Paper Review

Do Wide and Deep Networks Learn the Same Things? Uncovering How Neural Network Representations Vary with Width and Depth

Thao Nguyen, et al. Google Research

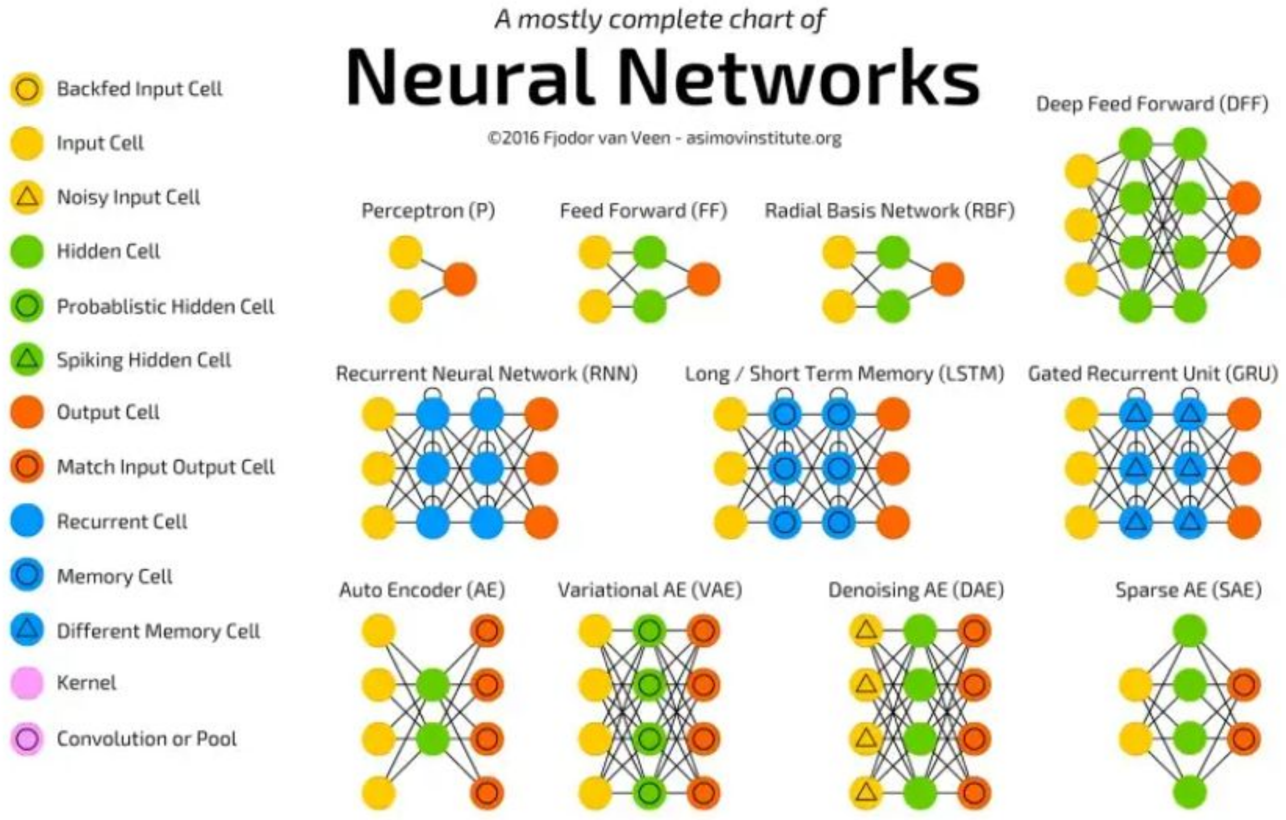
ICLR 2021

R&D Center (Industrial AI Research), POSCO ICT

Susang Kim

Contents

1.Introduction - Neural Network Design Challenges



Data (Scale, Variance)
Objective Function
Learning Algorithm
Model Architecture
Representations
(Hidden & Distributed)
and so on….

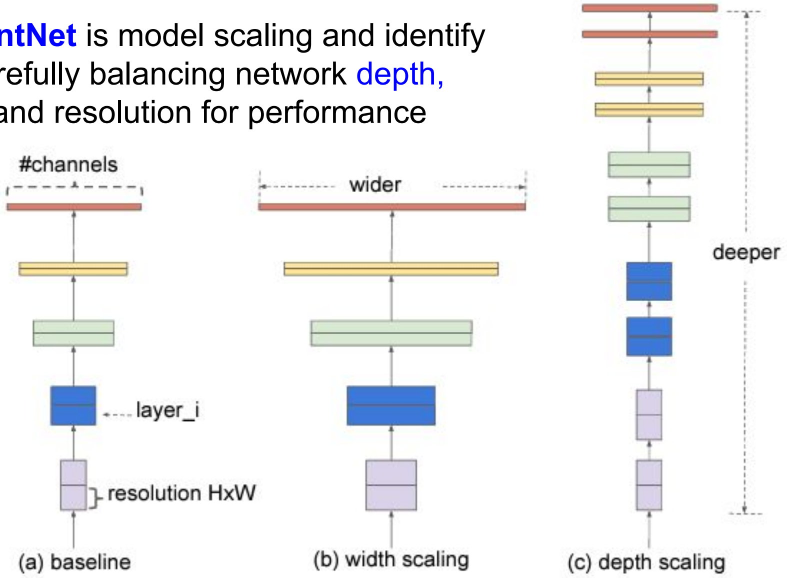
Scaling Models.
-ResNet-18,31,50,101
-ViT-Tiny, Small, Base

https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

1.Introduction - Motivation

Limited understanding how to affect scaling Models by varying **Depth and Width.**
How to design scaling models to improve **performance** by varying depth and width.
Do these different model architectures **learn different intermediate features** (hidden layer)?
How do depth and width **affect final learned representations**?
How varying depth and width affects finding a **redundancy**?

EfficientNet is model scaling and identify that carefully balancing network depth, width, and resolution for performance



group name	output size	block type = $B(3,3)$
conv1	32×32	$[3\times3, 16]$
conv2	32×32	$\begin{bmatrix} 3\times3, 16\times k \\ 3\times3, 16\times k \end{bmatrix} \times N$
conv3	16×16	$\begin{bmatrix} 3\times3, 32\times k \\ 3\times3, 32\times k \end{bmatrix} \times N$
conv4	8×8	$\begin{bmatrix} 3\times3, 64\times k \\ 3\times3, 64\times k \end{bmatrix} \times N$
avg-pool	1×1	$[8 \times 8]$

Decrease depth and increase width of residual networks. **Wide Residual Networks (WRNs)**

Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." ICML 2019.
Zagoruyko, Sergey, and Nikos Komodakis. "Wide residual networks." BMVC 2016.

1.Introduction

We develop a method based on **Centered Kernel Alignment (CKA)** to efficiently measure the similarity of the hidden representations of **wide and deep neural networks.**

1) Apply **CKA to different network architectures** to find difference between representations.

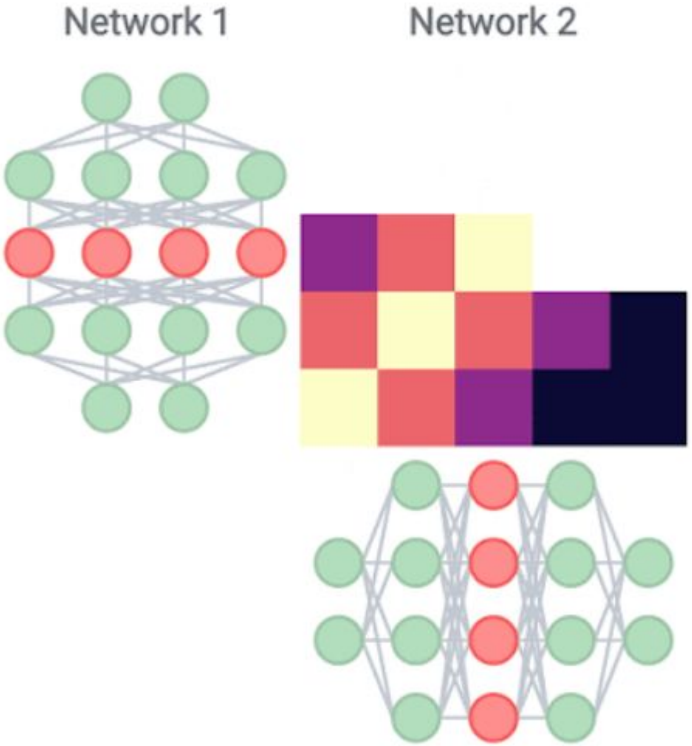
2) A **block structure appears in overparameterized models**.

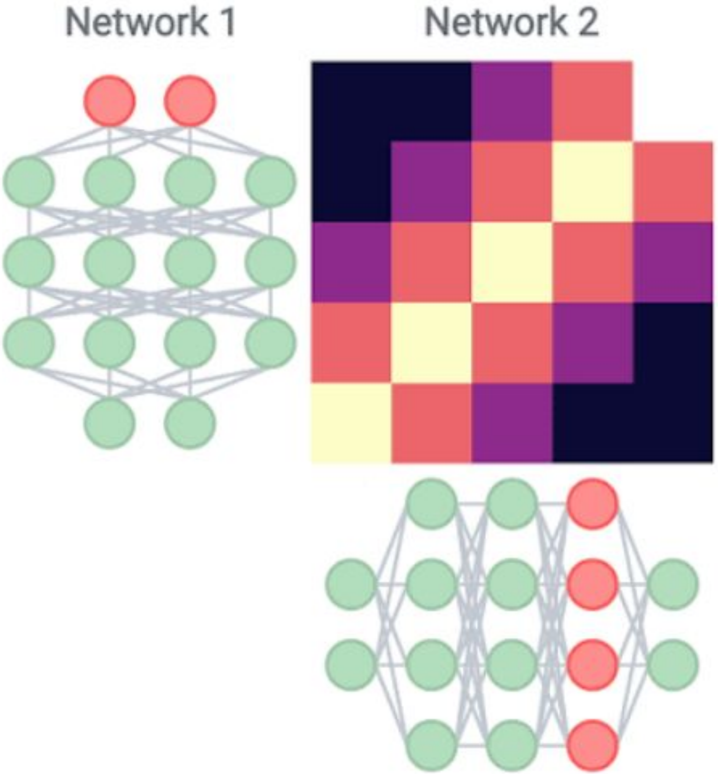
3) Find that the block structure corresponds to hidden representations having **a single principal component that explains the majority of the variance in the representation**.

4) We show that some hidden layers exhibiting **the block structure can be pruned with minimal impact on performance**.

5) We find that wide and deep models make systematically **different mistakes on ImageNe**t, even when these networks achieve similar overall accuracy. **(wide is scenes / deep is goods)**

2.Preliminaries - Comparing Neural Net Representation





2.Challenges in comparing representations

Euclidean Distance

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Cosine Similarity

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Dot Product Similarity

$$\langle \text{vec}(XX^T), \text{vec}(YY^T) \rangle = \text{tr}(XX^T YY^T) = ||Y^T X||_F^2.$$

Is it possible to compare neural network representations?
various representations having neurons or dimensions.
(Invariance to Invertible Linear Transformation, Orthogonal Transform, Isotropic scaling)

$$A = \sigma(\mathbf{w} \cdot \mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w} \cdot \mathbf{x} + b)}}$$

$$s(X, Y) = s(XA, YB)$$

2.Comparing Similarity Structures - CKA

One way to understand trained neural networks is by comparing their representations by CKA

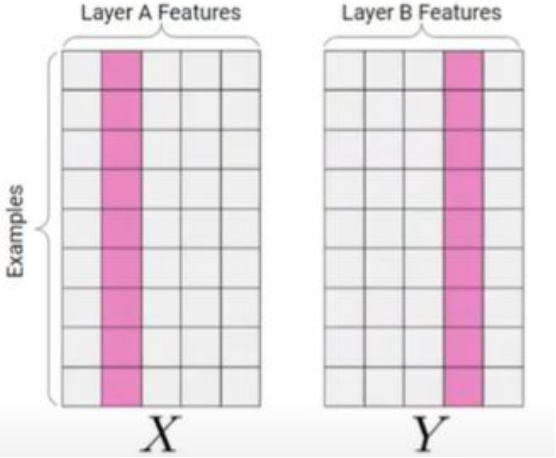
Centering Matrix is Idempotent matrix

$$C_n = I_n - \frac{1}{n} J_n \qquad A^2 = A.$$

$$C_1 = [0],$$

$$C_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

$$C_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{bmatrix}$$



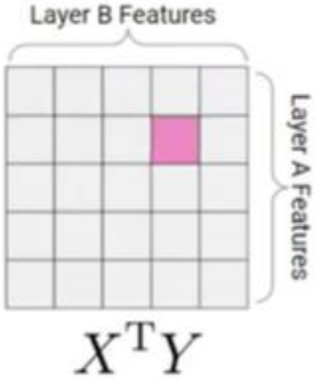
Dot Product based similarity, trace matrix

$$\langle a, b \rangle = \sum_{i=1}^{n} a_i b_i \qquad A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \mathrm{vec}(A) = \begin{bmatrix} a \\ c \\ b \\ d \end{bmatrix}$$

$$\langle \mathrm{vec}(XX^{\mathrm{T}}), \mathrm{vec}(YY^{\mathrm{T}}) \rangle = \mathrm{tr}(XX^{\mathrm{T}}YY^{\mathrm{T}}) = ||Y^{\mathrm{T}}X||_{\mathrm{F}}^{2}.$$

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}$$

$$\mathrm{vec}(A) = [a_{1,1}, \ldots, a_{m,1}, a_{1,2}, \ldots, a_{m,2}, \ldots, a_{1,n}, \ldots, a_{m,n}]^{\mathrm{T}}$$



Kornblith, Simon et al. “Similarity of Neural Network Representations Revisited.”, ICML 2019.

2.Comparing Similarity Structures - CKA

Centered Kernel Alignment (CKA) is a similarity metric designed to measure the similarity of between representations of features in neural networks.(summarizes measurements into a single scalar)

HSIC is the Hilbert-Schmidt independence criterion

$$H_n = I_n - \frac{1}{n}11^T.$$

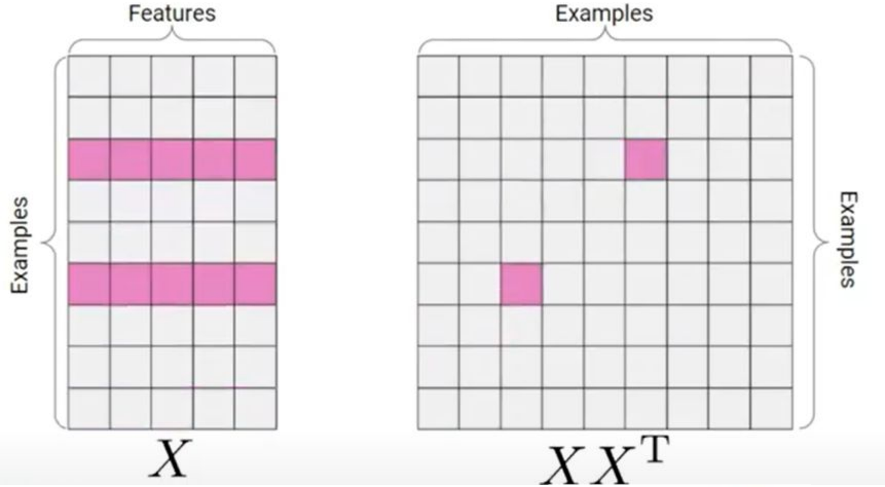
$$\mathrm{CKA}(\boldsymbol{K}, \boldsymbol{L}) = \frac{\mathrm{HSIC}(\boldsymbol{K}, \boldsymbol{L})}{\sqrt{\mathrm{HSIC}(\boldsymbol{K}, \boldsymbol{K})\mathrm{HSIC}(\boldsymbol{L}, \boldsymbol{L})}},$$

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top \quad \boldsymbol{L} = \boldsymbol{Y}\boldsymbol{Y}^\top$$

$$\mathrm{HSIC}(K, L) = \frac{1}{(n-1)^2}\mathrm{tr}(KHLH),$$

$$\text{Let } K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \text{ and } L_{ij} = l(\mathbf{y}_i, \mathbf{y}_j)$$

HSIC = 0 implies independence. where K and L are two kernels.

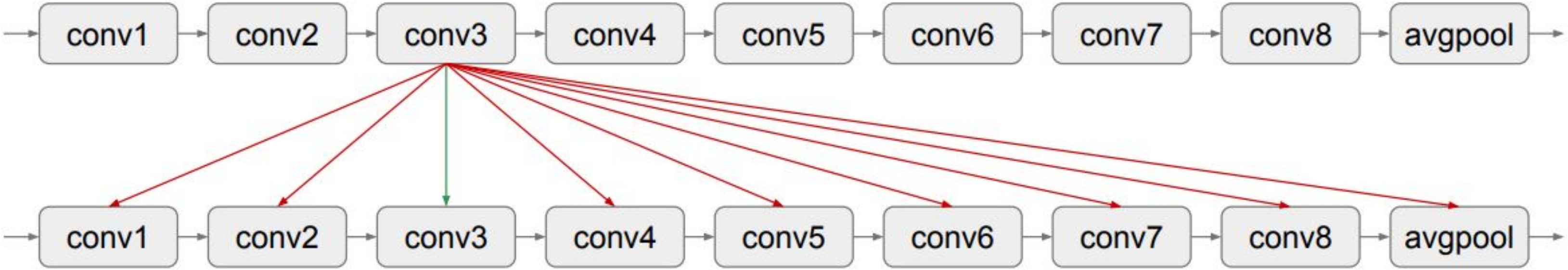


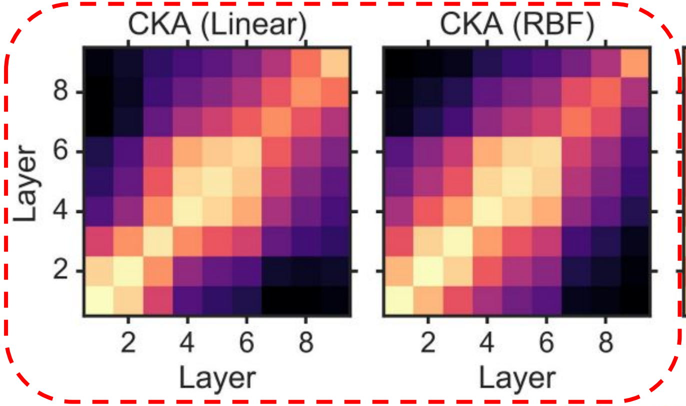
Gram matrices reflects the similarities. $G = V^T V$

Kornblith, Simon et al. “Similarity of Neural Network Representations Revisited.”, ICML 2019.

2.To understand trained neural networks

Architecturally identical networks A and B **trained from different random initializations**, a layer from net A should be most similar to the architecturally corresponding layer in net B





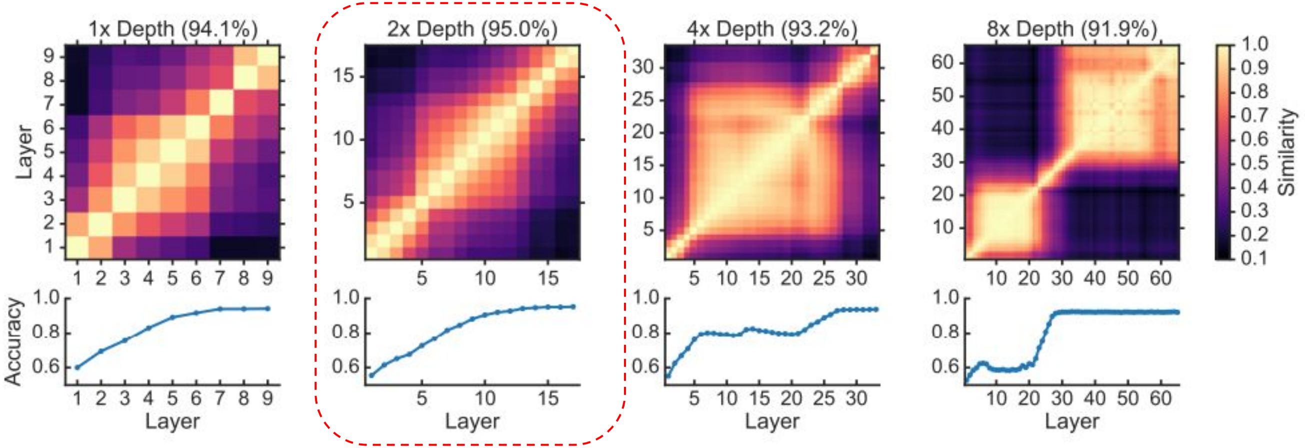
Index	Accuracy
CCA ($\bar{\rho}$)	1.4
CCA (R^2_{CCA})	10.6
SVCCA ($\bar{\rho}$)	9.9
SVCCA (R^2_{CCA})	15.1
PWCCA	11.1
Linear Reg.	45.4
Linear HSIC	22.2
CKA (Linear)	**99.3**
CKA (RBF 0.2)	80.6
CKA (RBF 0.4)	**99.1**
CKA (RBF 0.8)	**99.3**

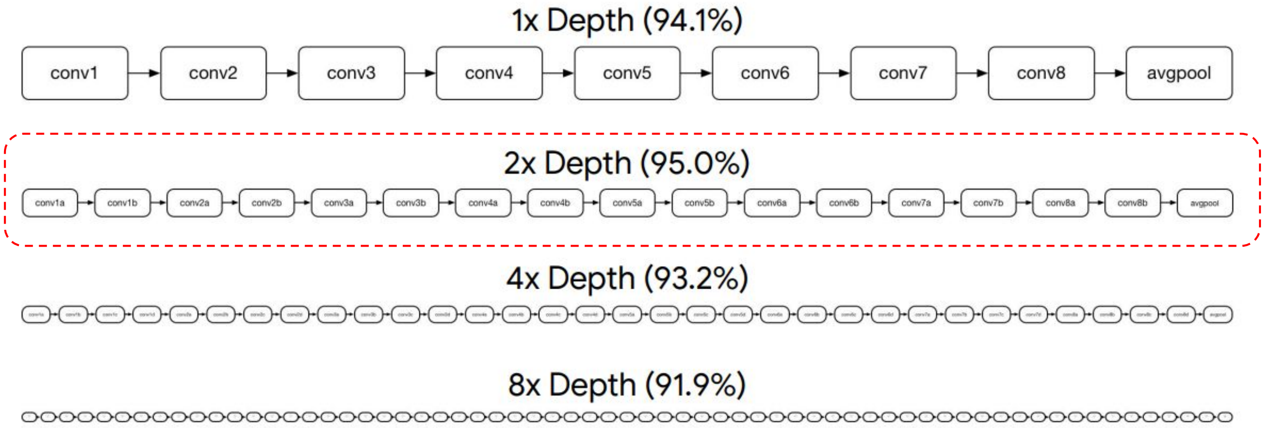
Radial Basis Function (RBF) kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-||\mathbf{x}_i - \mathbf{x}_j||_2^2/(2\sigma^2)).$

Kornblith, Simon et al. "Similarity of Neural Network Representations Revisited.", ICML 2019.

2.CKA Reveals Network Pathology

CKA between layers of individual networks of different depths on the CIFAR-10 test set

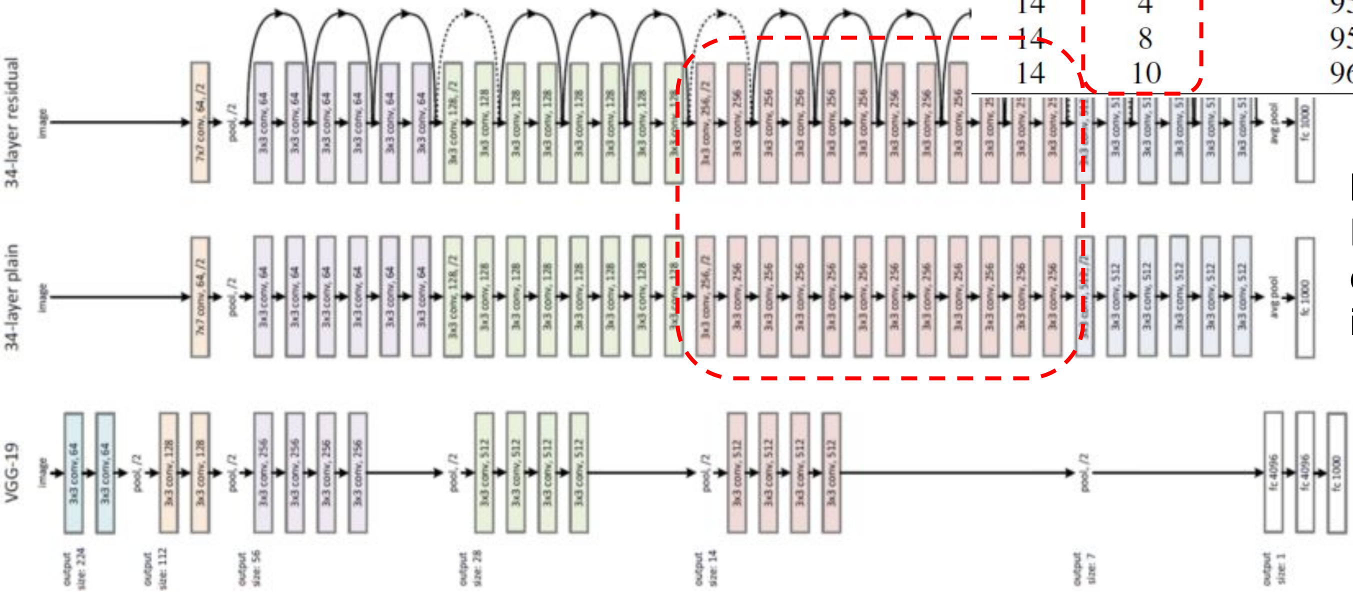




3.Methods - Width and Depth

CIFAR-10, CIFAR-100
width = 1, 2, 4, 8, 10
depth = 14, 20, 26, 38
SGD,cosine decay learning, batch size of 128, to train each model for 300 epochs

Depth	Width	CIFAR-10 Test Accuracy (%)	CIFAR-100 Test Accuracy (%)
32	1	93.5	71.2
44	1	94.0	72.0
56	1	94.2	73.3
110	1	94.3	74.0
164	1	94.4	73.9
14	1	92.0	67.8
14	2	94.1	72.9
14	4	95.4	77.0
14	8	95.9	80.0
14	10	96.0	80.2



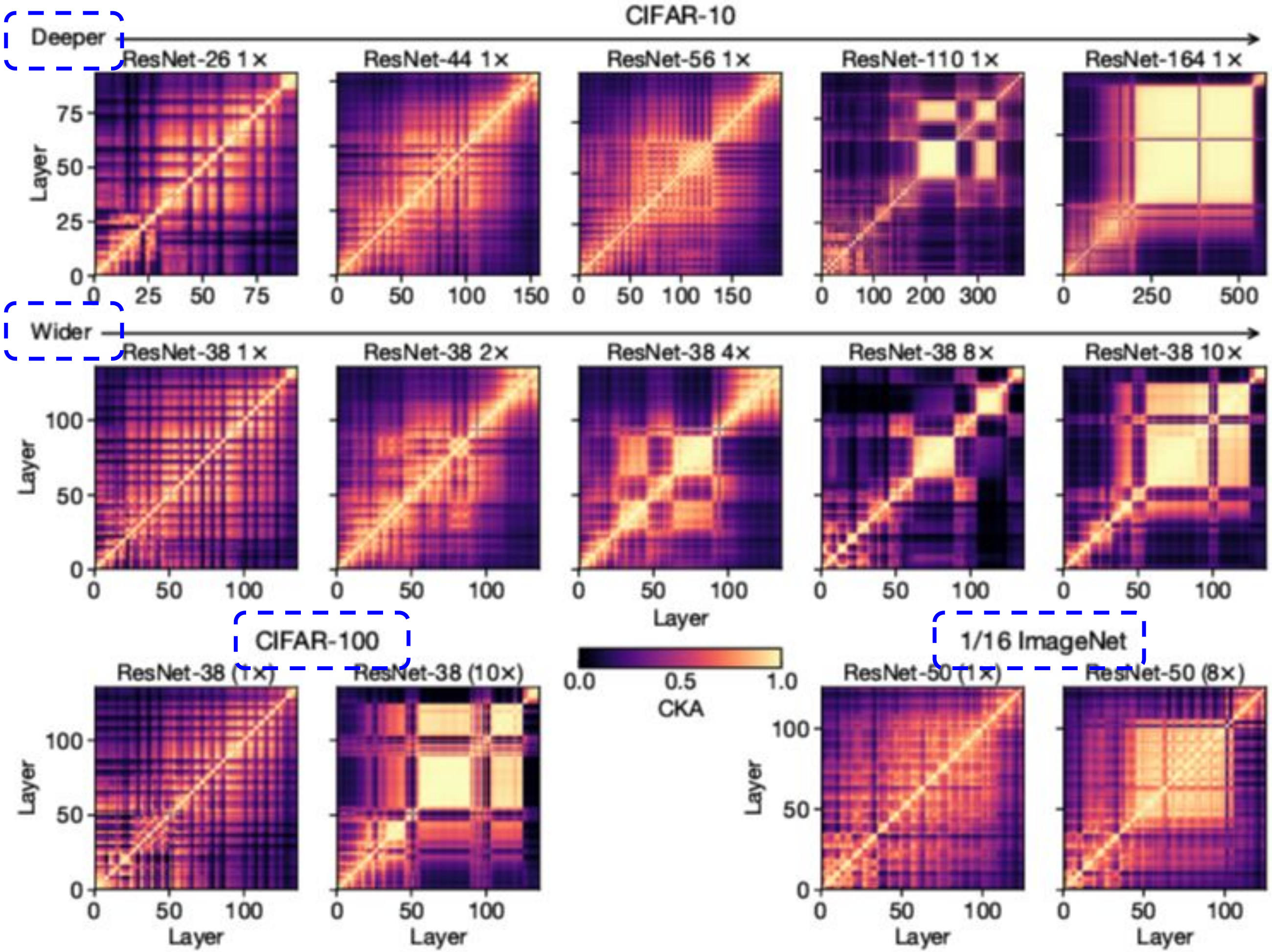
ImageNet
ResNet-50 increase depth or width in the third stage only
120 epochs batch size of 256

He, Kaiming, et al. "Deep residual learning for image recognition." CVPR 2016.

3.Emergence of the block structure with increasing width or depth.

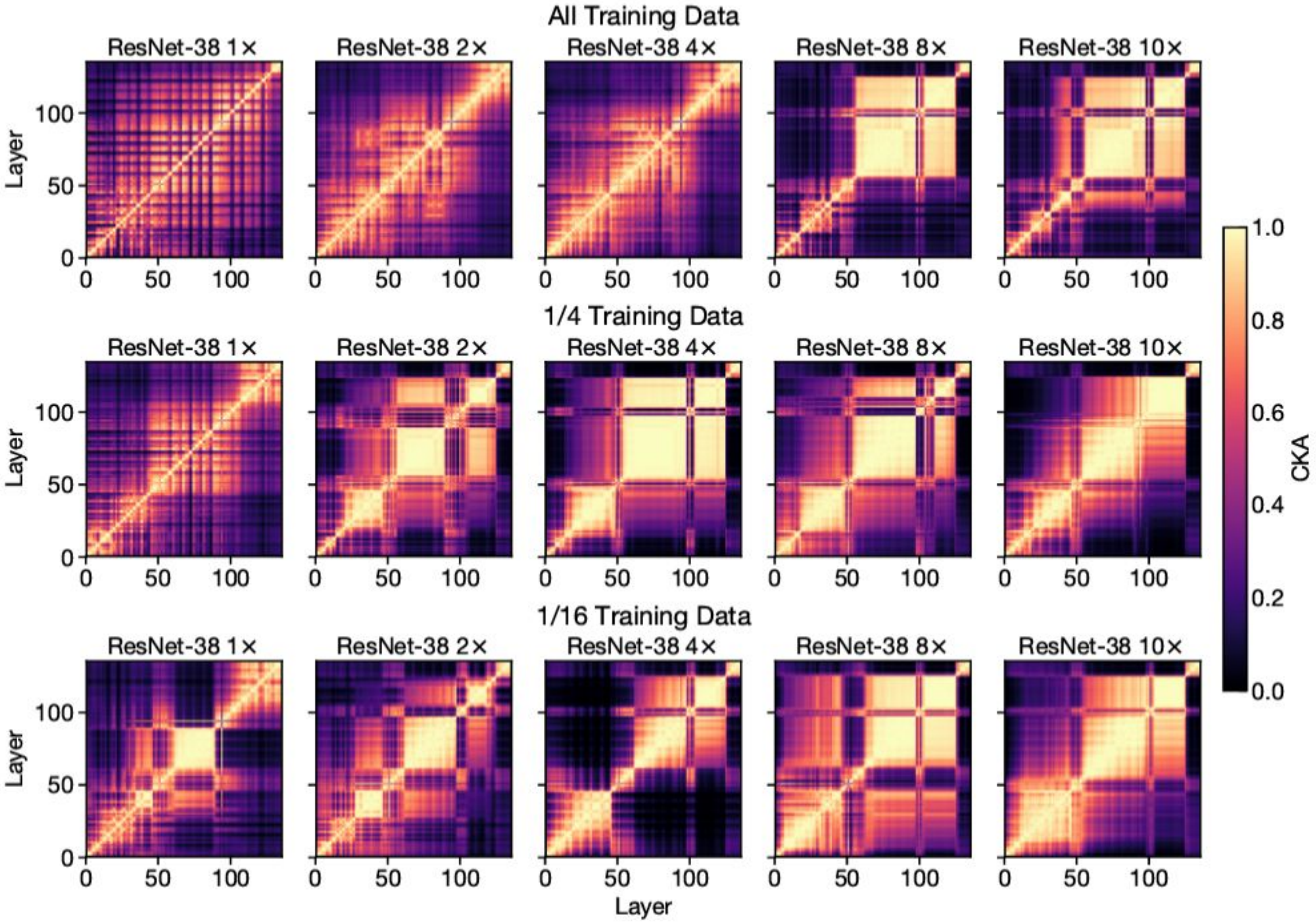
As the model gets wider or deeper,we see the emergence of a distinctive **block structure**.

This block structure mostly **appears in the later layers** (the last two stages) of the network.

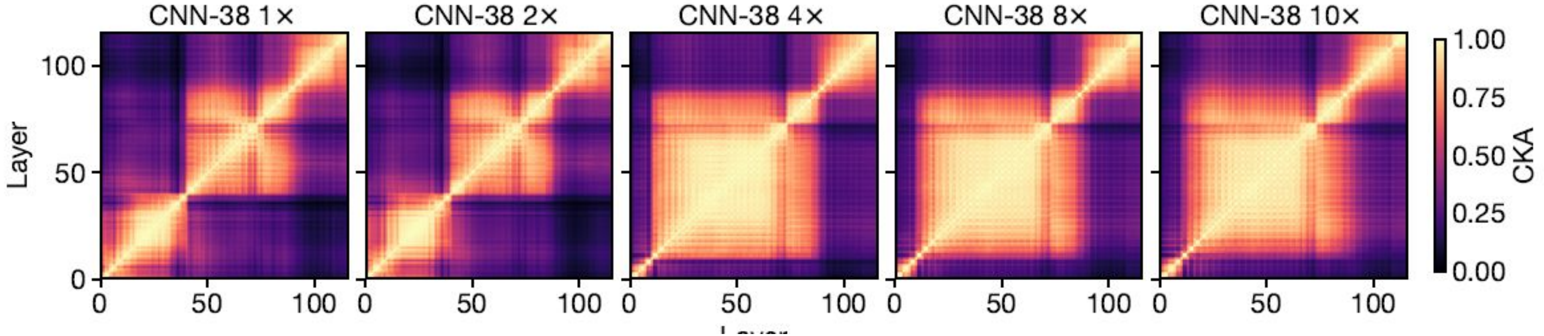


3.Block structure with narrower networks when trained on less data.

Smaller dataset size, smaller (narrower) models now also exhibit the block structure

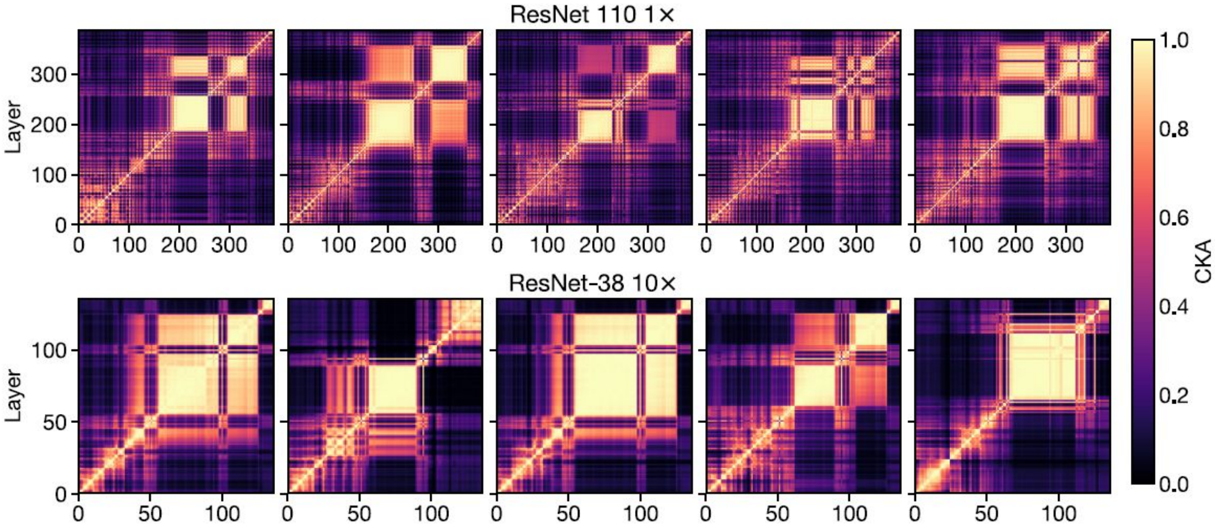


3.Block structure without residual connections & Random initializations



Block structure also **appears in models without residual connections** (Removed Residual Connections)

Block structure varies **across random initializations**



3.The First Principal Component

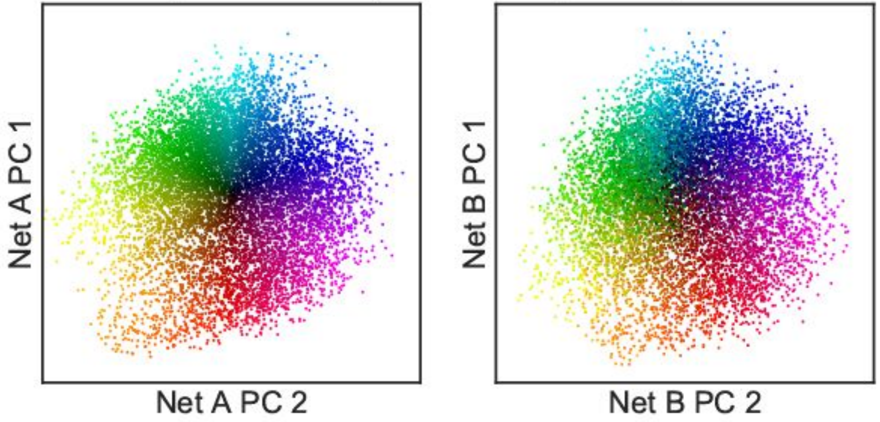
For centered matrices of activations $\boldsymbol{X} \in \mathbb{R}^{n \times p_1}$, $\boldsymbol{Y} \in \mathbb{R}^{n \times p_2}$, linear CKA may be written as:

$$\mathrm{CKA}(XX^\mathrm{T}, YY^\mathrm{T}) = \frac{\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \lambda_X^i \lambda_Y^j \langle \mathbf{u}_X^i, \mathbf{u}_Y^j \rangle^2}{\sqrt{\sum_{i=1}^{p_1} (\lambda_X^i)^2} \sqrt{\sum_{j=1}^{p_2} (\lambda_Y^j)^2}}$$

where $\boldsymbol{u}_X^i \in \mathbb{R}^n$ and $\boldsymbol{u}_Y^i \in \mathbb{R}^n$ are the i^{th} normalized principal components of $\boldsymbol{X}$ and $\boldsymbol{Y}$

Let the i^{th} eigenvalue of XX^T (squared singular value of X) be indexed as λ_X^i.

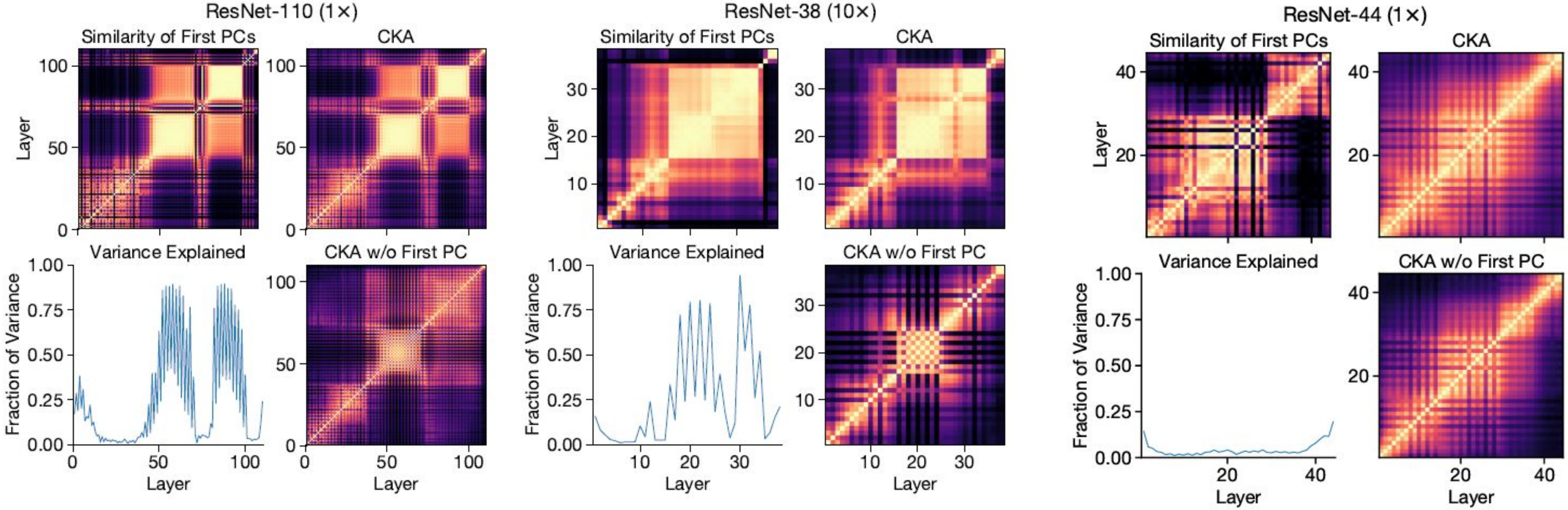
Examples Colored By Net A Principal Components



CIFAR-10 Test (first two PCA in intermediate layer

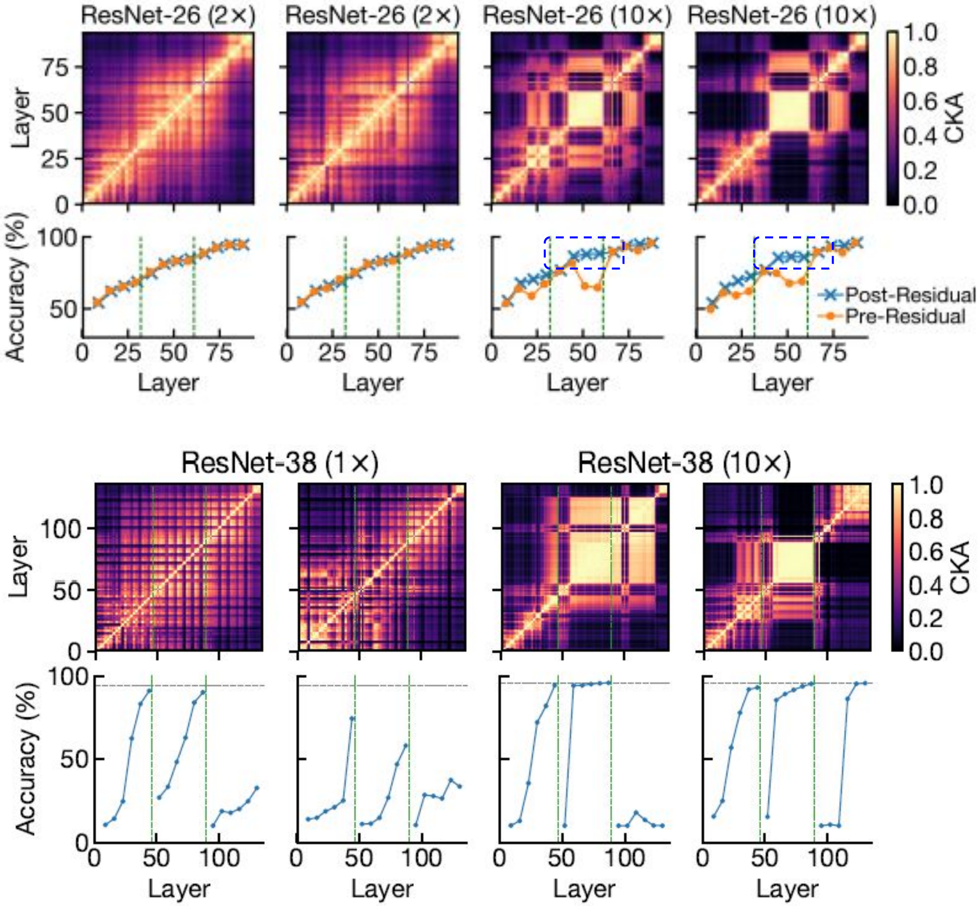
Kornblith, Simon, et al. "Similarity of neural network representations revisited." ICML 2019.

3.Block structure & Principal component



This principal component is also preserved throughout the block structure,
Variance measure is significantly higher where the block structure is present.

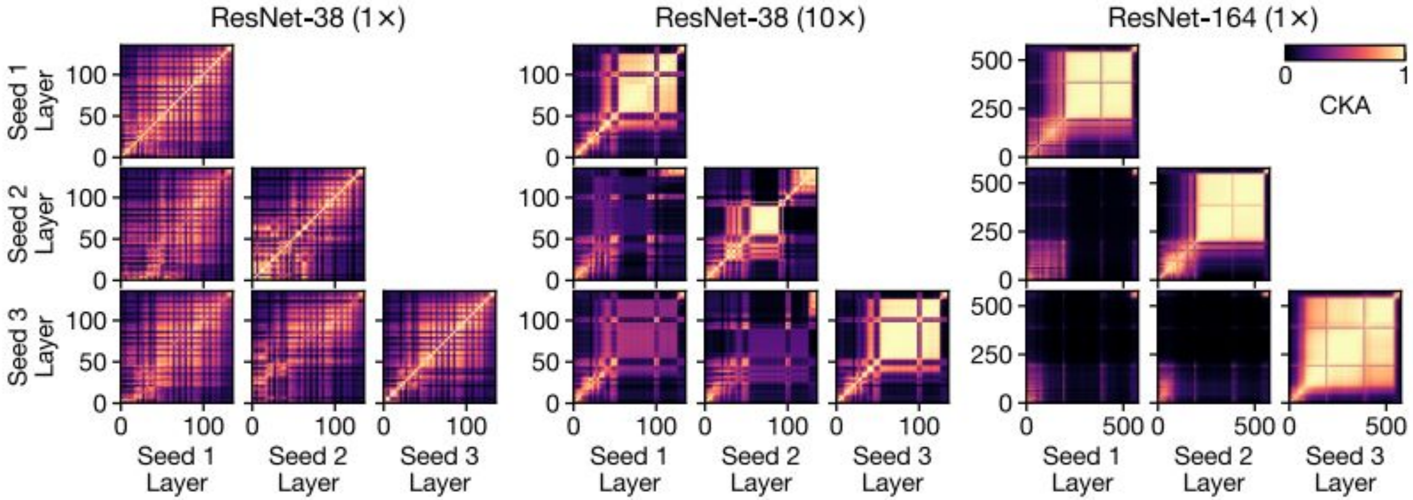
3.Accuracy related with linear probe & block structure

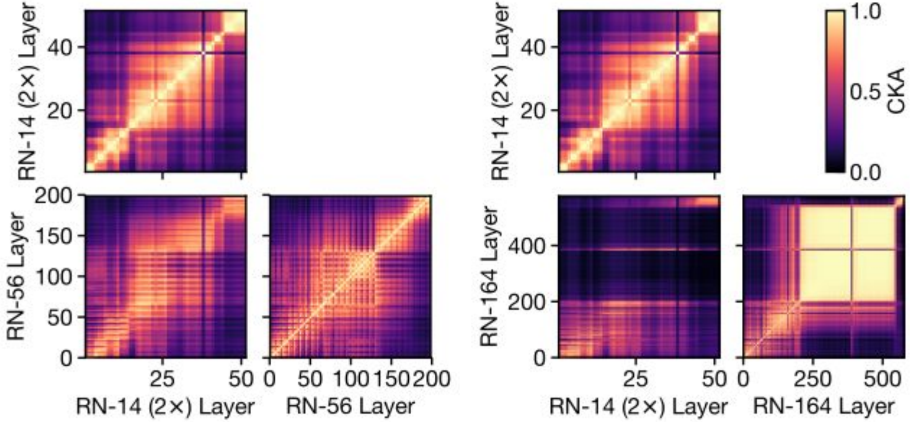


Without the block structure monotonic increase in accuracy throughout the network, with the block structure linear probe accuracy shows little improvement inside the block structure. Comparing the accuracies of probes for layers pre- and post-residual connections play an important role in preserving representations in the block structure.

Proceed to pruning blocks one-by-one from the end of each residual stage,
This result suggests that block structure could be an indication of redundant modules in model design, and that the similarity of its constituent layer representations could be leveraged for model compression.

3. Different initializations & model capacity

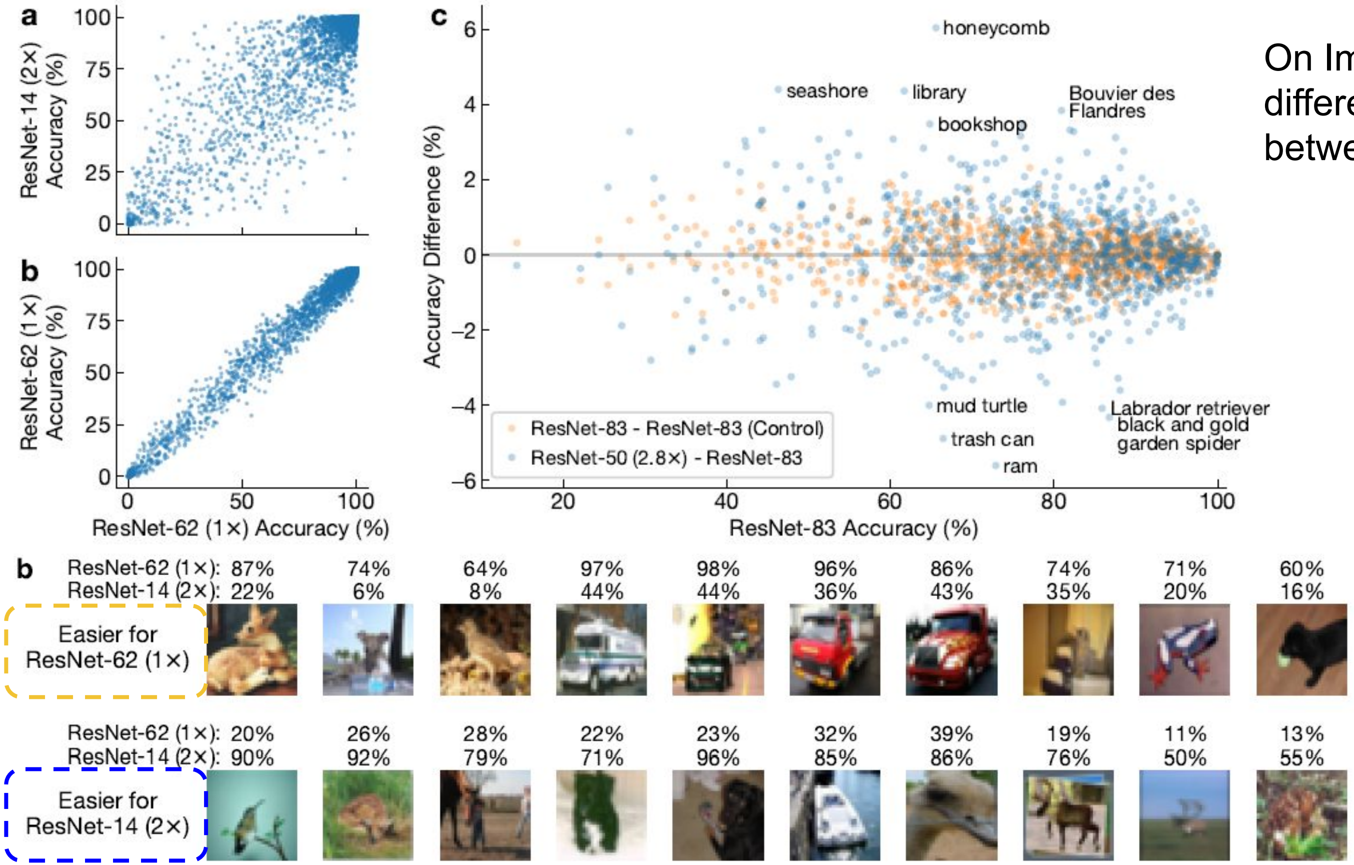




Representations across models

3.Depth and Width affects on Model prediction

CIFAR-10 **ImageNet**



On ImageNet there are statistically differences in class-level error rates between wide and deep models.

Width -> Scene
Depth -> Object

Cifar-10 : highest accuracy differences between the two types of models

3.Comparison of accuracy of wide and deep

Class	# Classes	Wide Acc.	Deep Acc.	Diff.	p-value
entity	1000	78.0 ± 0.01	78.0 ± 0.01	-0.03	0.89
physical entity	997	78.0 ± 0.01	78.0 ± 0.01	-0.03	0.89
object	958	78.1 ± 0.01	78.1 ± 0.01	-0.04	0.76
whole	949	78.2 ± 0.02	78.2 ± 0.01	-0.05	0.48
artifact	522	73.8 ± 0.02	73.8 ± 0.02	-0.01	1
living thing	410	83.5 ± 0.02	83.6 ± 0.02	-0.10	**0.023**
organism	410	83.5 ± 0.02	83.6 ± 0.02	-0.10	**0.023**
animal	398	83.3 ± 0.02	83.4 ± 0.02	-0.09	**0.032**
container	100	72.7 ± 0.05	72.7 ± 0.04	0.00	1
covering	90	72.0 ± 0.05	72.2 ± 0.05	-0.19	0.13
conveyance	72	83.5 ± 0.04	83.4 ± 0.05	0.13	0.65
vehicle	67	83.2 ± 0.04	83.1 ± 0.05	0.11	0.76
hunting dog	63	81.2 ± 0.05	81.2 ± 0.05	0.01	1
commodity	63	72.2 ± 0.06	72.6 ± 0.07	-0.42	$\mathbf{5.1 \times 10^{-5}}$
consumer goods	62	72.3 ± 0.06	72.7 ± 0.07	-0.41	$\mathbf{6.7 \times 10^{-5}}$
invertebrate	61	83.6 ± 0.05	83.8 ± 0.04	-0.16	0.37
bird	59	92.5 ± 0.04	92.7 ± 0.05	-0.21	**0.0018**
structure	58	75.9 ± 0.06	75.5 ± 0.07	0.42	$\mathbf{5.7 \times 10^{-5}}$
matter	50	77.6 ± 0.05	77.4 ± 0.05	0.17	0.74

P-values are computed using a t-test with multiple testing (Holm-Sidak) correction.

4.Conclusion

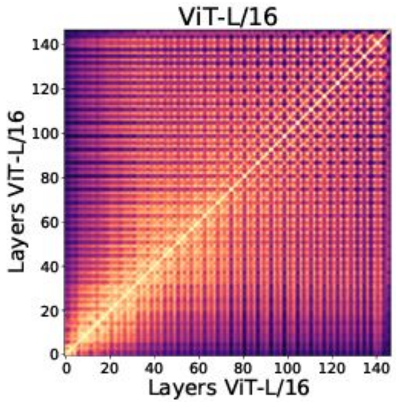
[Contribution]
Guiding researchers to **design networks.**(design wide and depth network for performance)
Similarity of constituent layer representations could be **leveraged for model compression. (Block Structure)**
Statistically significant differences in **class-level error rates** between wide and deep models.

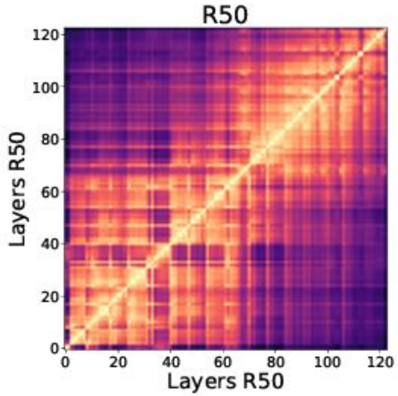
[Limitation]
Small dataset.(Cifar10 or Cifar100) more explore on Imagenet 1K.
Other Architecture (CNN, GAN and Transformer…)

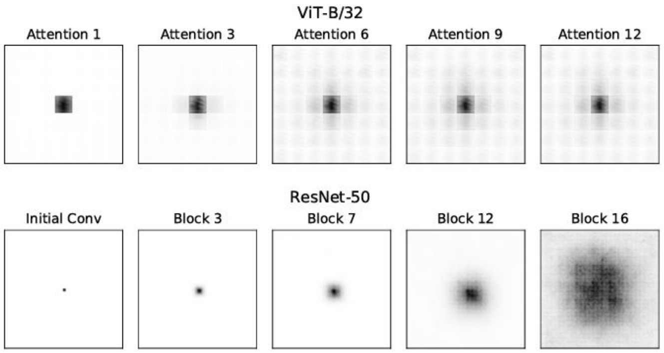
[Future Work]
How to design block transformer per stage. (ViT)
How does it related with Param and FLOPS.
Suppress block structure on training time.
Generalize to other Domain and Vision tasks(NLP, Detection).
Contrastive learning for feature similarity? (CKA).

Do Vision Transformers See Like Convolutional Neural Networks?

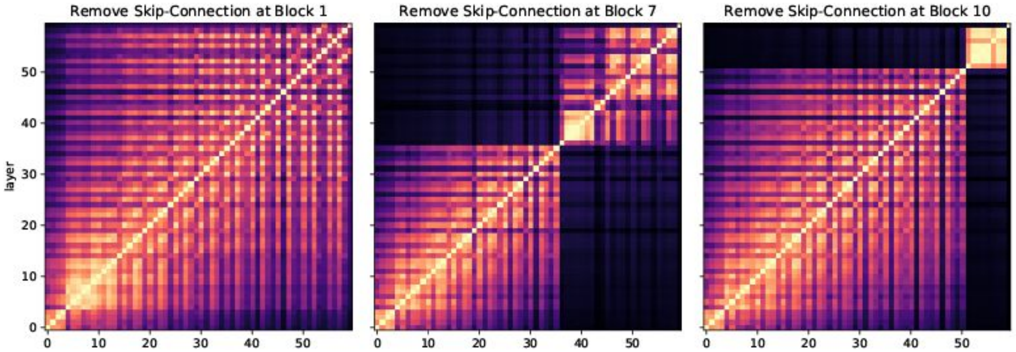
Analyzing the internal representation of ViTs and CNNs on image classification, we find differences between the two architectures, such as ViT having more uniform representations across all layers

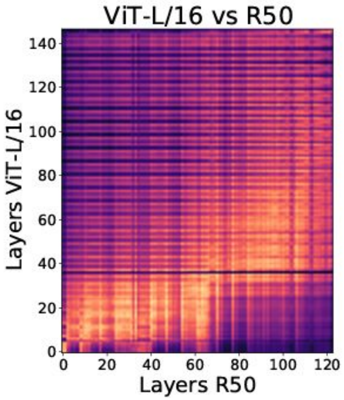






ViT models without skip connection ⇒ 4% drop



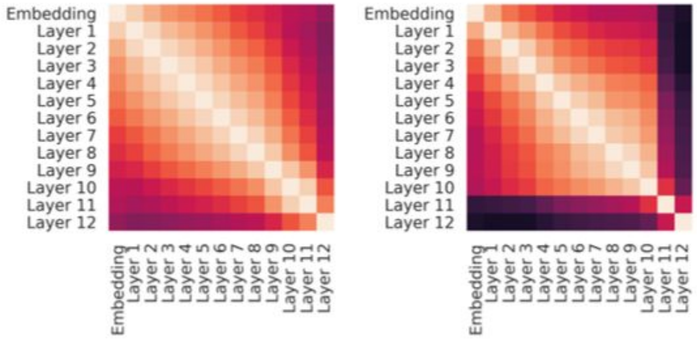


A good paper on a timely topic. All reviewers recommend acceptance. Could be a spotlight presentation.

Raghu, Maithra, et al. "Do vision transformers see like convolutional neural networks?." NeurIPS 2021

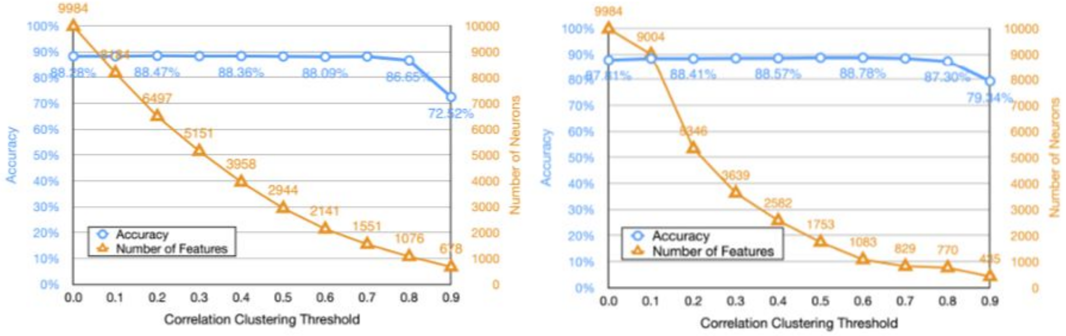
Analyzing Individual Neurons in Pre-trained Language Models

General Redundancy and Task-specific Redundancy. We dissect two popular pretrained models, **BERT and XLNet, studying how much redundancy** they exhibit at a representation-level and at a more fine-grained neuron-level



(a) BERT (b) XLNet

Figure 1: Pairwise Similarity between the layers. Brighter colors indicate higher similarity.



(a) BERT (b) XLNet

General neuron-level redundancy in BERT and XLNet; comparing the average reduction of neurons for different number of features

Adjacent layers are most redundant in the network, with **lower layers having greater redundancy with adjacent layers**. Comparing models, **XLNet is more redundant than BERT.**

Durrani, Nadir, et al. "Analyzing individual neurons in pre-trained language models." *EMNLP 2020.*

Openreview (ICLR 2021)

Neural networks with different architectures (width and depth learn similar representations). All reviewers agree that the investigations are thorough and the experimental discoveries are convincing and well explained.

Official Blind Review #1 (Rating 6: Marginally above acceptance threshold)
- I wonder if the **block structure arises dependent to the residual blocks**. I want to see more experiments with other network architectures. I expect to see an modified network architecture or a method to **balance the network size and accuracy** . However, just about theoretical analysis based on experiment phenomenon.

Official Blind Review #2 (Rating 8: Top 50% of accepted papers, clear accept)
- The most interesting and somewhat surprising finding is that even though two networks with different number of parameters and layers but with the same accuracy make very different mistakes, and there is a pattern to it. **The weakest part is the similarity analysis, which does not seem to reveal much new**. I propose lower score only due to the unclear choice of similarity function, as described above.

Official Blind Review #3 (Rating 6: Marginally above acceptance threshold)
- This is an interesting method and characterization of resnet behavior, with thorough experiments that tie together different aspects of the approach. **CKA is used to show a type of blockwise similarity**, much of which is subsequently explained, and related experimentally to classification performance using linear probes through the layers.

Official Blind Review #4 (Rating 7: Good paper, accept)
- In my humble opinion, the paper is very clearly written, presenting at the beginning of each section the scientific question they try to answer. Do the authors have solid reasons to believe that **their findings generalize to other neural models** (other ConvNets, recurrent, generative,...) and problems (regression, dense prediction,...?

https://openreview.net/forum?id=KJNcAkY8tY4

Thanks
Any Questions?

You can send mail to
Susang Kim(healess1@gmail.com)