	$\sigma_2(S/\sqrt{d})$	$X-\min X$	$f_\alpha(x)$		Z	57.1

1 Develop a generic pooling framework.
2 Formulate methods as instantiations.
3 Discuss the properties.
4 Derive SimPool.
5 Benchmark.

References

[1] He et al. Deep residual learning for image recognition *CVPR*, 2016.
[2] Tolias et al. Particular object retrieval with integral max-pooling of CNN activations *ICLR*, 2016.
[3] Radenović et al. Fine-Tuning CNN Image Retrieval with No Human Annotation *PAMI*, 2018.
[4] Pinheiro et al. From image-level to pixel-level labeling with convolutional networks *CVPR*, 2015.
[5] Tolias et al. Learning and aggregating deep local descriptors for instance-level recognition *ECCV*, 2020.
[6] Mialon et al. A trainable optimal transport embedding for feature aggregation and its relationship to attention *arXiv*, 2020.
[7] Locatello et al. Object-centric learning with slot attention *NeurIPS*, 2020.
[8] Hu et al. Squeeze-and-Excitation Networks *CVPR*, 2018.
[9] Woo et al. Cbam: Convolutional block attention module *ECCV*, 2018.
[10] Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale *ICLR*, 2021.
[11] Touvron et al. Going deeper with Image Transformers *ICCV*, 2021.
[12] Lee et al. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree *AISTATS*, 2016.
[13] Hu et al. Gather-excite: Exploiting feature context in convolutional neural networks *NeurIPS*, 2018.
[14] Siméoni et al. Localizing Objects with Self-Supervised Transformers and no Labels *BMVC*, 2021.
[15] Caron et al. Emerging properties in self-supervised vision transformers *ICCV*, 2021.
[16] He et al. Deep residual learning for image recognition *CVPR*, 2016.
[17] Liu et al. A convnet for the 2020s *CVPR*, 2022.
[18] Deng et al. Imagenet: A large-scale hierarchical image database *CVPR*, 2009.
[19] Wah et al. The Caltech-UCSD Birds-200-2011 Dataset *CIT*, 2011.
[20] Everingham et al. The pascal visual object classes (voc) challenge *IJCV*, 2009.
[21] Vo et al. Toward unsupervised, multi-object discovery in large-scale image collections *ECCV*, 2020.

SimPool: A simple attention-based universal pooling



- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by W_Q (W_K) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2\left(K^\top\mathbf{q}/\sqrt{d}\right)$.
- Global representation: $\mathbf{u} = \pi_{\mathrm{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Property: Universal (networks & settings)

Method	Ep	ResNet-50	ConvNeXt-S	ViT-S
Baseline	100	77.4	81.1	72.7
CaiT [11]	100	77.3	81.2	72.6
Slot [7]	100	77.3	80.9	72.9
GE [13]	100	77.6	81.3	72.6
SimPool	100	**78.0**	**81.7**	**74.3**
Baseline	300	78.1†	83.1	77.9
SimPool	300	**78.7**†	**83.5**	**78.7**

Image classification accuracy on ImageNet-1k [18]. Left: Supervised; Below: self-supervised [15]. Baseline: GAP for convolutional, [CLS] for transformers.

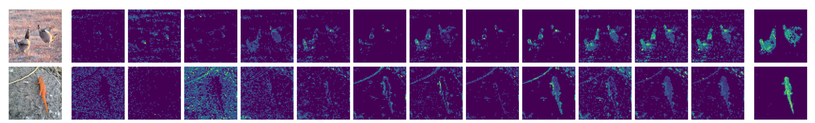
Method	Ep	ResNet-50		ConvNeXt-S		ViT-S	
		k-NN	Prob	k-NN	Prob	k-NN	Prob
Baseline	100	61.8	63.0	65.1	68.2	68.9	71.5
SimPool	100	**63.8**	**64.4**	**68.8**	**72.2**	**69.8**	**72.8**

Property: High-Quality Attention Maps from Transformers

Attention maps of ViT-S trained on ImageNet-1k. For baseline, we use the mean attention map of the [CLS] token. For SimPool, we use the attention map $\mathbf{a}$.

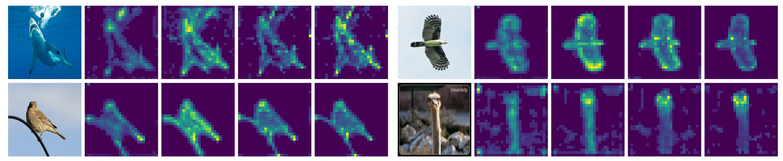
Property: Resolving the Attention Deficit

[CLS] vs. SimPool. Attention maps of ViT-T trained on ImageNet-1k for 100 epochs under supervision. For [CLS], we use the mean attention map of the [CLS] token of each block. For SimPool, we use the attention map $\mathbf{a}$.



Property: High-Quality Attention Maps from CNNs

Attention maps of ResNet-50 [16] and ConvNeXt-S [17] trained on ImageNet-1k under supervision and self-supervision [15]. We use the attention map $\mathbf{a}$.



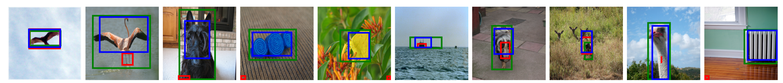
Property: Localization

Left: Object localization MaxBoxAccV2 on CUB [19] test and ImageNet-1k validation set with ViT-S.
Right: Object discovery CorLoc on VOC [20] trainval set and COCO 20k [21] with self-supervised ViT-S.

Method	Supervised		Self-Supervised	
	CUB	ImageNet	CUB	ImageNet
Baseline	63.1	53.6	82.7	62.0
SimPool	**77.9**	**64.4**	**86.1**	**66.1**
Baseline@20	62.4	50.5	65.5	52.5
SimPool@20	**74.0**	**62.6**	**72.5**	**58.7**

Method	DINO-seg [14, 15]			LOST [14]		
	VOC07	VOC12	COCO	VOC07	VOC12	COCO
Baseline	30.8	31.0	36.7	55.5	59.4	46.6
SimPool	**53.2**	**56.2**	**43.4**	**59.8**	**65.0**	**49.4**
Baseline@20	14.9	14.8	19.9	50.7	56.6	40.9
SimPool@20	**49.2**	**54.8**	**37.9**	**53.9**	**58.8**	**46.1**

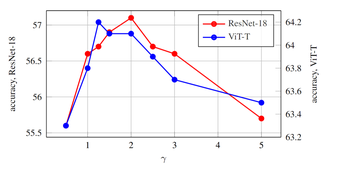
Object localization on ImageNet-1k with supervised ViT-S. Green: ground-truth; red: baseline, blue: SimPool.



Efficiency & Ablation

Network	Pooling	Depth	Init	Accuracy	#params
Base	GAP	12	12	73.3	22.1M
Base	[CLS]	12	0	72.7	22.1M
Base + 1		13	0	73.2	23.8M
Base + 2		14	0	73.7	25.6M
Base + 3		15	0	73.8	27.4M
Base + 4		16	0	73.9	29.2M
Base + 5		17	0	**74.6**	30.9M
Base	SimPool	12	12	**74.3**	22.3M
Base − 1		11	11	73.9	20.6M
Base − 2		10	10	73.6	18.7M
Base − 3		9	9	72.5	17.0M

Left: Trade-off between performance and parameters. Supervised ViT-S on ImageNet-1k. Base+b (Base−b): b blocks added to (removed from) the network.
Below: Image classification accuracy on ImageNet-1k vs. exponent $\gamma = (1-\alpha)/2$ for ResNet-18 and ViT-T.



Conclusion

- Improves performance of convolutional networks and transformers under supervised or self-supervised setting.
- Outperforms the other pooling methods.
- Produces high-quality attention maps that delineate object boundaries.
- Presents strong localization properties.
- Discards the traditional [CLS] from vision transformers.