How Do Vision Transformers Work?

Abstract

fundamental explanations to help better understand the nature of MSAs
following properties of MSAs and Vision Transformers (ViTs)

1 MSAs improve not only accuracy but also generalization by flattening the loss landscapes
2 MSAs and Convs exhibit opposite behaviors
3 Multi-stage neural networks behave like a series connection of small individual models

AlterNet
Conv blocks at the end of a stage are replaced with MSA blocks

Introduction

MSA의 성공 방식?
weak inductive bias and capture of long-range dependencies
over-flexibility
tendency to overfit - small data에 불리 - 사실 아님

RELATED WORK

Self-attentions
weak inductive bias of MSA가 좋을까?
appropriate constraints may actually help - local MSAs

MSAs intriguing properties
1 MSAs improve the predictive performance of CNNs
2 ViTs are robust against data corruptions, image occlusions (Naseer et al., 2021), and adversarial attacks
3 MSAs closer to the last layer significantly improve predictive performance

의문점
1 What properties of MSAs do we need to better optimize NNs? Do the long-range dependencies of MSAs help NNs learn?
2 Do MSAs act like Convs? If not, how are they different?
3 How can we harmonize MSAs with Convs? Can we just leverage their advantages?

MSA = a trainable spatial smoothing of feature maps

$$\boldsymbol{z}_j = \sum_i \texttt{Softmax}\left(\frac{\boldsymbol{QK}}{\sqrt{d}}\right)_i \boldsymbol{V}_{i,j}$$

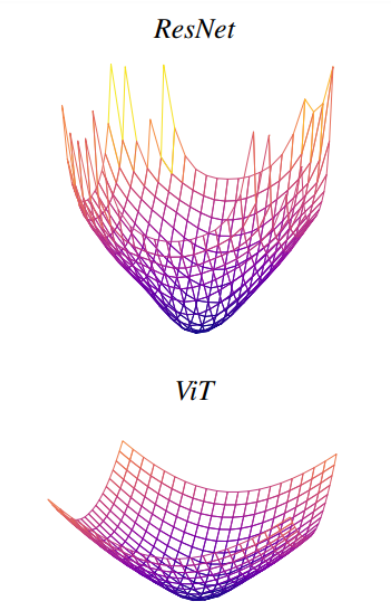
spatial smoothings not only improve accuracy but also robustness by spatially ensembling feature map points and flattening the loss landscapes

CONTRIBUTION

1 What properties of MSAs do we need to improve optimization?
Their weak inductive bias disrupts NN training

a key feature of MSAs is their data specificity, not long-range dependency.
local MSAs with a 3 × 3 receptive field outperforms global MSA because they reduce unnecessary degrees of freedom

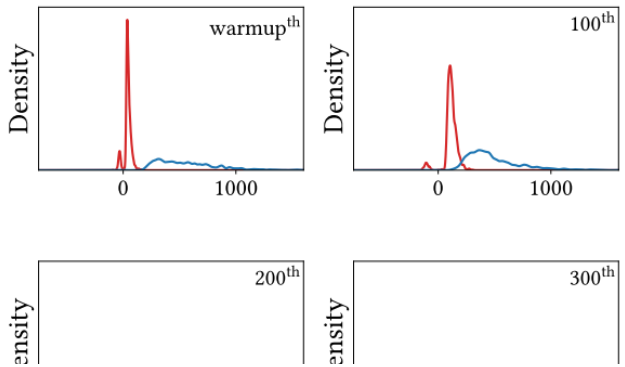
장점

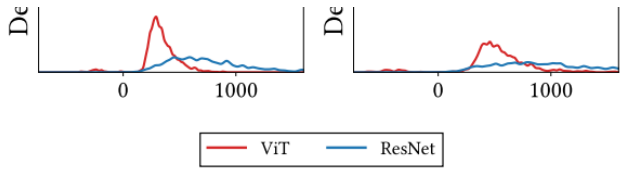


(a) Loss landscape visualizations

improve not only accuracy but also robustness in large data regimes

단점





(b) Hessian max eigenvalue spectra

MSAs allow negative Hessian eigenvalues in small data regimes.

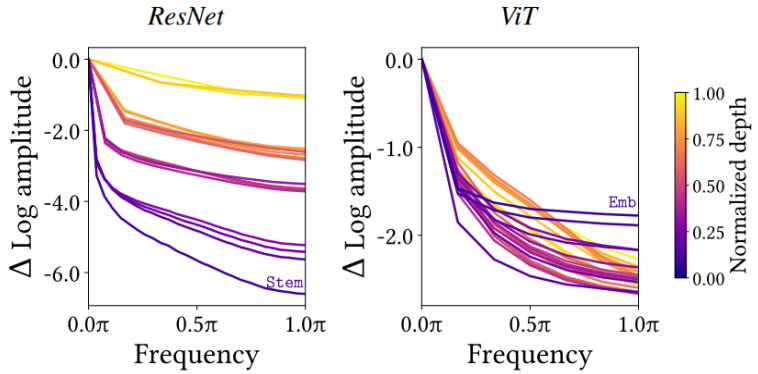
loss landscapes of MSAs are non-convex

non-convexity disturbs NN optimization

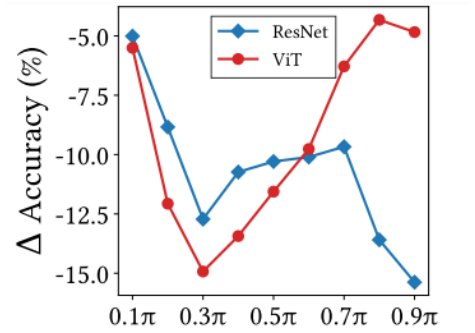
2 Do MSAs act like Convs?

exhibit opposite behaviors

MSAs aggregate feature maps, but Convs diversify them



(a) Relative log amplitudes of Fourier transformed feature maps.

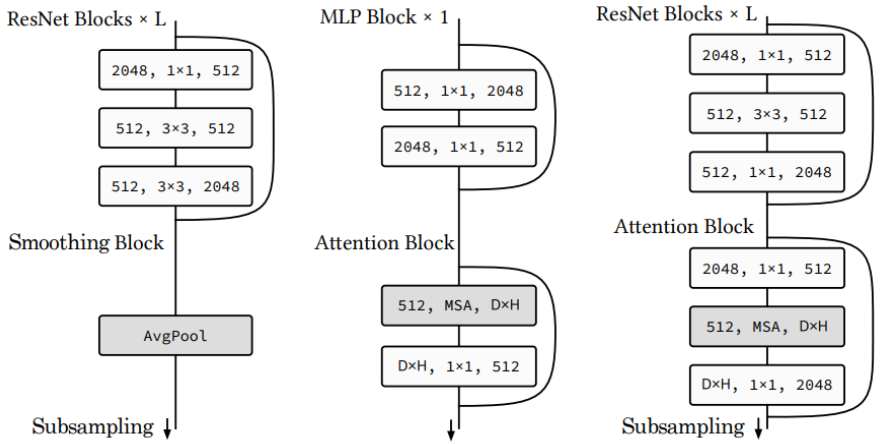


(b) Robustness for noise frequency

Convs are vulnerable to high-frequency noise but that MSAs are not

3 How can we harmonize MSAs with Convs?

multi-stage NNs behave like a series connection of small individual models



(a) Spatial smoothing (b) Canonical Transformer (c) Alternating pattern (*ours*)

MSAs are generalized spatial smoothings that complement Convs, not simply generalized Convs that replace conventional Convs

WHAT PROPERTIES OF MSAS DO WE NEED TO IMPROVE OPTIMIZATION?

vanilla ViT

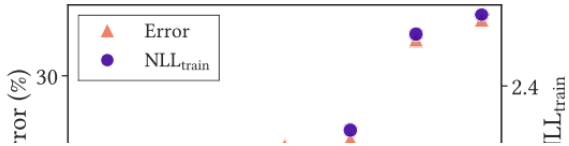
PiT : ViT + multi-stage

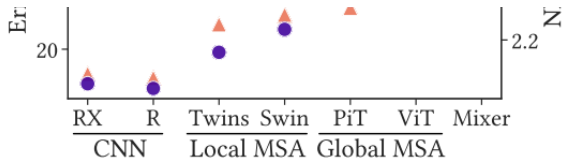
Swin : ViT + multi-stage + local MSA

inductive biases enable ViTs to learn strong representations

The stronger the inductive biases, the stronger the representations (not regularizations)

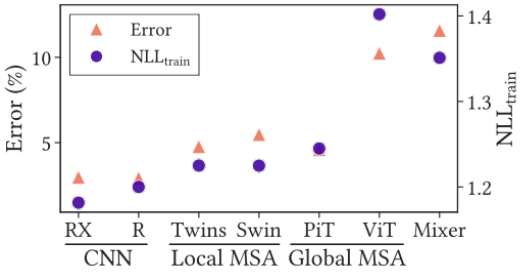
weak inductive biases -> overfit training datasets?



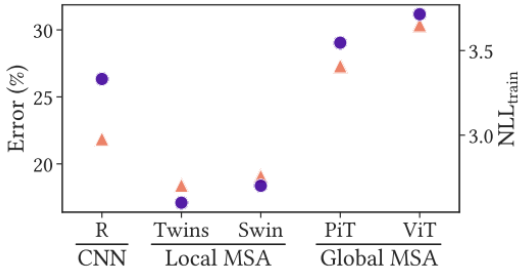


(a) Error and NLL_{train} for each model.

stronger the inductive bias, the lower both the test error and the training NLL

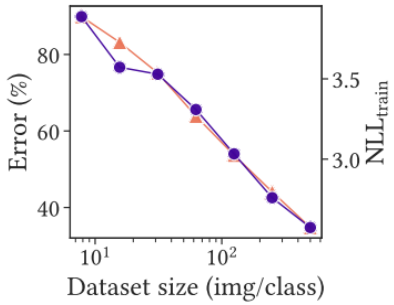


(a) CIFAR-10



(b) ImageNet

ViT does not overfit small training datasets.



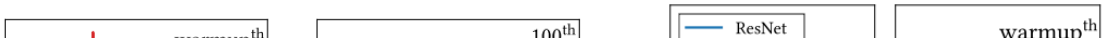
(b) Performance of ViT for dataset size

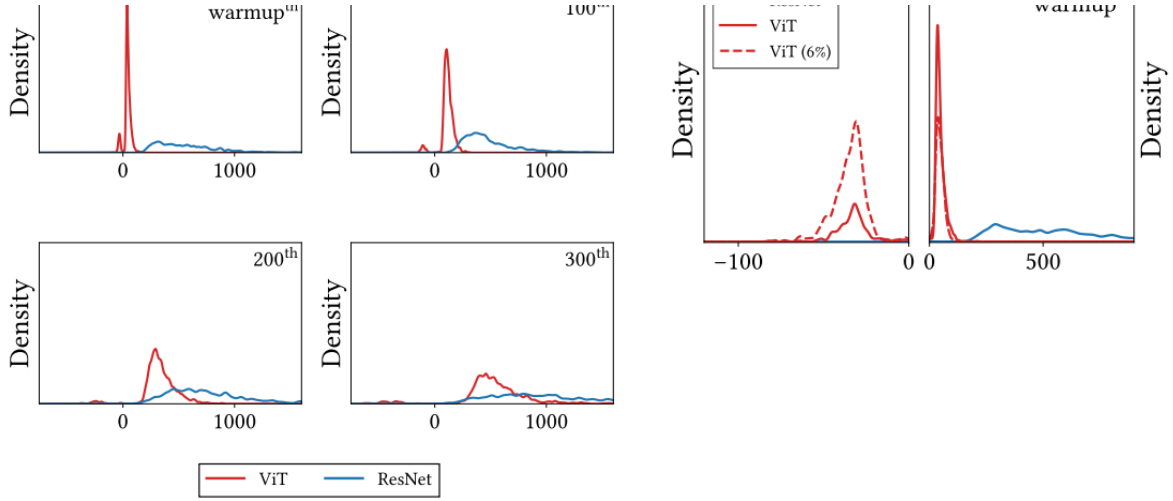
error & NLL 같이 오름

ViT's poor performance in small data regimes is not due to overfitting

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





(b) Hessian max eigenvalue spectra

Loss landscape smoothing methods aids in ViT training

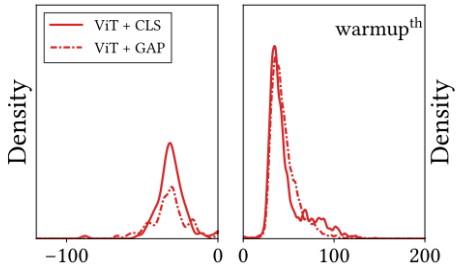


Figure 6. **GAP classifier suppresses negative Hessian max eigenvalues** in an early phase of training. We present Hessian max eigenvalue spectrum of ViT with GAP classifier instead of CLS token.

GAP classifier suppresses negative Hessian max eigenvalues, suggesting that GAP convexify the loss.

MSAs flatten the loss landscape

they reduce the magnitude of Hessian eigenvalues.

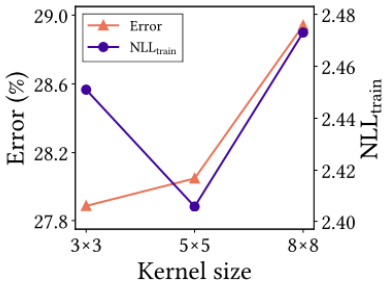
eigenvalues of ViT are significantly smaller than that of CNNs

While large eigenvalues impede NN training

In large data regimes, the negative Hessian eigenvalues disappears

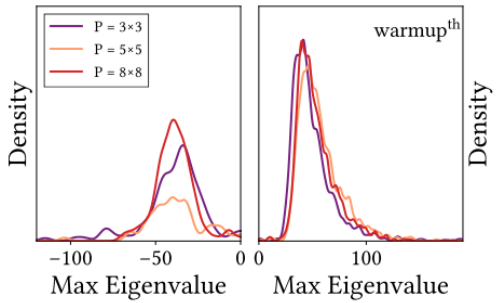
A key feature of MSAs is data specificity (not long-range dependency)
long-range dependency and data specificity

long-range dependency hinders NN optimization



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

strong locality inductive bias
reduce computational complexity as originally proposed
aid in optimization by convexifying the loss landscape



(b) Hessian negative and positive max eigenvalue spectra in early phase of training

5x5 kernel restricts unnecessary degrees of freedom -> less negative eigenvalues

data specificity improves NNs

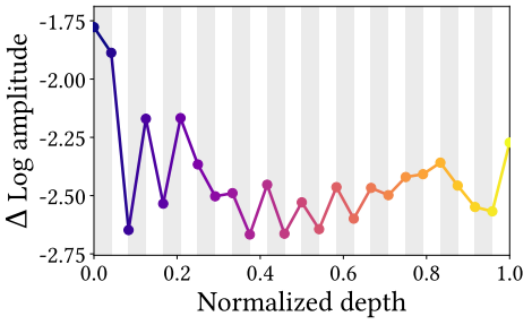
DO MSAS ACT LIKE CONVS?

Convs are data-agnostic and channel-specific in that they mix channel information without exploiting data information.

MSAs, on the contrary, are data-specific and channel-agnostic

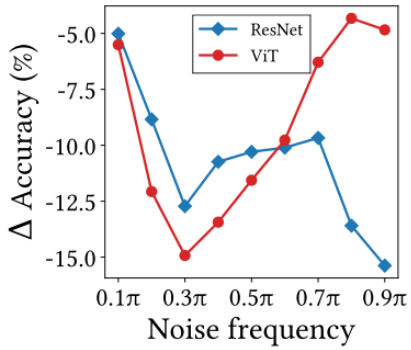
MSAs are low-pass filters, but Convs are highpass filters.

MSAs will tend to reduce high-frequency signals.



MSAs almost always decrease the high-frequency amplitude, and MLPs—corresponding to Convs— increase it.

The only exception is in the early stages of the model. In these stages, MSAs behave like Convs, i.e., they increase the amplitude



(b) Robustness for noise frequency

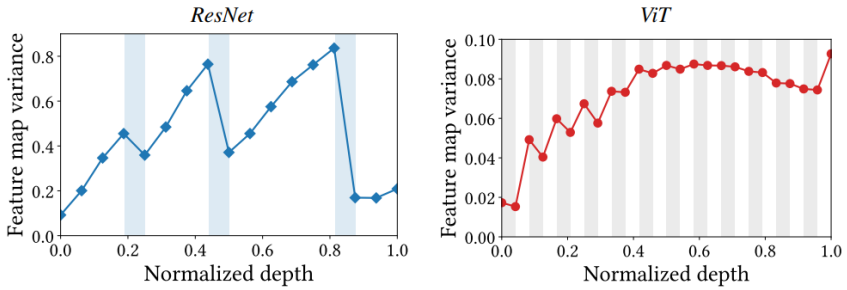
ViT and ResNet are vulnerable to low-frequency noise and high-frequency noise, respectively

Low-frequency signals and the high-frequency signals each correspond to the shape and the texture of images.

The results thus suggests that MSAs are shape-biased (Naseer et al., 2021), whereas Convs are texture-biased

MSAs aggregate feature maps, but Convs do not.

MSA - average feature maps - reduce variance of feature map points



MSAs (gray area) reduce the variance of feature map points, but Convs (white area) increase the variance.

reducing the feature map uncertainty helps optimization by ensembling and stabilizing the transformed feature maps

two additional patterns for feature map variance

1 the variance accumulates in every NN layer and tends to increase as the depth increases

2 the feature map variance in ResNet peaks at the ends of each stage

Therefore, we can improve the predictive performance of ResNet by inserting MSAs at the end of each stage

we also can improve the performance by using MSAs with a large number of heads in late stages.

HOW CAN WE HARMONIZE MSAS WITH CONVS?

leverages only the advantages of the two modules.

DESIGNING ARCHITECTURE

Multi-stage NNs behave like individual models

the pattern of feature map variance repeats itself at every stages

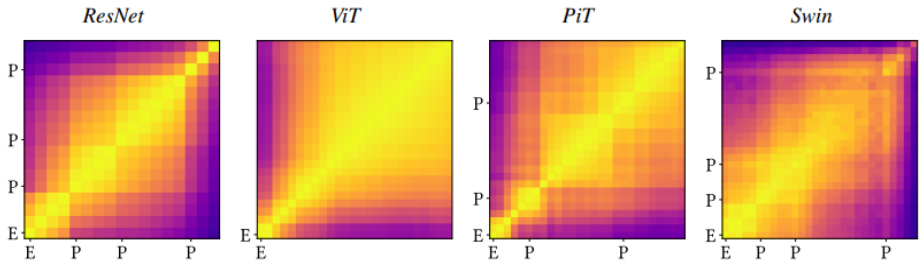
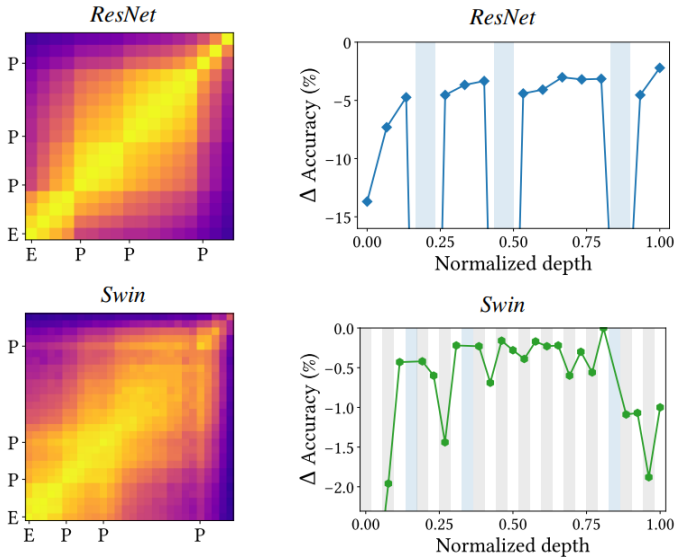


Figure D.3. **Multi-stage ViTs have block structures in representational similarities**. Block structures can be observed in all multi-stage NNs, namely, ResNet, PiT, and Swin. "E" is the

stem/embedding and "P" is the pooling (subsampling) layer.

feature map similarities of CNNs have a block structure

multi-stage ViTs, such as PiT and Swin, also have a block structure.



(a) Feature map similarity (b) Accuracy of one-unit-removed model.

Figure 10. **Multi-stage CNNs and ViTs behave like a series connection of small individual models.** *Left:* The feature map similarities show the block structure of ResNet and Swin. "E" stands for stem/embedding and "P" for pooling (subsampling) layer. *Right:* We measure decrease in accuracy after removing one unit from the trained model. Accuracy changes periodically, and this period is one stage. White, gray, and blue areas are Conv/MLP, MSA, and subsampling layers, respectively.

removing a layer at the beginning of a stage impairs accuracy more than removing a layer at the end of a stage

In ResNet, removing an early stage layers hurts accuracy more than removing a late stage layers

The case of Swin is even more interesting. At the beginning of a stage, removing an MLP hurts accuracy. At the end of a stage, removing an MSA seriously impairs the accuracy.

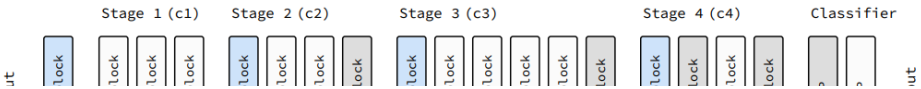
MSAs closer to the end of a stage to significantly improve the performance

Build-up rule. AlterNet

replace Conv blocks with MSA blocks from the end of a baseline CNN model

If the added MSA block does not improve predictive performance, replace a Conv block located at the end of an earlier stage with an MSA block

Use more heads and higher hidden dimensions for MSA blocks in late stages



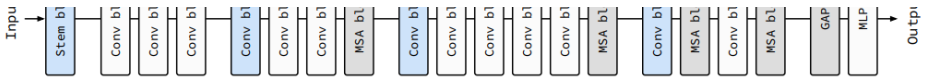
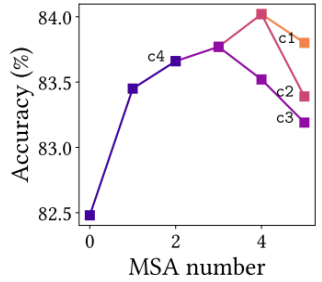
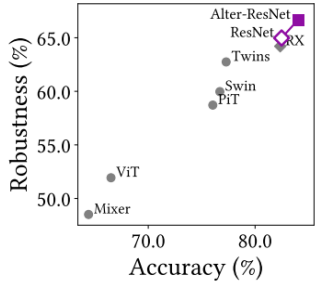


Figure 11. **Detailed architecture of Alter-ResNet-50 for CIFAR-100.** White, gray, and blue blocks mean Conv, MSA, and subsampling blocks. All stages (except stage 1) end with MSA blocks. This model is based on pre-activation ResNet-50. Following Swin, MSAs in stages 1 to 4 have 3, 6, 12, and 24 heads,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

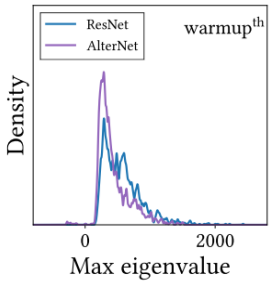
PERFORMANCE



(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

DISCUSSION

MSAs are not merely generalized Convs

generalized spatial smoothings that complement Convs.

help NNs learn strong representations by ensembling feature map points and flattening the loss landscape

MSA to be able to significantly improve the results in dense prediction tasks by ensembling feature maps

strong data augmentation for MSA training harms uncertainty calibration