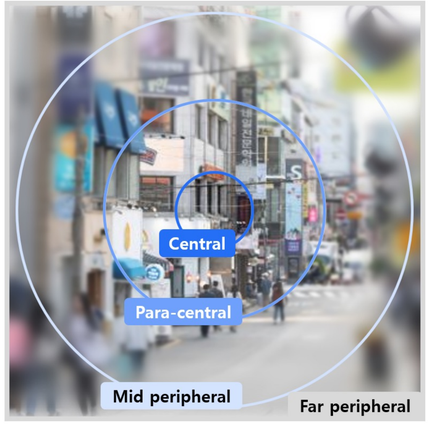
Peripheral Vision Transformer

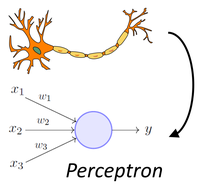
Juhong Min[1] Yucheng Zhao[2,3] Chong Luo[2] Minsu Cho[1]
POSTECH[1] MSRA[2] USTC[3]

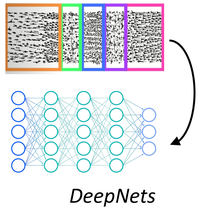
Code | arXiv | Project page

Introduction

Human vision { **Foveal vision** for *local* details in high-resolution
Peripheral vision for *global* perception of the view







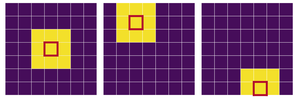
Imitations of biological design have shown their efficacy in ML

What do we miss in learning visual representation?

Peripheral vision

Feature transform in machine vision

Convolution: dominant for the last decade – *local & static*



Pros: requires less training data

Cons: local & static transform

Self-attention: rising transform in vision – *global & dynamic*

Pros: global & dynamic transform

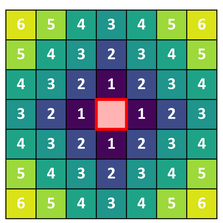
Cons: require more training data

Modeling peripheral vision naturally resolves the both limitations

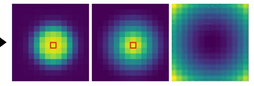
Proposed approach

Modelling peripheral vision: a Roadmap

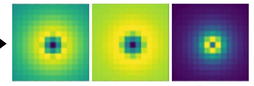
: feature to transform, *i.e.*, query at position $\mathbf{q} \in \mathbb{R}^2$



single-layer(1) $\Phi_{\mathrm{p}}^{(h)}(\mathbf{R}) = \sigma\left[\mathbf{R}\mathbf{W}_{\mathrm{p}}^{(h)}\right]$



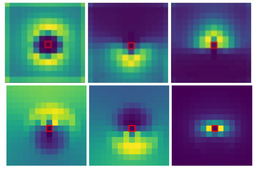
multi-layer(2) $\Phi_{\mathrm{p}}^{(h)}(\mathbf{R}) = \sigma\left[\mathrm{Linear}(\mathrm{ReLU}(\mathrm{Linear}(\mathbf{R};\mathbf{W}_{\mathrm{p1}}))\mathbf{W}_{\mathrm{p2}}^{(h)}\right)$



1-6: relative distances between query and keys ($\|\mathbf{q}-\mathbf{k}\|_2$ where $\mathbf{q},\mathbf{k} \in \mathbb{R}^2$)

(1) & (2): not desirable; most real-world objects are **not rotational symmetric**

Peripheral projections



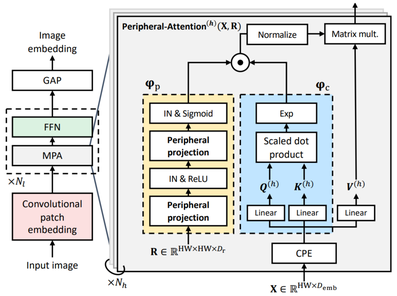
effectively ***breaks rotational symmetric property*** while ***preserving peripheral design*** to a sufficient extent

$$\Phi_{\mathrm{p}}^{(h)}(\mathbf{R})_{\mathbf{q},\mathbf{k},:} := \sigma\left[\sum_{\mathbf{n}\in\mathcal{N}(\mathbf{k})} \mathrm{ReLU}\left(\sum_{\mathbf{m}\in\mathcal{N}(\mathbf{q})} \mathbf{R}_{\mathbf{q},\mathbf{m},:}\mathbf{W}_{\mathrm{p1}\ \mathbf{m}-\mathbf{k},:,:}\right)\mathbf{W}_{\mathrm{p2}\ \mathbf{n}-\mathbf{k},:}^{(h)}\right]$$

Peripheral Vision Transformer (PerViT)

$$\text{Peripheral-Attention}^{(h)}(\mathbf{X},\mathbf{R}) := \text{Normalize}\left[\Phi_{\mathrm{c}}^{(h)}(\mathbf{X}) \odot \Phi_{\mathrm{p}}^{(h)}(\mathbf{R})\right]\mathbf{V}^{(h)}$$

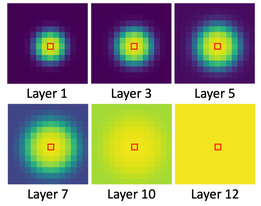
$$\mathrm{MPA}(\mathbf{X}) := \underset{h\in[N_h]}{\mathrm{concat}}\left[\text{Peripheral-Attention}^{(h)}(\mathbf{X},\mathbf{R})\right]\mathbf{W}_{\mathrm{out}} + \mathbf{b}_{\mathrm{out}}$$



Peripheral initialization

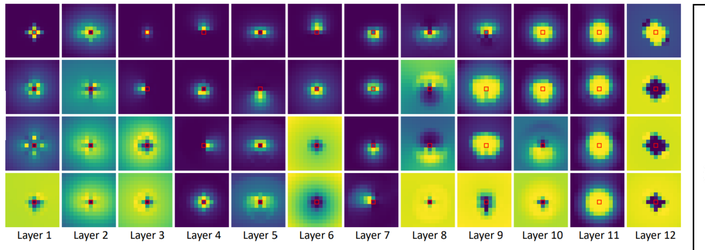
shallow layers: local attention

deep layers: global attention



Experimental results & analyses

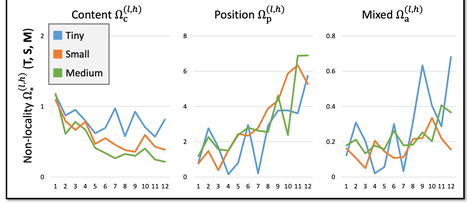
Analyses on learned position- and content-based attentions

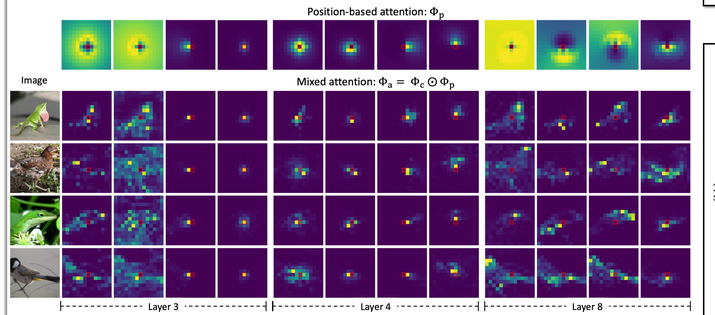


Many of learned attention focuses on central region, processing details

Non-locality measure

$$\Omega_{\mathrm{p}}^{(l,h)} := \frac{1}{|\mathcal{P}|^2}\sum_{(\mathbf{q},\mathbf{k})\in\mathcal{P}\times\mathcal{P}} \Phi_{\mathrm{p}\ \mathbf{q},\mathbf{k}}^{(l,h)}\mathbf{R}_{\mathbf{q},\mathbf{k}}^{\mathrm{euc}}$$



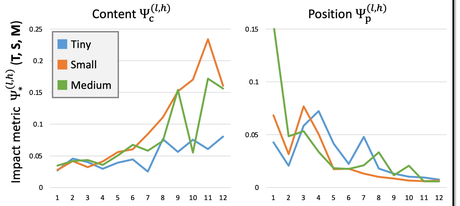


Performs local/static & global/dynamic transforms ***in a single-shot***

PerViT ***exploits benefits of both convolution and self-attention***

Impact measure

$$\Psi_{\mathrm{p}}^{(l,h)} := [||\Phi_{\mathrm{a}}^{(l,h)} - \Phi_{\mathrm{p}}^{(l,h)}||_F]^{-1}$$



Model evaluation on ImageNet-1K

	Model	Size (M)	FLOPs (G)	Top-1 (%)
Columnar Vision Transformers (*single-resolution*)	DeiT-T (ICML'21)	5.7	1.3	72.2
	XCiT-T12/16 (NeurIPS'21)	7.0	1.2	77.1
	PerViT-T	7.6	1.6	78.8
	DeiT-S (ICML'21)	22	4.6	79.8
	T2T-ViT$_t$-14 (ICCV'21)	22	6.1	81.7
	XCiT-S12/16 (NeurIPS'21)	26	4.8	82.0
	PerViT-S	21	4.4	82.1
	DeiT-B (ICML'21)	86	18	81.8
	T2T-ViT$_t$-24 (ICCV'21)	64	15	82.6
	XCiT-S24/16 (NeurIPS'21)	48	9.1	82.6
	PerViT-M	44	9.0	82.9

Analyses on main components of PerViT

Φ_{p}	Φ_{c}	C-stem	CPE	Top-1	Top-5
✓	✓	✓	✓	78.8	94.3
✗	✓	✓	✓	77.3	94.1
✓	✗	✓	✓	76.8	93.5
✓	✓	✗	✓	77.8	94.0
✓	✓	✓	✗	78.1	94.0
✗	✓	✗	✓	76.3	93.2
✗	✓	✓	✗	76.7	93.3
✓	✓	✗	✗	76.5	93.4
✗	✓	✗	✗	72.3	93.4