How to train your ViT? Data Augmentation and Regularization in Vision Transformers

School of Industrial and Management Engineering, Korea University

Eun Ji Koh

KOREA UNIVERSITY

DMQA

Contents

Research Purpose

❖ How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers (arXiv, 2021)

- Google Research, Brain Team에서 연구하였으며 2021년 09월 2일 기준으로 인용 수 없음

How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers

Andreas Steiner*, Alexander Kolesnikov*, Xiaohua Zhai*
Ross Wightman†, Jakob Uszkoreit, Lucas Beyer*

Google Research, Brain Team; †independent researcher

{andstein,akolesnikov,xzhai,usz,lbeyer}@google.com, rwightman@gmail.com

Abstract

Vision Transformers (ViT) have been shown to attain highly competitive performance for a wide range of vision applications, such as image classification, object detection and semantic image segmentation. In comparison to convolutional neural networks, the Vision Transformer's weaker inductive bias is generally found to cause an increased reliance on model regularization or data augmentation ("AugReg" for short) when training on smaller training datasets. We conduct a systematic empirical study in order to better understand the interplay between the amount of training data, AugReg, model size and compute budget. [1] As one result of this study we find that the combination of increased compute and AugReg can yield models with the same performance as models trained on an order of magnitude more training data: we train ViT models of various sizes on the public ImageNet-21k dataset which either match or outperform their counterparts trained on the larger, but not publicly available JFT-300M dataset.

DMQA

Research Purpose

❖ How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers

- Vision Transformer는 vision 분야에서 CNN의 대안으로 떠오르고 있음
- CNN은 translational equivariance한 특징이 있으나, ViT는 없기 때문에 이를 극복하기 위해서는 다량의 training dataset 또는 strong augmentation and regularization schemes이 필요함
- 본 논문은 50,000개의 ViT 모델을 동일 조건 하에서 학습하고 성능을 비교하였으며, dataset 사이즈를 비롯해 여러 조건을 변경하며 computational cost와 model performance에 대해 연구함

Data Augmentation and Regularization in Vision Transformers

❖ Scope of the study

- ViT를 학습시키기 위해 대규모 데이터셋을 pre-training한 모델을 fine-tuning하는 방식이 주로 사용
- 모델의 computational and sample efficiency를 characterize하기 위한 다양한 방법이 있음
  - ➢ 1) Overall computational and sample cost를 측정하는 방법으로, 일반적으로 pre-training cost의 비중이 큼
  - ➢ 2) Pre-train된 모델을 사용하여 fine-tuning에 집중하는 경우에는 fine-tuning에 소모되는 cost를 측정하는 것이 중요
  - ➢ 3) Training cost는 모두 중요하지 않고 trained model을 통한 inference cost가 중요한 경우
- 본 논문에서는 주로 실무/현업에서 중요하게 생각하는 fine-tuning cost, inference cost에 집중함

Experiments

❖ Experiment set-up: **Model**

- 4 different configuration: ViT-Ti, ViT-s, ViT-B, ViT-L
- Patch size: 16x16 (additionally 32x32 for ViT-S, ViT-B)
- 기존의 ViT MLP head는 2개의 layer로 되어있는데, 이를 하나의 layer로 변경함
  - ➢ empirically 성능에 영향을 미치지 않고, optimization instability의 원인이 되기 때문
- Hybrid model도 실험함
- 이외의 모델 구조와 관련한 조건은 아래 표에 기재

Table 1: Configurations of ViT models.

Model	Layers	Width	MLP	Heads	Params
ViT-Ti [34]	12	192	768	3	5.8M
ViT-S [34]	12	384	1536	6	22.2M
ViT-B [10]	12	768	3072	12	86M
ViT-L [10]	24	1024	4096	16	307M

Table 2: ResNet+ViT hybrid models.

Model	Resblocks	Patch-size	Params
R+Ti/16	[]	8	6.4M
R26+S/32	[2, 2, 2, 2]	1	36.6M
R50+L/32	[3, 4, 6, 3]	1	330.0M

Experiments

❖ Experiment set-up: **Regularization and data Augmentations**

- Dropout 및 stochastic depth를 사용함
- Mixup, RandAugment를 사용
- 총 28개의 configuration을 생성: Dropout(2가지)*Augmentation(7가지)*Weight decay(2가지)
  - ➢ No dropout / no stochastic depth or dropout with prob. 0.1 and stochastic depth with maximal layer dropping prob. 0.1
  - ➢ 7 data augmentation setups for (l, m, α)*: none(0, 0, 0), light1(2, 0, 0), light2(2, 10, 0.2), medium1(2, 15, 0.2), medium2(2, 15, 0.5), strong1(2, 20, 0.5), strong2(2, 20, 0.8)
  - ➢ * α is a Mixup parameter. And l, m are number of augmentation layers and magnitude respectively in RandAugment
  - ➢ Weight decay: 0.1 or 0.3

Experiments

❖ Experiment

- **Fig1**
  - ➢ Regularization 및 image augmentation을 적절히 사용하면 training data 규모를 키우는 것과 유사한 성능을 얻을 수 있음을 보임. 그러나 너무 작은 데이터셋에는 잘 적용되지 않음
- **Fig2**
  - ➢ Transfer learning이 from scratch로 학습하는 것보다 성능이 좋음

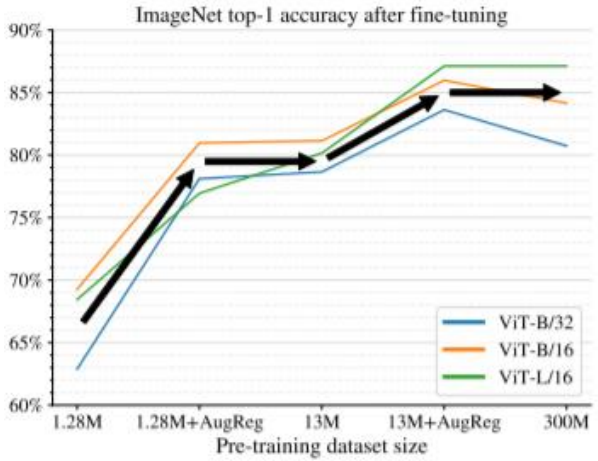


Fig. 1

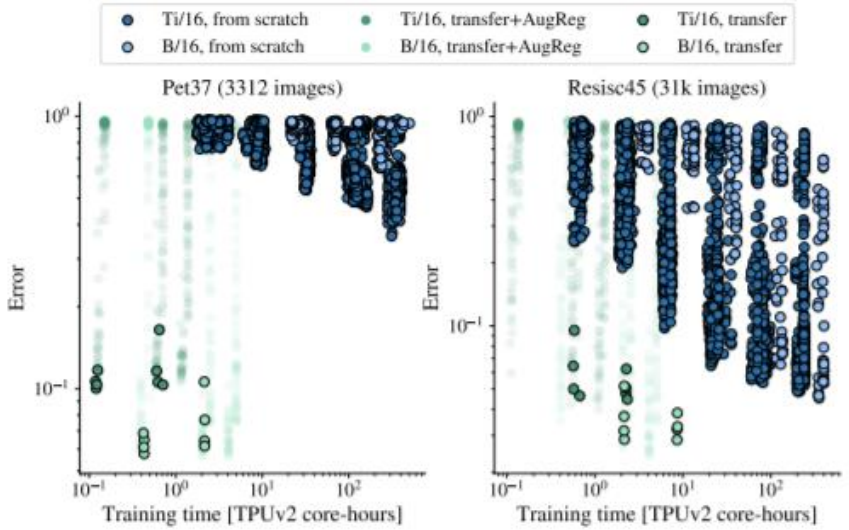


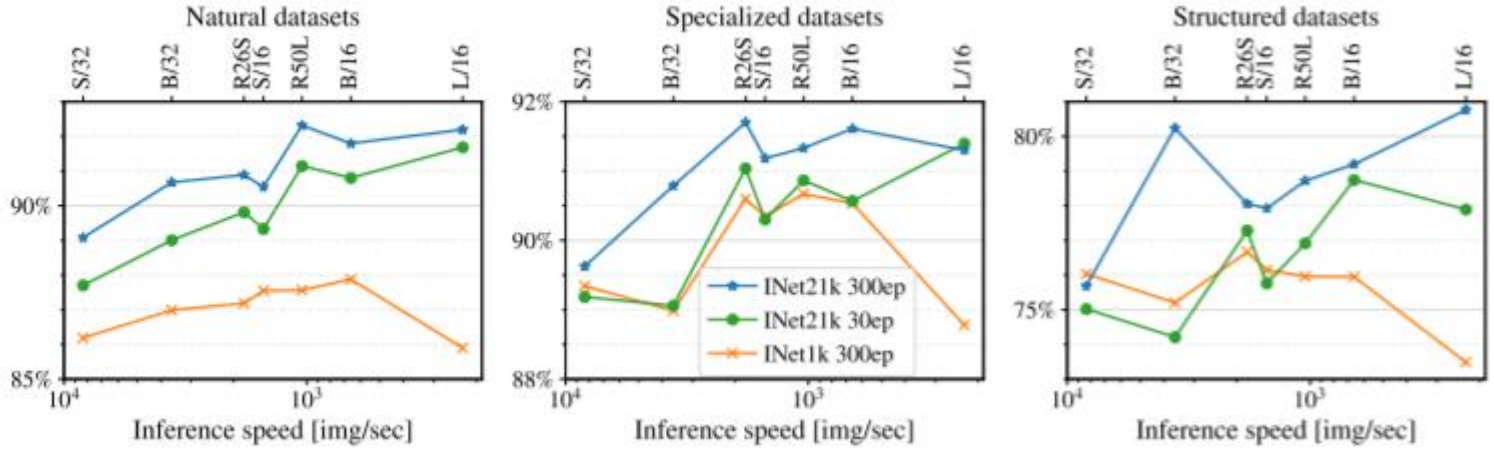
Fig. 2

DMQA

Experiments

❖ Experiment

- 앞의 장표에서 regularization과 augmentation을 적절히 사용하면 대규모 train set을 사용하는 모델과 유사한 성능을 낼 수 있다는 결과가 있었으나, transfer learning을 할 때에는 대규모 train set을 사용하는 모델이 더욱 좋은 성능을 냄
- 본 논문에서는 이 결과를 통해 "more data, more generic model"이라고 평가함

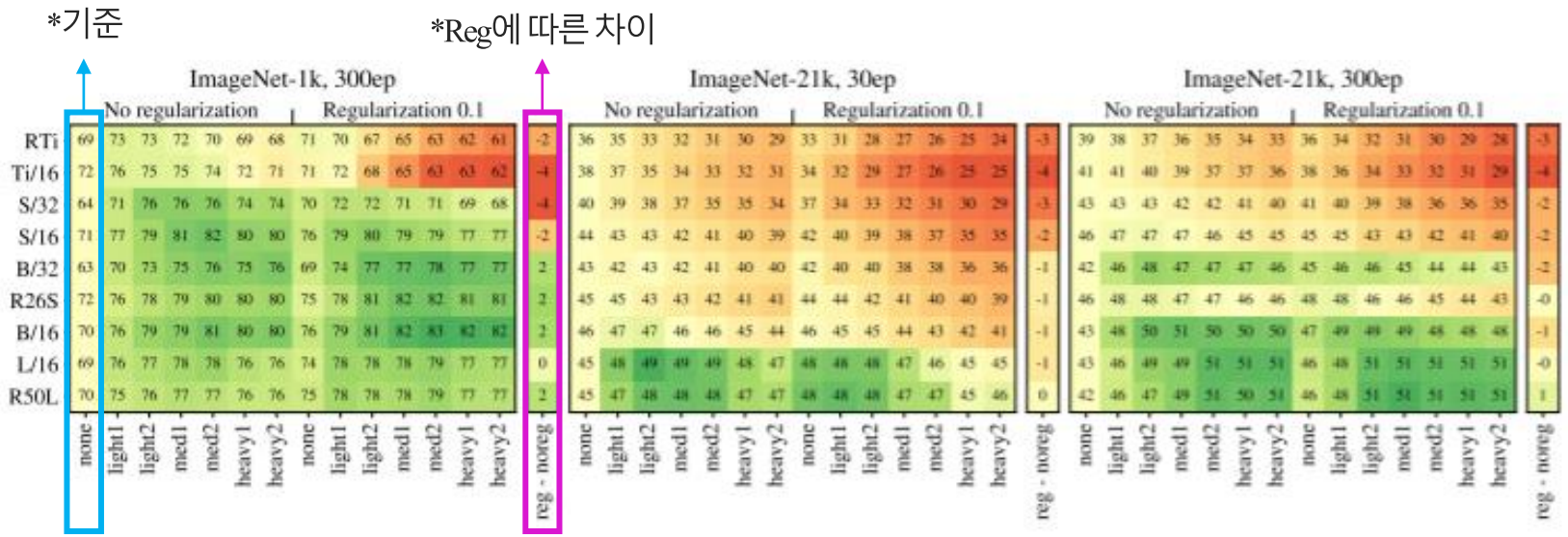


DMQA

Experiments

❖ Experiment

- Augmentation과 Regularization의 영향을 알아보고자 실험을 진행
- *기준보다 성능이 좋은 경우 녹색 계열, 성능이 나쁜 경우 붉은색 계열로 나타남
- *Regularization 여부에 따른 성능의 차를 의미함
- 상대적으로 데이터의 수가 적을 때 augmentation, regularization이 좋은 영향을 미치며, regularization에 비해 augmentation이 더욱 중요하게 영향을 미침

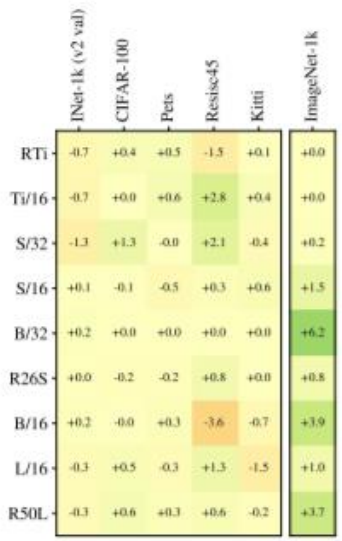


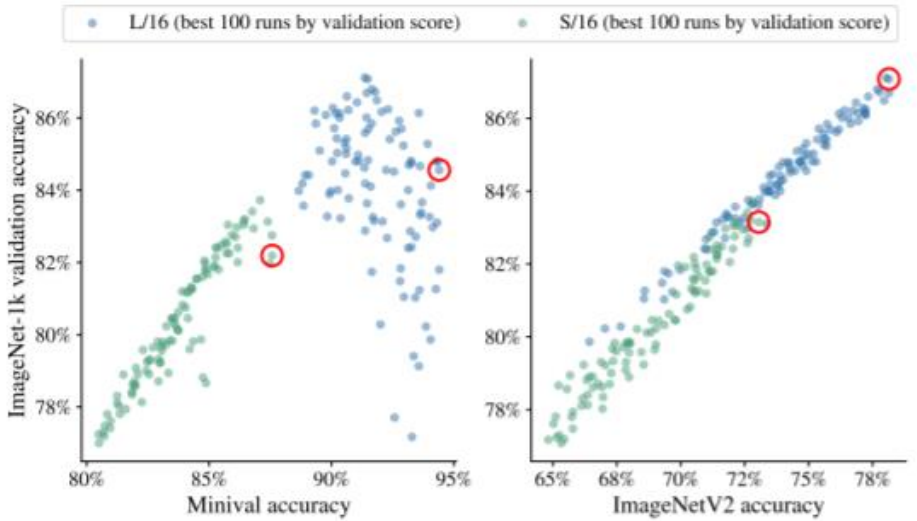
DMQA

Experiments

❖ Experiment

- 모델에 따라 적절한 regularization 및 augmentation이 달라짐
  - ➢ 따라서 "어떻게 application에 적합한 모델을 찾을 것인가?"에 대한 가이드가 필요
  - ➢ Way1) 가능한 모든 pre-trained model에 대해 실험을 진행하고 좋은 성능을 내는 모델 선정
  - ➢ Way2) Upstream의 validation accuracy가 가장 좋은 모델을 선정
- 실제 실험결과에 따르면 way2의 방식도 way1과 유사한 성능을 보임



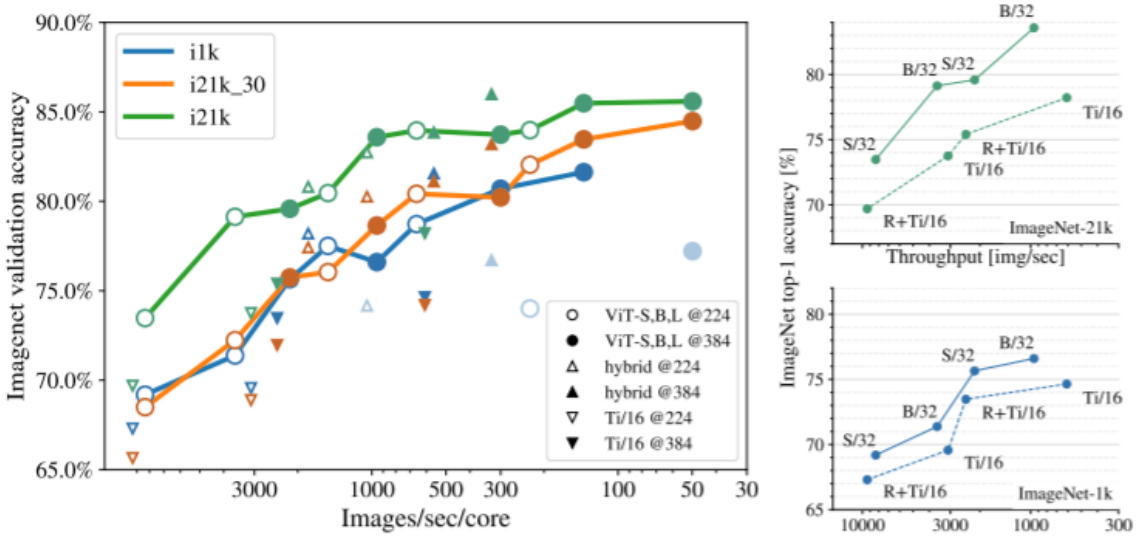


DMQA

Experiments

❖ Experiment

- Inference cost가 중요한 경우, 주로 tiny한 모델을 쓸 것으로 가정하고 실험을 진행
- 모델의 capacity를 줄이고 patch size를 작게 유지하는 것보다 모델의 capacity를 유지하고 patch size를 크게 하는 것이 더욱 좋은 성능을 냄
- 본 논문은 해당 결과를 통해 Parameter보다 sequence length가 speed와 capacity에 더욱 큰 영향을 미친다고 해석함



Conclusion

❖ Conclusion

- 본 논문은 vision Transformer pre-training시의 regularization, data augmentation, model size, and training data size간의 상호작용에 대해 연구함
- Pre-training시 사용한 데이터셋과 fine-tuning시 사용한 데이터셋 간의 관련이 적어도 transfer learning이 vision transformer를 학습시키기 위한 가장 좋은 방안임
- 또한, 유사한 성능을 내는 pre-trained model 중에서 training data가 많은 모델이 더 많은 data augmentation이 적용된 모델보다 더욱 합리적임

Reference

1. Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., & Beyer, L. (2021). How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. *arXiv preprint arXiv:2106.10270*.


Thank You