When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

School of Industrial and Management Engineering, Korea University

Yongwon Jo

KOREA UNIVERSITY

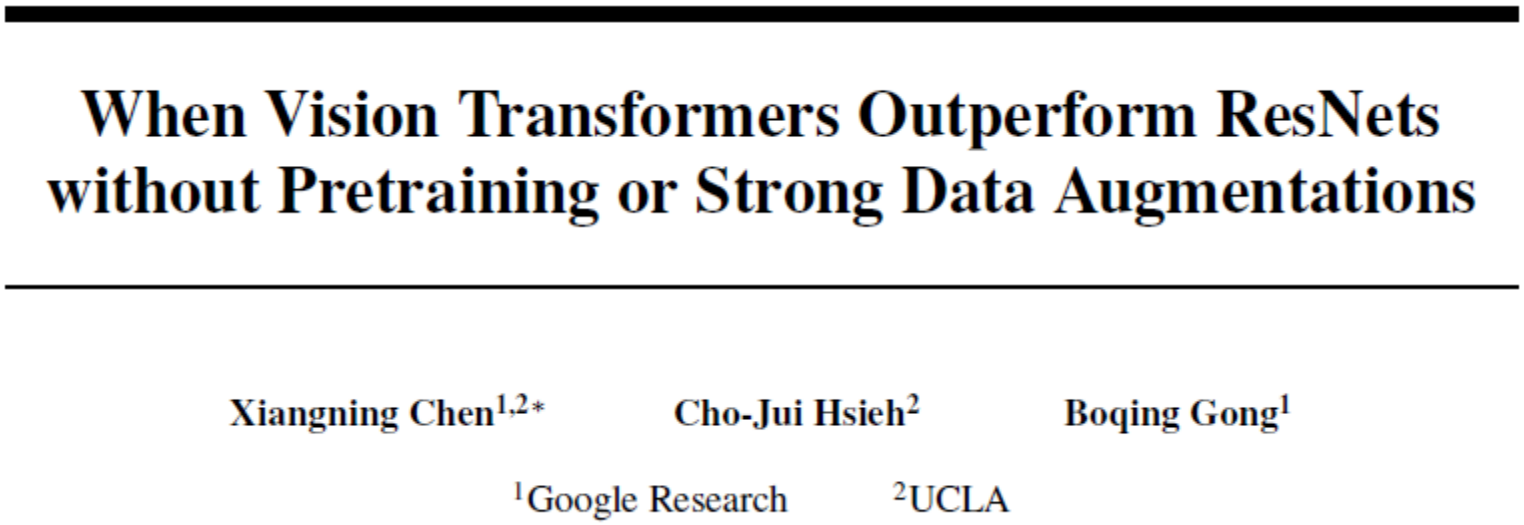
DMQA

Contents

Research Purpose

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations (arXiv, 2021)

- 저자들은 Google Research 그룹과 UCLA 소속이며 2021년 8월 20일 기준 인용 횟수는 0회



When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

Xiangning Chen[1,2]* **Cho-Jui Hsieh[2]** **Boqing Gong[1]**

[1]Google Research [2]UCLA

Research Purpose

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- Computer vision 분야에서 Vision transformer(ViT)와 Multi-layer perceptron(MLP)가 큰 화두
  - ➢ 이미지 내 미세한 특징 추출이 가능한 합성곱 신경망(Convolutional neural network, CNN)을 주로 사용
  - ➢ 미세한 특징 추출이 가능하다는 CNN의 Inductive bias와 같은 가정이 ViT와 MLP에는 존재×
  - ➢ 특정 가정이 없기에 두 모델은 많은 양의 데이터로 모델을 학습하거나 데이터 증강 기법이 반드시 필요
  - ➢ **'작은 데이터와 증강 기법 없이 CNN 기반 모델과 유사한 성능을 낼 수 없을까?'**

Research Purpose

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- ViT 와 MLP-Mixer의 문제점
  - ① ViT와 MLP-Mixer는 방대한 데이터나 강력한 데이터 증강 기법(Strong data augmentation) 적용 필요
  - ② 또한 모델 Hyperparameter에 매우 민감한 것으로 알려짐
  - ③ 기존 CNN 기반 분류기보다 In distribution 뿐만 아니라 Out of distribution에서 성능이 낮음

Table 1: Number of parameters, NTK condition number κ, Hessian dominate eigenvalue λ_{max}, accuracy on ImageNet, and accuracy/robustness on ImageNet-C. ViT and MLP-Mixer suffer divergent κ and converge to sharp regions of big λ_{max}; SAM rescues that and leads to better generalization.

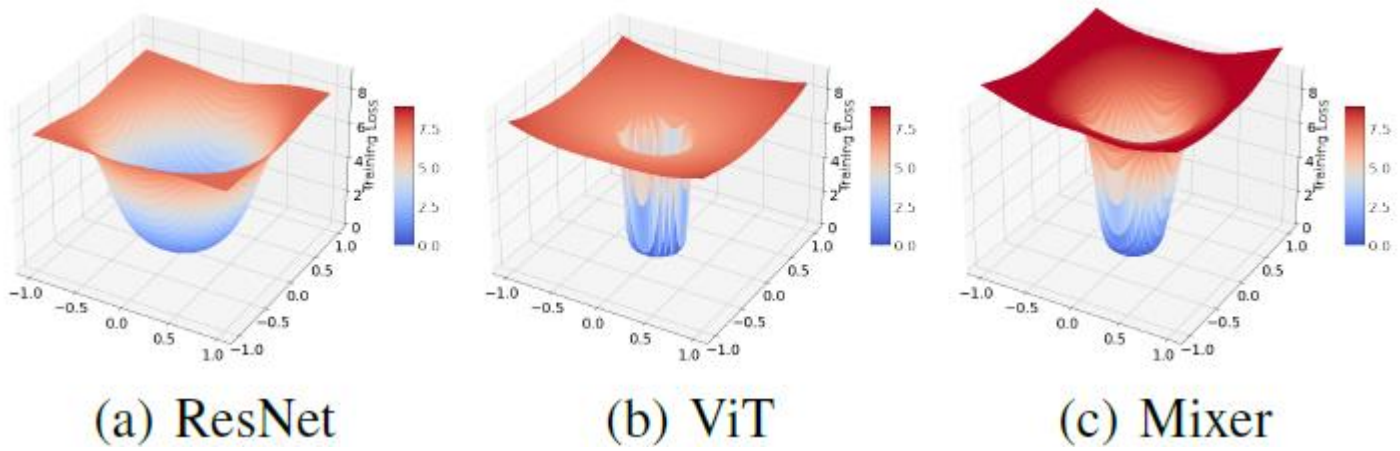
	ResNet-50	ResNet-152	ViT-B/16	ViT-B/16-SAM	Mixer-B/16	Mixer-B/16-SAM
#Params	25M	60M	87M		59M	
NTK κ	2801.6	2801.6	4205.3		14468.0	
Hessian λ_{max}	122.9	179.8	738.8	**20.9**	1644.4	**22.5**
ImageNet (%)	76.0	78.5	74.6	**79.9**	66.4	**77.4**
ImageNet-C (%)	44.6	50.0	46.6	**56.5**	33.8	**48.8**

ImageNet-C: ImageNet 내 이미지에 변형을 가해 Robustness/Generalization ability 측정 시 사용하는 데이터 셋 (Out of distribution) vs ImageNet (In distribution)

Research Purpose

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- 모델 파라미터들에 대한 Loss landscape를 시각화 하여 왜 ViT와 MLP 학습이 어려운지 설명
  - ➢ ResNet대비 ViT와 MLP-Mixer는 가파른 Local minima가 존재
  - ➢ 해당 minima에 빠질 시 빠져나가기 어렵고 일반화 성능이 낮은 것으로 알려짐
  - ➢ ResNet과 같이 평평한 minima의 경우 일반화 성능이 높음
  - ➢ Table 1 내 Hessian eigenvalue λ_{max}가 높을수록 날카로운 Loss landscape를 가진다고 할 수 있음

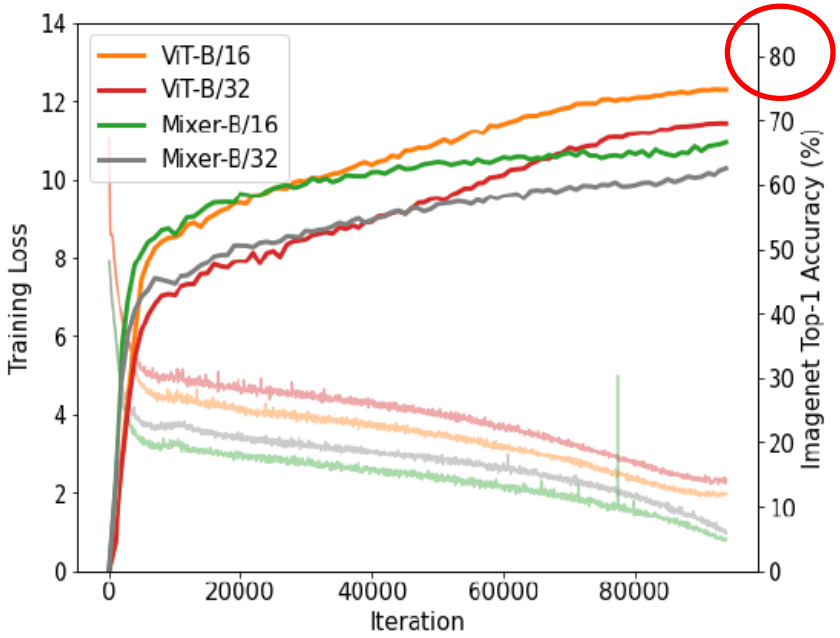


Cross entropy loss function을 사용해 학습한 각 모델 별 Landscape

Research Purpose

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- ViT와 MLP-Mixer의 학습 데이터 오차 비교
  - ➢ ViT 대비 MLP-Mixer 가 모델 파라미터 수도 적고 학습 데이터에 대한 성능은 높음
  - ➢ 하지만 테스트 데이터 셋에 대해서는 ViT 성능이 높기에 MLP-Mixer에서는 과적합이 발생한 것
  - ➢ Cross-token과 Self-attention 차이에서 발생하는 것으로 판단



Research Purpose

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- ViT 와 MLP-Mixer의 낮은 학습 가능성 (Trainability)
  - 학습 가능성을 입력 데이터와 가장 가까운 층의 가중치의 Jacobian matrix 이용해 정의
  - Neural tangent kernel(NTK):$\Theta(x, x') = J(x)J(x)^t$ 이며 J는 Jacobian matrix
  - NTK의 고유벡터를 산출하고 가장 큰 값을 가장 작은 값으로 나눈 값을 κ라 정의
  - ResNet 대비 ViT와 MLP-Mixer의 경우 매우 큰 κ이며 이 때문에 ViT와 MLP-Mixer 학습이 어려움

Table 1: Number of parameters, NTK condition number κ, Hessian dominate eigenvalue λ_{max}, accuracy on ImageNet, and accuracy/robustness on ImageNet-C. ViT and MLP-Mixer suffer divergent κ and converge to sharp regions of big λ_{max}; SAM rescues that and leads to better generalization.

	ResNet-50	ResNet-152	ViT-B/16	ViT-B/16-SAM	Mixer-B/16	Mixer-B/16-SAM
#Params	25M	60M	87M		59M	
NTK κ	2801.6	2801.6	4205.3		14468.0	
Hessian λ_{max}	122.9	179.8	738.8	**20.9**	1644.4	**22.5**
ImageNet (%)	76.0	78.5	74.6	**79.9**	66.4	**77.4**
ImageNet-C (%)	44.6	50.0	46.6	**56.5**	33.8	**48.8**

ImageNet-C: ImageNet 내 이미지에 변형을 가해 Robustness/Generalization ability 측정 시 사용하는 데이터 셋 (Out of distribution) vs ImageNet (In distribution)

Principled Optimizer for Convolution-Free Vision Architectures

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- First-order optimizer(ex. SGD, Adam)처럼 단순히 학습 데이터 오차만 줄이는 학습 방식 문제 제기
  - ➢ 일반적인 Deep neural network 최적화 문제는 Non-convex 이며 최적화하기 어려움
  - ➢ First-order optimizer 사용 시 손실 함수 Landscape 상에서 낮은 점을 찾을 수 있으나 일반화 성능↓
  - ➢ ViT와 MLP-Mixer와 유사한 Local minima가 학습 데이터에 과적합 되어 있고 일반화 성능↓
  - ➢ **가정: Loss landscape를 평평하게 만드는 Optimizer는 없을까?**
  - ➢ **Sharpness-Aware Minimization for Efficiently Improving Generalization**
  - ➢ 저자들은 Google brain 소속이며 8월 23일 기준 47회 인용 (2021 ICLR)

SHARPNESS-AWARE MINIMIZATION FOR EFFICIENTLY IMPROVING GENERALIZATION

Pierre Foret *
Google Research
pierre.pforet@gmail.com

Ariel Kleiner
Google Research
akleiner@gmail.com

Hossein Mobahi
Google Research
hmobahi@google.com

Behnam Neyshabur
Blueshift, Alphabet
neyshabur@google.com

Principled Optimizer for Convolution-Free Vision Architectures

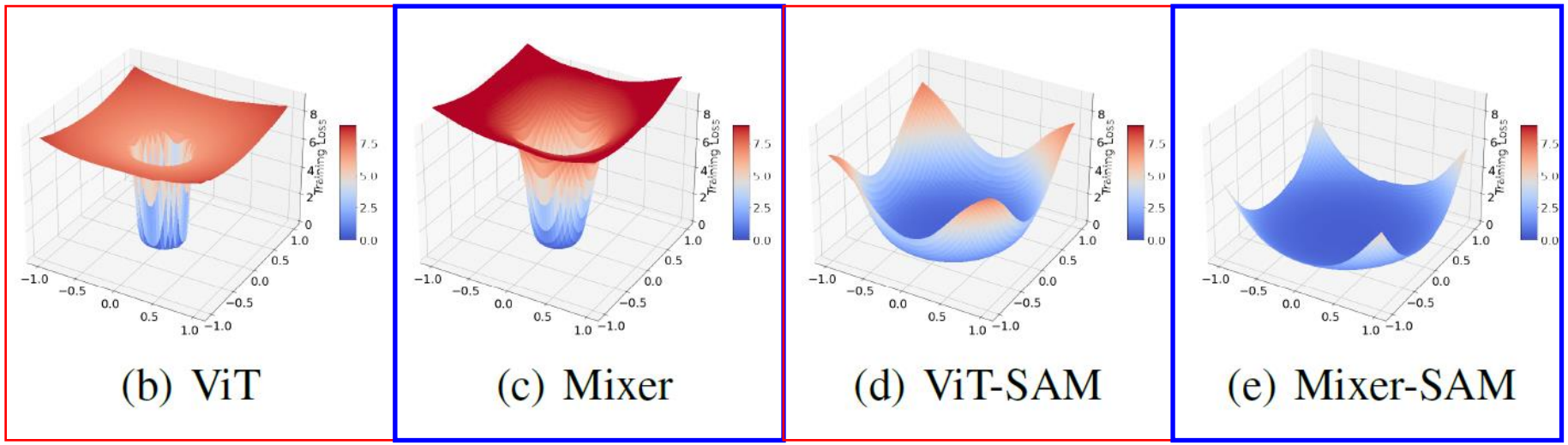
❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- First-order optimizer(ex. SGD, Adam)처럼 단순히 학습 데이터 오차만 줄이는 학습 방식 문제 제기
  - ➢ **가정: Loss landscape를 평평하게 만드는 Optimizer는 없을까?**
  - ➢ **Sharpness-Aware Minimization (SAM) for Efficiently Improving Generalization**
  - ➢ SAM은 (Local/ Global) minima 탐색 및 해당 포인트 주변도 낮은 Loss를 가지도록 하는 Optimization 기법
  - ➢ SAM 관련된 상세 내용은 해당 논문 참고 부탁드리며 이번엔 실험에 초점을 맞추어 설명

Principled Optimizer for Convolution-Free Vision Architectures

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- SAM을 사용해 ViT와 MLP-Mixer를 학습하고 Loss landscape 시각화 결과
  - ➢ 많은 데이터를 사용하거나 Strong data augmentation을 적용하지 않음
  - ➢ 단순히 ImageNet만 사용해서 학습하고 성능 평가 진행
  - ➢ 기존 Loss landscape 대비 Loss 값이 낮은 지점 주변이 완만해진 것을 확인 가능



SAM으로 모델 파라미터 업데이터 진행 후 모델 별 Landscape

Principled Optimizer for Convolution-Free Vision Architectures

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- SAM을 사용해 ViT와 MLP-Mixer를 학습하고 Loss landscape 시각화 결과
  - ➢ In distribution과 Out of distribution 데이터셋에 대한 정량적인 지표로 성능 향상 확인 가능

Table 2: Accuracy and robustness of ResNets, ViTs, and MLP-Mixers trained from scratch on ImageNet with SAM (improvement over the models trained using vanilla SGD is shown in the parentheses). We use the Inception-style preprocessing (with resolution 224) rather than a combination of strong data augmentations. ViTs achieve better accuracy and robustness than ResNets of similar size and throughput (calculated following [53]), and MLP-Mixers become on par with ResNets.

Model	#params	Throughput (img/sec/core)	ImageNet	Real	V2	ImageNet-R	ImageNet-C
ResNet							
ResNet-50-SAM	25M	2161	76.7 (+0.7)	83.1 (+0.7)	64.6 (+1.0)	23.3 (+1.1)	46.5 (+1.9)
ResNet-101-SAM	44M	1334	78.6 (+0.8)	84.8 (+0.9)	66.7 (+1.4)	25.9 (+1.5)	51.3 (+2.8)
ResNet-152-SAM	60M	935	79.3 (+0.8)	84.9 (+0.7)	67.3 (+1.0)	25.7 (+0.4)	52.2 (+2.2)
ResNet-50x2-SAM	98M	891	79.6 (+1.5)	85.3 (+1.6)	67.5 (+1.7)	26.0 (+2.9)	50.7 (+3.9)
ResNet-101x2-SAM	173M	519	80.9 (+2.4)	86.4 (+2.4)	69.1 (+2.8)	27.8 (+3.2)	54.0 (+4.7)
ResNet-152x2-SAM	236M	356	81.1 (+1.8)	86.4 (+1.9)	69.6 (+2.3)	28.1 (+2.8)	55.0 (+4.2)
Vision Transformer							
ViT-S/32-SAM	23M	6888	70.5 (+2.1)	77.5 (+2.3)	56.9 (+2.6)	21.4 (+2.4)	46.2 (+2.9)
ViT-S/16-SAM	22M	2043	78.1 (+3.7)	84.1 (+3.7)	65.6 (+3.9)	24.7 (+4.7)	53.0 (+6.5)
ViT-S/14-SAM	22M	1234	78.8 (+4.0)	84.8 (+4.5)	67.2 (+5.2)	24.4 (+4.7)	54.2 (+7.0)
ViT-S/8-SAM	22M	333	81.3 (+5.3)	86.7 (+5.5)	70.4 (+6.2)	25.3 (+6.1)	55.6 (+8.5)
ViT-B/32-SAM	88M	2805	73.6 (+4.1)	80.3 (+5.1)	60.0 (+4.7)	24.0 (+4.1)	50.7 (+6.7)
ViT-B/16-SAM	87M	863	79.9 (+5.3)	85.2 (+5.4)	67.5 (+6.2)	26.4 (+6.3)	56.5 (+9.9)
MLP-Mixer							
Mixer-S/32-SAM	19M	11401	66.7 (+2.8)	73.8 (+3.5)	52.4 (+2.9)	18.6 (+2.7)	39.3 (+4.1)
Mixer-S/16-SAM	18M	4005	72.9 (+4.1)	79.8 (+4.7)	58.9 (+4.1)	20.1 (+4.2)	42.0 (+6.4)
Mixer-S/8-SAM	20M	1498	75.9 (+5.7)	82.5 (+6.3)	62.3 (+6.2)	20.5 (+5.1)	42.4 (+7.8)
Mixer-B/32-SAM	60M	4209	72.4 (+9.9)	79.0 (+10.9)	58.0 (+10.4)	22.8 (+8.2)	46.2 (12.4)
Mixer-B/16-SAM	59M	1390	77.4 (+11.0)	83.5 (+11.4)	63.9 (+13.1)	24.7 (+10.2)	48.8 (+15.0)
Mixer-B/8-SAM	64M	466	79.0 (+10.4)	84.4 (+10.1)	65.5 (+11.6)	23.5 (+9.2)	48.9 (+16.9)

ImageNet-C & ImageNet-R: ImageNet 내 이미지에 변형을 가해 Robustness/Generalization ability 측정 시 사용하는 데이터 셋

Principled Optimizer for Convolution-Free Vision Architectures

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- SAM을 사용해 ViT와 MLP-Mixer의 Loss landscape를 의미하는 λ_{max} 값의 급격한 하락
- Loss landscape가 완만해지고 ViT&MLP-Mixer 기존 문헌 보다 좋은 Local minima를 찾은 것
- 모델 내 요소의 λ_{max} 값 계산을 통해 Self-attention, Cross-token 부분에서 큰 λ_{max} 감소 확인

Table 3: Dominant eigenvalue λ_{max} of the sub-diagonal Hessians for different network components, and norm of the model parameter w and the post-activation a_k of block k. Each ViT block consists of a MSA and a MLP, and MLP-Mixer alternates between a token MLP a channel MLP. Shallower layers have larger λ_{max}. SAM smooths every component.

Model	λ_{max} of diagonal blocks of Hessian							$\|w\|_2$	$\|a_1\|_2$	$\|a_6\|_2$	$\|a_{12}\|_2$
	Embedding	MSA/Token MLP	MLP/Channel MLP	Block1	Block6	Block12	Whole				
ViT-B/16	300.4	179.8	281.4	44.4	32.4	26.9	738.8	269.3	104.9	104.3	138.1
ViT-B/16-SAM	3.8	8.5	9.6	1.7	1.7	1.5	20.9	353.8	117.0	120.3	97.2
Mixer-B/16	1042.3	95.8	417.9	239.3	41.2	5.1	1644.4	197.6	96.7	135.1	74.9
Mixer-B/16-SAM	18.2	1.4	9.5	4.0	1.1	0.3	22.5	389.9	110.9	176.0	216.1

Principled Optimizer for Convolution-Free Vision Architectures

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- SAM의 효과로 기존 대비 객체를 정확히 인식하는 것을 확인 가능

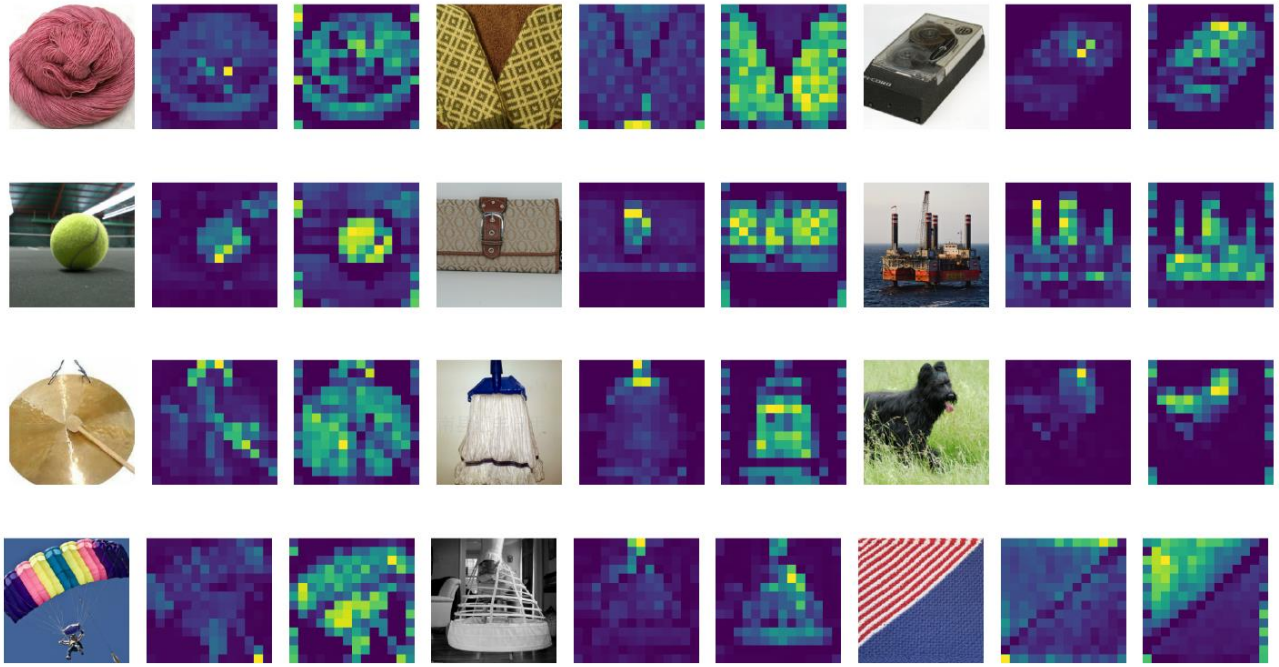


Figure 3: Raw images (**Left**) and attention maps of ViT-S/16 with (**Right**) and without (**Middle**) sharpness-aware optimization. ViT-S/16 with less sharp local optimum contains perceptive segmentation information in its attention maps.

Principled Optimizer for Convolution-Free Vision Architectures

❖ When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations

- 학습 데이터 개수를 감소시키며 성능 변화를 확인하는 실험
- 단순히 Cross entropy만 사용한 경우 ViT와 MLP-Mixer 성능 하락은 ResNet보다 더 큼
- 하지만 SAM을 사용해 학습한다면 성능 하락 속도를 줄여줄 수 있음
- Strong augmentation(ex. RandAugment, MixUp) 사용하여 성능 향상보다 SAM 성능 향상이 더큼

Table 4: Data augmentation, SAM, and their combination applied to different model architectures trained on ImageNet and its subsets.

Training Set	#Images	ResNet-152		ViT-B/16				Mixer-B/16			
		Vanilla	SAM	Vanilla	SAM	AUG	SAM + AUG	Vanilla	SAM	AUG	SAM + AUG
ImageNet	1,281,167	78.5	79.3	74.6	79.9	79.6	81.5	66.4	77.4	76.5	78.1
i1k (1/2)	640,583	74.2	75.6	64.9	75.4	73.1	75.8	53.9	71.0	70.4	73.1
i1k (1/4)	320,291	68.0	70.3	52.4	66.8	63.2	65.6	37.2	62.8	61.0	65.8
i1k (1/10)	128,116	54.6	57.1	32.8	46.1	38.5	45.7	21.0	43.5	43.0	51.0

Conclusion

❖ Conclusion

- ViT와 MLP-Mixer 학습이 왜 어려운지 Loss landscape 관점에서 해석
- 날카로운 Loss landscape가 아닌 평평한 landscape를 위해 SAM 방식으로 두 모델 학습
- In distribution과 Out of distribution에 대해 기존 대비 성능 향상 성공
- Robustness 뿐만 아니라 다양한 방식으로 SAM 학습 방식이 ViT와 MLP-Mixer 학습에 적합하다는 것을 증명

❖ 본 논문 읽은 뒤 나의 생각

- Optimization에 대해 다시 한번 생각하게 되는 기회
- Loss landscape를 직접 시각화 해보고 필자가 하고 있는 연구에 적용 가능성 검토
- 방대한 양의 실험을 체계적으로 한 것 같다는 생각

References

- Chen, X., Hsieh, C. J., & Gong, B. (2021). When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations. arXiv preprint arXiv:2106.01548.
- Foret, P., Kleiner, A., Mobahi, H., & Neyshabur, B. (2020). Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412.


Thank You