Brain-inspired L_p-Convolution benefits large kernels and aligns better with visual cortex

Jea Kwon[1*], Sungjun Lim[2], Kyungwoo Song[2†], C. Justin Lee[3†]

[1]Max Planck Institute [2]Yonsei University [3]Institute for Basic Science

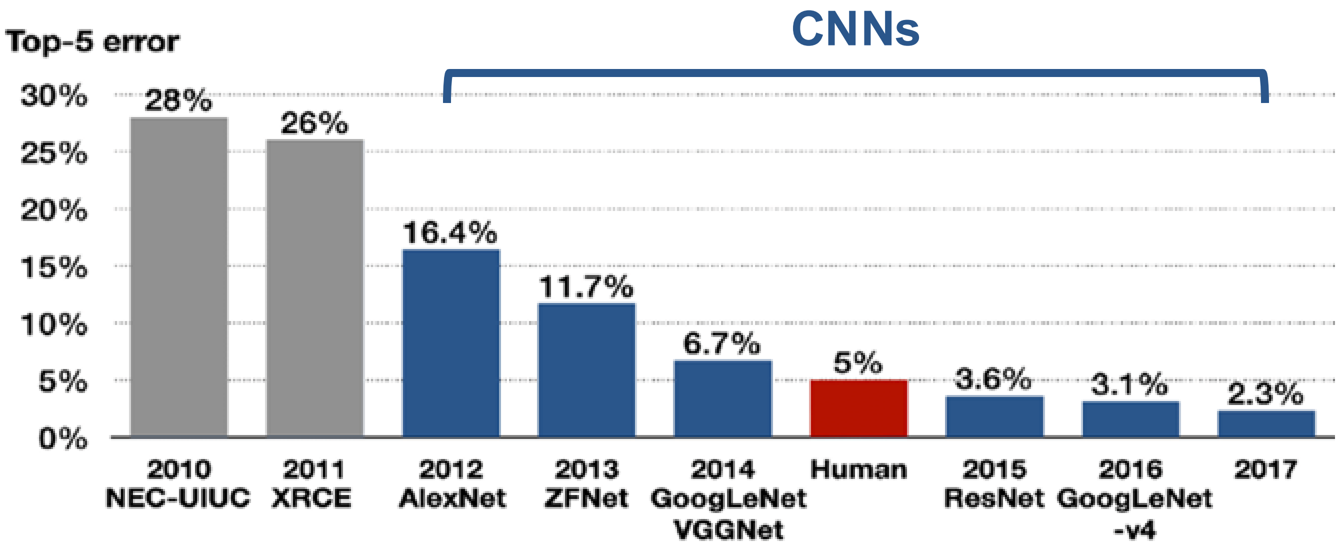
*: This work was conducted at the Institute for Basic Science.

†: Co-corresponding authors

Thirteenth International Conference on Learning Representations

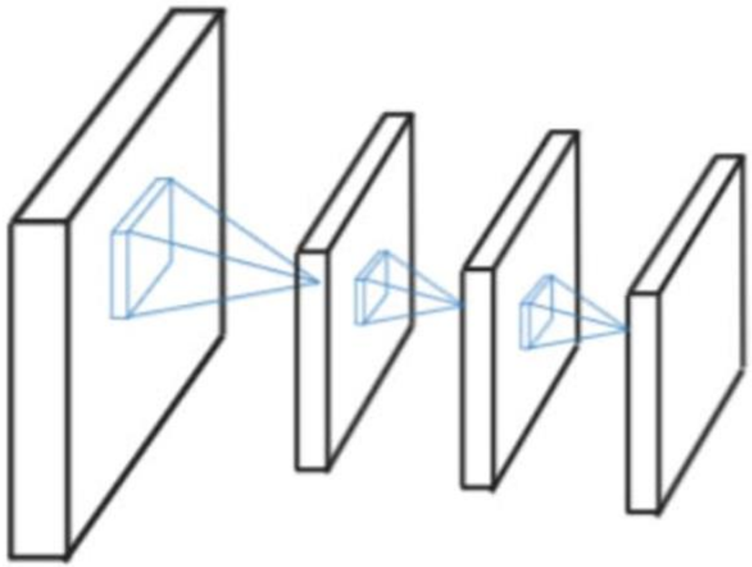
Convolutional Neural Networks (CNNs) revolutionized the machine vision

ImageNet Competition Winners



Kang et al. (2020)

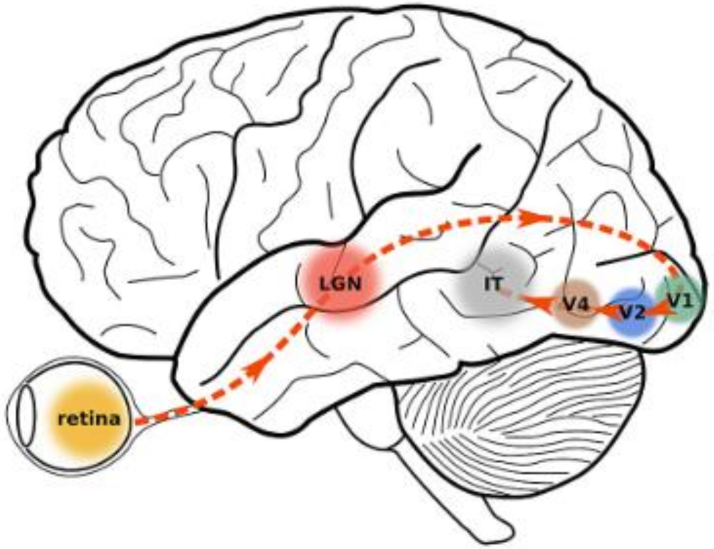
What made CNN successful?



1. Hierarchical structures
2. Local connectivity
3. Parameter sharing

Image Source: https://sgfin.github.io/2020/06/22/Induction-Intro/

CNN benefits from the strong inductive bias by mirroring the brain's structure

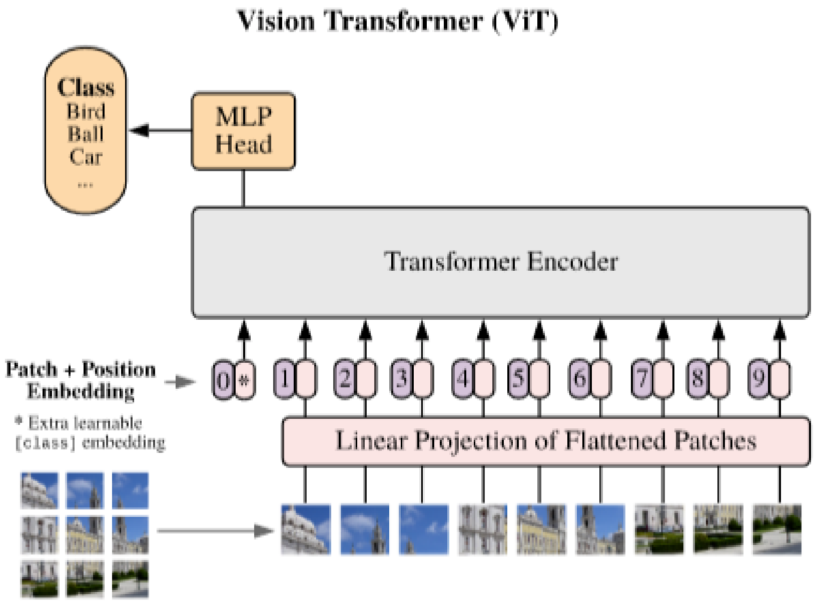


Strong Inductive Bias

1. Hierarchical structures - Hubel & Wiesel (1962, 1965)
2. Local connectivity - Fukushima (1980)
3. Parameter sharing - LeCun et al. (1989)

Image Source: https://commons.wikimedia.org/wiki/File:Human_Brain_sketch_with_eyes_and_cerebrellum.svg

Vision Transformers (ViTs) with *less inductive bias* outperforms CNNs



Dosovitskiy et al. (2020)

ImageNet-1k Benchmark

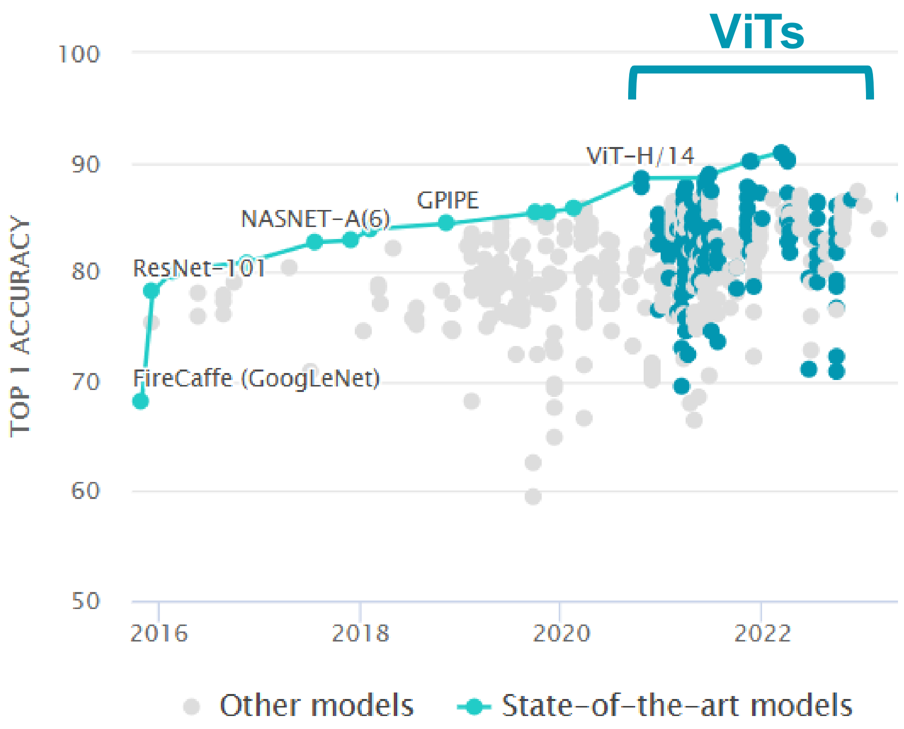


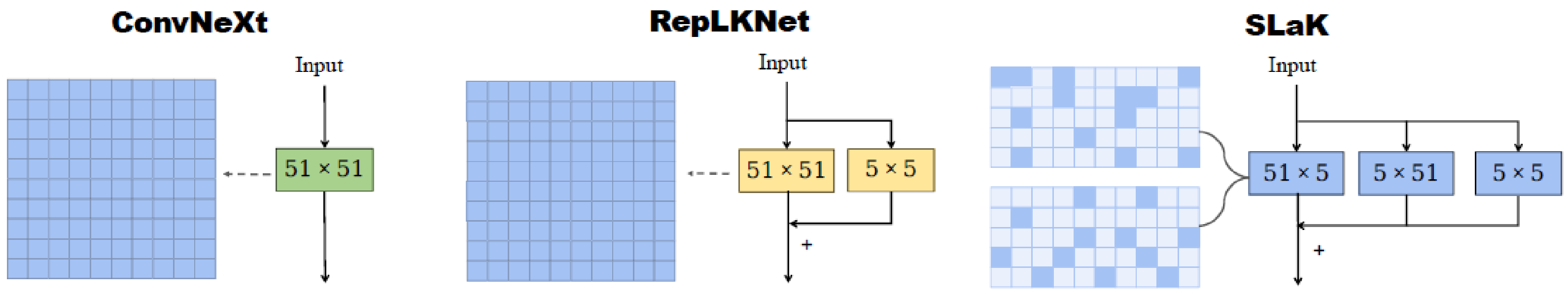
Image Source: https://paperswithcode.com/sota/image-classification-on-imagenet

CNNs outperform ViTs on smaller datasets thanks to strong inductive bias

	CIFAR-10	CIFAR-100	SVHN
ViT	81.36	54.31	95.17
ResNet18	**92.8**	**70.7**	**95.78**

"Understanding Why ViT Trains Badly on Small Datasets: An Intuitive Perspective."
Zhu et al. (2023)

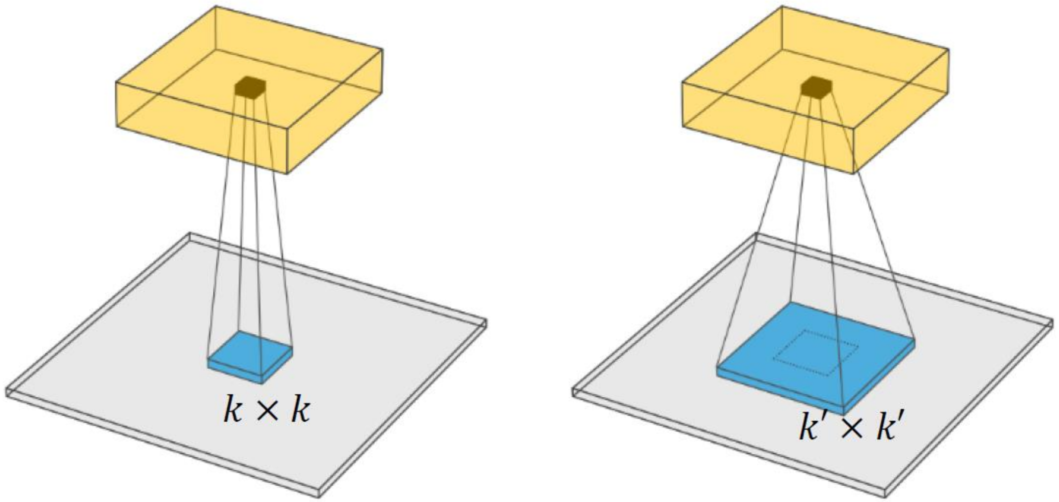
Recently, large kernel CNNs even show comparable performance to ViTs



Model	Image Size	#Param.	FLOPs	Top-1 Accuracy (%)
ViT-Base/16 [19]	224×224	87M	17.6G	77.9
DeiT-Base/16 [80]	224×224	87M	17.6G	81.8
RepLKNet-31B [17]	224×224	79M	15.3G	83.5
Swin-B [55]	224×224	88M	15.4G	83.5
ConvNeXt-B [56]	224×224	89M	15.4G	83.8
SLaK-B	224×224	95M	17.1G	**84.0**

Liu et al. *ICLR* (2023)

However, traditional CNNs do not benefit from simply increasing the kernel size. Why?



Kernel	CIFAR-100 AlexNet	CIFAR-100 VGG-16
(Base)	66.05 ± 0.33	70.26 ± 0.29
(Large)	***54.53 ± 0.65	**64.82 ± 2.92

Parameter # increase,
Performance decrease.

→ Large kernel problem

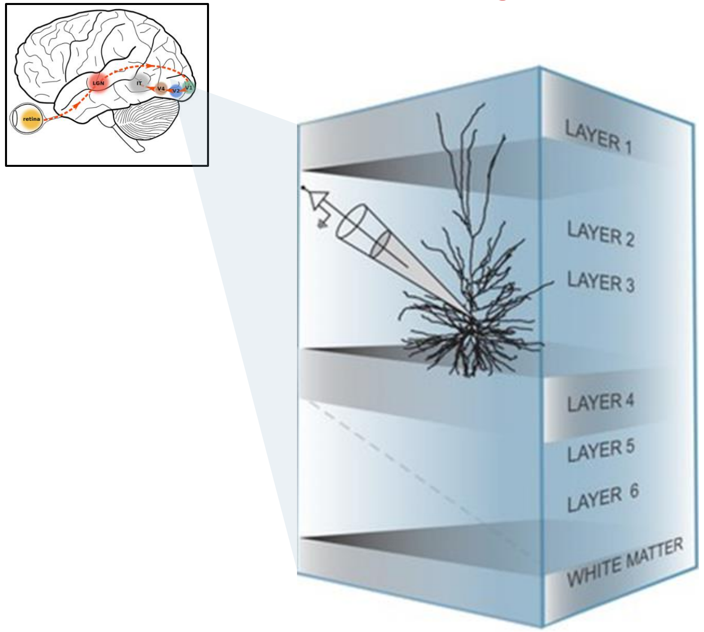
Introduction Summary

1. CNNs outperform ViTs in small datasets due to strong inductive biases originated from the brain

2. Modern CNNs show comparable performance with large kernels while traditional CNNs do not.

CNN is not exact replica of brain

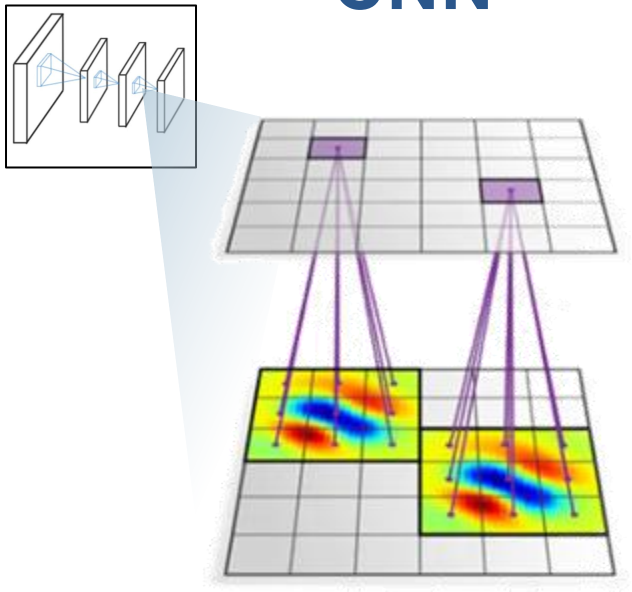
Brain



Natalia et al. (2018)

Sparse & Diverse

CNN



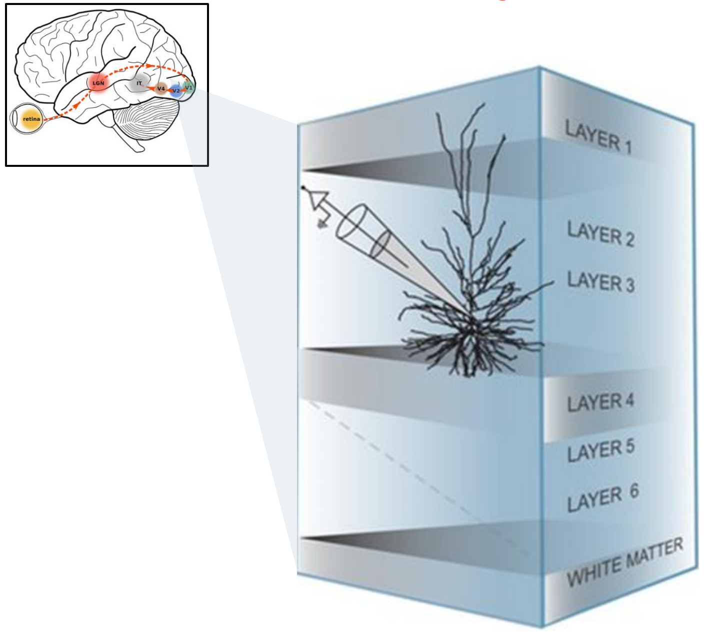
Dense & Uniform

Different Connectivity Patterns

Image Source: https://maurice-weiler.gitlab.io/blog_post/cnn-book_2_conventional_cnns/

Brain's local connectivity follows Gaussian Sparsity

Brain

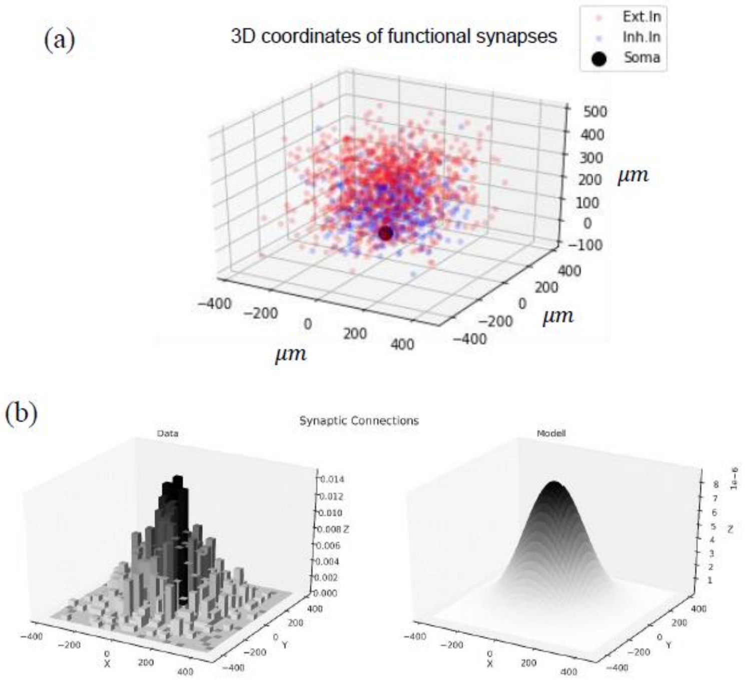


Natalia et al. (2018)

Sparse & Diverse

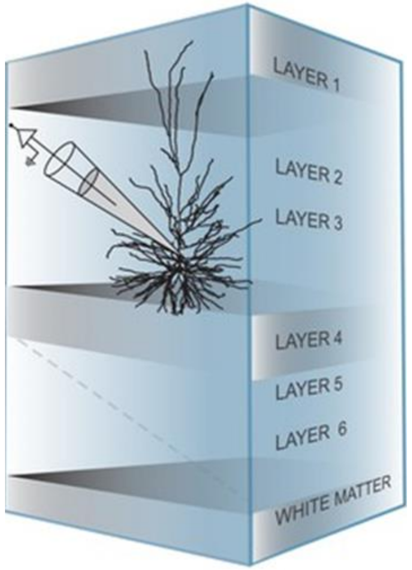
Functional synapse distribution

analyzed data from Rossi et al. *Nature* (2020)

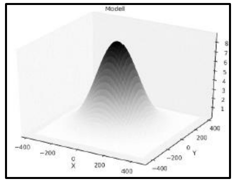


Gaussian sparsity as a new inductive bias?

RQ: Can brain-inspired Gaussian sparsity benefit large kernel CNN and aligns better with Brain?



Natalia et al. (2018)



Brain-inspired
Gaussian sparsity

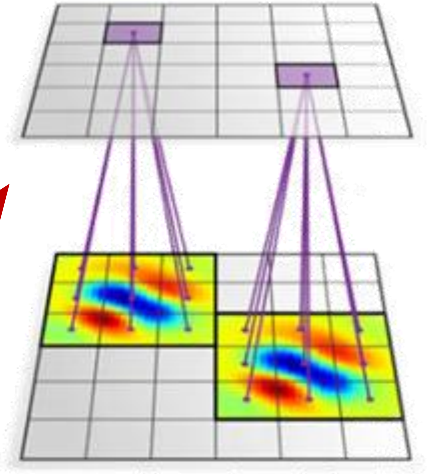


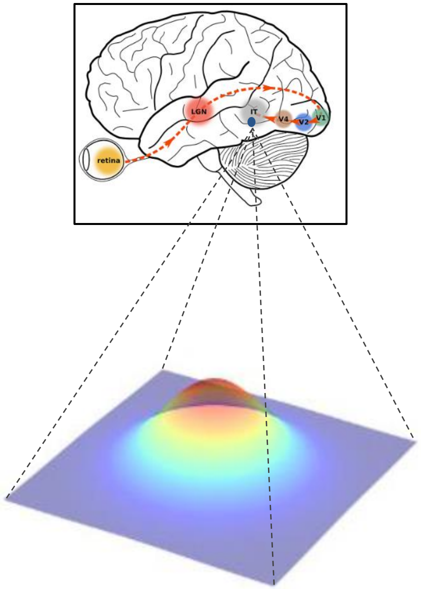
Image Source: https://maurice-weiler.gitlab.io/blog_post/cnn-book_2_conventional_cnns/

Multivariate p-generalized Normal Distribution (MPND) to bridge *biological* and *artificial* local connectivity

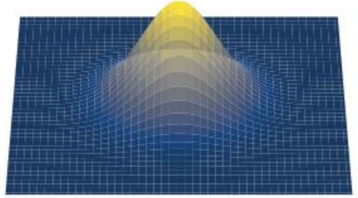
MPND

Goodman & Kotz (1973)

biological ← → *artificial*



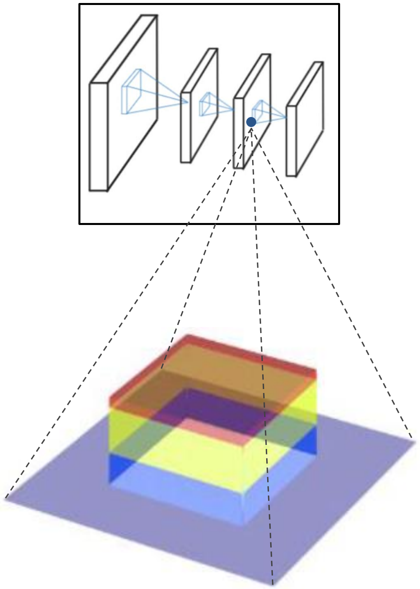
Gaussian, Sparse
$(p \to 2)$



$\mathbf{p} = 2.0 \qquad \mathbf{C} = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$

$\alpha = 1.0 \qquad \theta = 0.0 \qquad \gamma = 1.0$

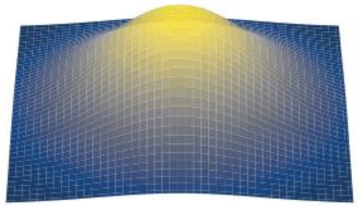
$$\beta \exp\left(-\|\mathbf{C}\Delta\mathbf{s}\|_p^p\right)$$



Uniform, Dense
$(p \to \infty)$

Conformational diversity of MPND with p and C

Diverse shape

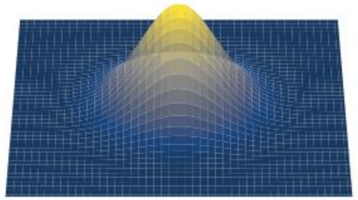


$\mathbf{p} = 2.0$ $\quad \mathbf{C} = \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{bmatrix}$

$\alpha = 0.5 \quad \theta = 0.0 \quad \gamma = 1.0$

MPND

Goodman & Kotz (1973)

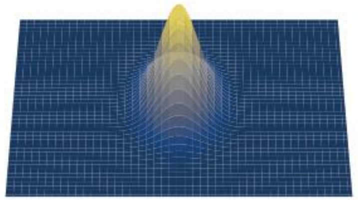


$\mathbf{p} = 2.0$ $\quad \mathbf{C} = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$

$\alpha = 1.0 \quad \theta = 0.0 \quad \gamma = 1.0$

$$\beta \exp\left(-\|\mathbf{C}\Delta\mathbf{s}\|_p^p\right)$$

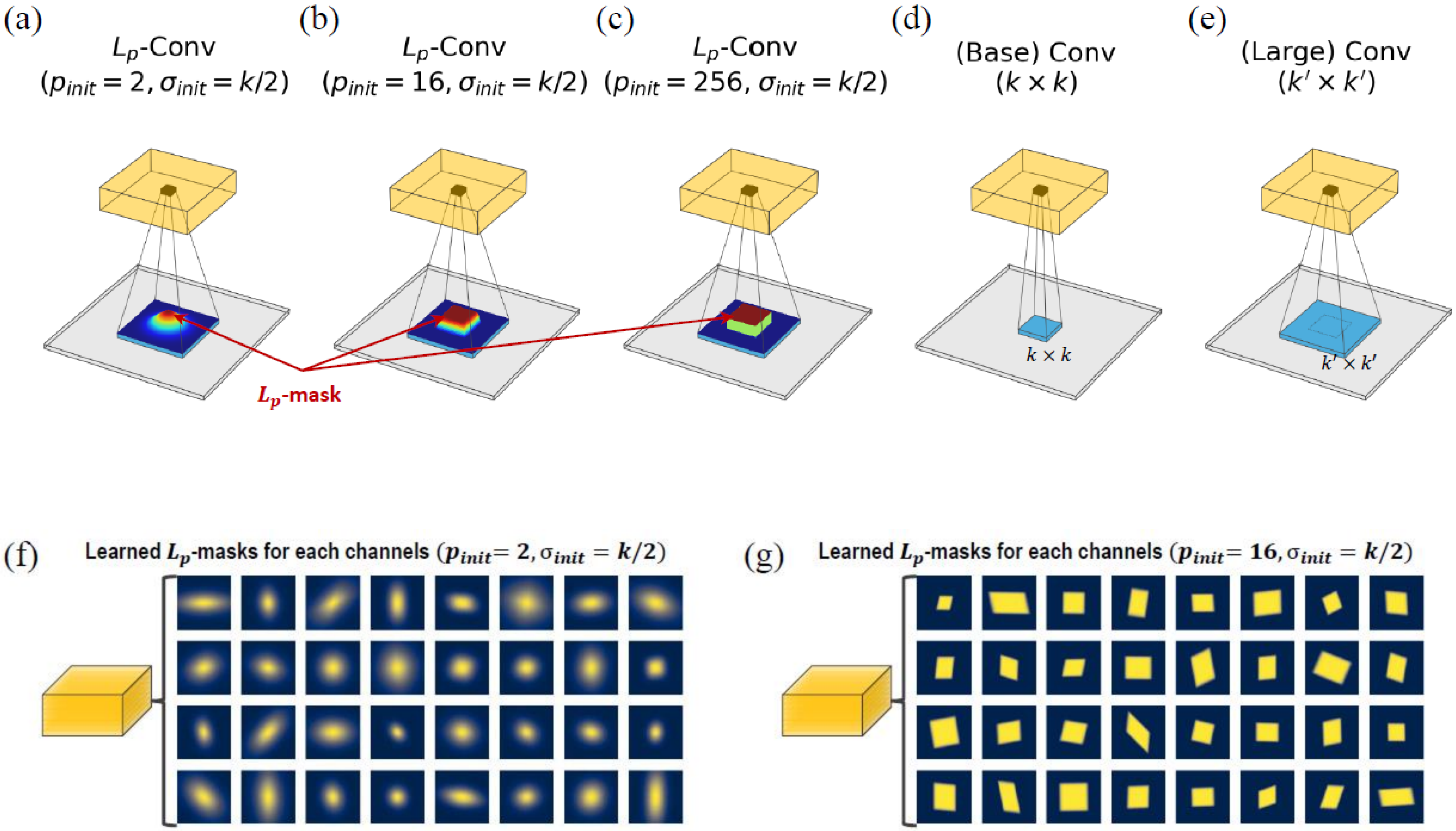
Diverse size



$\mathbf{p} = 2.0$ $\quad \mathbf{C} = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 2.0 \end{bmatrix}$

$\alpha = 2.0 \quad \theta = 90.0 \quad \gamma = 2.0$

L_p-Convolution: Introducing MPND in Convolution



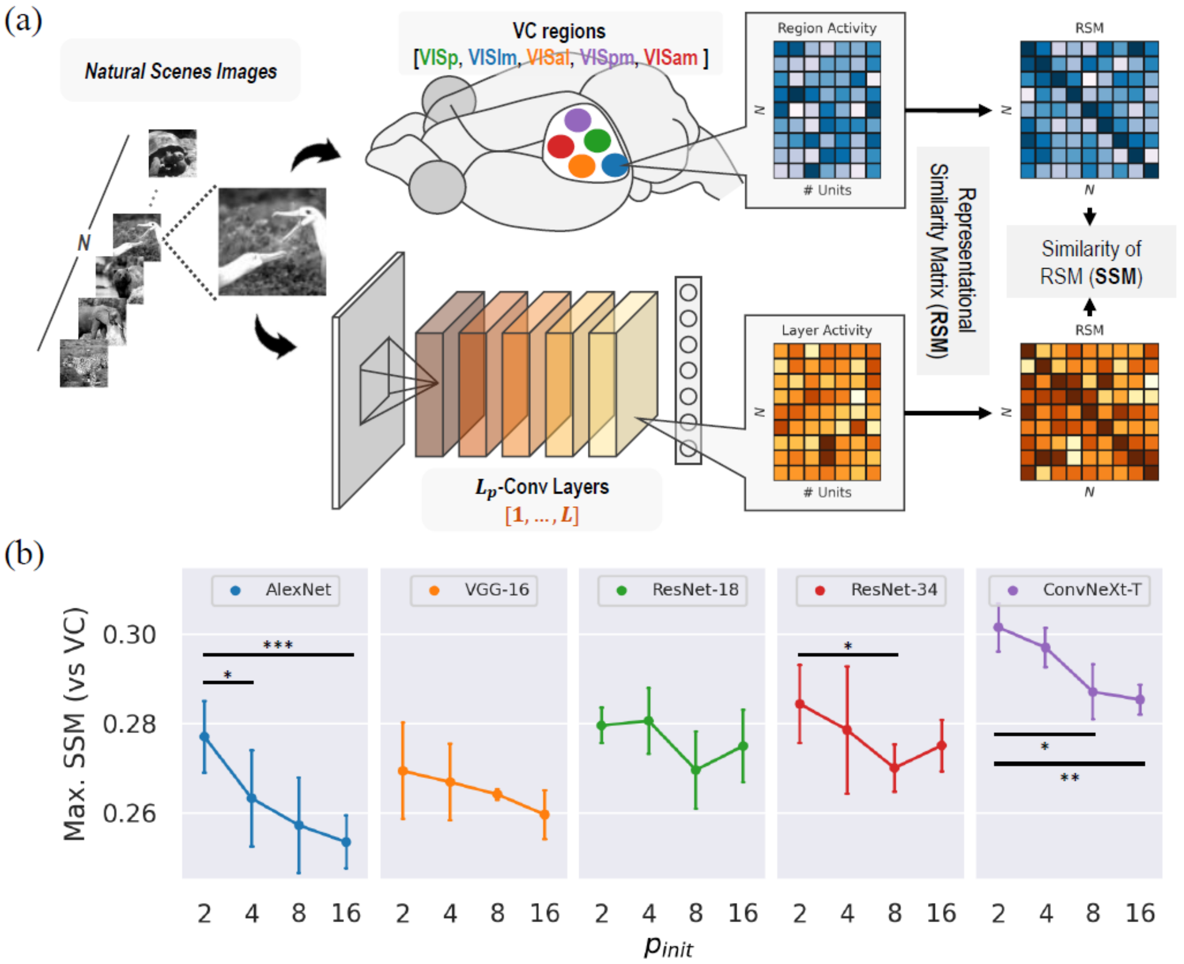
L_p-Convolution benefits traditional CNNs with large kernels in small datasets

CIFAR-100							
Layer	Kernel	p_{init}	AlexNet	VGG-16	ResNet-18	ResNet-34	ConvNeXt-T
(Base) Conv	$k \times k$	-	66.05 ± 0.33	70.26 ± 0.29	71.22 ± 0.18	72.47 ± 0.23	58.36 ± 6.48
(Large) Conv	$l \times l$	-	$^{***}54.53 \pm 0.65$	$^{**}64.82 \pm 2.92$	$^{***}72.80 \pm 0.27$	$^{***}73.52 \pm 0.11$	$^{ns}54.13 \pm 1.14$
$^{\dagger}L_p$-Conv		256	$^{ns}65.95 \pm 0.32$	$^{**}71.03 \pm 0.38$	$^{ns}71.24 \pm 0.23$	$^{ns}72.61 \pm 0.27$	$^{ns}60.34 \pm 2.80$
L_p-Conv		16	$^{**}\mathbf{67.12 \pm 0.37}$	$^{**}70.87 \pm 0.23$	$^{***}72.35 \pm 0.30$	$^{***}73.32 \pm 0.23$	$^{ns}61.30 \pm 1.71$
L_p-Conv	$l \times l$	8	$^{**}66.85 \pm 0.18$	$^{**}71.14 \pm 0.29$	$^{***}72.26 \pm 0.28$	$^{***}73.37 \pm 0.15$	$^{ns}59.94 \pm 5.04$
L_p-Conv		4	$^{*}66.68 \pm 0.28$	$^{***}71.71 \pm 0.36$	$^{***}73.00 \pm 0.15$	$^{***}74.07 \pm 0.22$	$^{ns}59.34 \pm 7.53$
L_p-Conv		2	$^{ns}66.13 \pm 0.33$	$^{***}\mathbf{72.88 \pm 0.30}$	$^{***}\mathbf{73.86 \pm 0.14}$	$^{***}\mathbf{74.95 \pm 0.11}$	$^{ns}\mathbf{62.61 \pm 3.03}$

TinyImageNet							
Layer	Kernel	p_{init}	AlexNet	VGG-16	ResNet-18	ResNet-34	ConvNeXt-T
(Base) Conv	$k \times k$	-	52.25 ± 0.35	67.75 ± 0.07	66.63 ± 0.51	69.22 ± 0.11	70.25 ± 0.45
(Large) Conv	$l \times l$	-	$^{***}35.52 \pm 0.46$	$^{ns}66.96 \pm 1.50$	$^{***}68.33 \pm 0.19$	$^{ns}69.46 \pm 0.36$	$^{ns}68.66 \pm 1.50$
$^{\dagger}L_p$-Conv		256	$^{ns}52.60 \pm 0.12$	$^{ns}67.72 \pm 0.18$	$^{ns}66.37 \pm 0.55$	$^{ns}69.27 \pm 0.27$	$^{ns}70.45 \pm 0.44$
L_p-Conv		16	$^{***}53.98 \pm 0.50$	$^{***}69.29 \pm 0.25$	$^{**}67.72 \pm 0.43$	$^{**}70.00 \pm 0.33$	$^{ns}70.62 \pm 0.30$
L_p-Conv	$l \times l$	8	$^{**}54.07 \pm 0.91$	$^{***}69.72 \pm 0.16$	$^{*}67.63 \pm 0.45$	$^{***}69.81 \pm 0.23$	$^{ns}70.52 \pm 0.36$
L_p-Conv		4	$^{***}\mathbf{54.30 \pm 0.48}$	$^{***}69.79 \pm 0.30$	$^{**}68.20 \pm 0.50$	$^{**}69.99 \pm 0.44$	$^{ns}\mathbf{70.74 \pm 0.37}$
L_p-Conv		2	$^{***}54.13 \pm 0.53$	$^{***}\mathbf{69.96 \pm 0.45}$	$^{***}\mathbf{68.45 \pm 0.36}$	$^{***}\mathbf{70.43 \pm 0.24}$	$^{ns}70.72 \pm 0.31$

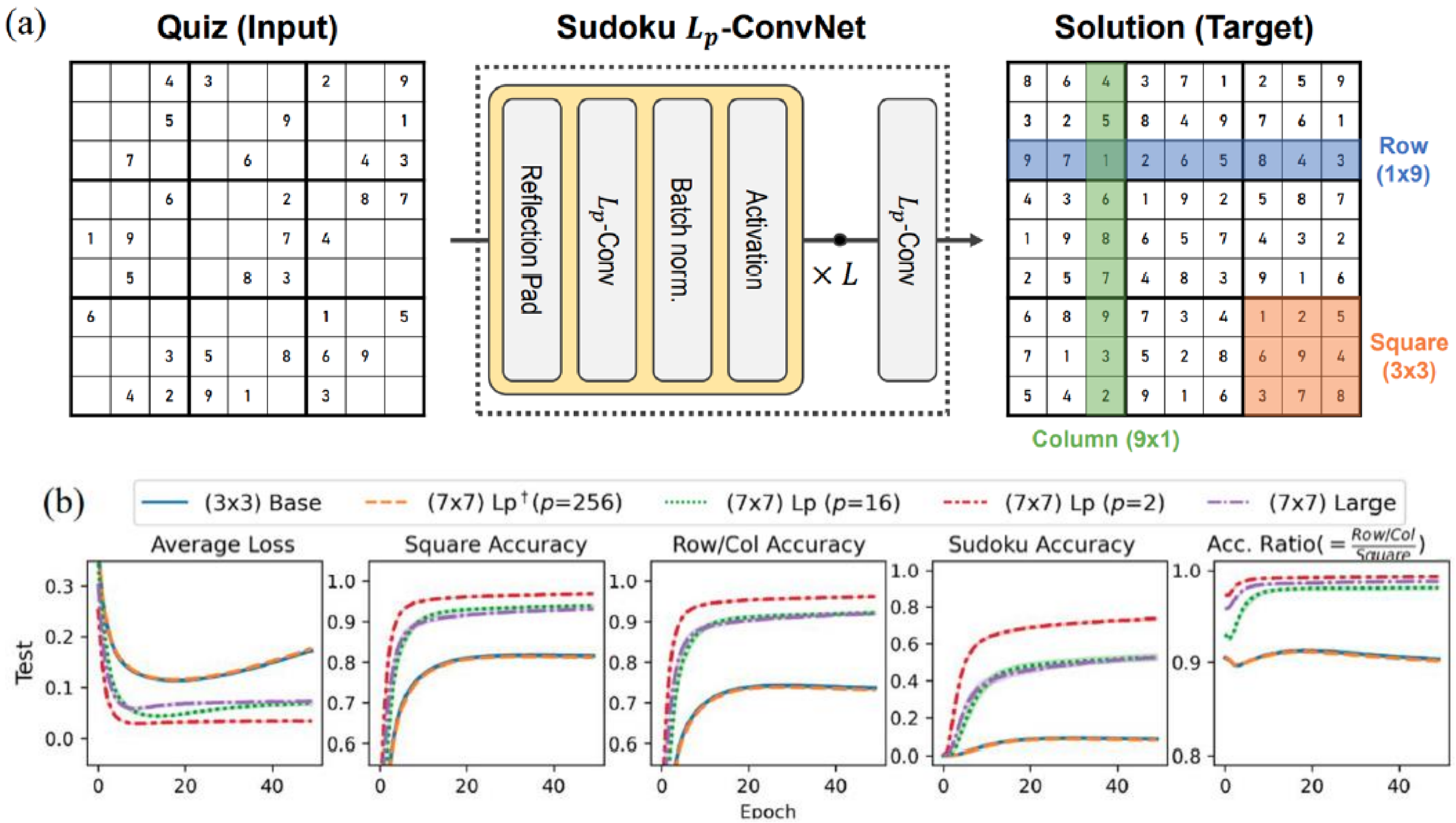
biological $(p \to 2)$ **>** ***artificial*** $(p \to \infty)$

CNNs with Gaussian sparsity aligns better with brain's representation

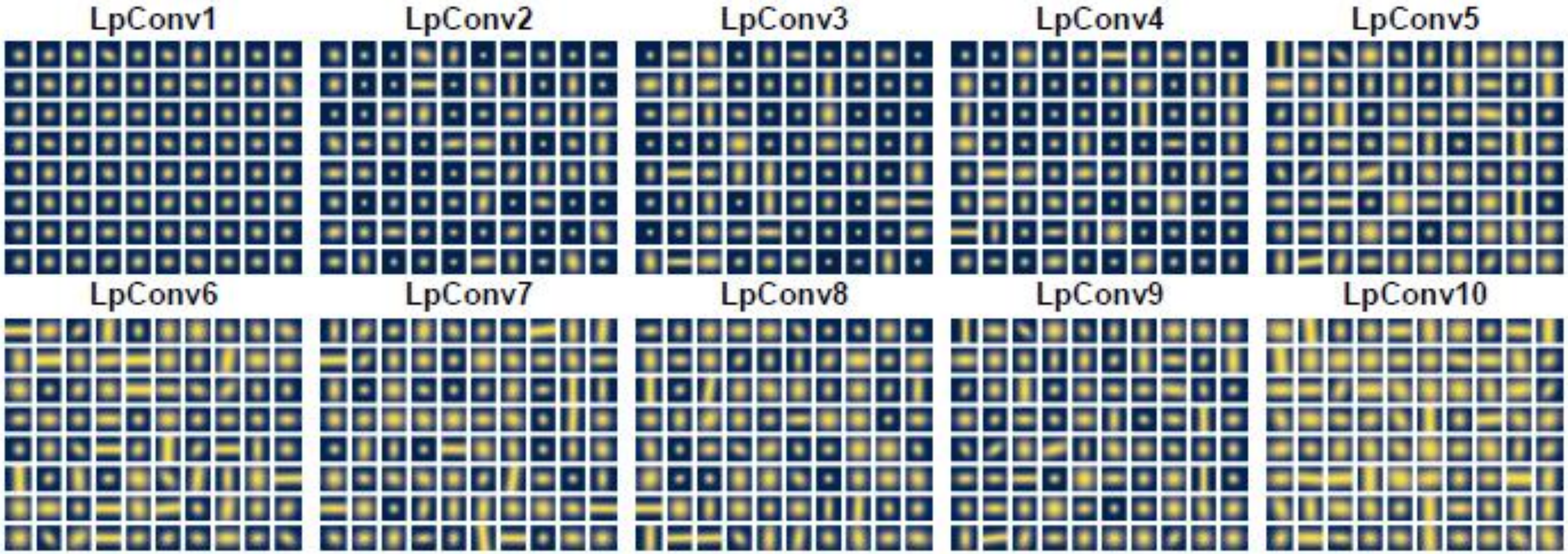


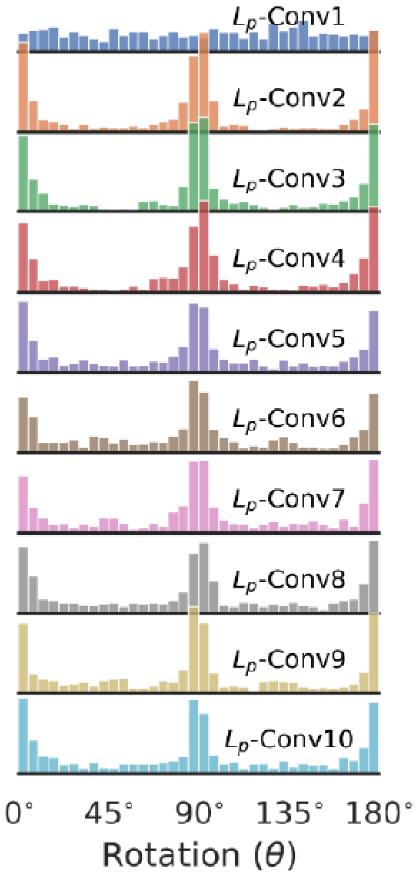
biological $(p \to 2)$ **>** ***artificial*** $(p \to \infty)$

Using Sudoku challenge to investigate conformational changes in L_p-Masks

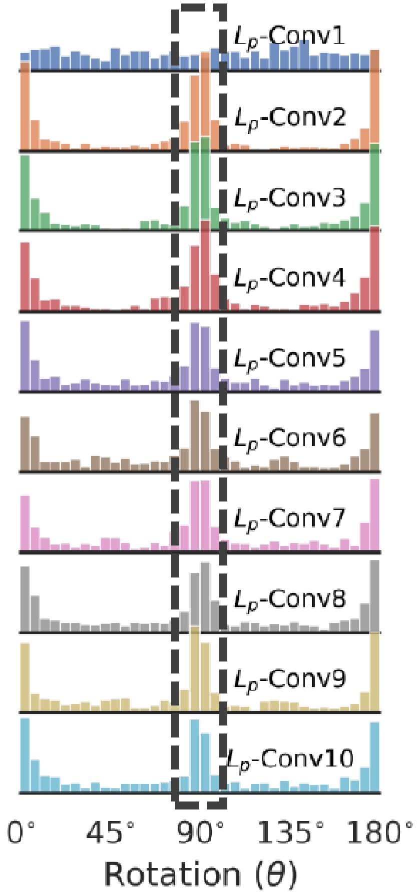


Learned L_p-Masks exhibit vertical and horizontal orientations

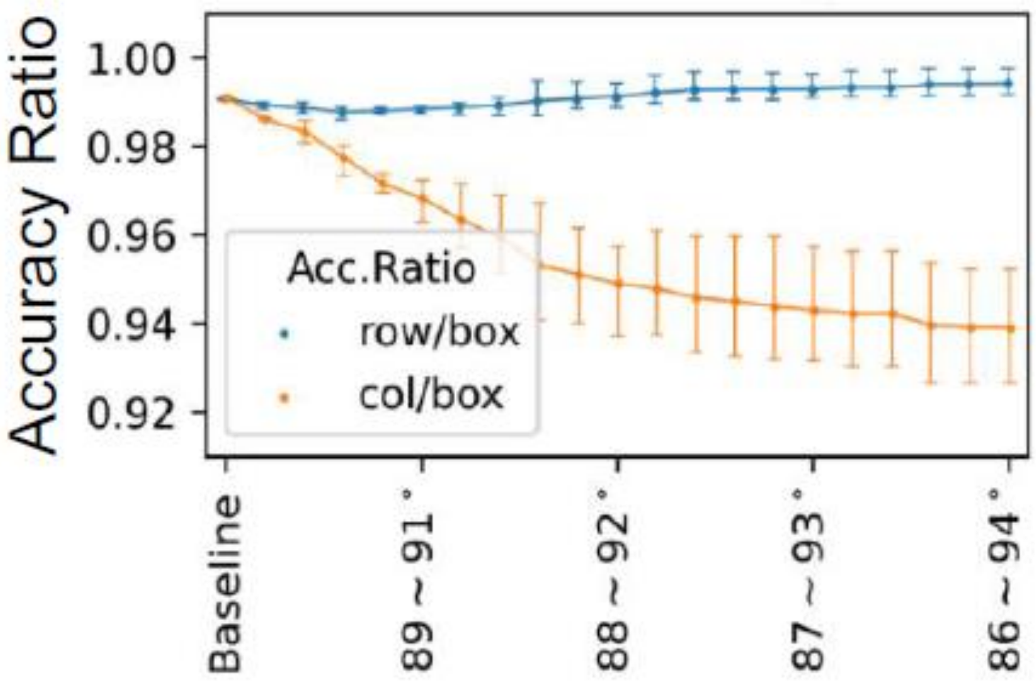




Sudoku analysis reveals task-dependent conformational adaptability of L_p-Masks

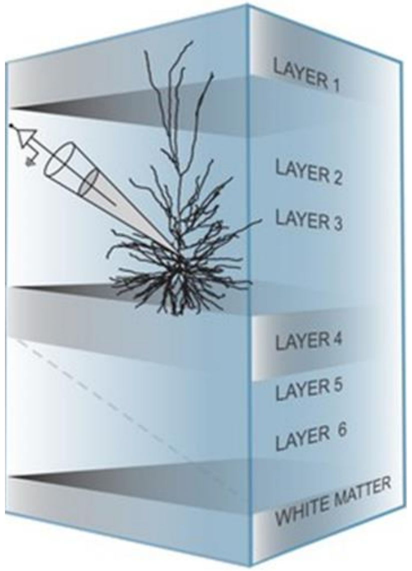


Vertical Convolutional Filter Ablation

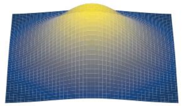


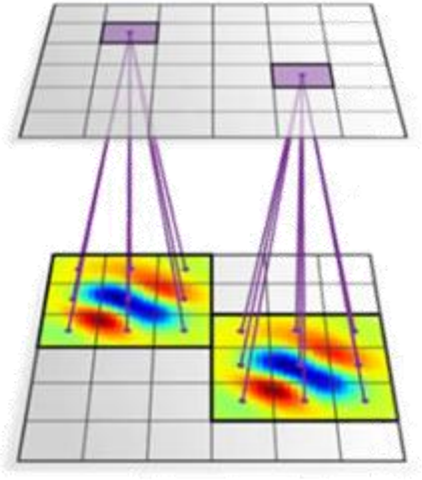
Summary: We propose a new brain-inspired inductive bias which can further benefit CNN

Brain



CNN





L_p-Convolution
(with Conformational Adaptability)

CNN's inductive bias

1. Hierarchical structures
2. Local connectivity
3. Parameter sharing
4. Gaussian sparsity