How Do Vision Transformers Work? (ICLR 2022)

Group #3

Gunho Park, Jihoon Lee and Junseo Jo

Department of Electrical Engineering

POSTECH, Korea

Introduction

Empirical observations from prior works

✓ 1. Multi-head self-attentions (MSAs) improve the predictive performance of CNNs

✓ 2. ViTs are robust against data corruptions, image occlusions, and adversarial attacks

✓ 3. MSAs closer to the last layer significantly improve predictive performance

Three key questions

✓ 1. Multi-head self-attentions (MSAs) improve the predictive performance of CNNs

→ What properties of MSAs do we need to improve optimization?

✓ 2. ViTs are robust against data corruptions, image occlusions, and adversarial attacks

→ Do MSAs act like Convs?

✓ 3. MSAs closer to the last layer significantly improve predictive performance

→ How can we harmonize MSAs with Convs?

Q1. What properties of MSAs do we neet to improve optimization?

Group #3

Gunho Park, Jihoon Lee and Junseo Jo

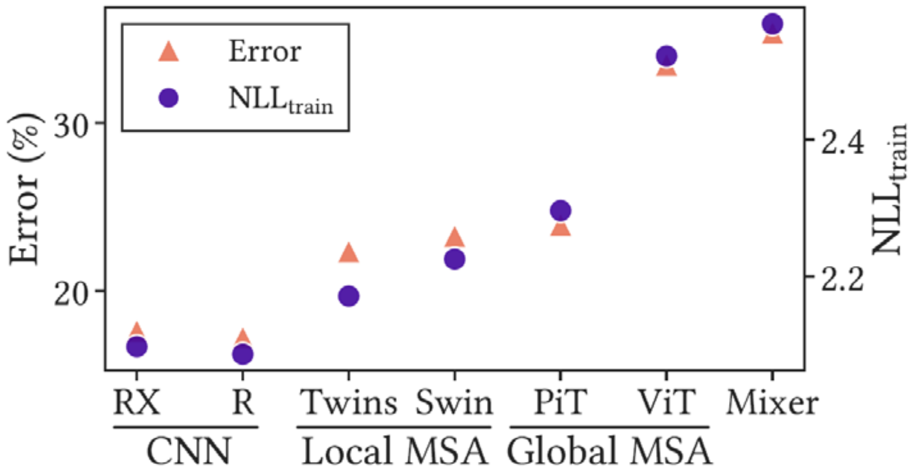
Department of Electrical Engineering

POSTECH, Korea

What properties of MSAs do we need to improve optimization?

The stronger the inductive biases, the stronger the representations.

- ✓ Contrary to our expectations, the stronger the inductive bias, the lower both test error and the training negative log-likelihood (NLL)
- ✓ Weak inductive biases disrupt NN training
- ✓ Models with strong inductive biases (CNNs) show better performance compared to models with weak inductive biases (MSAs)

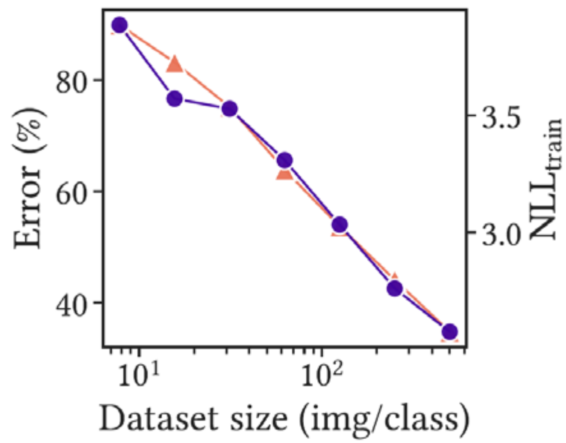


(a) Error and NLL$_{train}$ for each model.

What properties of MSAs do we need to improve optimization?

ViT does not overfit small training datasets.

- ✓ As the size of the dataset decreases, not only the error but also NLL increases
- ✓ If ViT is overfitted to small training datasets, NLL of train dataset should not increase.
- ✓ ViT's poor performance in small data regimes is not due to overfitting



(b) Performance of ViT for dataset size.

What properties of MSAs do we need to improve optimization?

What makes ViT show poor performance in small data regimes?

→ **ViT's non-convex losses lead to poor performance**

✓ The loss function of ViT is non-convex, while that of ResNet is strongly (near-)convex.

✓ ViT has a number of negative Hessian eigenvalues, while ResNet only has a few in the early stage of training

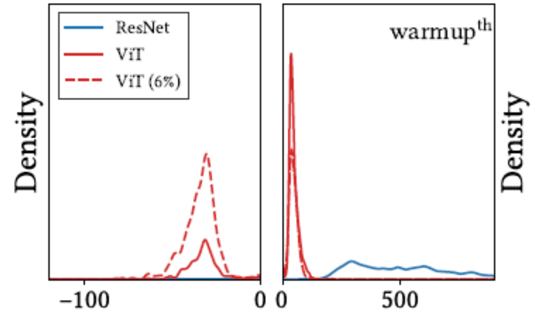


Figure 4: **Hessian max eigenvalue spectra show that MSAs have their advantages and disadvantages.** The dotted line is the spectrum of ViT using 6% dataset for training. *Left*: ViT has a number of negative Hessian eigenvalues, while ResNet only has a few. *Right*: The magnitude of ViT's positive Hessian eigenvalues is small. See also Fig. 1c for more results.

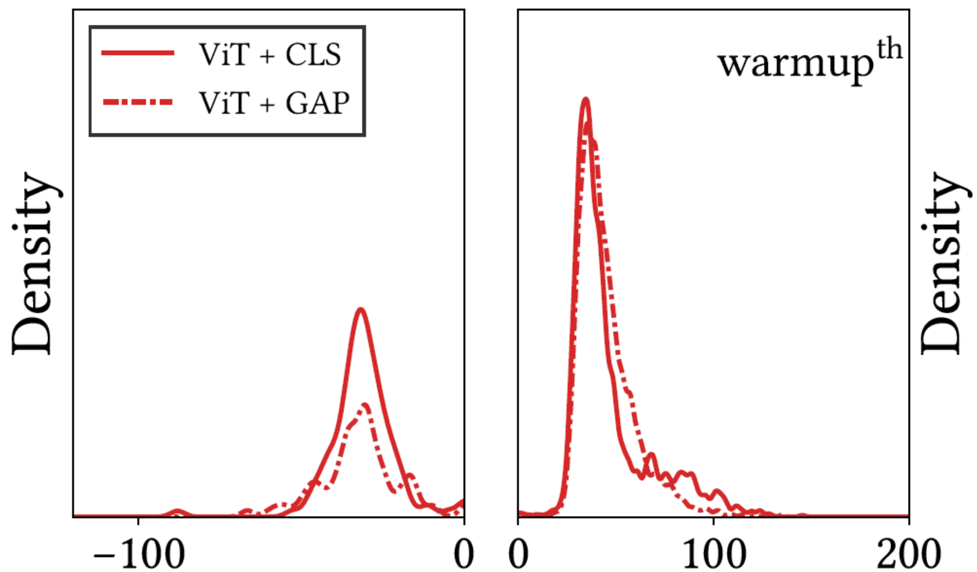
What properties of MSAs do we need to improve optimization?

Loss landscape smoothing methods aids in ViT training.

✓ Replace class (CLS) token to Global average pooling (GAP) classifier

✓ GAP classifier suppresses negative Hessian max eigenvalues in an early phase of training

- GAP classifier improves the accuracy by +2.7%

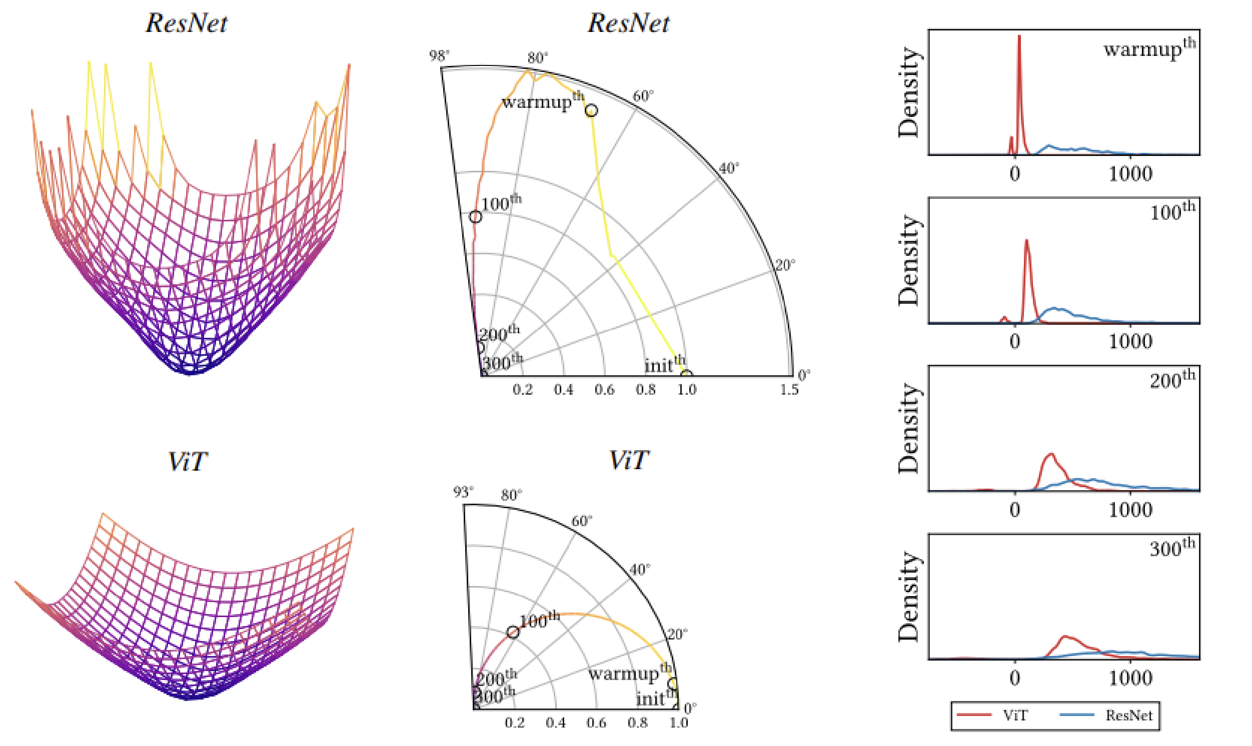


What properties of MSAs do we need to improve optimization?

MSAs flatten the loss landscape.

✓ MSAs reduce the magnitude of Hessian eigenvalues.

- Helps NNs learn better representations



(a) Loss landscape visualizations (b) Trajectories in polar coordinates (c) Hessian max eigenvalue spectra

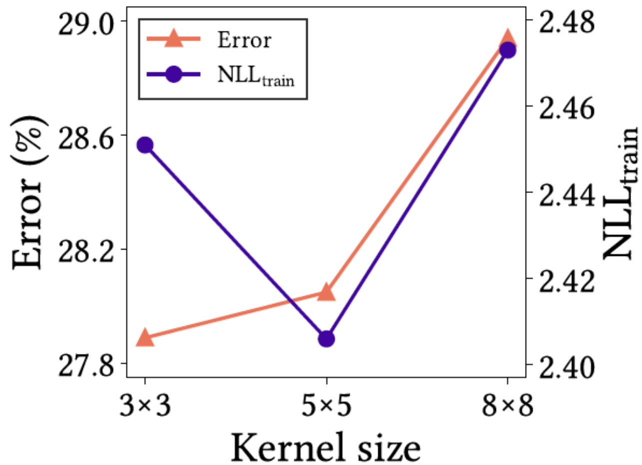
What properties of MSAs do we need to improve optimization?

A key feature of MSAs is data specificity (not long-range dependency).

✓ The two distinguishing features of MSAs are long-range dependency and data specificity

✓ The long-rage dependency hinders NN optimization

- 5 x 5 kernel (Local MSA) outperforms 8 x 8 kernel (Global MSA)
- 3 x 3 is worse than 5 x 5 but better than 8 x 8 kernel



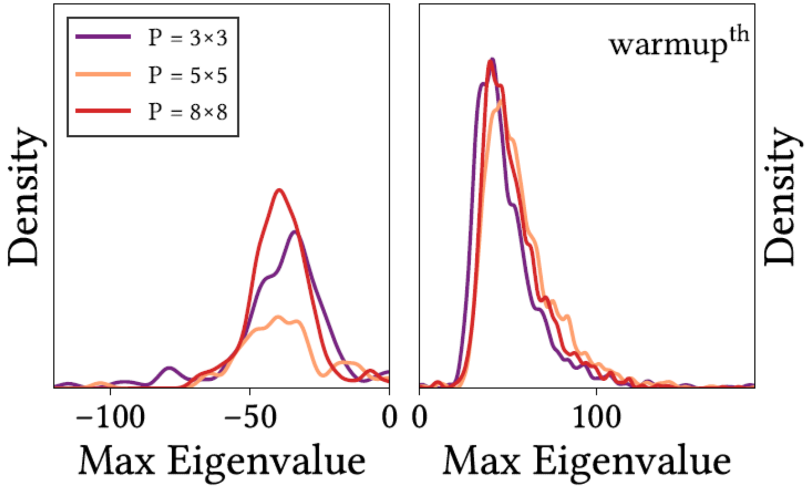
(a) Error and NLL_{train} of ViT with local MSA for kernel size

What properties of MSAs do we need to improve optimization?

A key feature of MSAs is data specificity (not long-range dependency).

✓ The strong locality inductive bias not only reduce computational complexity but also aid in optimization by convexifying the loss landscape.

- 5 x 5 is better than 8 x 8 (unnecessary degrees of freedom)
- 5 x 5 is better than 3 x 3 (ensembles a larger number of feature map points)



(b) Hessian negative and positive max eigen-value spectra in early phase of training

Q2 & Q3

Group #3

Gunho Park, Jihoon Lee and Junseo Jo

Department of Electrical Engineering

POSTECH, Korea

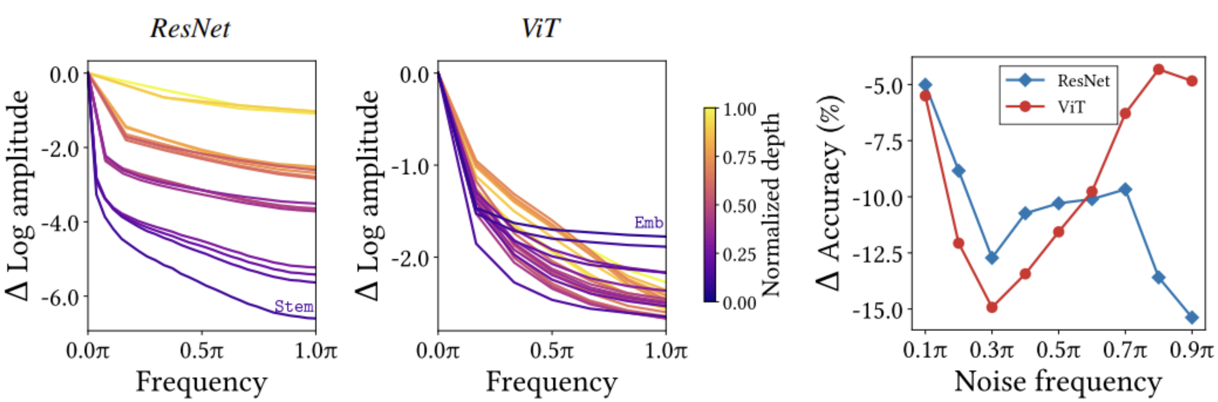
Do MSAs act like Convs?

MSAs and Convs exhibit opposite behaviors

✓ MSAs reduce high-frequency signals, while Convs amplifies high frequency components

- MSAs : low-pass filter / Convs : high-pass filters

MSAs and Convs are complementary



(a) Relative log amplitudes of Fourier transformed feature maps. (b) Robustness for noise frequency

How can we harmonize MSAs with Convs?

Applying spatial smoothing at the end of a stage improves accuracy

The authors propose an alternating pattern of Convs and MSAs network (AlterNet)

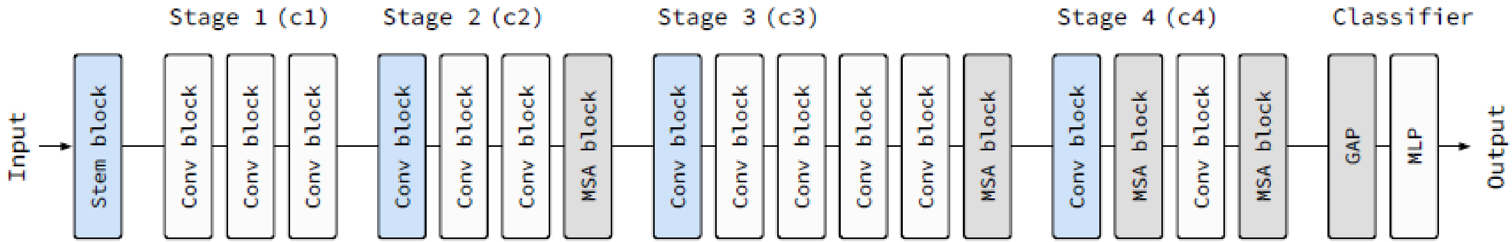
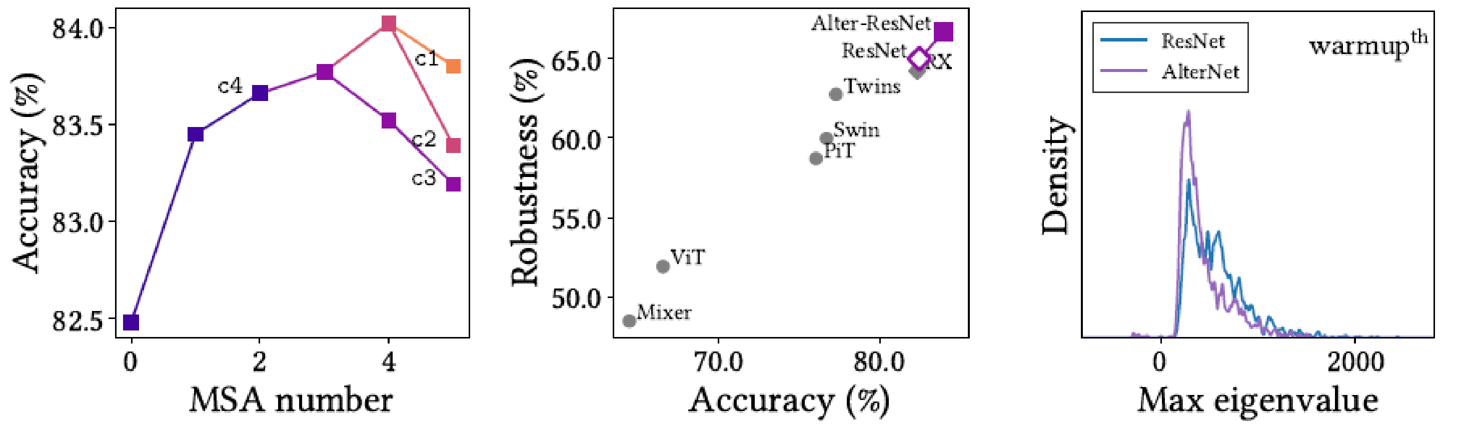


Figure 11: **Detailed architecture of Alter-ResNet-50 for CIFAR-100.** White, gray, and blue blocks mean Conv, MSA, and subsampling blocks. All stages (except stage 1) end with MSA blocks. This model is based on pre-activation ResNet-50. Following Swin, MSAs in stages 1 to 4 have 3, 6, 12, and 24 heads, respectively.

How can we harmonize MSAs with Convs?

AlterNet outperforms CNNs not only on large datasets but also on small datasets



(a) Accuracy of AlterNet for MSA number

(b) Accuracy and robustness in a small data regime (CIFAR-100)

(c) Hessian max eigenvalue spectra in an early phase of training

Figure 12: **AlterNet outperforms CNNs and ViTs.** *Left:* MSAs in the late of the stages improve accuracy. We replace Convs of ResNet with MSAs one by one according to the build-up rules. `c1` to `c4` stands for the stages. Several MSAs in `c3` harm the accuracy, but the MSA at the end of `c2` improves it. *Center:* AlterNet outperforms CNNs even in a small data regime. Robustness is mean accuracy on CIFAR-100-C. "RX" is ResNeXt. *Right:* MSAs in AlterNet suppress the large eigenvalues; i.e., AlterNet has a flatter loss landscape than ResNet in the early phase of training.

Final project plan

Group #3

Gunho Park, Jihoon Lee and Junseo Jo

Department of Electrical Engineering

POSTECH, Korea

Our project experiment plan

We want to check

✓ Effect of flattening loss landscape on ViT

✓ Robustness of ViT in large dataset

by comparing

✓ Training speed and final accuracy

of ResNet and ViT in

✓ different optimizers

✓ and their hyperparameter settings

Thank you