How Do Vision Transformers Work?

Namuk Park, Songkuk Kim
namuk.park@yonsei.ac.kr

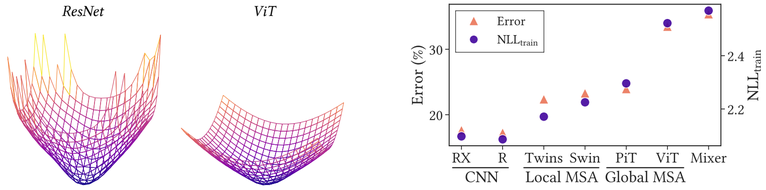
How Do Vision Transformers Work?

Namuk Park[1,2], Songkuk Kim[1]
[1]Yonsei University, [2]NAVER AI Lab

Paper: https://arxiv.org/abs/2202.06709
Code: https://github.com/xxxnell/how-do-vits-work

What Properties of Self-Attentions Do We Need?

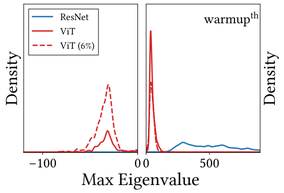
MSAs (multi-head self-attention) have flat but non-convex losses. In contrast, Convs have convex but sharp losses.



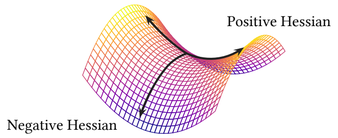
(a) Loss landscape visualization

(b) Test error and training NLL for each model

Figure 1: *Left:* **Loss landscape visualization (*Left*)** show that ViT has a flatter loss than ResNet. *Right*: Weak inductive bias (e.g. long-range dependency) disturbs NN optimization.

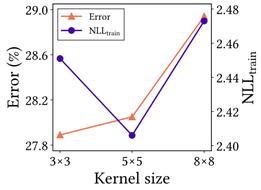


(a) Hessian max eigenvalue spectrum

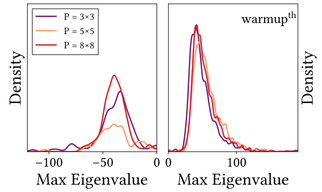


(b) Local geometry of non-convex (saddle) points

Figure 2: **Hessian max eigenvalue spectra** show that MSAs have their pros and cons. ViT has a number of negative Hessian eigenvalues, while ResNet only has a few. The magnitude of ViT's positive Hessian eigenvalues is small.



(a) Error and NLL of ViT with local MSA for kernel size



(b) Hessian max eigenvalue spectrum for kernel size

Figure 3: The key feature of MSA is data specificity, not long-range dependency. ***Left***: Convolutional ViT demonstrates that locality constraint improves ViT. ***Right***: Locality inductive bias suppresses the negative Hessian eigenvalues.

Do Self-Attentions Act Like Convs?

MSAs are low-pass filter, but Convs are high-pass filter. It suggests that MSAs are shape-biased, whereas Convs are texture-biased.

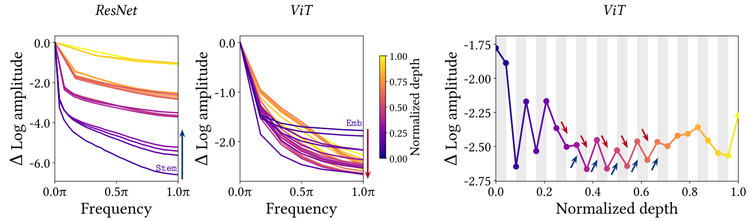
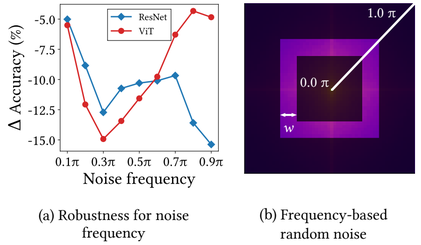


Figure 4: Relative log amplitudes of **Fourier transformed feature map** show that ViT tend to reduces high-frequency signals, while ResNet amplify them. ***Left***: In ViT, MSAs (gray area) generally reduce the high-frequency (1.0π) component of feature map, and Conv/MLPs (white area) amplify it.



(a) Robustness for noise frequency

(b) Frequency-based random noise

Figure 5: We measure the **decrease in accuracy against frequency-based random noise**. ViT is robust against high-frequency noise, while ResNet is vulnerable to them.

It suggests that low-frequency signals and high-frequency signals are informative to MSAs and Convs.

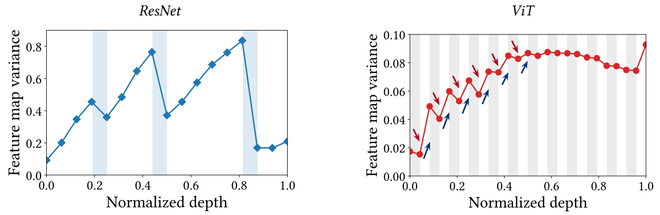


Figure 6: MSAs (gray area) reduce **the variance of feature map points**, but Convs/MLPs (white area) increase the variance. The blue area is subsampling layer. The results implies that MSAs aggregate feature maps, and Convs convert them.

How Can We Harmonize Self-Attentions with Convs?

MSAs closer to the end of a stage (not a model) and Convs at the beginning of a stage significantly improve the performance.

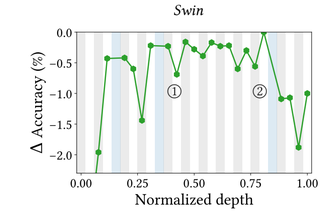
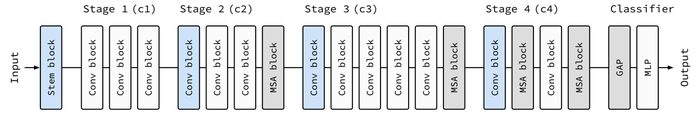
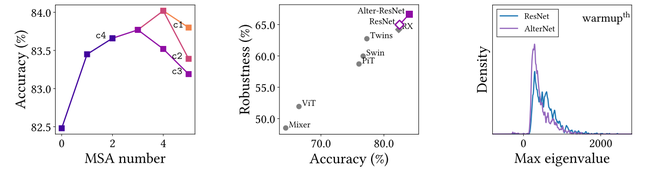


Figure 7: We measure **decrease in accuracy after removing one unit** from the trained model. Accuracy changes periodically, and this period is one stage. In Swin, ① Convs (white area) play an important role at the beginning of a stage, and ② MSAs (gray area) play an important role at the end of a stage.



(a) Architecture of Alter-ResNet-50 for CIFAR-100



(b) Accuracy of AlterNet for MSA number

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

(d) Hessian max eigenvalue spectra

Figure 8: We propose ***AlterNet***, a model in which Conv blocks at the end of *a stage* are replaced with MSA blocks. AlterNet outperforms CNNs even in small data regimes.

In summary, appropriate inductive biases improves NN optimization, and self-attentions have a spatial smoothing inductive bias.

	Self-Attention	Convolution
Loss Landscape	Flat but non-convex	Convex but sharp
Fourier Analysis	Low-pass filter (shape-biased)	High-pass filter (texture-biased)
Best Practice	The end of a stage	The beginning of a stage

Three Key Questions

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

▸

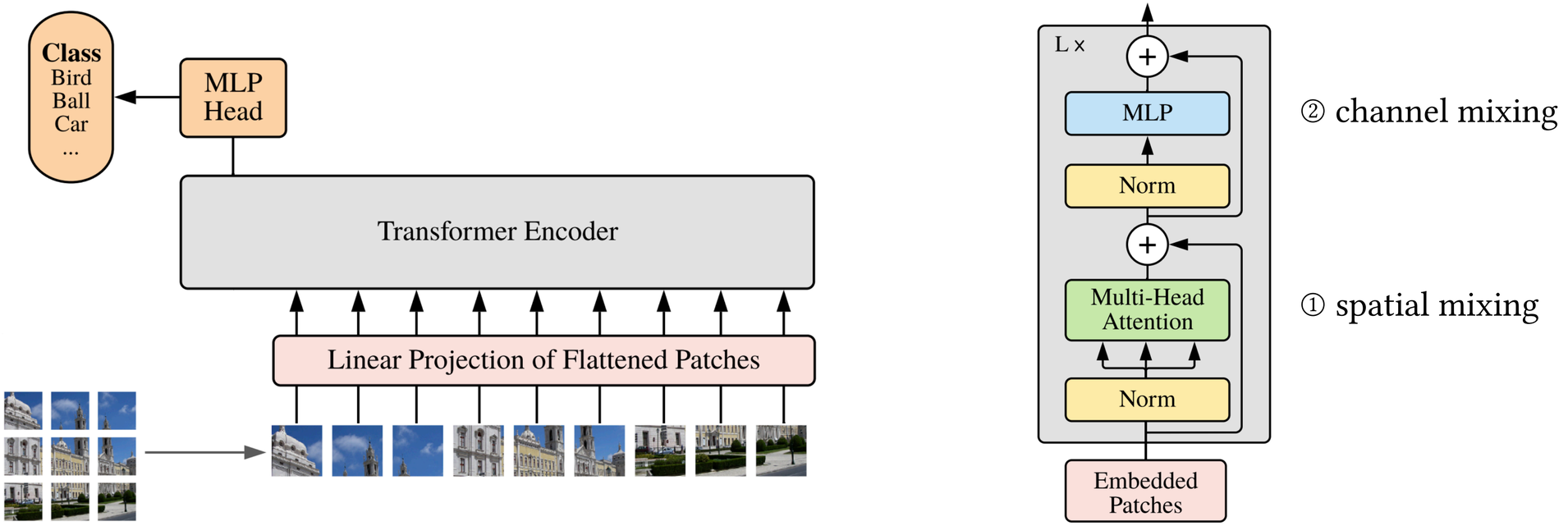
Q2. Do MSAs act like Convs?

▸

Q3. How can we harmonize MSAs with Convs?

▸

Vision Transformers Patchify Images



(a) Patch embedding

(b) Transformer module

Overview of ViTs. ***Left***: Vision Transformers (ViTs) split image into multiple patches. This patch embedding corresponds to the `stem` module of CNNs. ***Right***: A module of ViTs (a Transformer) consists of one multi-head self-attention (MSA) block and one MLP block.

Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." ICLR (2020).

MSAs Are Convs With Large & Data-Specific Kernels

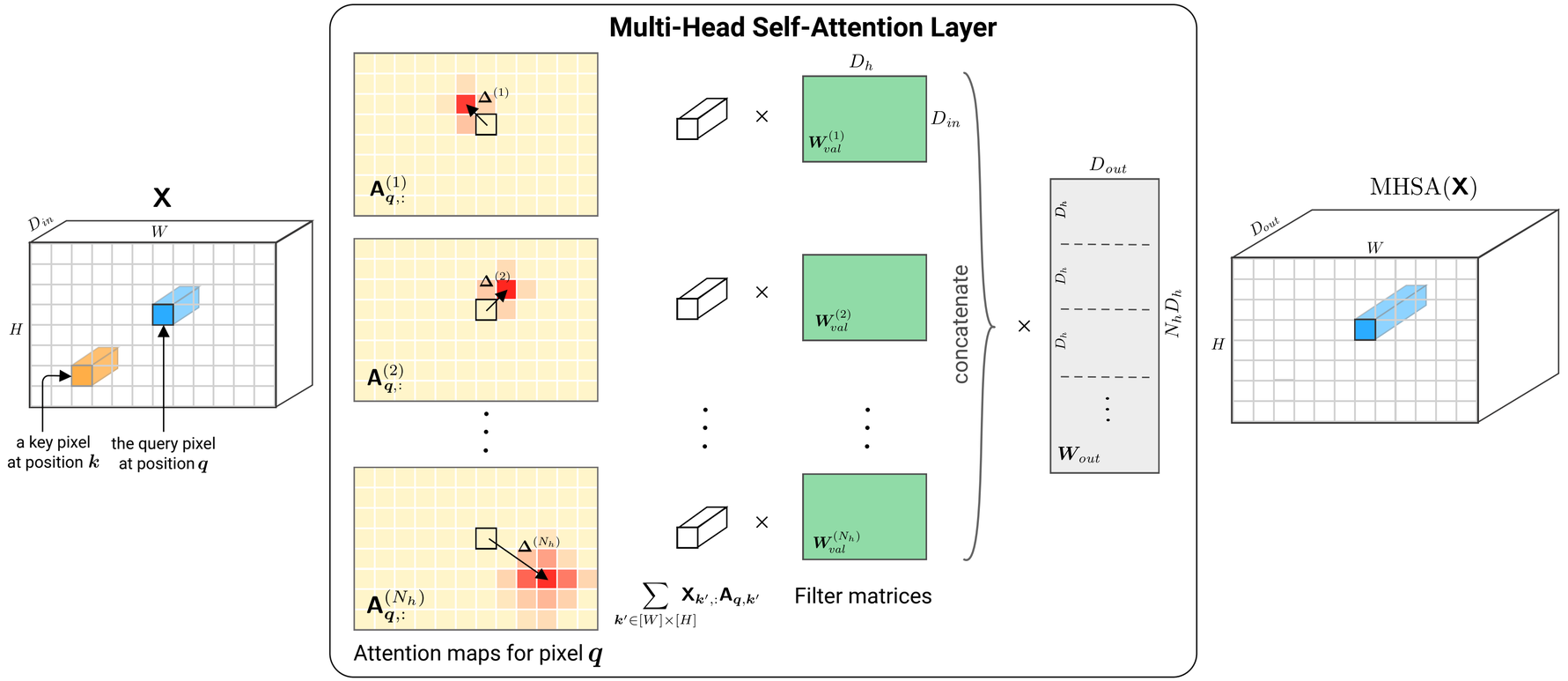


Illustration of a ***"multi-head self-attention"*** (MSA) layer applied to a tensor image $\boldsymbol{X}$.
We represent self-attention maps (red boxes)—similarities between a query pixel and a key pixel.

Cordonnier, Jean-Baptiste, Andreas Loukas, and Martin Jaggi. "On the Relationship between Self-Attention and Convolutional Layers." ICLR (2020).

Three Key Questions

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

- Modeling long-range dependencies improve the accuracy and generalization.

Q2. Do MSAs act like Convs?

- MSAs are at least as expressive as Convs, so MSAs behave like generalized Convs.

Q3. How can we harmonize MSAs with Convs?

- MSAs closer to the end of a model improve accuracy, since they capture abstractions.

Three Key Questions

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

- ~~Modeling long-range dependencies improve the accuracy and generalization~~.

Q2. Do MSAs act like Convs?

- ~~MSAs are at least as expressive as Convs, so MSAs behave like generalized Convs~~.

Q3. How can we harmonize MSAs with Convs?

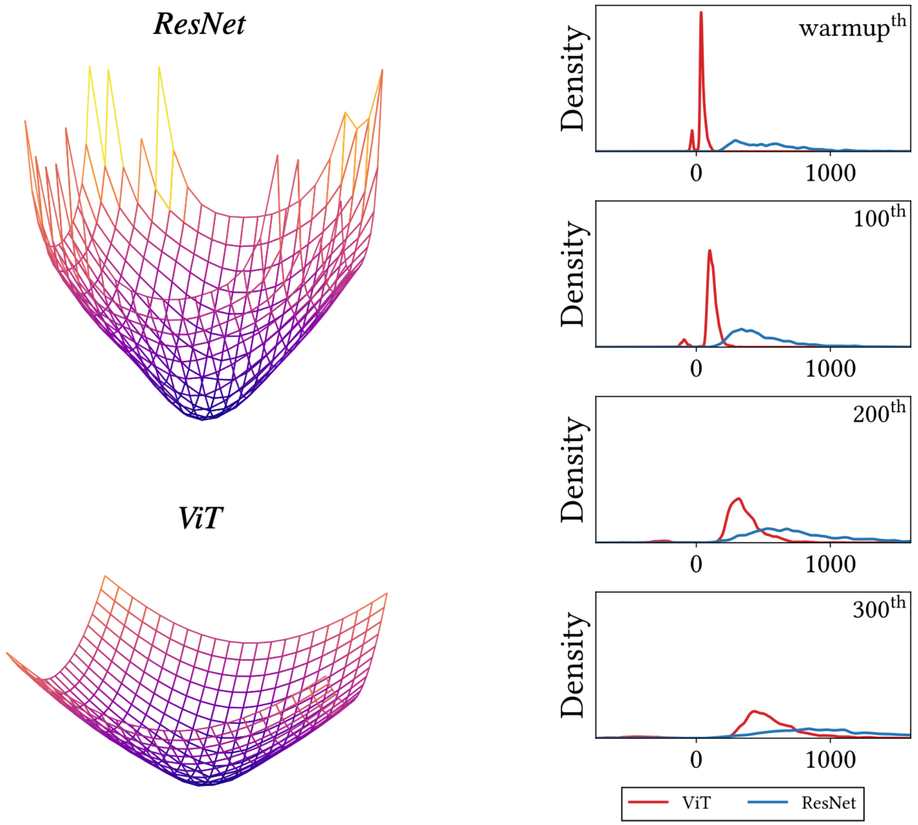
- ~~MSAs closer to the end of a model improve accuracy, since they capture abstractions~~.

All explanations and intuitions are poorly supported!

Question One

What Properties of MSAs Do We Need To Improve Optimization?

MSAs Flatten Loss Landscapes

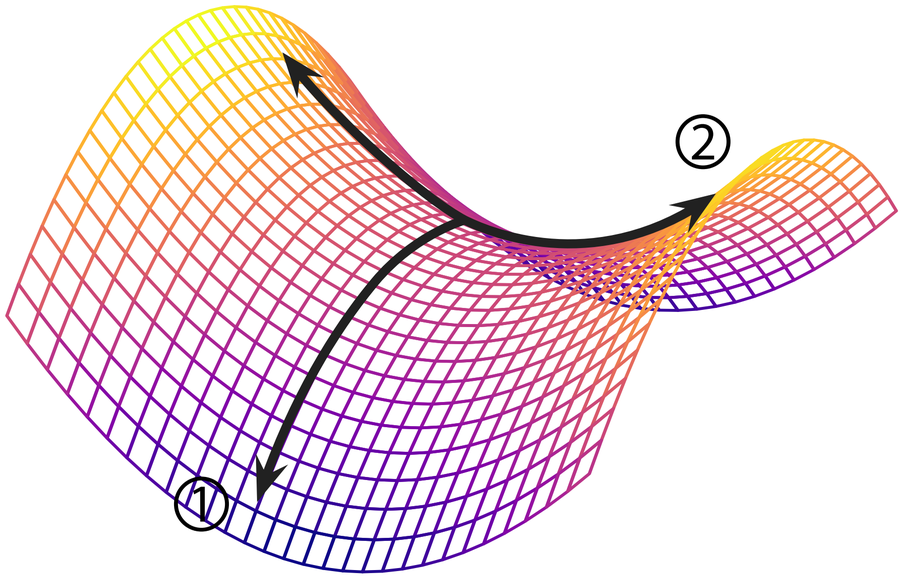


Two different aspects consistently show that MSAs flatten loss landscape. ***Left***: Loss landscape visualizations show that ViT has a flatter loss than ResNet. ***Right***: The magnitude of the Hessian eigenvalues of ViT is smaller than that of ResNet during training phases.

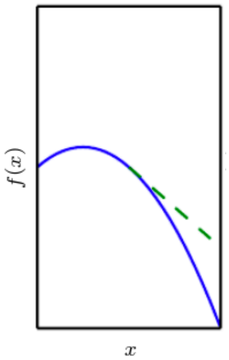
For loss landscape visualization, see arxiv.org/abs/1712.09913.

Local Geometry in NN Landscapes

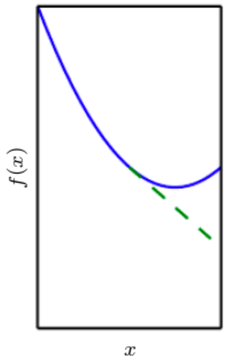
Saddle point
(non-convex)



① Negative curvature
(concave)



② Positive curvature
(convex)

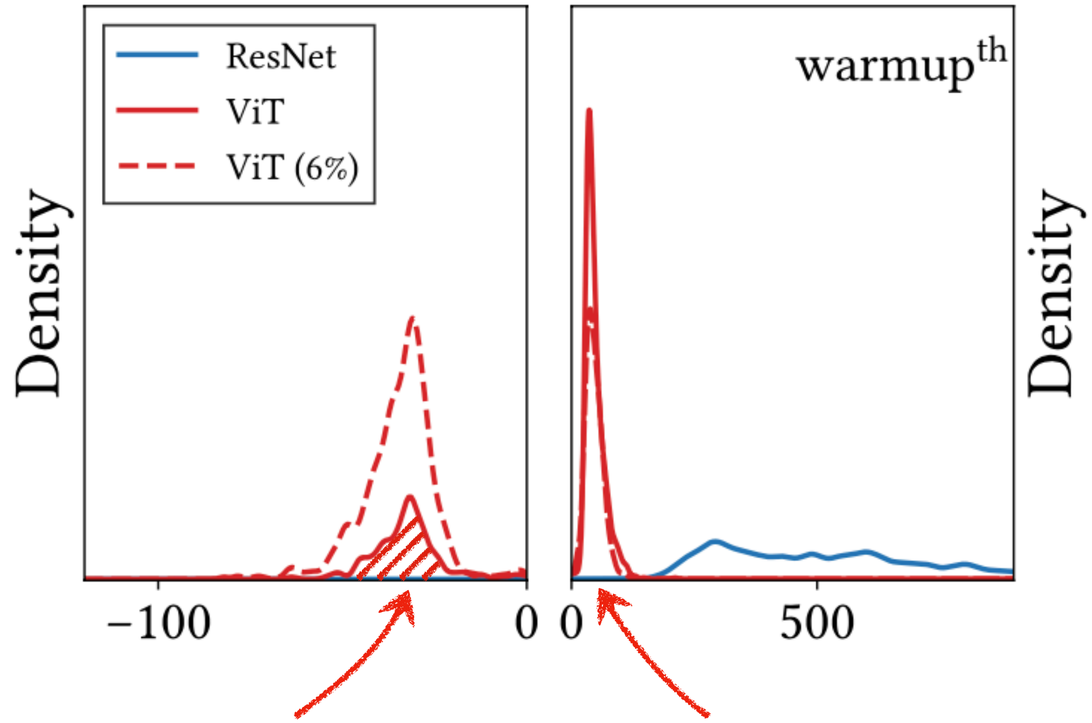


Hessian represents the local geometry of the loss landscapes.

- If Hessian is positive: convex
- If Hessian is indefinite: non-convex (e.g., saddle point)

Source: https://rohanvarma.me/optimizing-dnn/

Criteria For Sharpness and Non-convexity



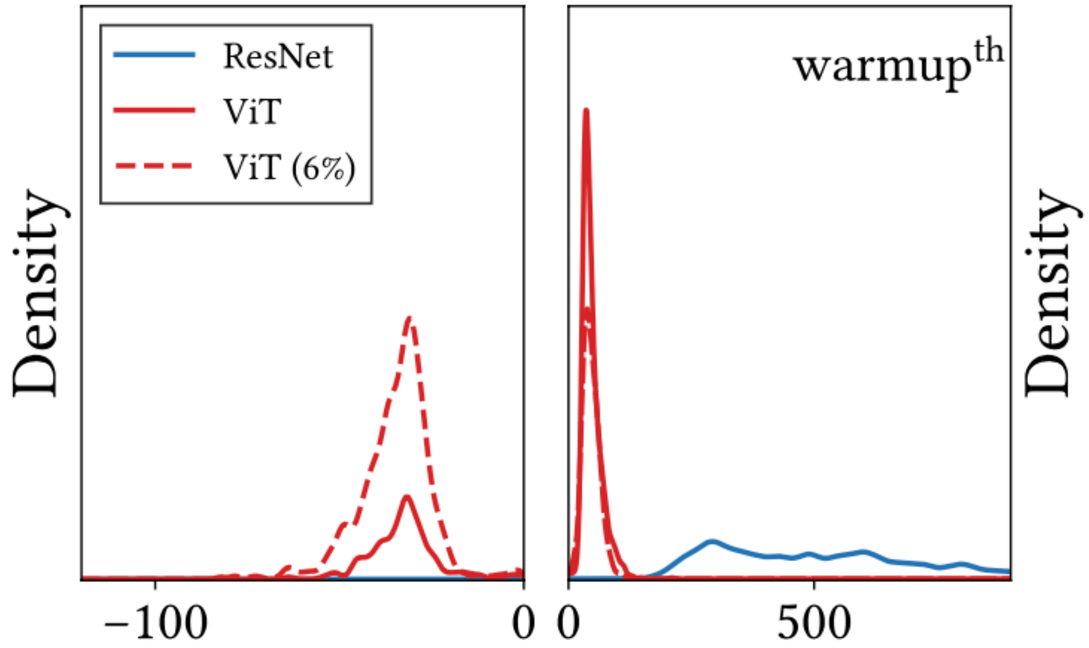
Non-convexity

Negative max Eigenvalue proportion

Sharpness

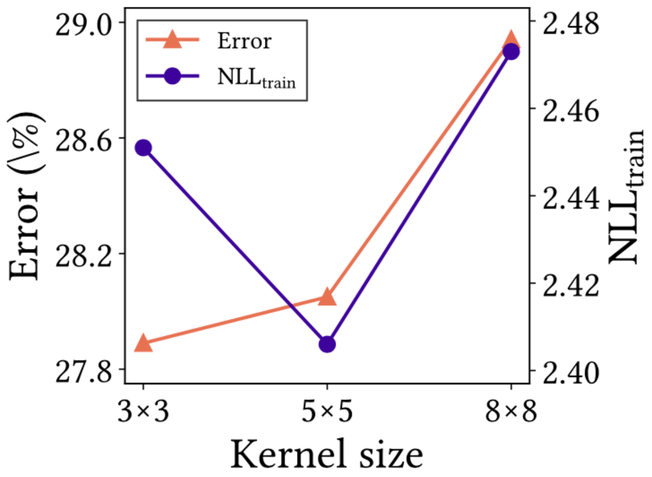
Average of Positive max Eigenvalues

ViT Has a Flat but Non-convex Loss Landscape

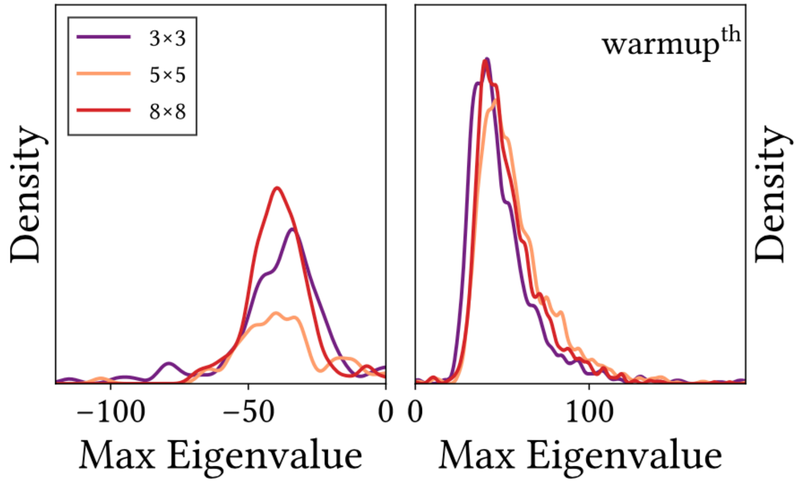


Hessian max eigenvalue spectra show that MSAs have their pros and cons. The dotted line is the spectrum of ViT using 6% dataset for training. ***Left***: ViT has a number of negative Hessian eigenvalues, while ResNet only has a few. ***Right***: The magnitude of ViT's positive Hessian eigenvalues is small.

A Key Feature Is *Not* Long-Range Dependency



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

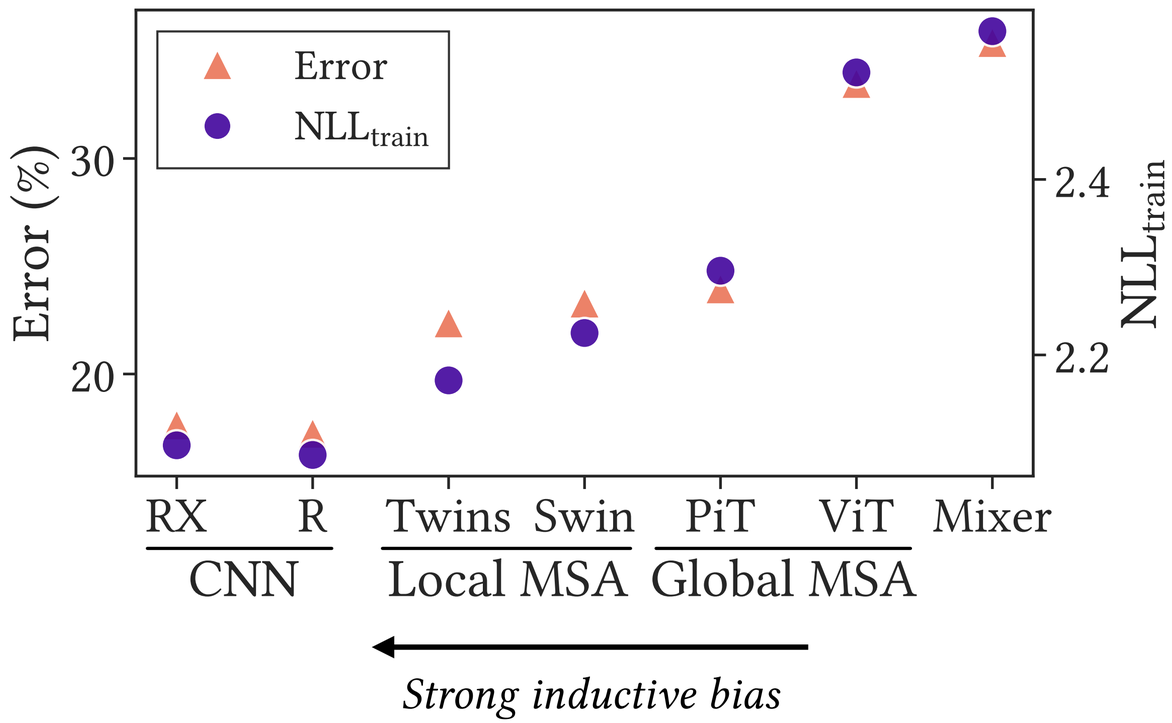


(b) Hessian negative and positive max eigenvalue spectra in early phase of training

Locality constraint improves the performance of ViT. We analyze the ViT with convolutional MSAs. Convolutional MSA with 8 × 8 kernel is 'global MSA'. ***Left***: Local MSAs learn stronger representations than global MSA. ***Right***: Locality inductive bias suppresses the negative Hessian eigenvalues, i.e., local MSAs have convex losses.

The Stronger the Inductive Biases, the Stronger the Reprs

NOT Regularizations!

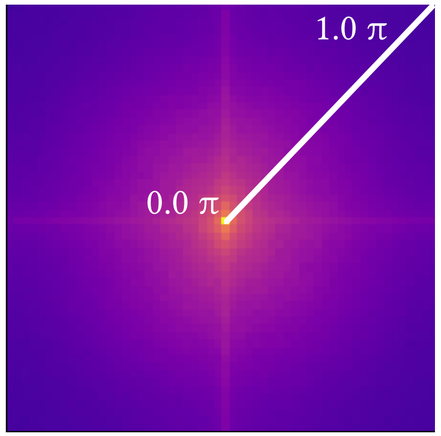


ViT does not overfit training datasets. The lower the NLL_{train}, the lower the error.
It implies that "weak inductive bias (e.g. long-range dependency) disturbs NN optimization".

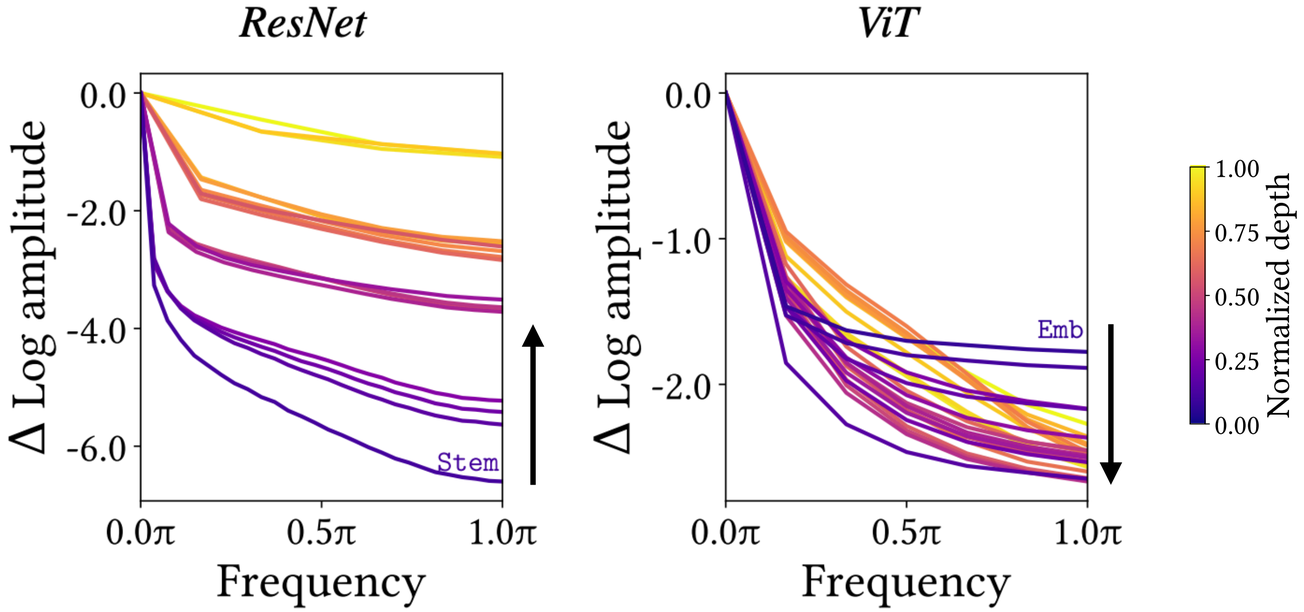
Question Two

Do MSAs Act Like Convs?

MSAs Are Low-Pass Filters, but Convs Are High-Pass Filters



(a) Fourier transformed feature maps

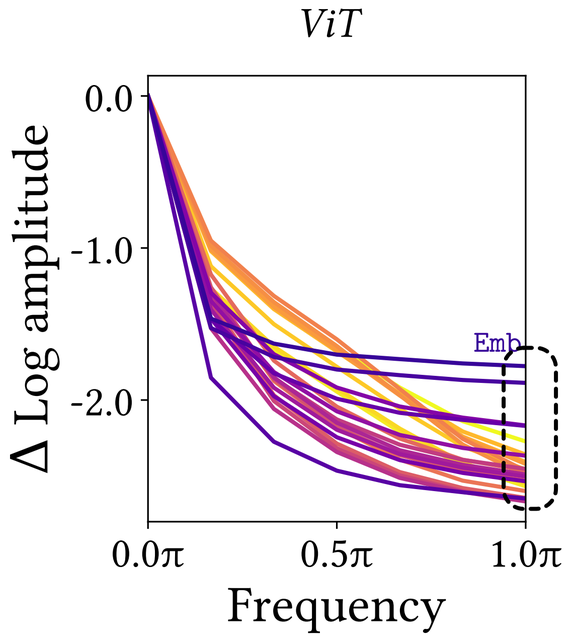


(b) Relative log amplitudes of Fourier transformed feature maps

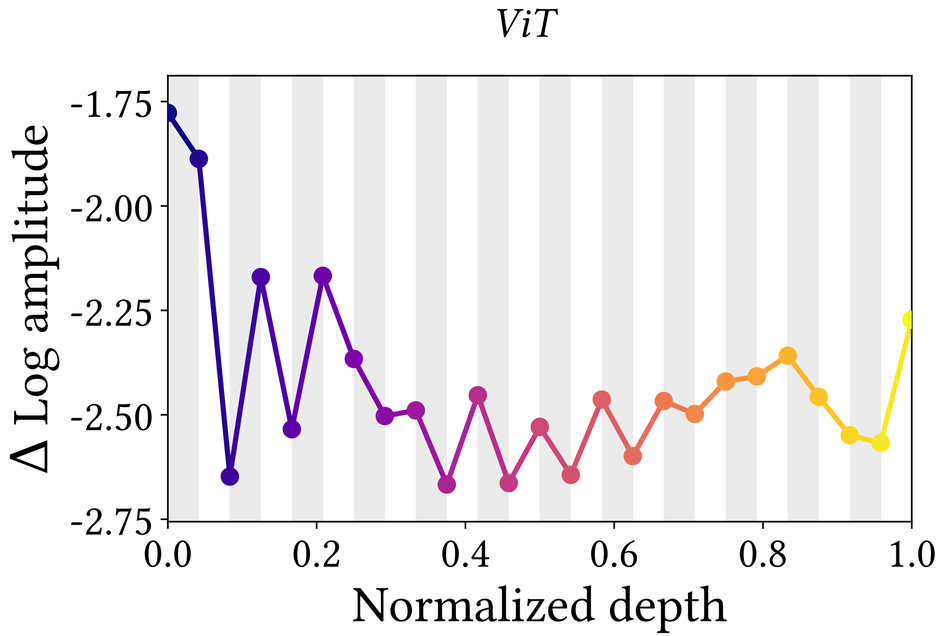
The Fourier analysis shows that MSAs do not act like Convs. Relative log amplitudes of Fourier transformed feature map show that ViT tends to reduce high-frequency signals, while ResNet amplifies them.

Fourier analysis of feature maps is **first proposed** in these works. Compare to arxiv.org/abs/1906.08988 & arxiv.org/abs/2105.03824.

MSAs Are Low-Pass Filters, but Convs Are High-Pass Filters



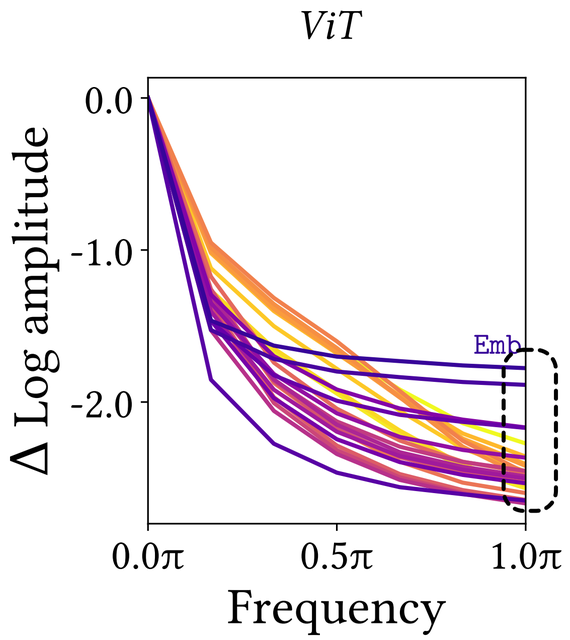
(a) Δ Log amplitude for frequency.



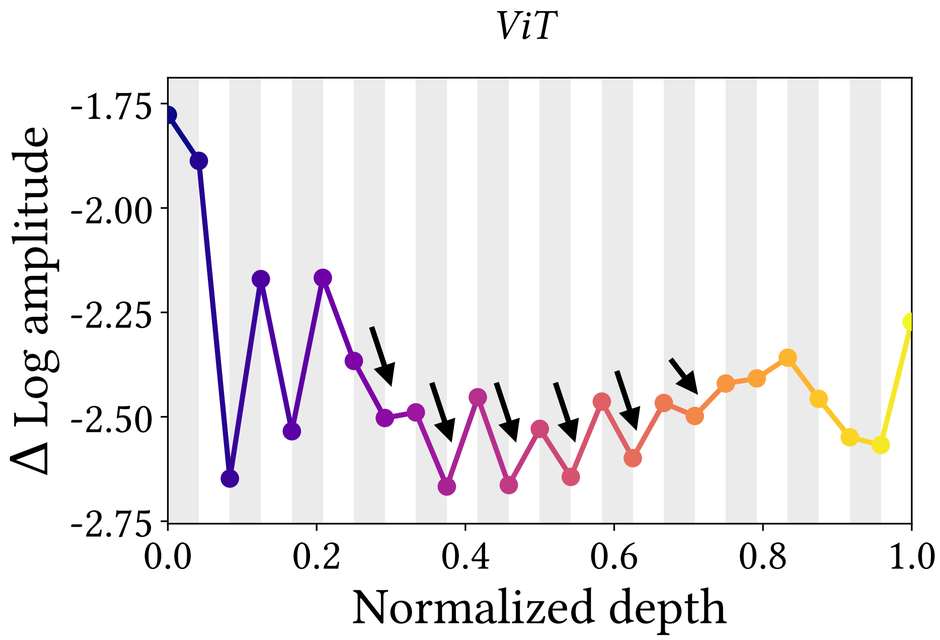
(b) Δ Log amplitude at high-frequency 1.0 π.

MSAs (gray area ▯) generally reduce the high-frequency component, and MLPs (white area ▯) amplify it. It implies that low-frequency signals and high-frequency signals are informative to MSAs and Convs, respectively.

MSAs Are Low-Pass Filters, but Convs Are High-Pass Filters



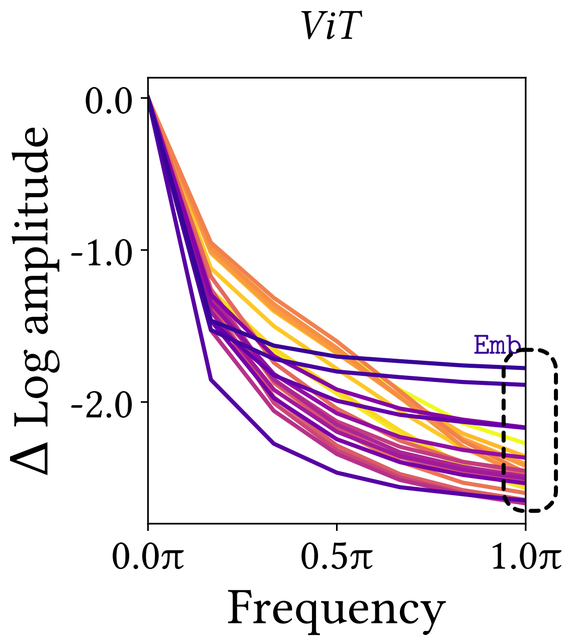
(a) Δ Log amplitude for frequency.



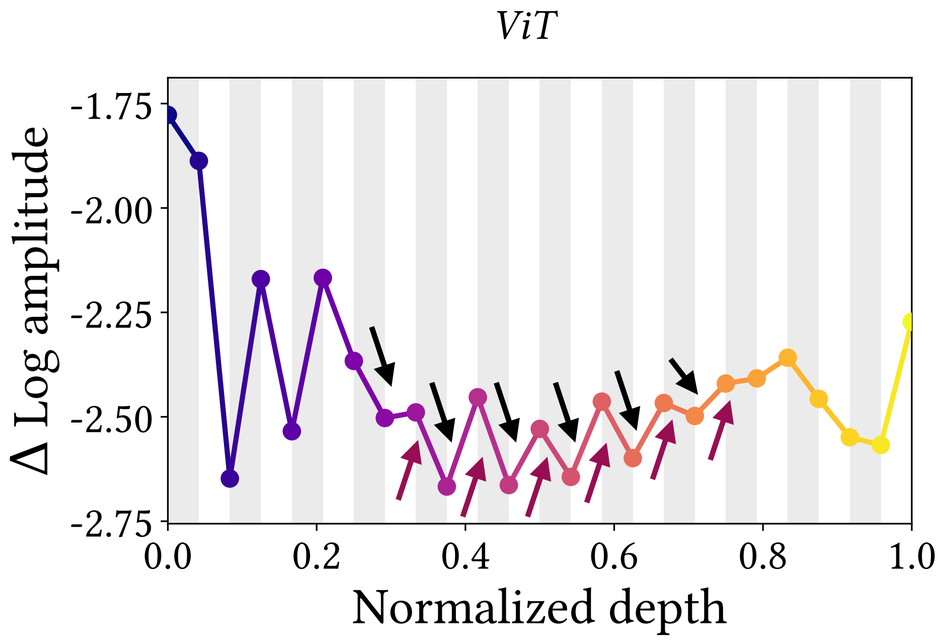
(b) Δ Log amplitude at high-frequency 1.0 π.

MSAs (gray area ▯) generally reduce the high-frequency component, and MLPs (white area ▯) amplify it. It implies that low-frequency signals and high-frequency signals are informative to MSAs and Convs, respectively.

MSAs Are Low-Pass Filters, but Convs Are High-Pass Filters



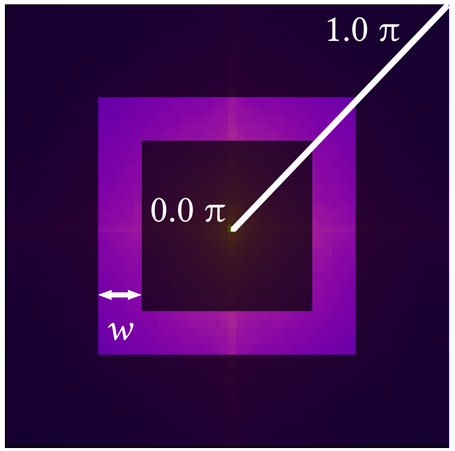
(a) Δ Log amplitude for frequency.



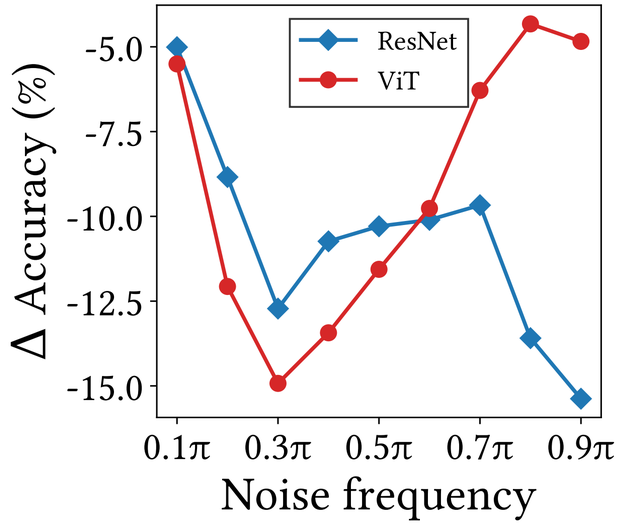
(b) Δ Log amplitude at high-frequency 1.0 π.

MSAs (gray area ▯) generally reduce the high-frequency component, and MLPs (white area ▯) amplify it. It implies that low-frequency signals and high-frequency signals are informative to MSAs and Convs, respectively.

ViT Is Robust Against High-Frequency Noise



(a) Frequency-based random noise



(b) Robustness for noise frequency

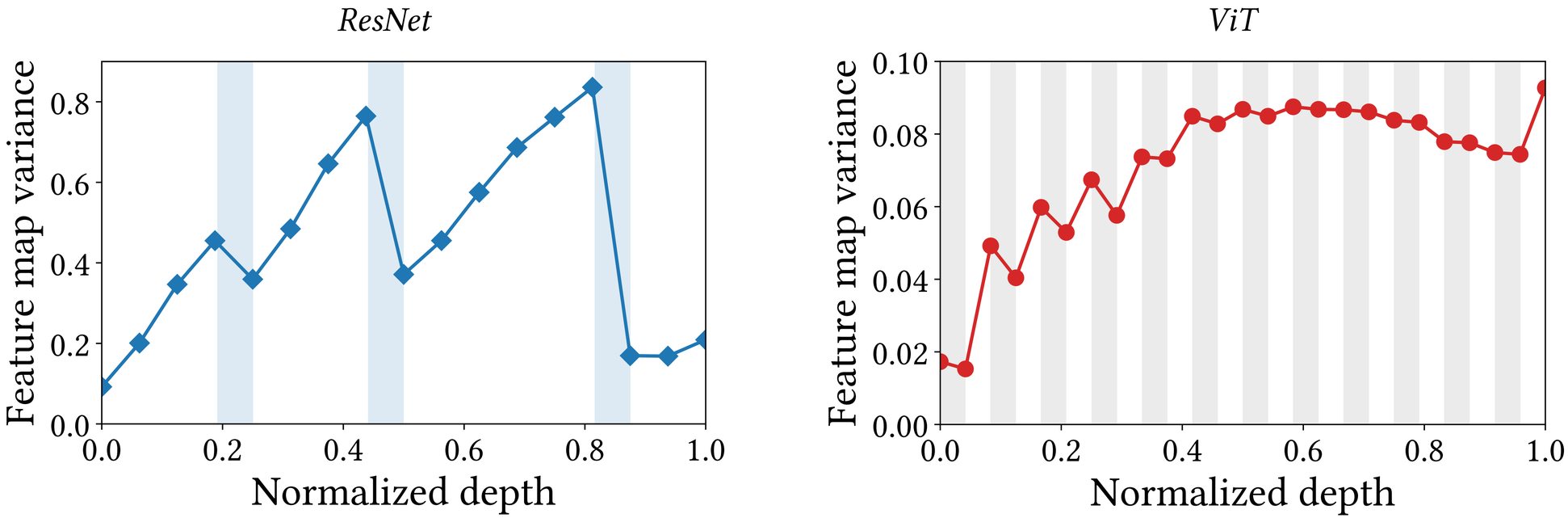
The Fourier analysis shows that MSAs do not act like Convs. We measure the decrease in accuracy against frequency-based random noise. ResNet is vulnerable to high-frequency noise, while ViT is robust against them.

arxiv.org/abs/2103.15670 also figured out that ViTs are robust against high-frequency adversarial attacks.

MSA is shape-biased.

Conv is texture-biased.

MSAs Aggregate Feature Maps, but Convs Do Not

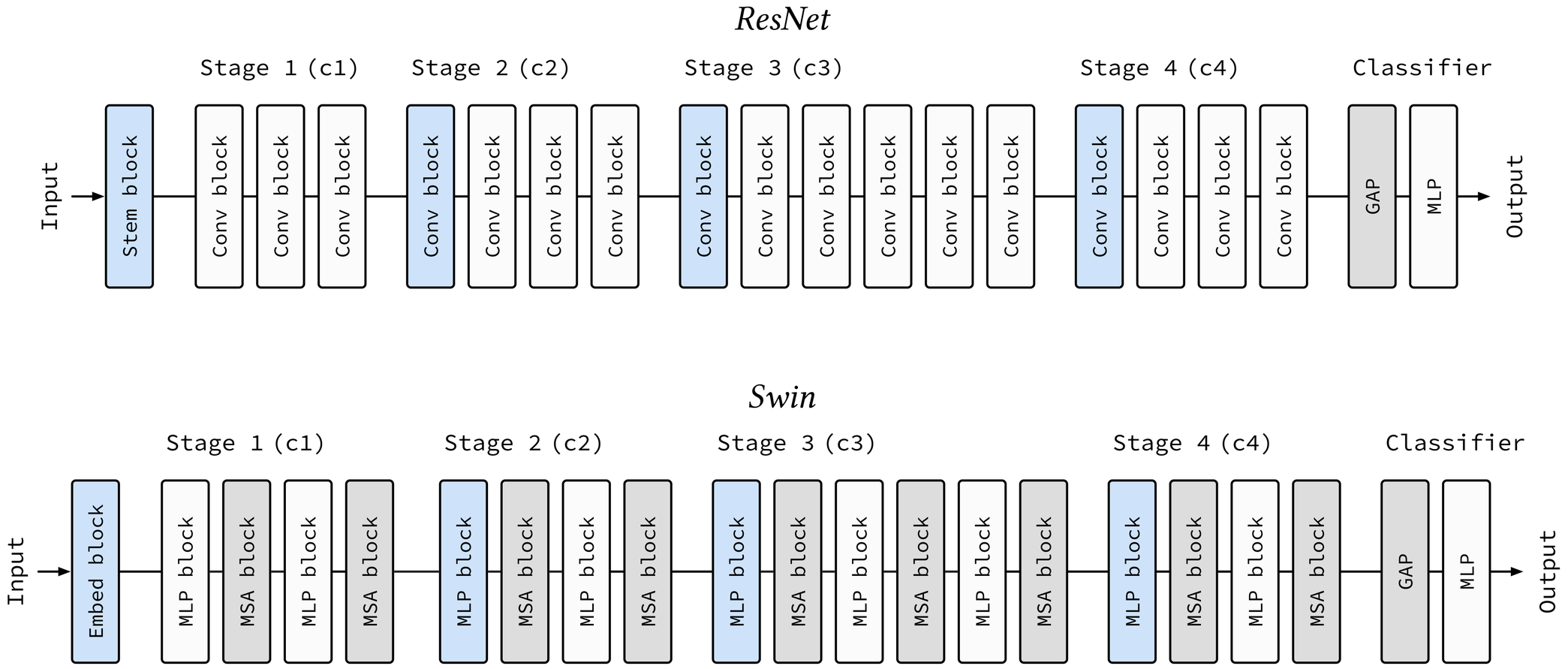


MSAs (gray area ▯) reduce the variance of feature map points, but Convs/MLPs (white area ▯) increase the variance. The blue area (▯) is subsampling layer.
The results implies that MSAs aggregate feature maps, and Convs transform them.

Question Three

How Can We Harmonize MSAs With Convs?

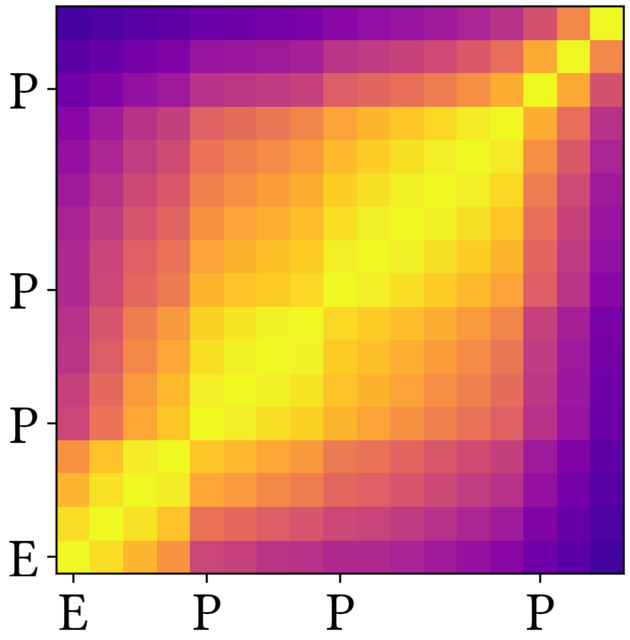
Architecture of Swin Transformer



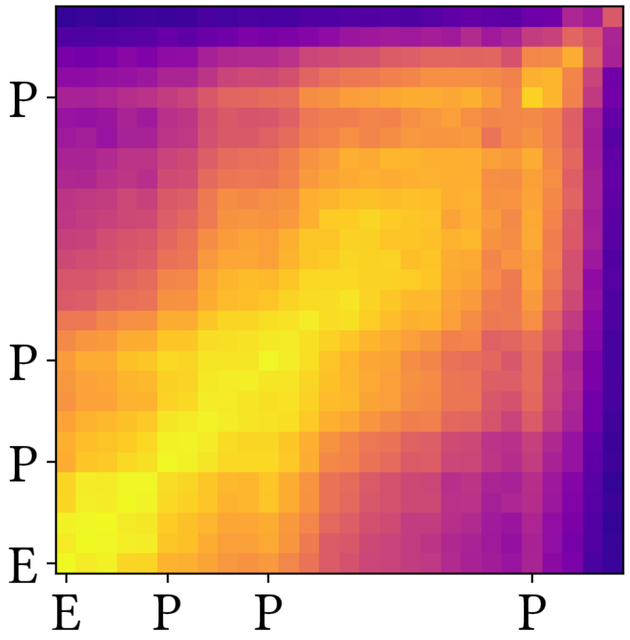
See arxiv.org/abs/2103.14030 for more details on Swin Transformer.

Multi-Stage NNs Have Representational Block Structures

ResNet



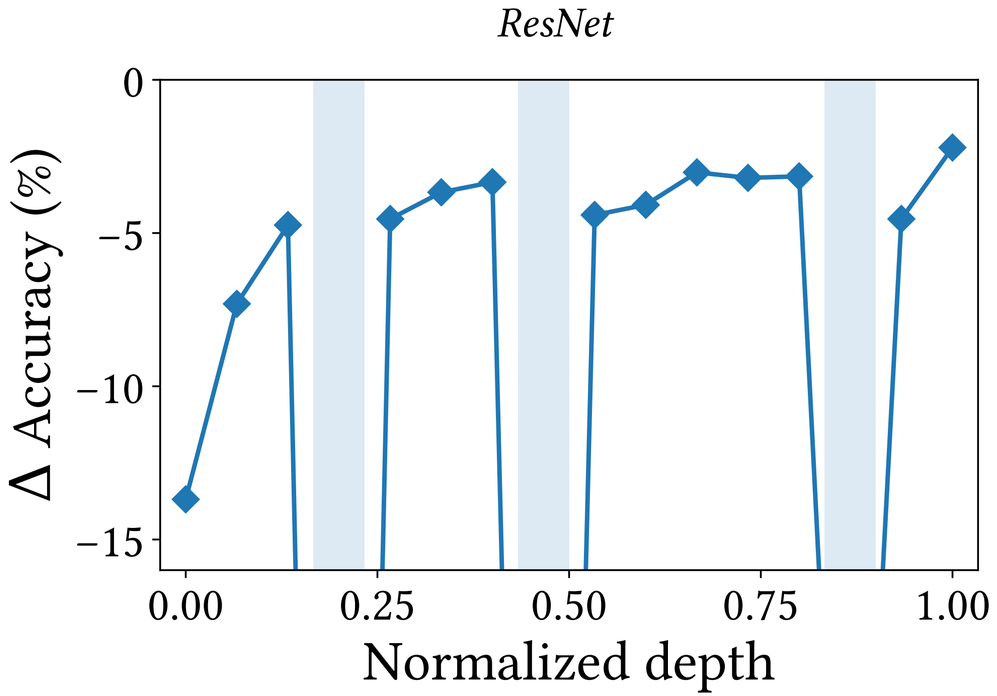
Swin

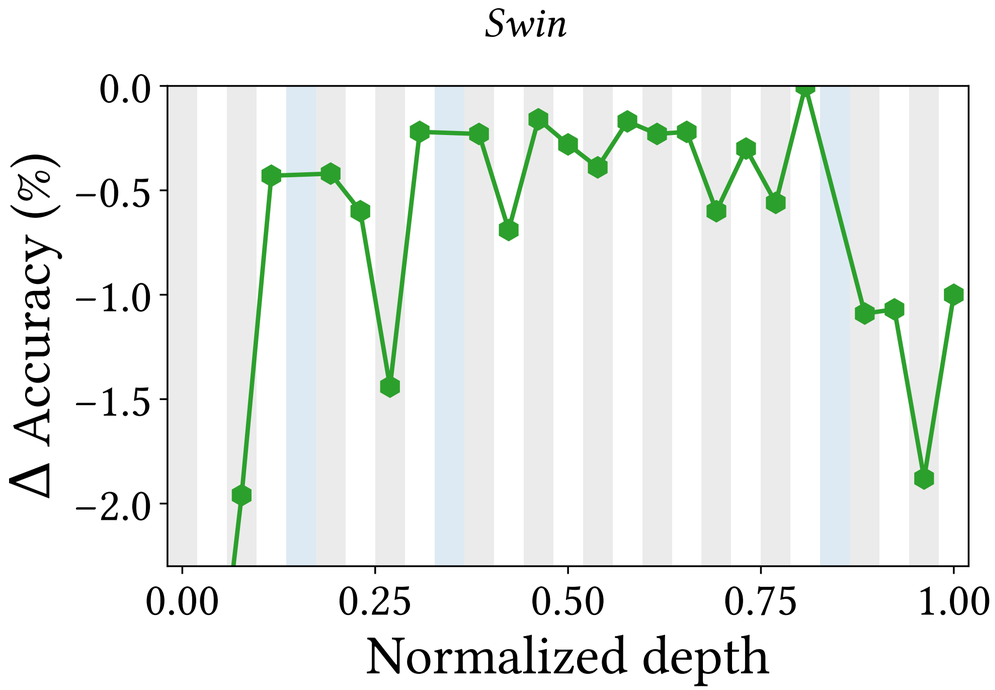


The feature map similarities show the block structure of ResNet and Swin. "E" stands for stem/ embedding and "P" for pooling (sub- sampling) layer. Since vanilla ViT does not have this structure the structure is an intrinsic characteristic of multi-stage architectures.

See arxiv.org/abs/2010.15327 for more details on representational similarity.

Lesion Studies Show the Periodicity of the Accuracy Drop

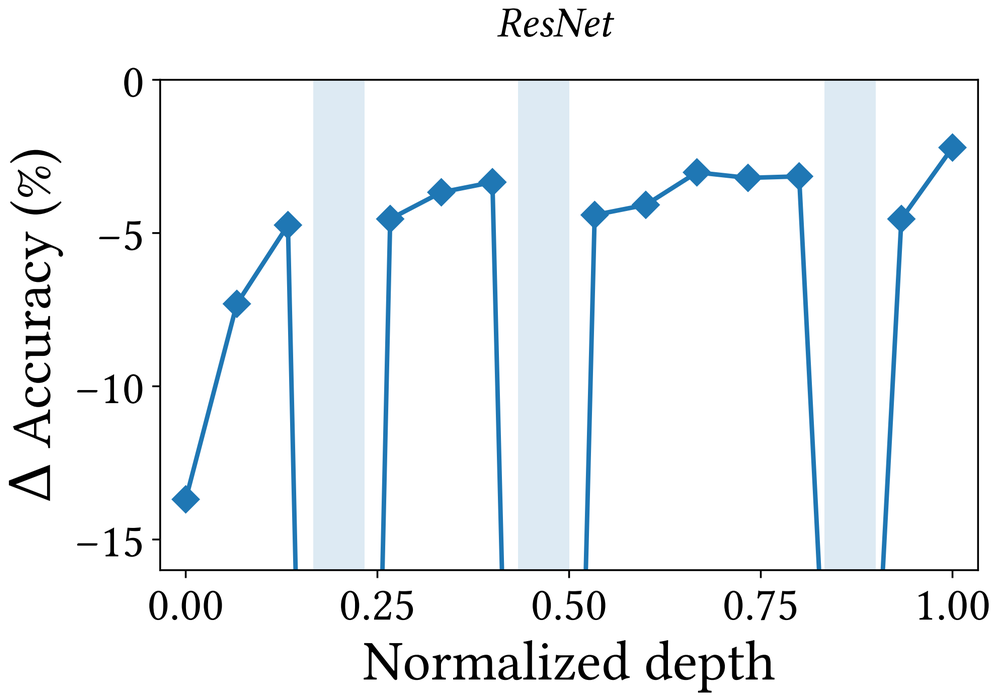


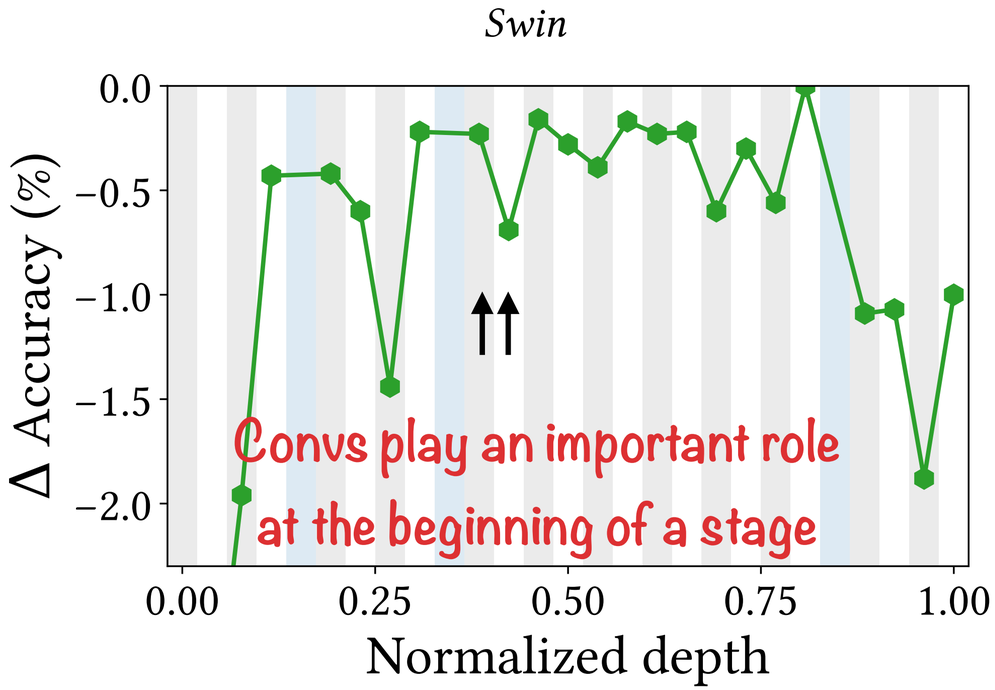


MSAs closer to the end of a stage significantly improve the performance. We measure decrease in accuracy after removing one unit from the trained model. Accuracy changes periodically, and this period is one stage.

https://arxiv.org/abs/2103.14586 is first reported the results of lesion studies on ViTs. We expand the analysis.

Lesion Studies Show the Periodicity of the Accuracy Drop

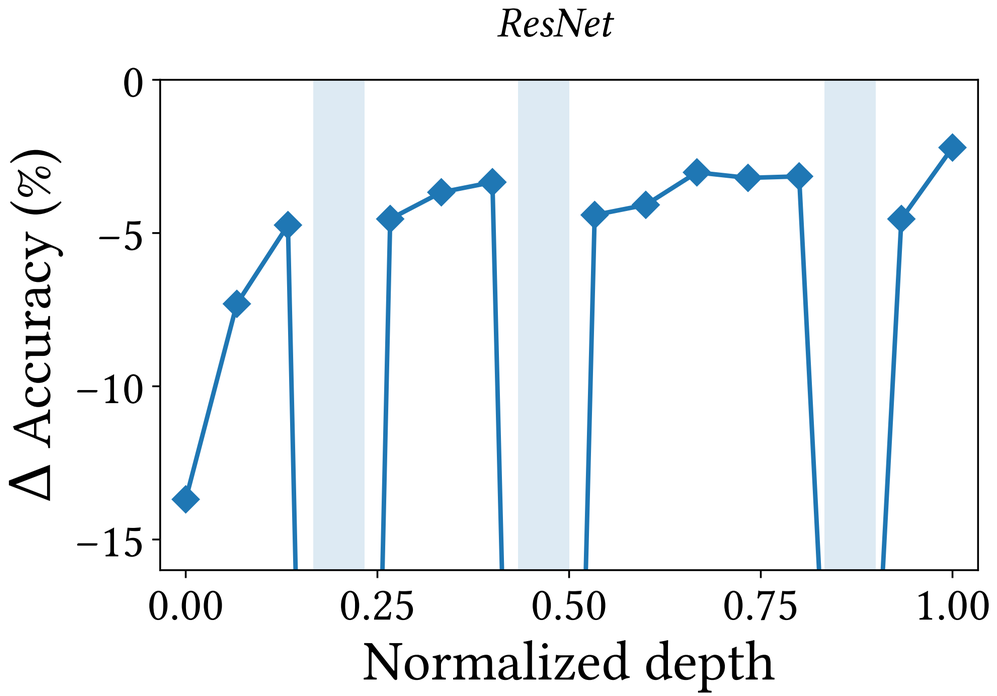


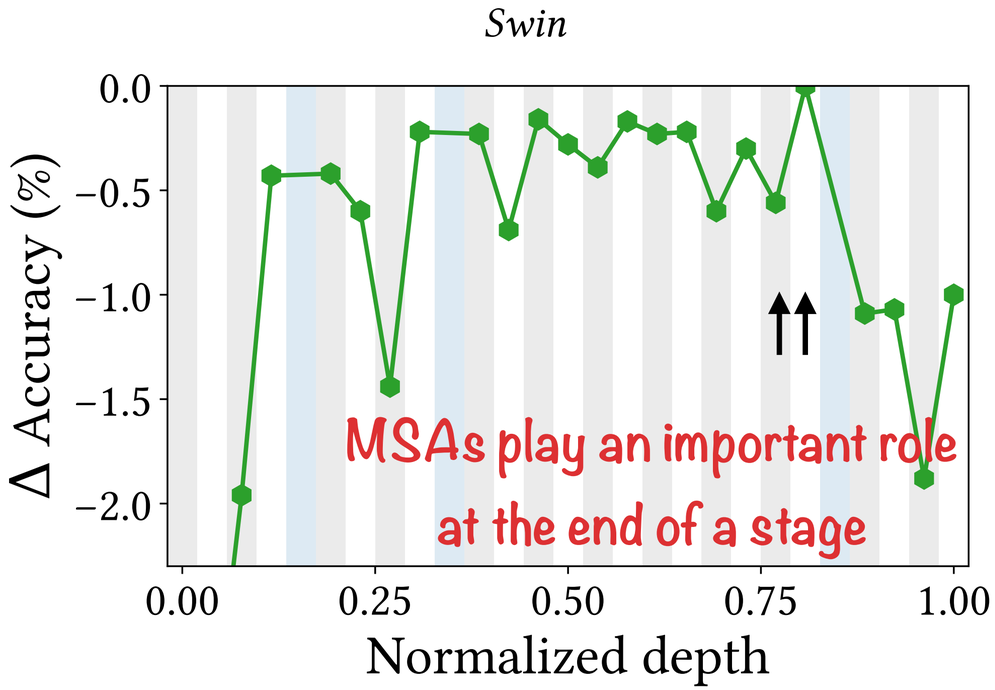


MSAs closer to the end of a stage significantly improve the performance. We measure decrease in accuracy after removing one unit from the trained model. Accuracy changes periodically, and this period is one stage.

https://arxiv.org/abs/2103.14586 is first reported the results of lesion studies on ViTs. We expand the analysis.

Lesion Studies Show the Periodicity of the Accuracy Drop

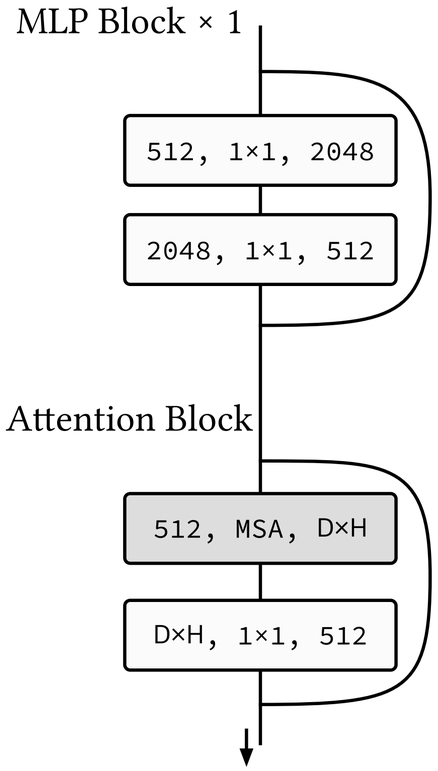




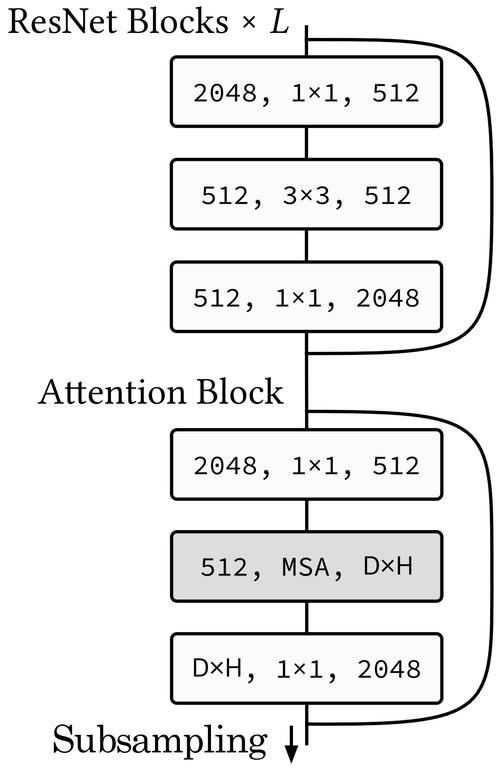
MSAs closer to the end of a stage significantly improve the performance. We measure decrease in accuracy after removing one unit from the trained model. Accuracy changes periodically, and this period is one stage.

https://arxiv.org/abs/2103.14586 is first reported the results of lesion studies on ViTs. We expand the analysis.

Alternating Pattern of Convs and MSAs



(a) Canonical Transformer

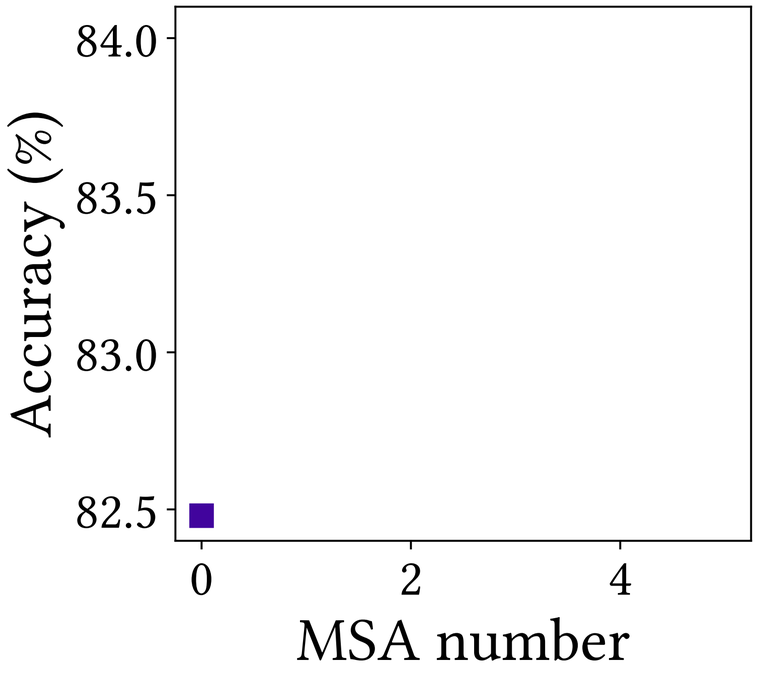


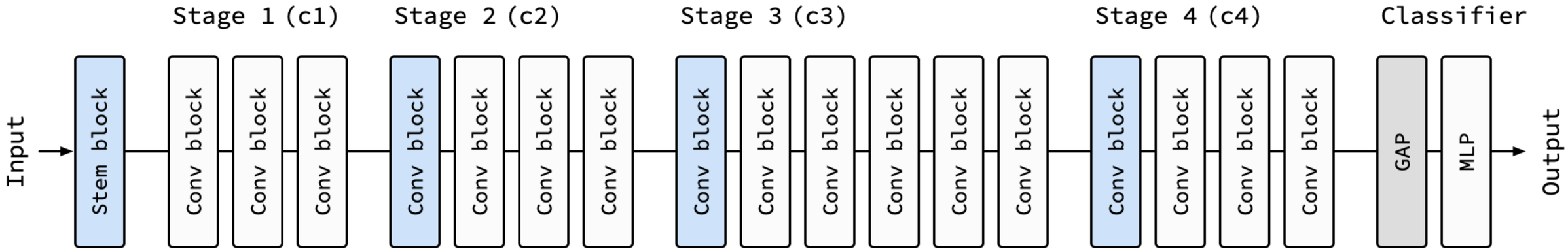
(b) Alternating pattern

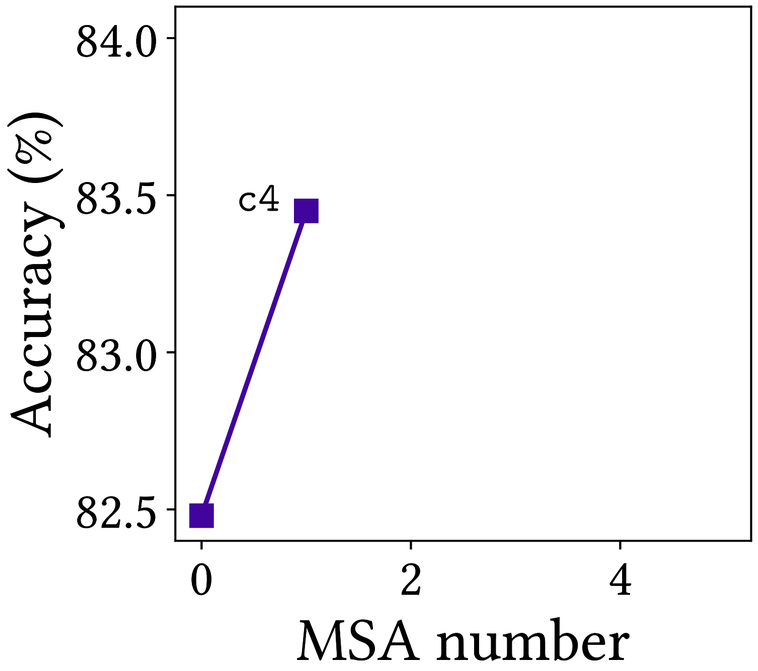
Alternating pattern replace Conv blocks at the end of a stage with MSA blocks. ***Left***: The stages of ViTs consist of repetitions of canonical Transformers. “D” is the hidden dimension and “H” is the number of heads. ***Right***: The stages using alternating pattern consists of a number of CNN blocks and anMSA block

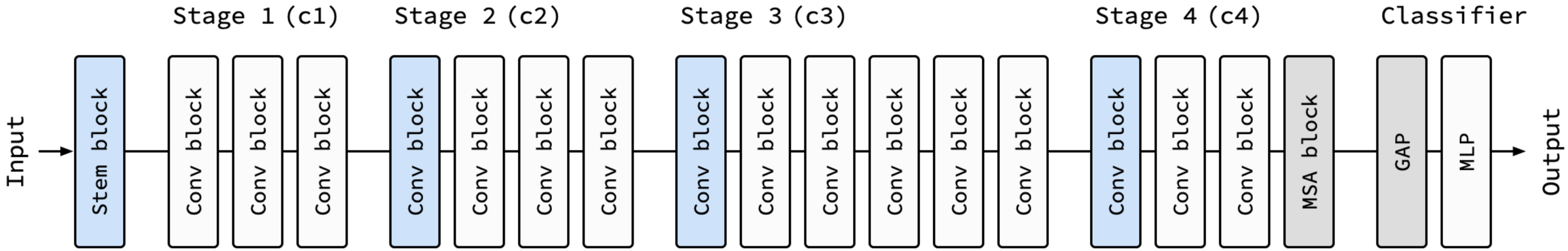
Build-Up Rule: How to Apply MSA to Your Own CNNs

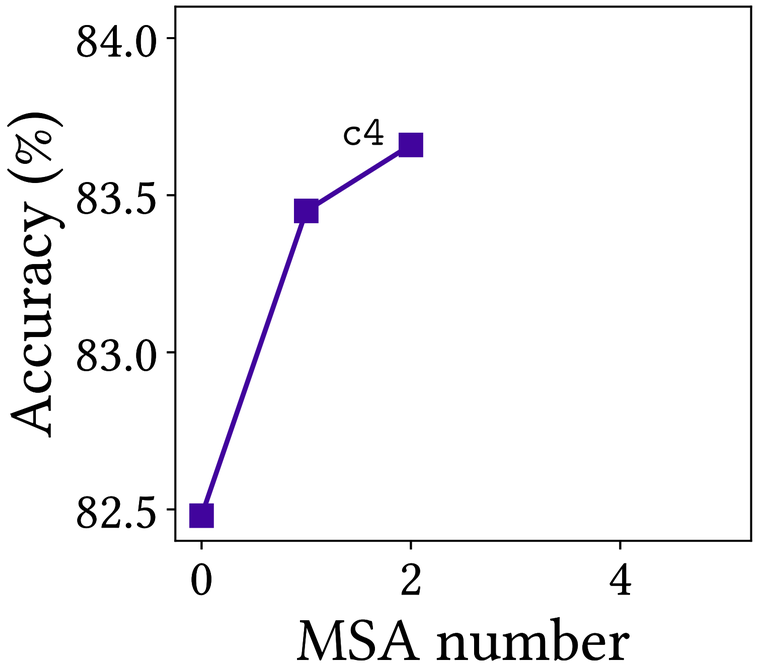
- Alternately replace Conv blocks with MSA blocks from the end of a baseline CNN model.
- If the added MSA block does not improve predictive performance, replace a Conv block located at the end of an earlier stage with an MSA block .
- Use more heads and higher hidden dimensions for MSA blocks in late stages.

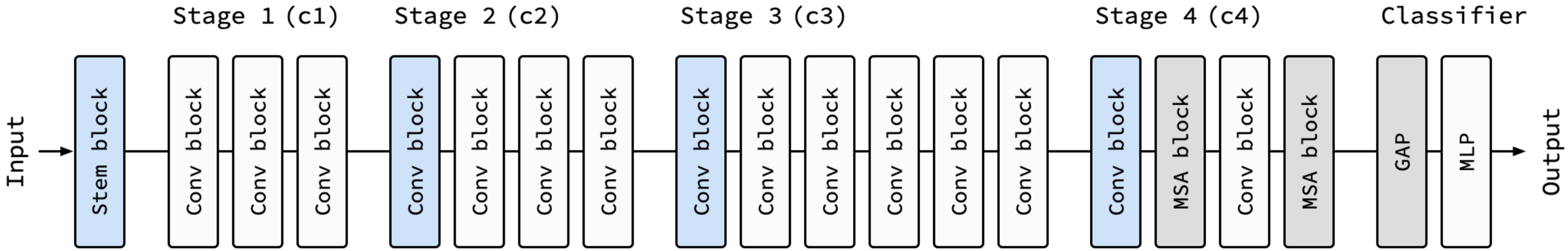

84.0
83.5
83.0
82.5
Accuracy (%)
0
2
4
MSA number

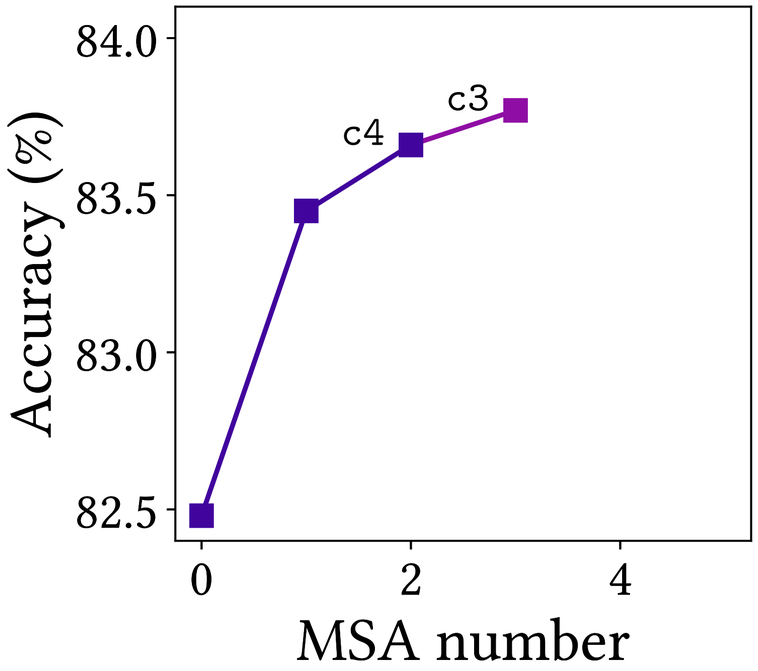

Stage 1 (c1)
Stage 2 (c2)
Stage 3 (c3)
Stage 4 (c4)
Classifier
Input
Stem block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
GAP
MLP
Output

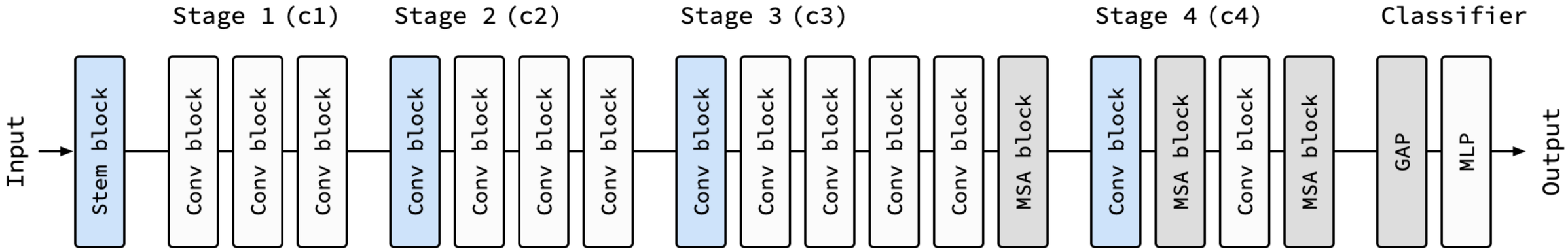

84.0
83.5
83.0
82.5
Accuracy (%)
c4
0
2
4
MSA number

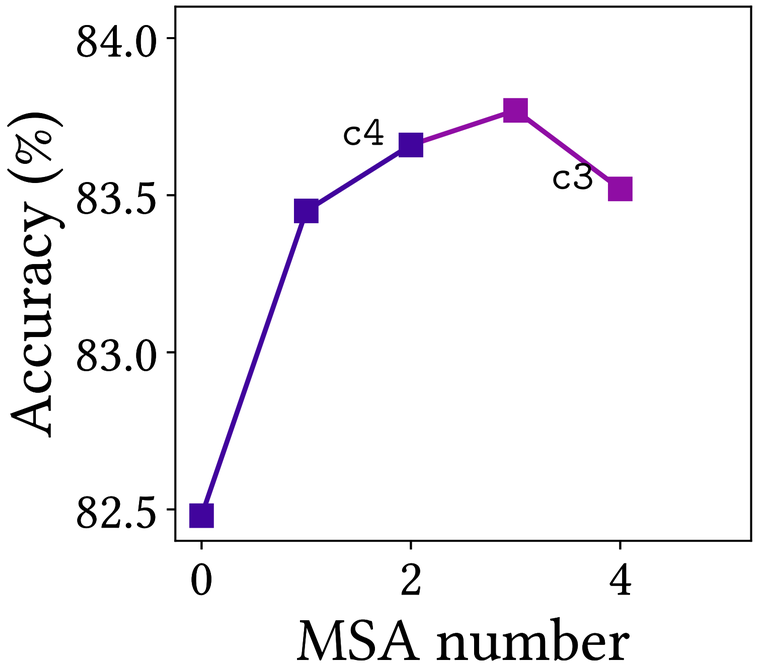

Stage 1 (c1)
Stage 2 (c2)
Stage 3 (c3)
Stage 4 (c4)
Classifier
Input
Stem block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
MSA block
GAP
MLP
Output


84.0
83.5
83.0
82.5
Accuracy (%)
c4
0
2
4
MSA number

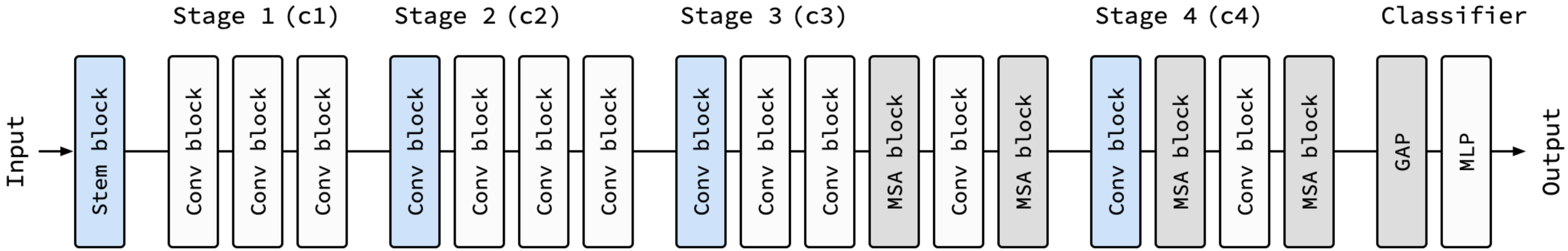

Stage 1 (c1)
Stage 2 (c2)
Stage 3 (c3)
Stage 4 (c4)
Classifier
Input
Stem block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
MSA block
Conv block
MSA block
GAP
MLP
Output

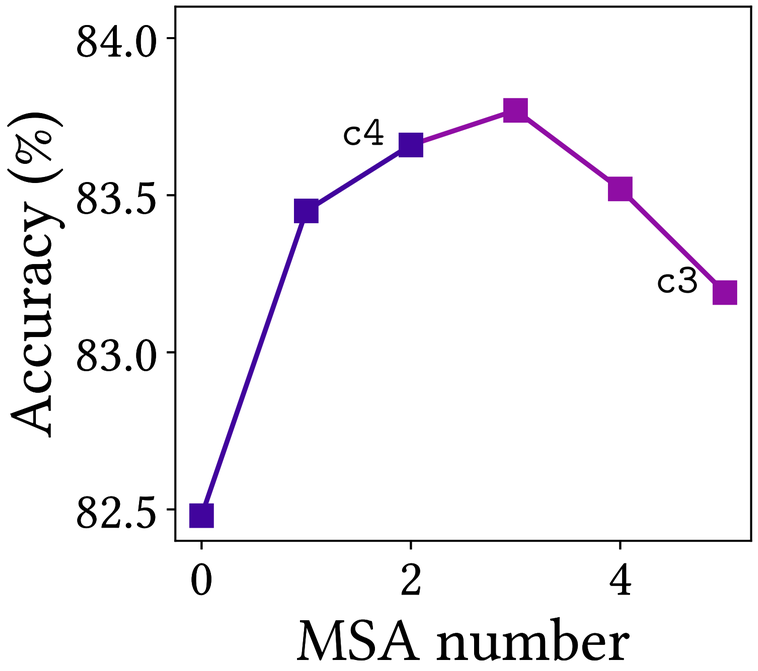

84.0
83.5
83.0
82.5
Accuracy (%)
c4
c3
0
2
4
MSA number


Stage 1 (c1)
Stage 2 (c2)
Stage 3 (c3)
Stage 4 (c4)
Classifier
Input
Stem block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
MSA block
Conv block
MSA block
Conv block
MSA block
GAP
MLP
Output

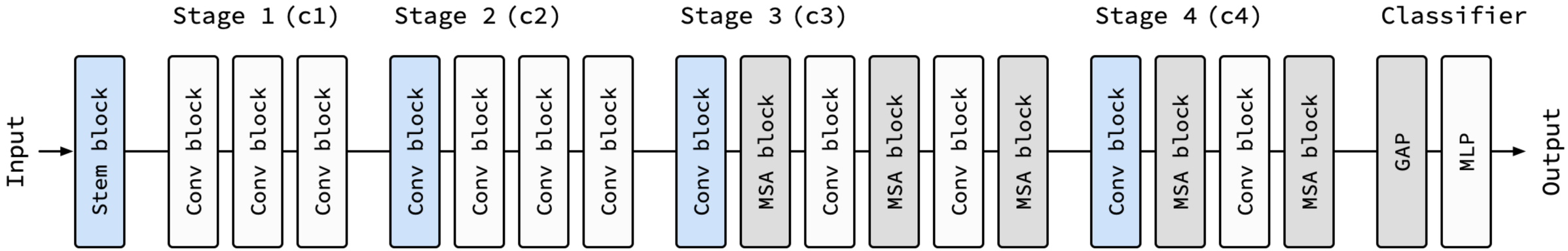


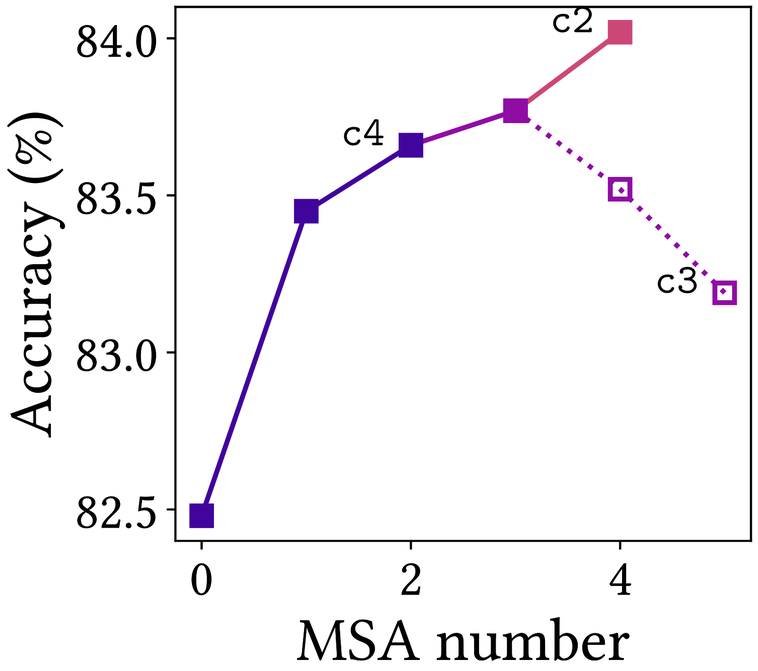
As expected,
MSAs in c3 harm accuracy.



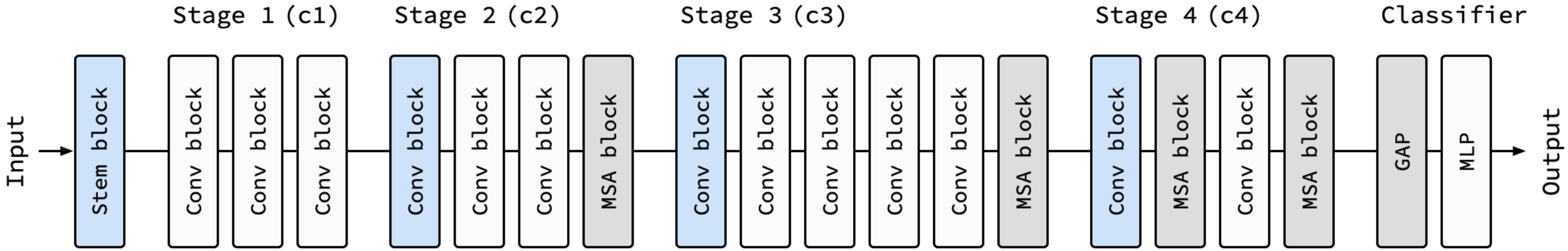


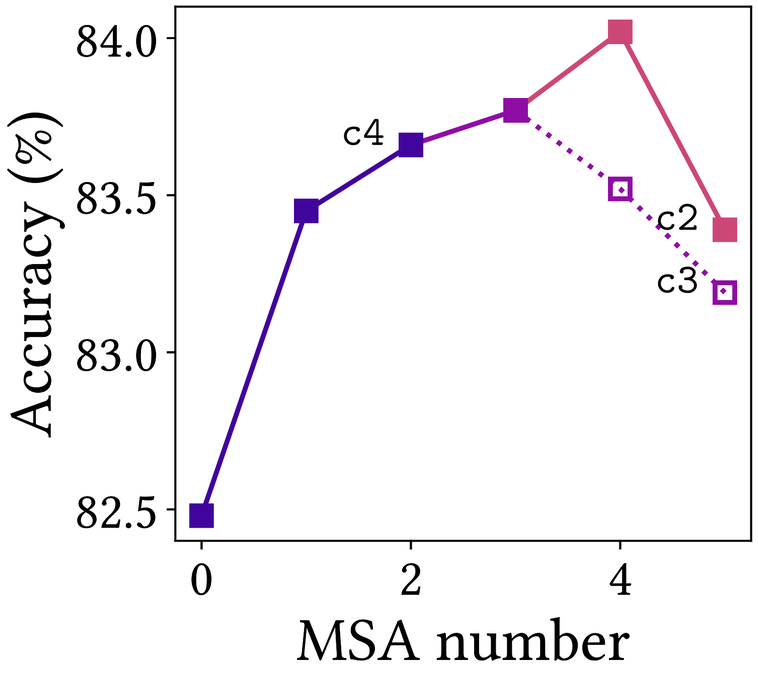
As expected,
MSAs in c3 harm accuracy.

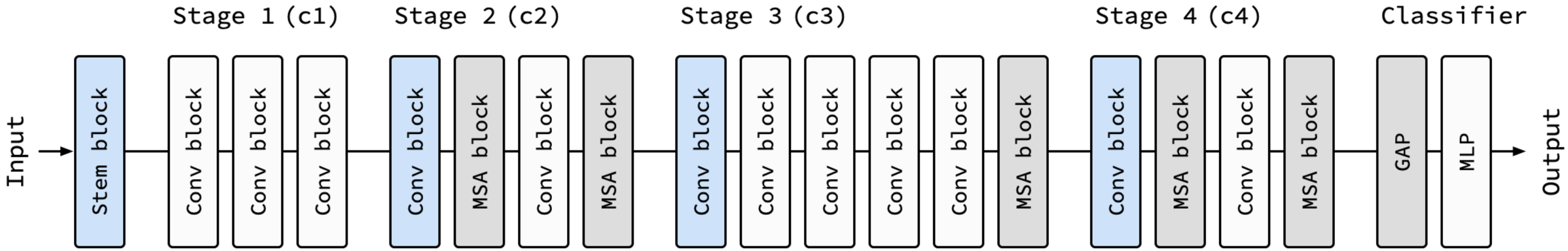


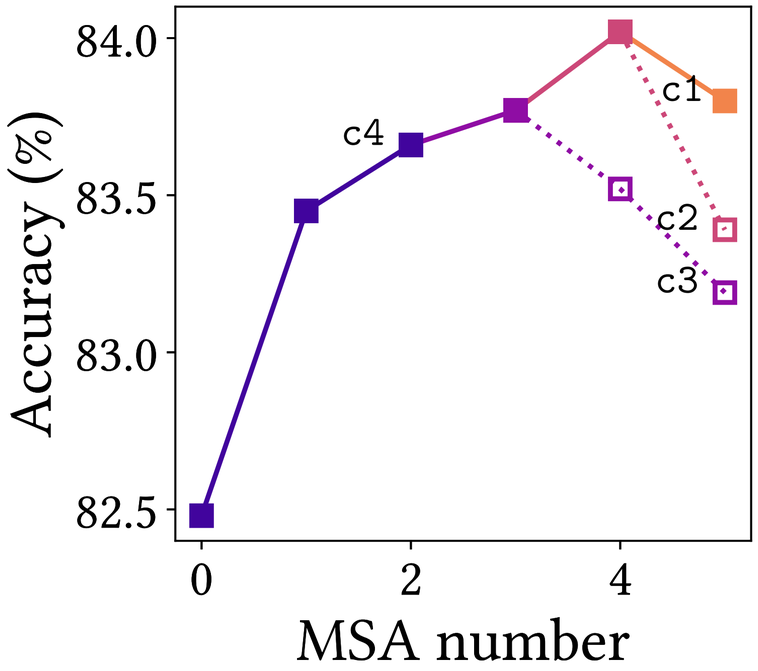


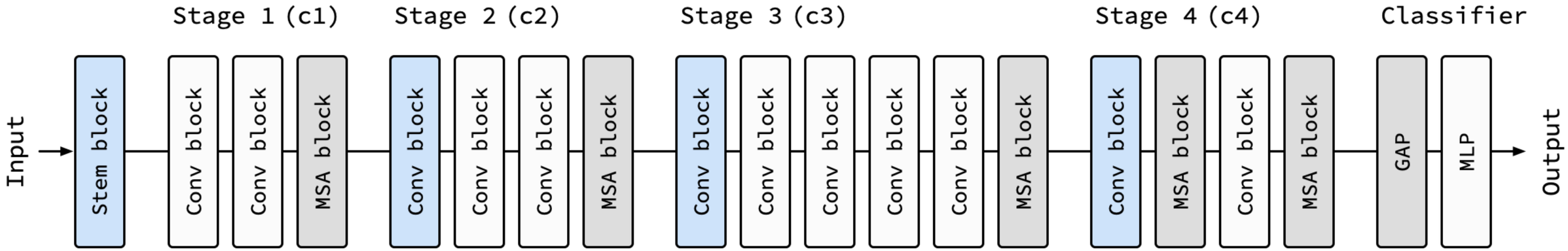
Surprisingly,
a MSA in c2 improves accuracy!

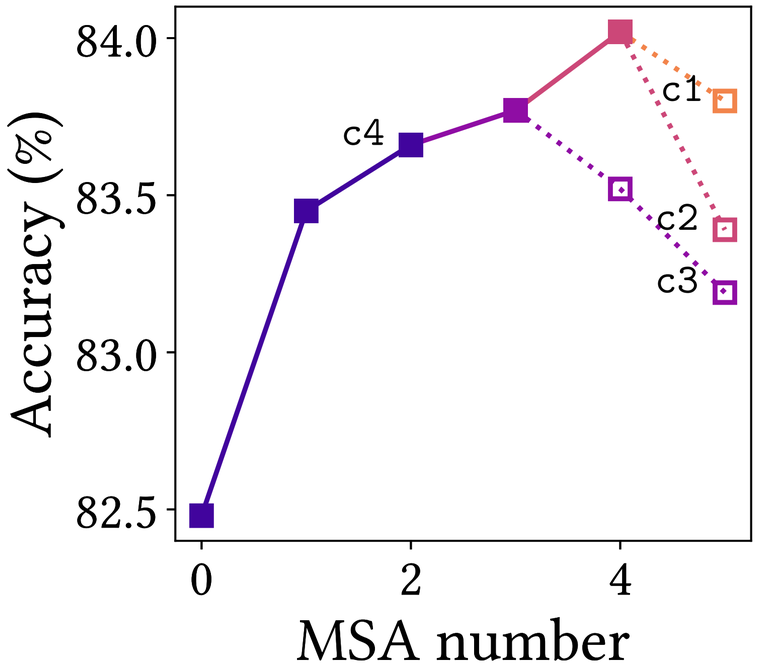


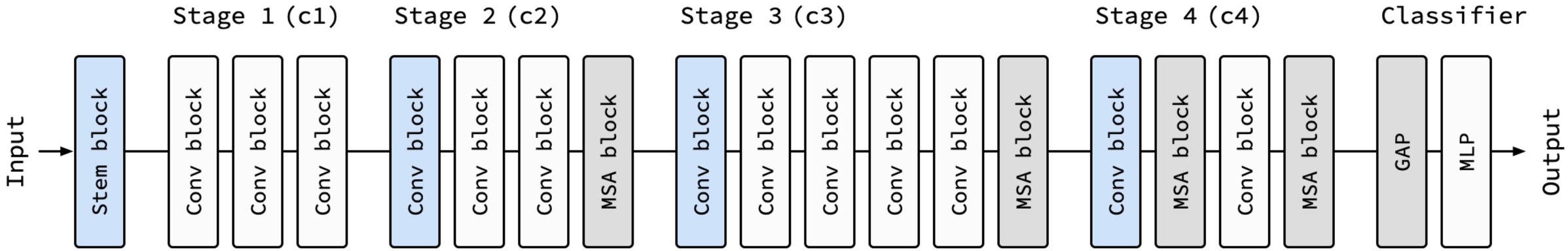

84.0
83.5
83.0
82.5
Accuracy (%)
0
2
4
MSA number
c4
c2
c3


Stage 1 (c1)
Stage 2 (c2)
Stage 3 (c3)
Stage 4 (c4)
Classifier
Input
Output
Stem block
Conv block
MSA block
GAP
MLP

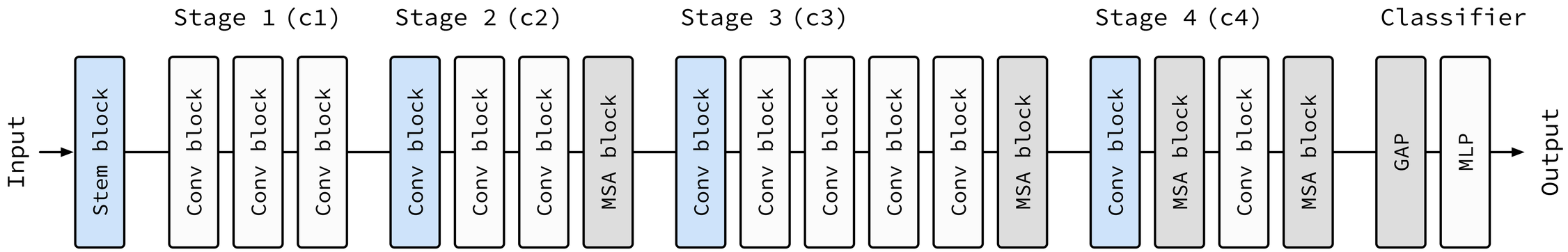

84.0
83.5
83.0
82.5
Accuracy (%)
0
2
4
MSA number
c4
c1
c2
c3


Stage 1 (c1)
Stage 2 (c2)
Stage 3 (c3)
Stage 4 (c4)
Classifier
Input
Stem block
Conv block
Conv block
MSA block
Conv block
Conv block
Conv block
MSA block
Conv block
Conv block
Conv block
Conv block
Conv block
MSA block
Conv block
MSA block
Conv block
MSA block
GAP
MLP
Output


84.0
83.5
83.0
82.5
Accuracy (%)
0
2
4
MSA number
c4
c1
c2
c3

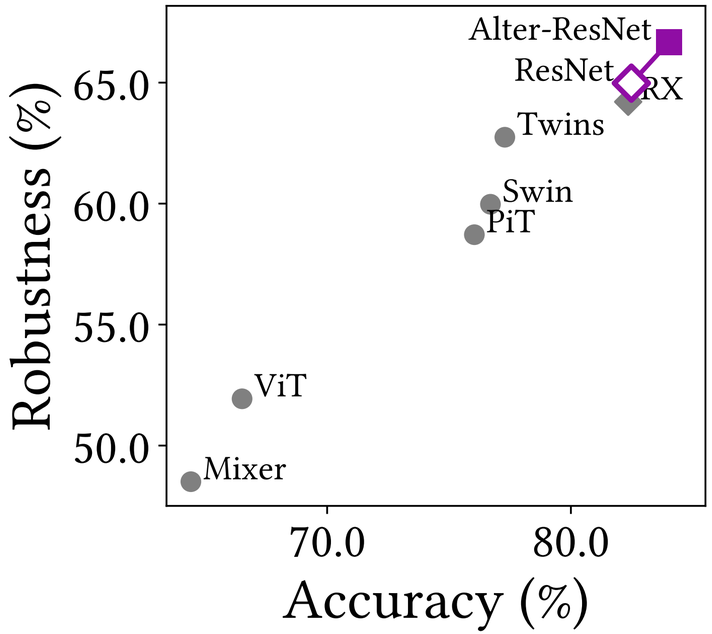

Stage 1 (c1)
Stage 2 (c2)
Stage 3 (c3)
Stage 4 (c4)
Classifier
Input
Stem block
Conv block
Conv block
Conv block
Conv block
Conv block
Conv block
MSA block
Conv block
Conv block
Conv block
Conv block
Conv block
MSA block
Conv block
MSA block
Conv block
MSA block
GAP
MLP
Output

AlterNet Based on ResNet-50

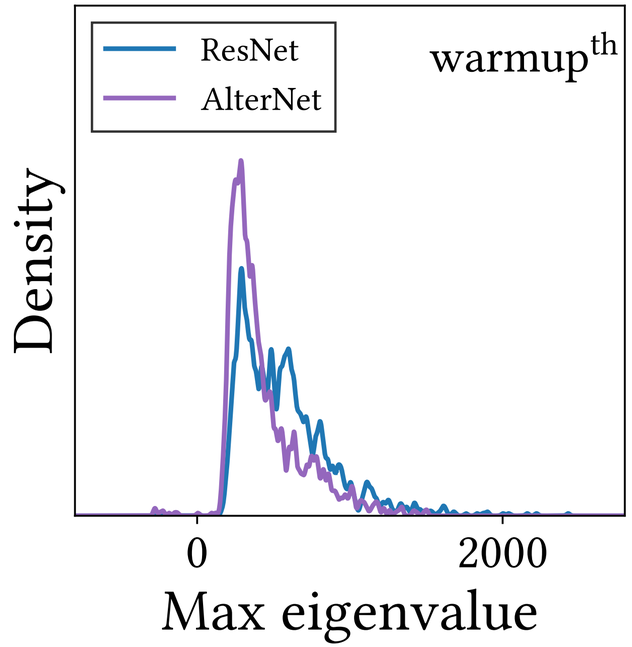


Detailed architecture of Alter-ResNet-50 for CIFAR-100. White, gray, and blue blocks mean Conv, MSA, and subsampling blocks. All stages (except stage 1) end with MSA blocks. This model is based on pre-activation ResNet-50. Following Swin, MSAs in stages 1 to 4 have 3, 6, 12, and 24 heads.

AlterNet Outperforms CNNs By Flattening Losses



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



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

AlterNet outperforms CNNs. Left: AlterNet outperforms CNNs even in a small data regime. "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.

How Do Vision Transformers Work?

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

- MSAs flatten loss landscapes. Long-range dependency disrupts NN optimization.

Q2. Do MSAs act like Convs?

- MSAs and Convs exhibit opposite behaviors; e.g., MSAs are low-pass filters.

Q3. How can we harmonize MSAs with Convs?

- MSAs *at the end of a stage* play a key role in prediction.

In summary, appropriate inductive biases improves NN optimization, and self-attentions have a spatial smoothing inductive bias, which has the following properties.

	Self-Attention	Convolution
Loss Landscape	Flat but non-convex	Convex but sharp
Fourier Analysis	Low-pass filter (shape-biased)	High-pass filter (texture-biased)
Best Practice	The end of a stage	The beginning of a stage

Finale

So, Why Do Vision Transformers Work That Way?

MSAs Are Trainable Spatial Smoothings

$$p(\boldsymbol{z}_j|\boldsymbol{x}_j, \mathcal{D}) \simeq \sum_i \overbrace{\pi(\boldsymbol{x}_i|\boldsymbol{x}_j)}^{\text{Importance}} \overbrace{p(\boldsymbol{z}_j|\boldsymbol{x}_i, \boldsymbol{w}_i)}^{\text{Prediction}}$$

where $\pi(\boldsymbol{x}_i|\boldsymbol{x}_j) = \begin{cases} 1/4, & \text{if } \boldsymbol{x}_i \text{ is a neighbour of } \boldsymbol{x}_j \\ 0, & \text{otherwise} \end{cases}$

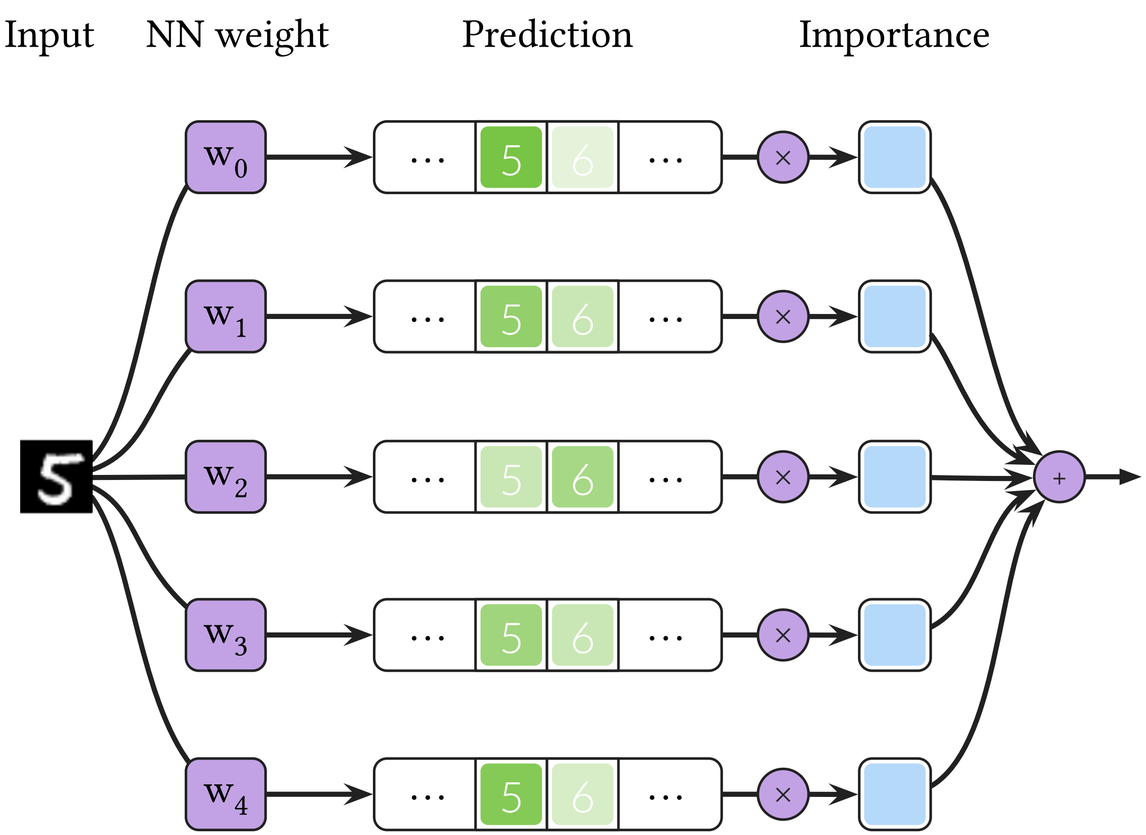
(box blur)

$$\boldsymbol{z}_j = \sum_i \underbrace{\texttt{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}}{\sqrt{d}}\right)_i}_{\text{Importance}} \underbrace{\boldsymbol{V}_{i,j}}_{\text{Prediction}}$$

(self-attention)

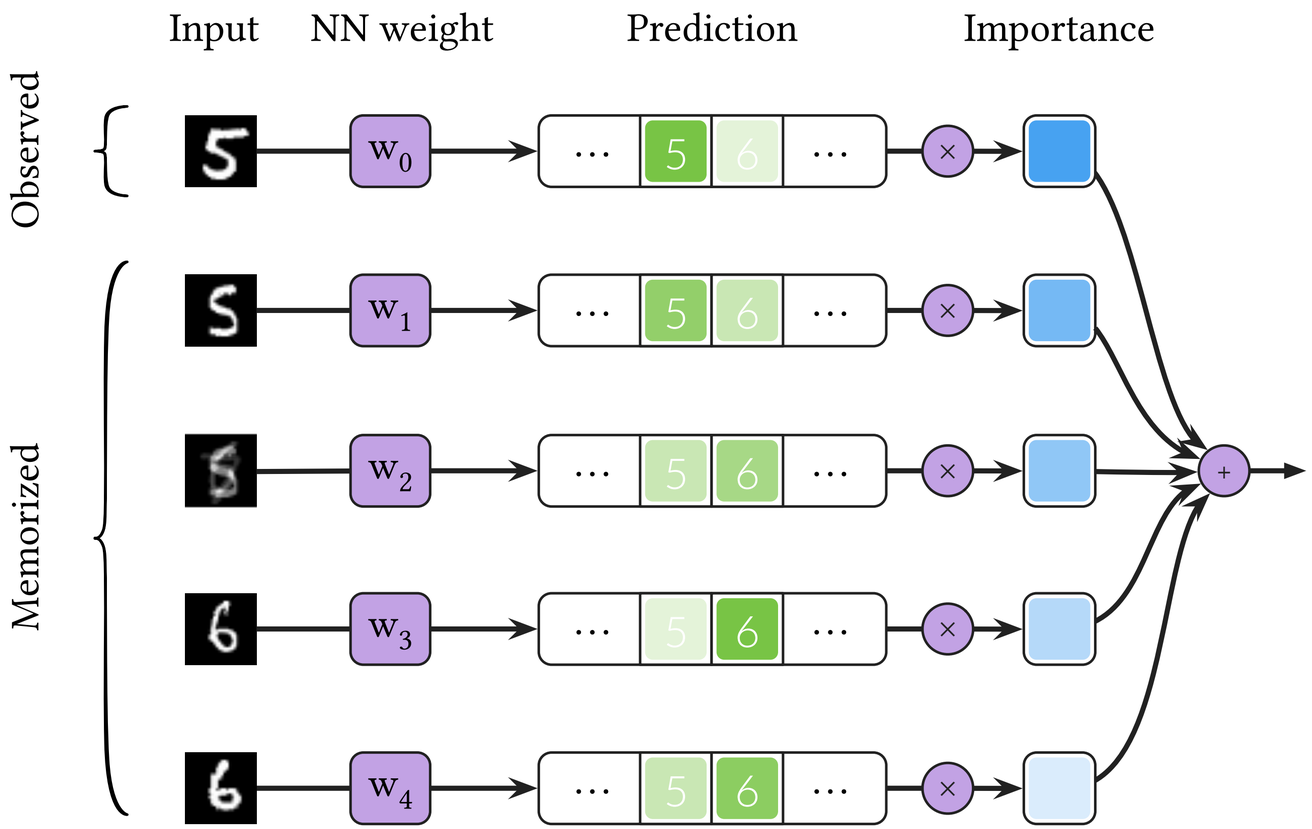
Blur is a spatial ensemble

Ensemble Average Require Multiple NN Executions



BNN inference is ensemble average of NN predictions for one observed data point. Using *N* neural networks in the ensemble would requires *N times more* computational complexity than one NN execution. The NN weights are sampled from ***weight distribution***.

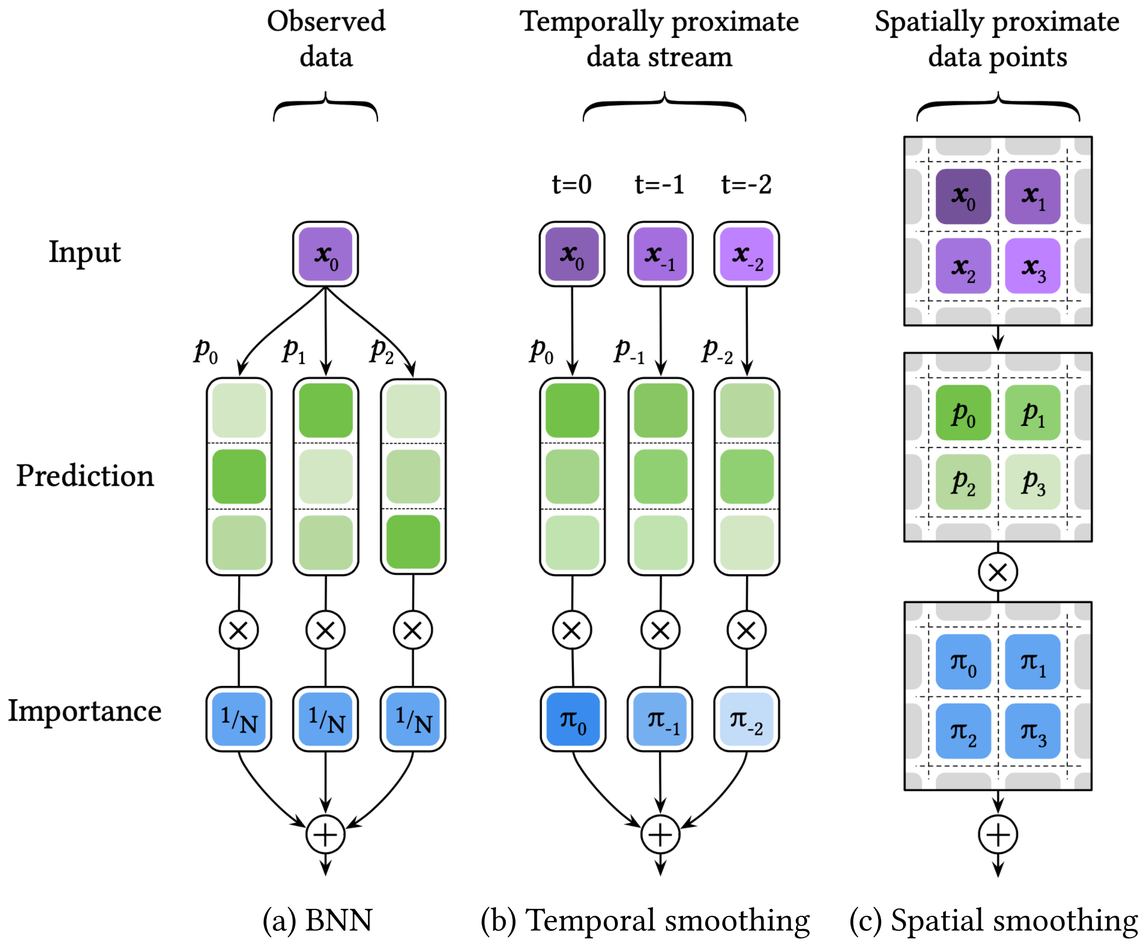
Memorized Results Can Speed up the Ensemble



We ensemble average the prediction for the observed input and memorized predictions with importance-weights. This method executes NN for an observed data only once, and complements the result with previously calculated predictions for other data.

Importance is similarity between the observed data and the memorized data. In other words, memorized data is sampled from ***data distribution***, and the probability is the importance.

Spatial Smoothing: Ensemble of Feature Maps



(a) BNN (b) Temporal smoothing (c) Spatial smoothing

Comparison of three different Bayesian neural network inferences: canonical BNN inference, temporal smoothing, and spatial smoothing. Spatial smoothing aggregates neighboring feature map points.

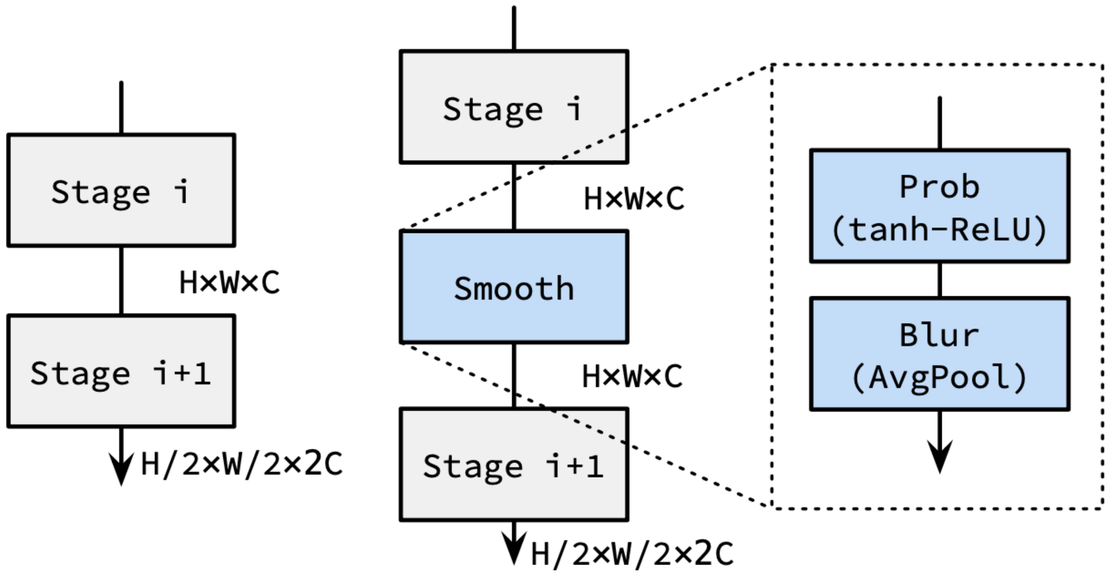
Ensemble Averaging for Proximate Data Points

proximate data points

= neighboring feature maps

$$p(\boldsymbol{z}_j|\boldsymbol{x}_j, \mathcal{D}) \simeq \sum_i \underbrace{\pi(\boldsymbol{x}_i|\boldsymbol{x}_j)}_{\text{Importance}} \underbrace{p(\boldsymbol{z}_j|\boldsymbol{x}_i, \boldsymbol{w}_i)}_{\text{Prediction}}$$

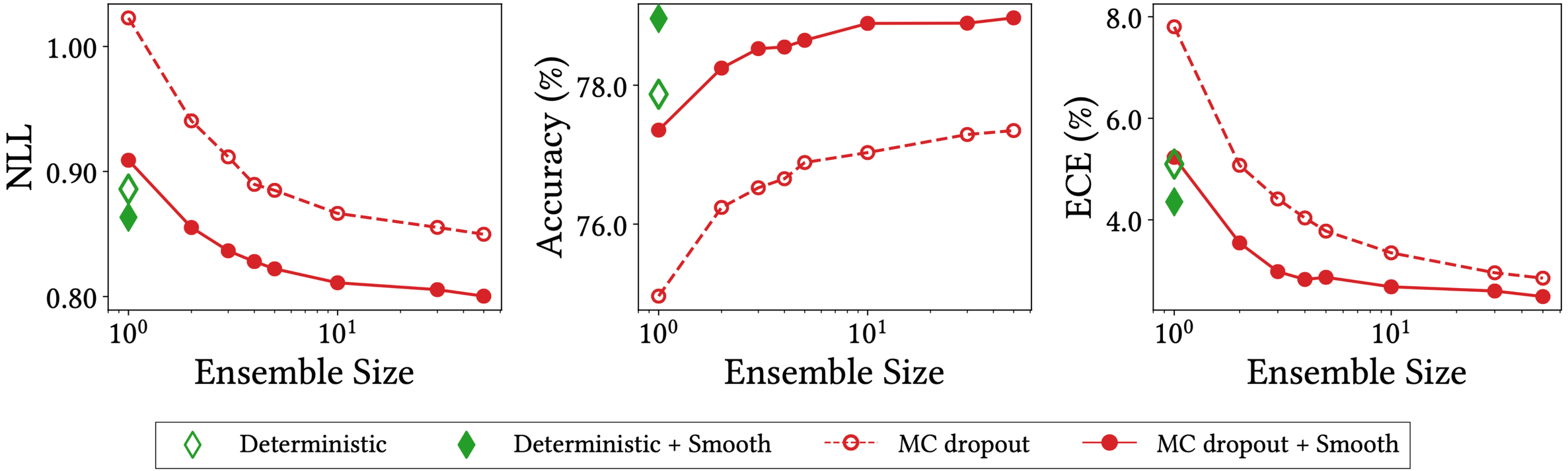
Module Architecture of Spatial Smoothing



(a) ResNet (b) ResNet + Smooth

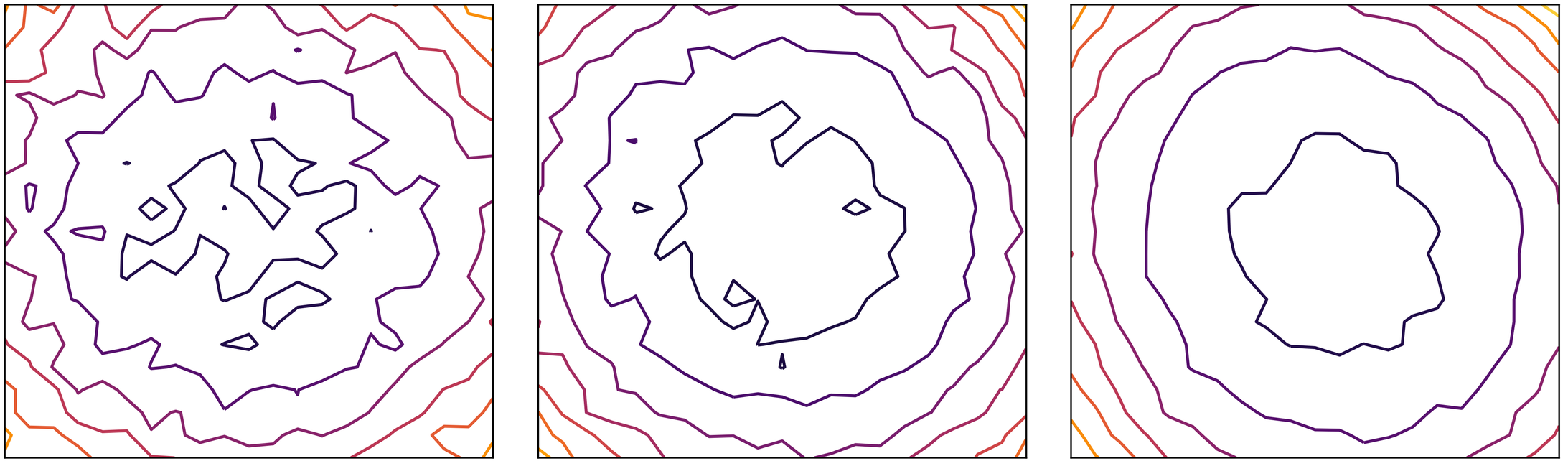
Stages of CNNs such as ResNet (left) and the stages incorporating spatial smoothing layer (right). "`Prob`" in spatial smoothing transforms a real-valued feature map into probability. "`Blur`" aggregates the probabilities from feature maps.

Spatial Smoothing Improves the Predictive Performances



Spatial smoothing improves accuracy and uncertainty estimation across a whole range of ensemble sizes. In particular, spatial smoothing achieve high predictive performance merely with a handful of ensembles. We report the predictive performance of ResNet-18 on CIFAR-100.

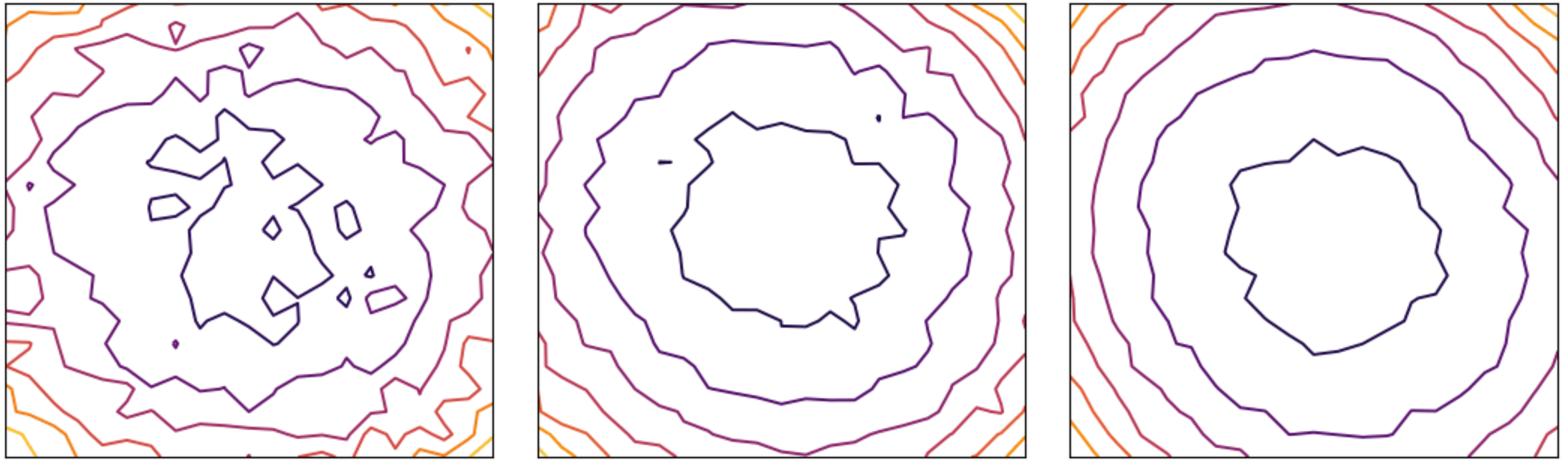
Spatial Smoothings Smoothen the Loss Landscapes



(a) MLP classifier (b) GAP classifier (*vanilla*) (c) GAP classifier + Smooth

The loss landscape of the MLP classifier is the most irregular. Both GAP and spatial smoothing flatten the loss landscapes. We present the loss landscape visualizations of ResNet-18 models with MC dropout on CIFAR-100.

Spatial Smoothings Smoothen the Loss Landscapes



(a) MLP classifier (b) GAP classifier (*vanilla*) (c) GAP classifier + Smooth

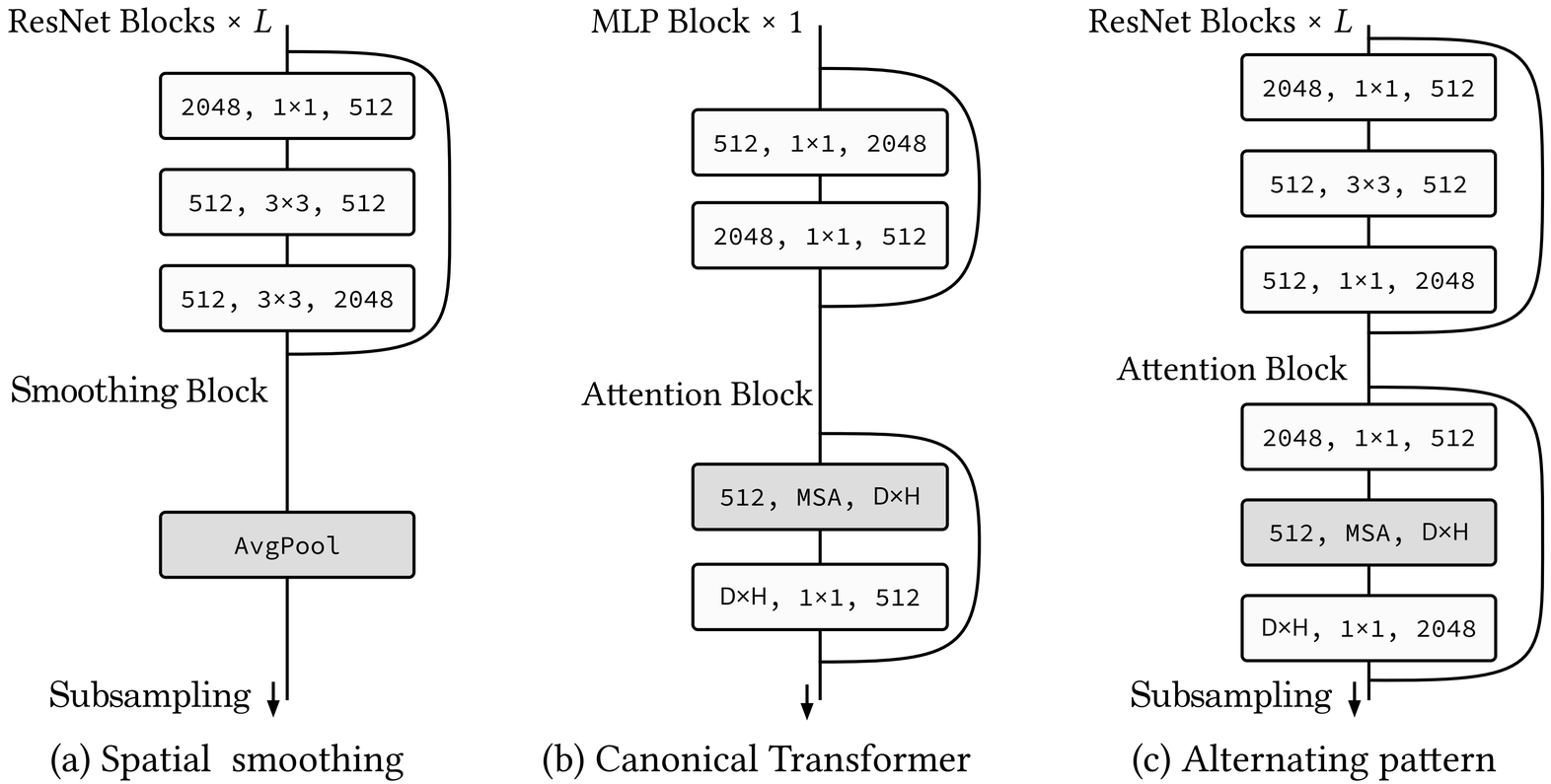
The loss landscape of the MLP classifier is the most irregular. Both GAP and spatial smoothing flatten the loss landscapes. We present the loss landscape visualizations of ResNet-18 models with MC dropout on CIFAR-100.

This is an animation! Please visit: https://github.com/xxxnell/spatial-smoothing/

MSAs Are a Trainable Spatial Smoothing

- Spatial smoothing helps in NN optimization by flattening the loss landscapes.
- Spatial smoothing is a low-pass filter. Spatial smoothing also improves the robustness against high-frequency noise significantly.
- Spatial smoothing is effective when applied at the end of a stage. This is because it aggregates the transformed feature map predictions.

Comparison of Three Different Repeating Patterns



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

Left: Spatial smoothings are located at the end of CNN stages. ***Middle***: The stages of ViTs consist of repetitions of canonical Transformers. "D" is the hidden dimension and "H" is the number of heads. ***Right***: The stages using alternating pattern consists of a number of CNN blocks and an MSA block.

Conclusion

MSAs are not merely generalized Convs, but rather **generalized spatial smoothings** that complement Convs. They inherit the properties of spatial smoothings:

- **Optimization:** MSAs flatten loss landscapes. It improves not only accuracy but also generalization. Such improvement is primarily attributable to their formulation including data specificity, ***NOT*** long-range dependency.
- **Behavior:** MSAs and Convs exhibit opposite behaviors. For example, MSAs are low-pass filters, but Convs are high-pass filters. MSAs aggregate feature maps, but Convs transform them.
- **Architecture:** Multi-stage neural nets behave like a series connection of small individual models. MSAs at the end of a stage play a key role in prediction.

In short, **self-attention formulation is an appropriate inductive bias** that complements Convs.

Further Information

- **Researcher Homepage**: https://www.namukpark.com
- **"How Do Vision Transformers Work?"** (paper): https://openreview.net/forum?id=D78Go4hVcxO
- **"How Do Vision Transformers Work?"** (code and summary): https://github.com/xxxnell/how-do-vits-work
- **"Blurs Behave Like Ensembles: Spatial Smoothings to Improve Accuracy, Uncertainty, and Robustness"** (prior work): https://arxiv.org/abs/2105.12639
- **"Blurs Behave Like Ensembles: Spatial Smoothings to Improve Accuracy, Uncertainty, and Robustness"** (code and summary): https://github.com/xxxnell/spatial-smoothing