Vision Transformer in Healthcare: Harnessing the Power and Unraveling the Trade-offs

Presented by Sadaf Farkhani

Danish Research Center for Magnetic Resonance

Aim

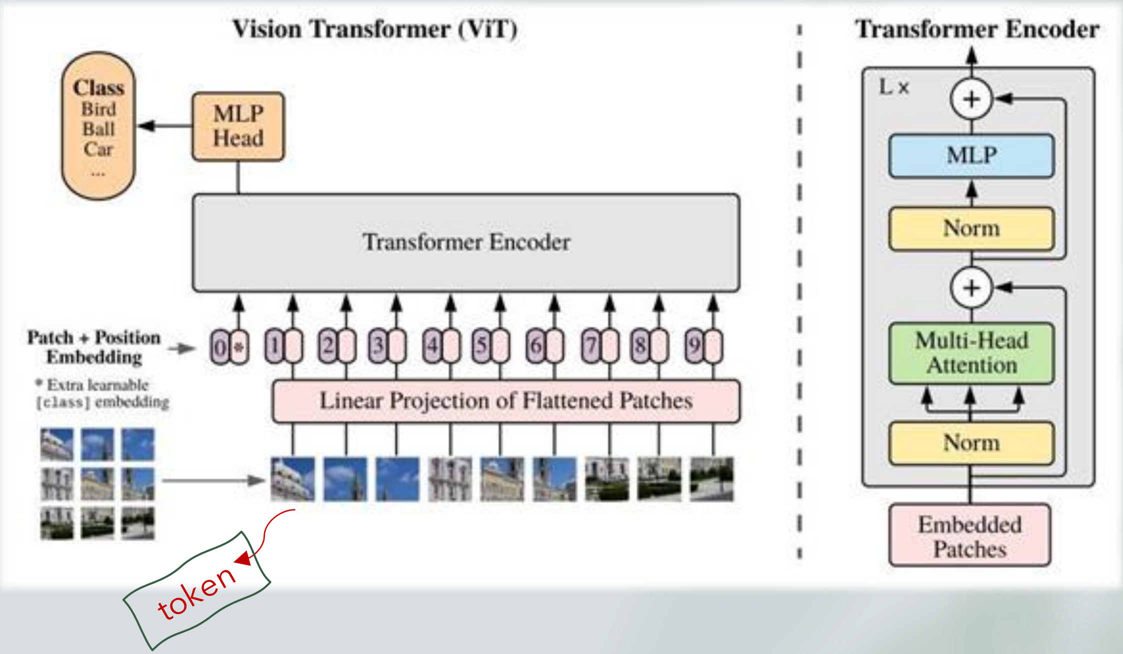
- Introduce the fundamental concepts of Vision Transformer (ViT)
- Explore the application of ViT in the context of 3D medical images
- Conducting a comparison between CNNs and ViT
- ViT or CNN? That is the question :)

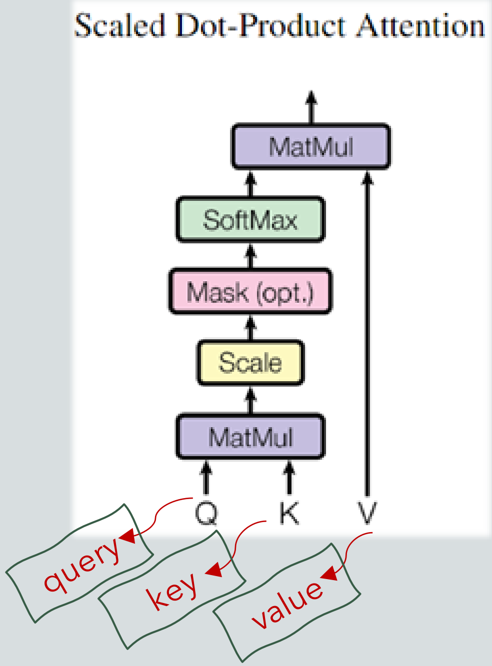
Vision Transformer (ViT)

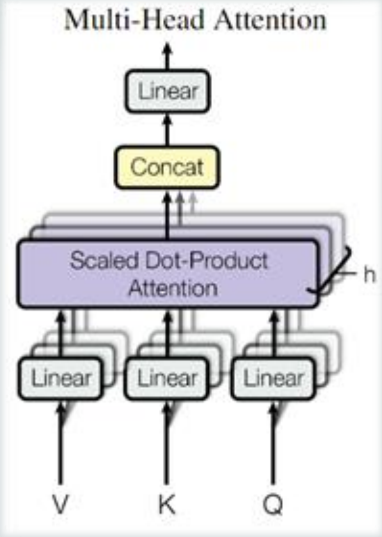
- Position embedding
- Multi-head attention

$$Attention = Softmax(\frac{QK^T}{d_k})$$

- Feed-forward layers (MLP)

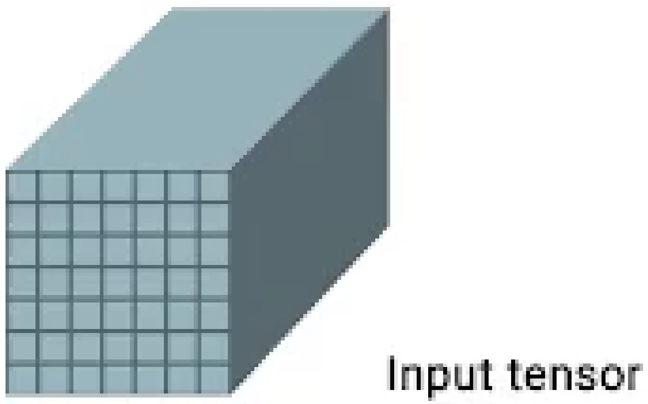






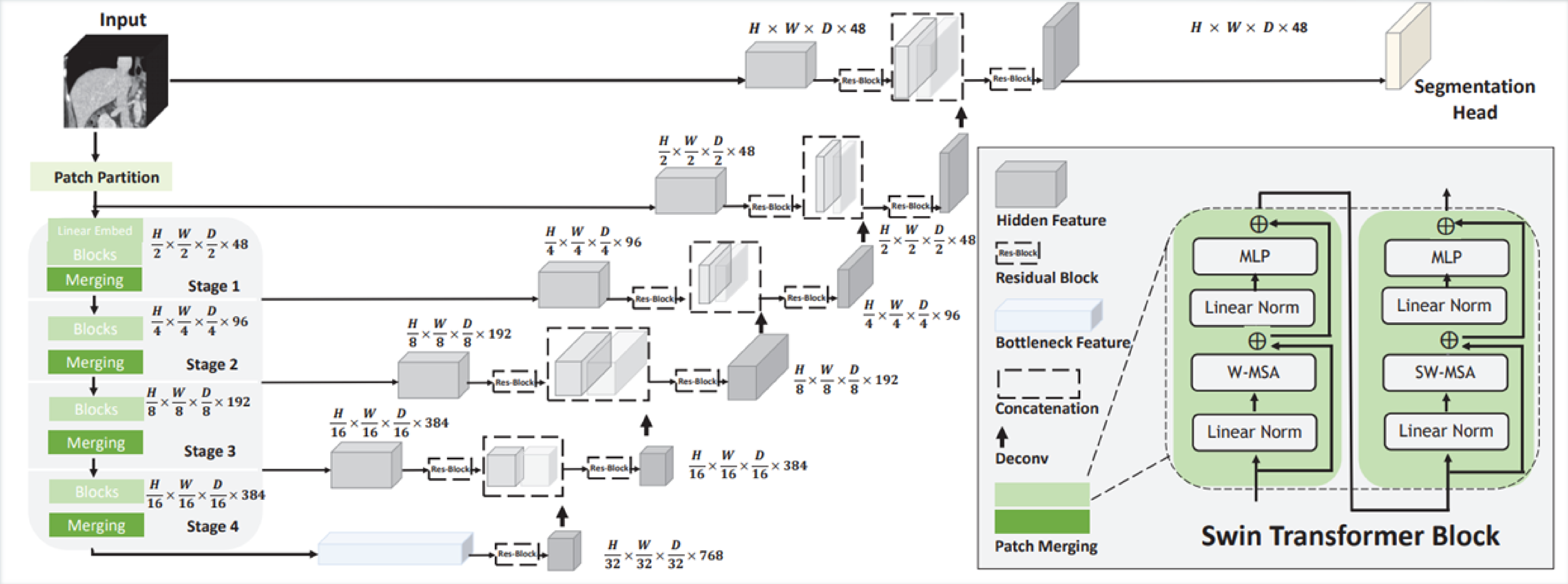
Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations. 2020

Multi-head Attention



https://ai.googleblog.com/2021/12/improving-vision-transformer-efficiency.html

3D Medical images (Swin-UNETR)

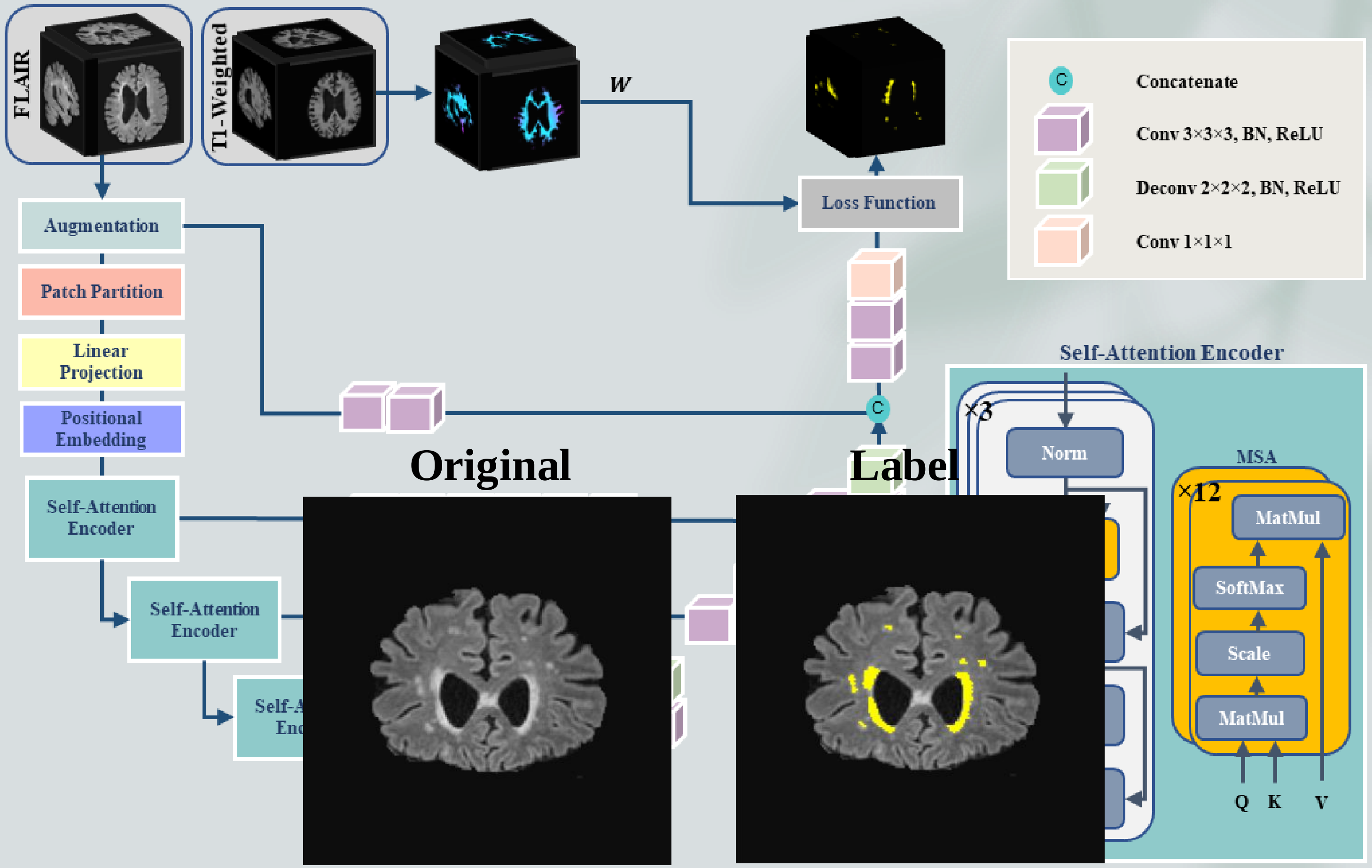


Tang, Yucheng, et al. "Self-supervised pre-training of swin transformers for 3d medical image analysis. " *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

3D Medical Images (VoSHT)

Task: brain lesion segmentation

- Sparse imbalanced lesions
- Multiple modalities are required
- More input modalities → more parameters
- Weighted loss function



S. Farkhani, et. al, "End-to-end Volumetric Segmentation of White Matter Hyperintensities: Effect of Data, Model, and Loss Function", under review.

Results

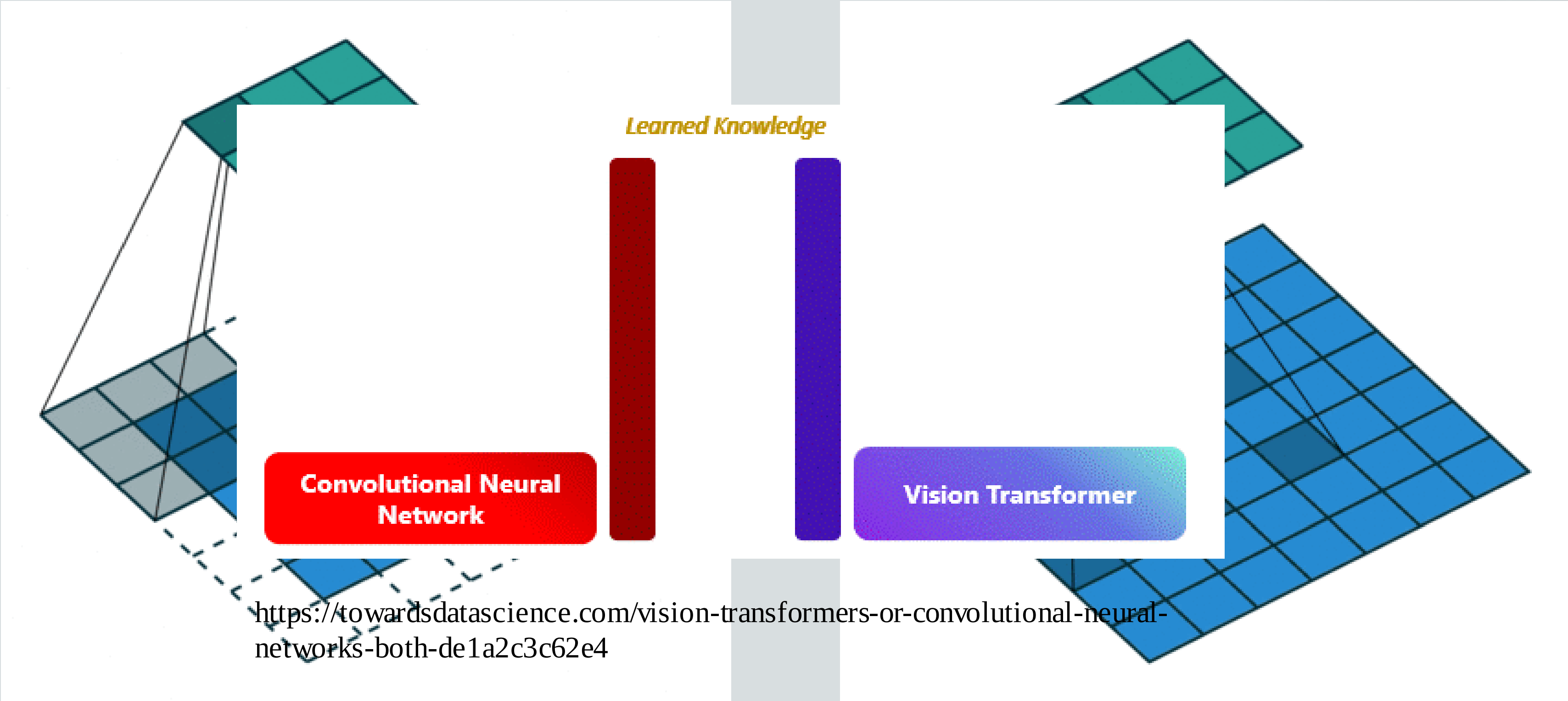
In-distribution Test set

Out-of-distribution Test set

METHOD	DSC	MSD	HD	RECALL	AUC-PR
LST	71.71	6.30	22.84	76.61	46.92
BIANCA	74.62	5.73	20.53	87.08	57.41
UNET	84.25	2.28	12.59	86.62	70.19
UNETR	81.88	2.76	15.61	80.83	65.76
TrUENet	80.94	2.61	13.79	76.63	66.50
UNET+	85.43	1.82	11.12	87.48	72.41
VoSHT	**85.59**	**1.76**	**10.30**	**87.48**	**72.67**

Metric	**DSC**	**HD**	**MSD**	**Recall**	**AUC-PR**
UNET	82.115	16.062	3.180	78.983	67.255
UNETR	78.811	13.709	3.254	74.489	62.374
TrUENet	83.453	15.834	3.570	78.890	64.720
UNET+	**85.533**	**9.157**	**1.711**	**83.736**	**72.376**
VoSHT	84.132	14.022	2.638	81.491	70.199

1. Global vs. Local



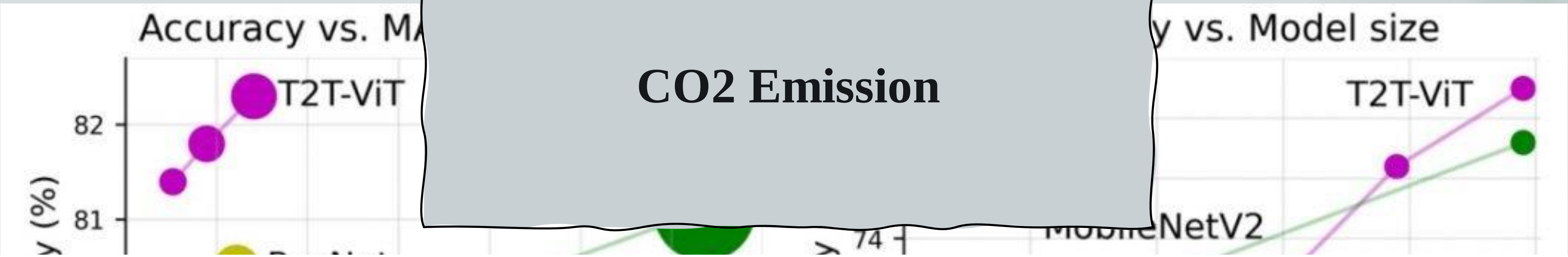
https://towardsdatascience.com/vision-transformers-or-convolutional-neural-networks-both-de1a2c3c62e4

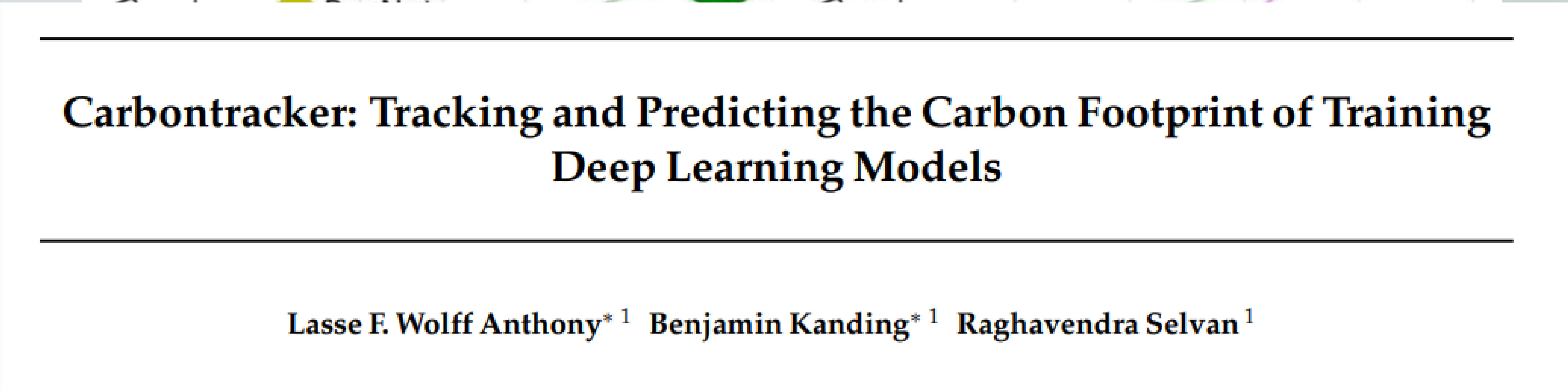
Conv. layer

Atrous Conv. layer

2. Parameters

CO2 Emission





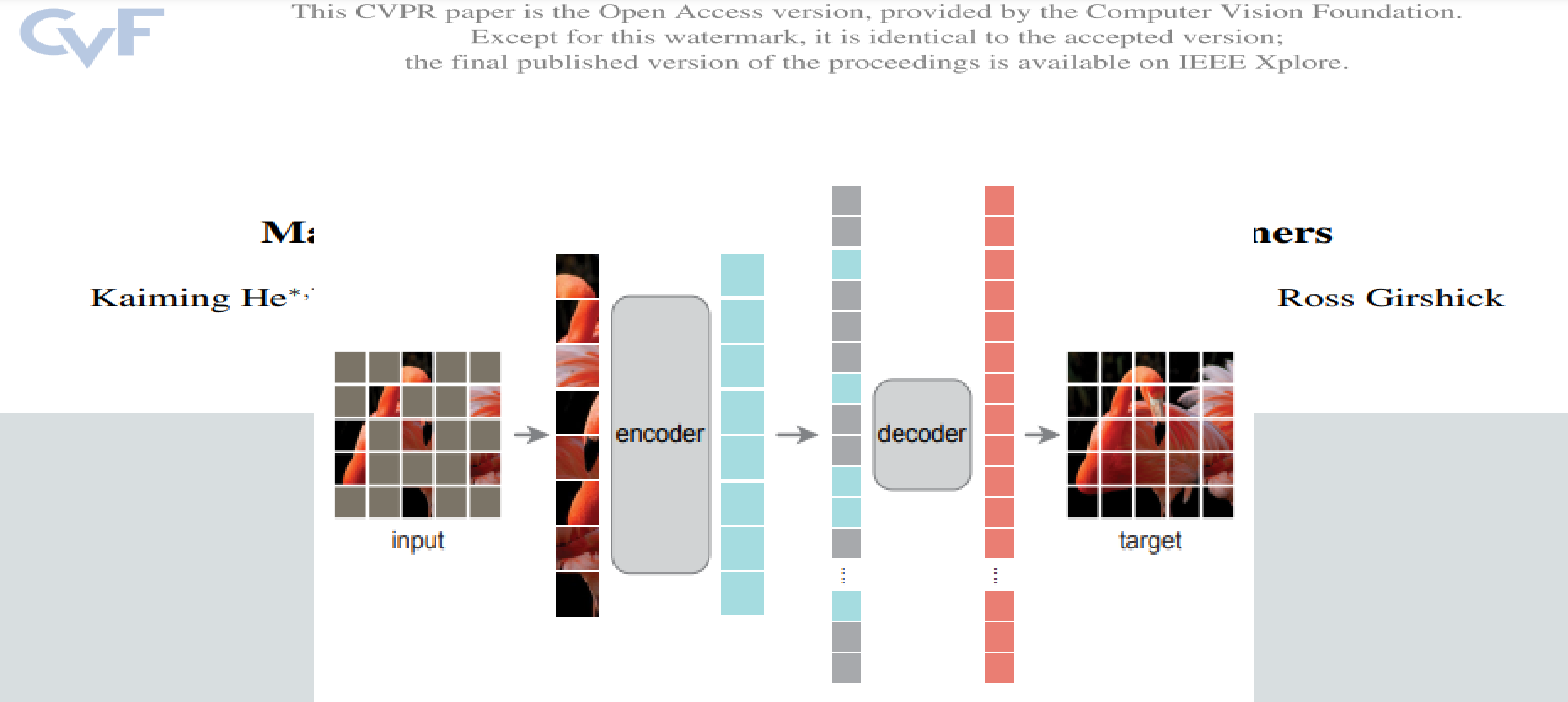
Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models

Lasse F. Wolff Anthony* [1] Benjamin Kanding* [1] Raghavendra Selvan [1]

Yuan, Li, et al. "Tokens-to-token vit: Training vision transformers from scratch on imagenet." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

Anthony, Lasse F. Wolff, Benjamin Kanding, and Raghavendra Selvan. "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models." *arXiv preprint arXiv:2007.03051* (2020).

2. Parameters



He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

2. Parameters

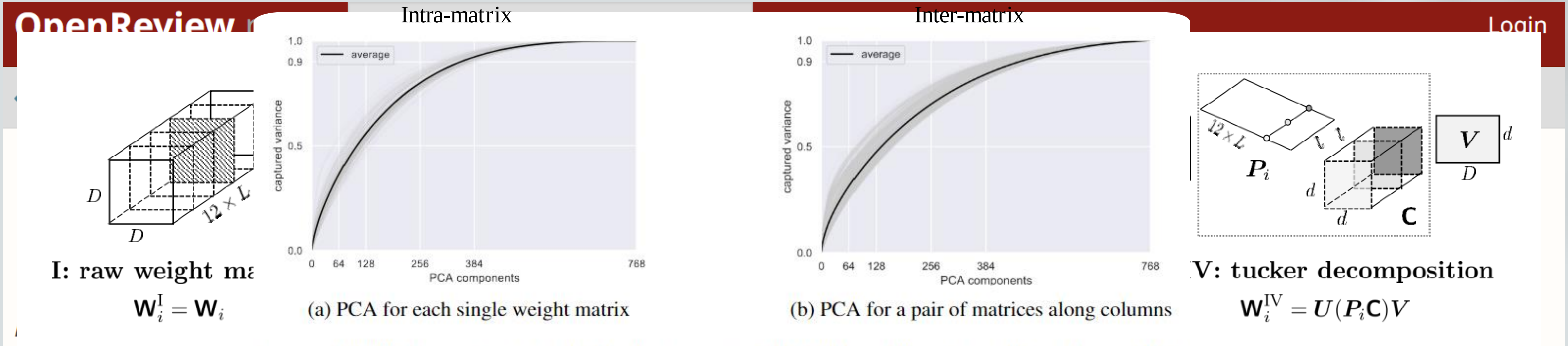


Figure 1: PCA for existing weight block matrices in BERT-base. We got nearly similar results in Fig. 5 for paired matrices along rows and columns, as shown in App. C.

Ren, Yuxin, et al. "Exploring extreme parameter compression for pre-trained language models." *arXiv preprint arXiv:2205.10036* (2022).

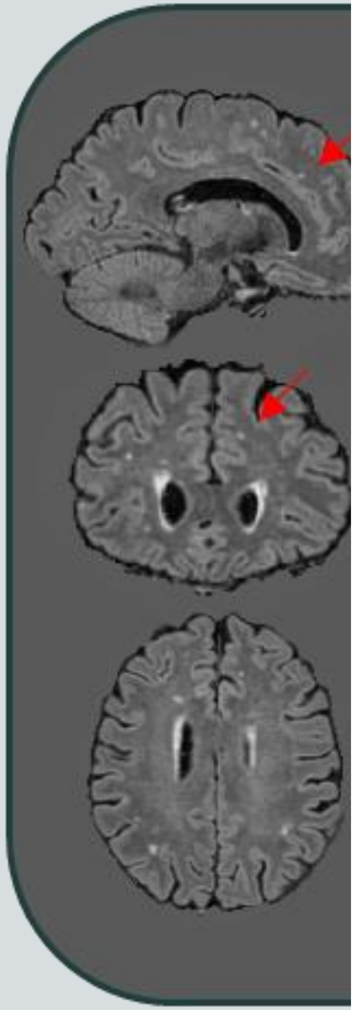
3. Generalization

In-di… n Test set

Sagittal

Coronal

Axial



Are Tr…

[1]Johns …
{ytongb…

Transform…
demonstrat…
recent work…
Neural Net…
are drawn …
compared …
In this pap…
Transforme…
With our …
Transforme…
surprisingl…
against adv…
While rega…
training on…
Transforme…
suggest su…
self-attenti…
hope this …
robustness …
at https:/…

Can CNNs Be More Robust Than Transformers?

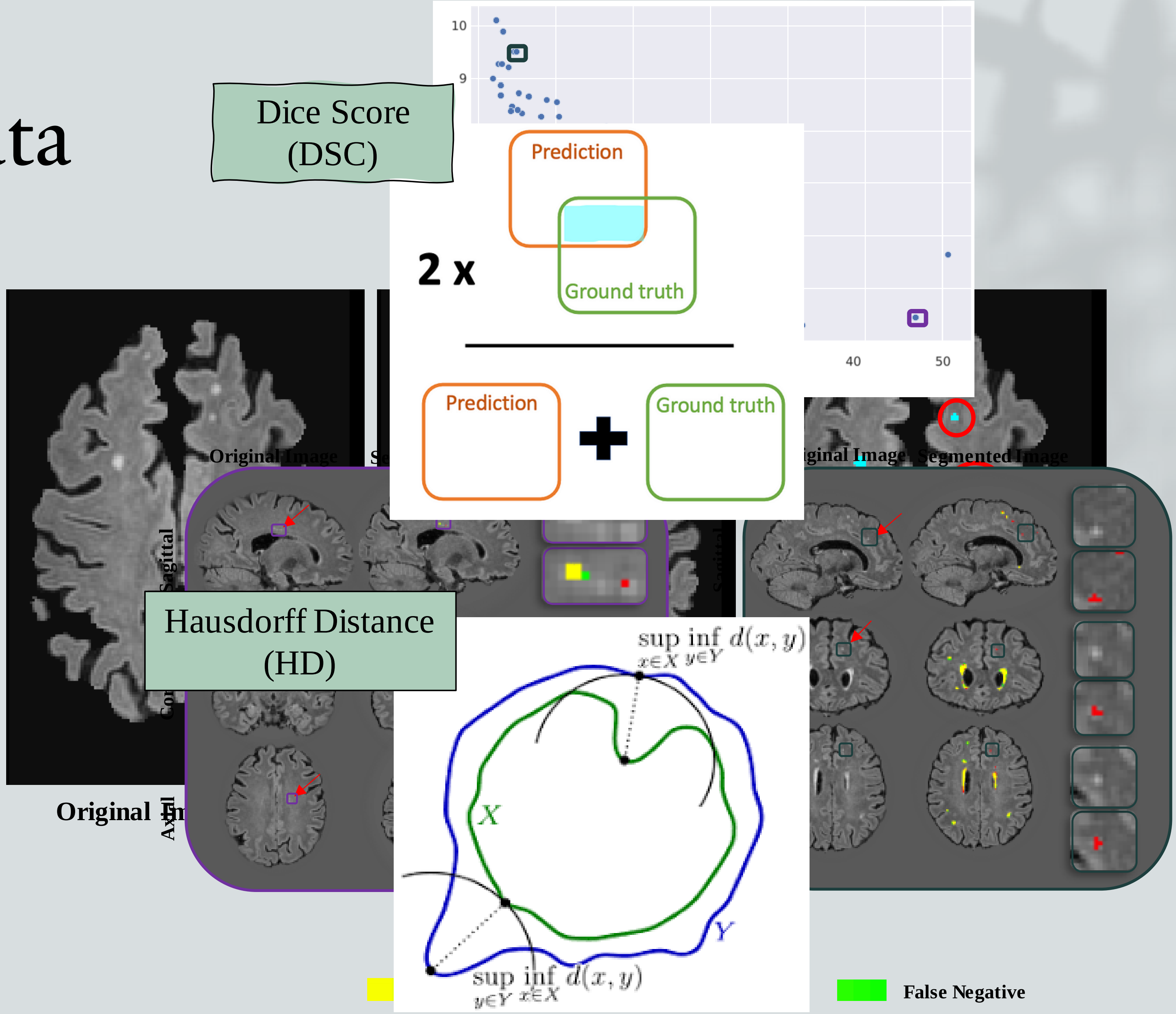
Zeyu Wang[1] **Yutong Bai**[2] **Yuyin Zhou**[1] **Cihang Xie**[1]
[1]UC Santa Cruz [2]Johns Hopkins University

Abstract

The recent success of Vision Transformers is shaking the long dominance of Convolutional Neural Networks (CNNs) in image recognition for a decade. Specifically, in terms of robustness on out-of-distribution samples, recent research finds that Transformers are inherently more robust than CNNs, regardless of different training setups. Moreover, it is believed that such superiority of Transformers should largely be credited to their *self-attention-like architectures per se*. In this paper, we question that belief by closely examining the design of Transformers. Our findings lead to three highly effective architecture designs for boosting robustness, yet simple enough to be implemented in several lines of code, namely a) patchifying input images, b) enlarging kernel size, and c) reducing activation layers and normalization layers. Bringing these components together, we are able to build pure CNN architectures without any attention-like operations that are as robust as, or even more robust than, Transformers. We hope this work can help the community better understand the design of robust neural architectures. The code is publicly available at https://github.com/UCSC-VLAA/RobustCNN.

4. Real-world Data

- Imperfect labels
- Imperfect metrics



CNN vs. ViT

ViT	CNN
- Globally-capturing dependencies - High number of parameters - High redundancy (within attention head and between layers of Transformers) - Higher demand for GPU/data accessibility - Higher CO2 emission - Perform better on remarkably big objects	- Locally-capturing dependencies - Lower number of parameters - High redundancy among kernels - Lower demand for GPU/data accessibility - Perform better on small objects - Better generalization performance

Summary

- Necessity of pret
- Achieving balan odules
- Environmental i ects of your model
- Priority on generalization: significance in real-world applications

Achieving Synergy: Harmonizing Between CNN and Transformer

Reference

[1] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations. 2020.

[2] Tang, Yucheng, et al. "Self-supervised pre-training of swin transformers for 3d medical image analysis. " *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[3] https://colab.research.google.com/drive/1hXIQ77A4TYS4y3UthWF-Ci7V7vVUoxmQ?usp=sharing#scrollTo=twSVFOM9SopW

[4] Yuan, Li, et al. "Tokens-to-token vit: Training vision transformers from scratch on imagenet." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[5] *He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.*

[6] *Ren, Yuxin, et al. "Exploring extreme parameter compression for pre-trained language models." arXiv preprint arXiv:2205.10036 (2022).*

[7] *S. Farkhani, N. Demnitz, CJ. Baroxbekk, H. Lundell, H. R. Siebner, E. T. Petersen, K. H. Madsen, "End-to-end Volumetric Segmentation of White Matter Hyperintensities: Effect of Data, Model, and Loss Function", submitted to Computer Methods and Programming in Biomedicine, 2023.*

[8] Bai, Yutong, et al. "Are transformers more robust than cnns?." Advances in neural information processing systems 34 (2021): 26831-26843.

[9] Wang, Zeyu, et al. "Can CNNs Be More Robust Than Transformers?." The Eleventh International Conference on Learning Representations. 2022.