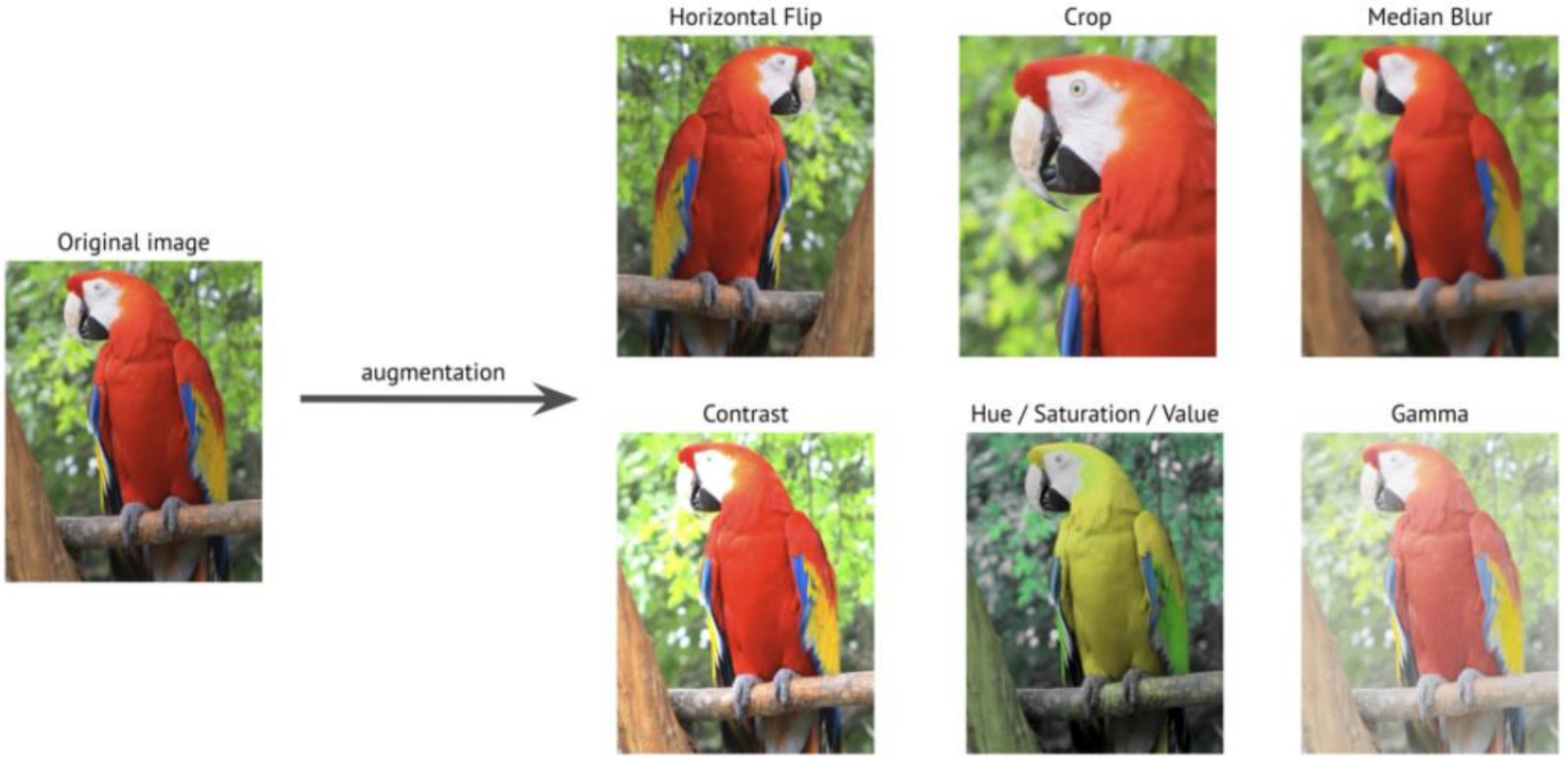
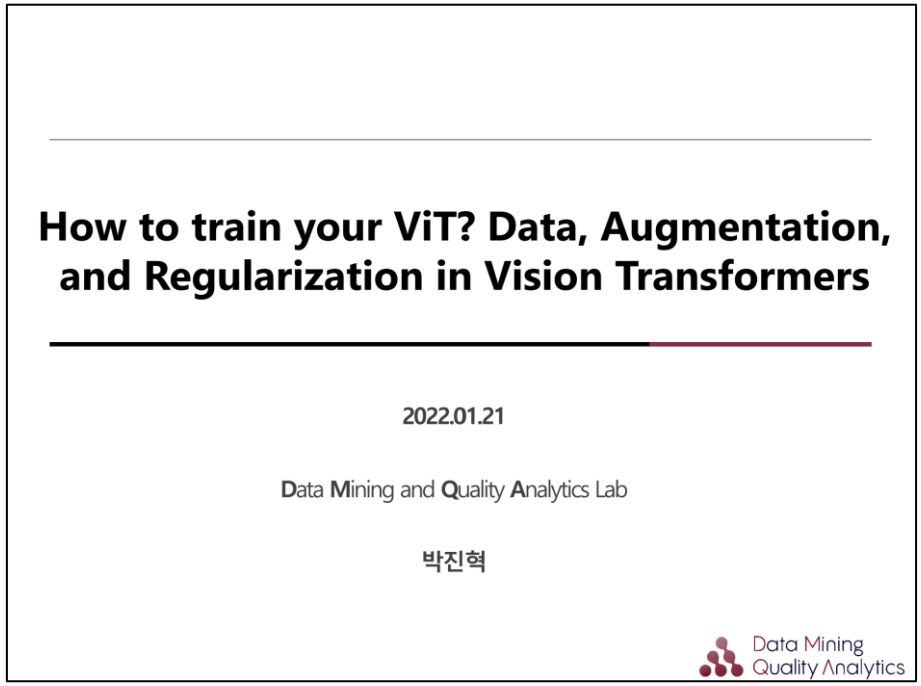
How to train your ViT?
: Mixup Augmentation

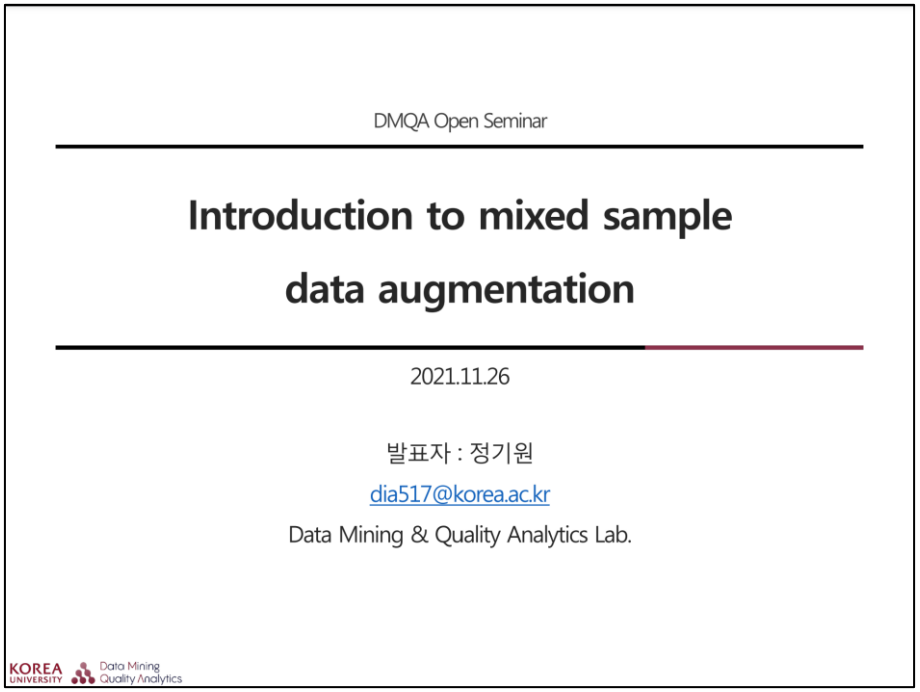
2023.09.12.

일반적으로 떠올릴 수 있는 이미지 데이터 증강 방법



출처


How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers
2022.01.21
Data Mining and Quality Analytics Lab
박진혁
Data Mining
Quality Analytics

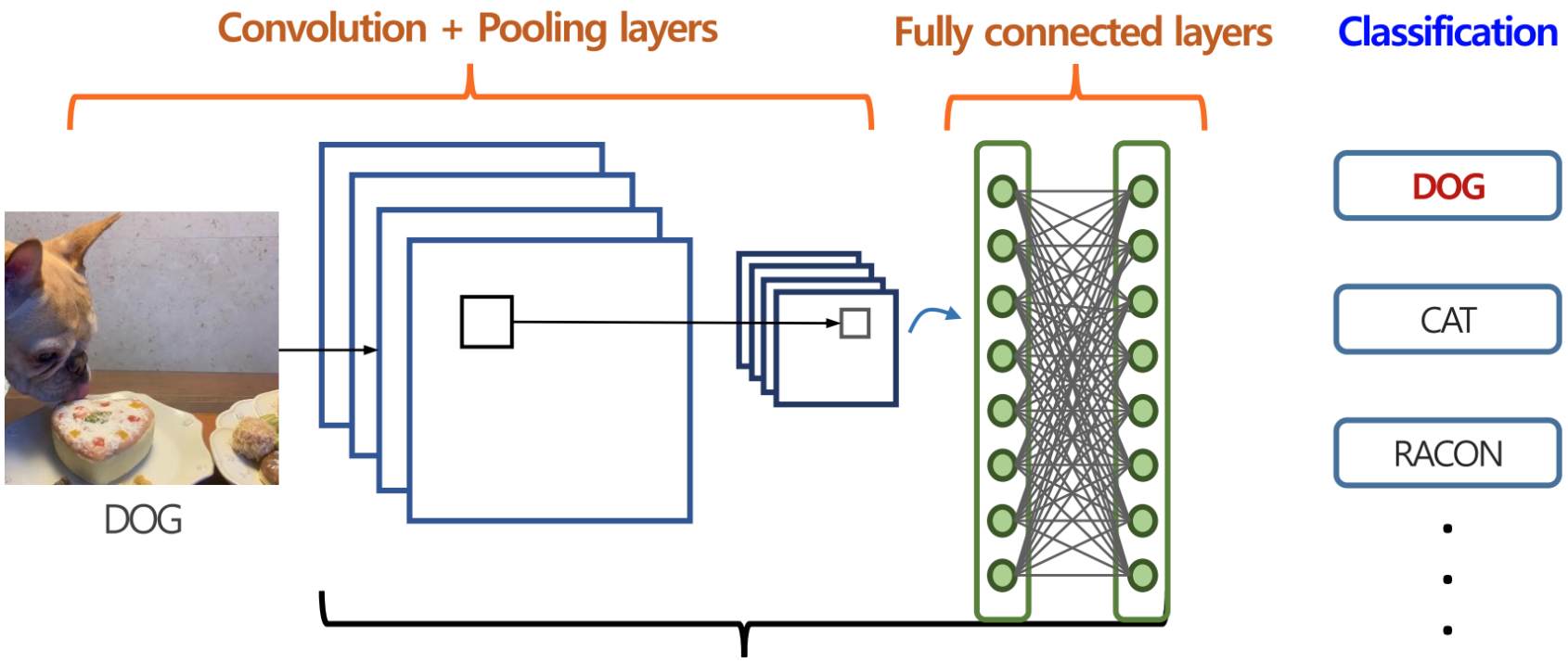

DMQA Open Seminar
Introduction to mixed sample data augmentation
2021.11.26
발표자 : 정기원
dia517@korea.ac.kr
Data Mining & Quality Analytics Lab.
KOREA UNIVERSITY
Data Mining
Quality Analytics

목차

1. Introduction

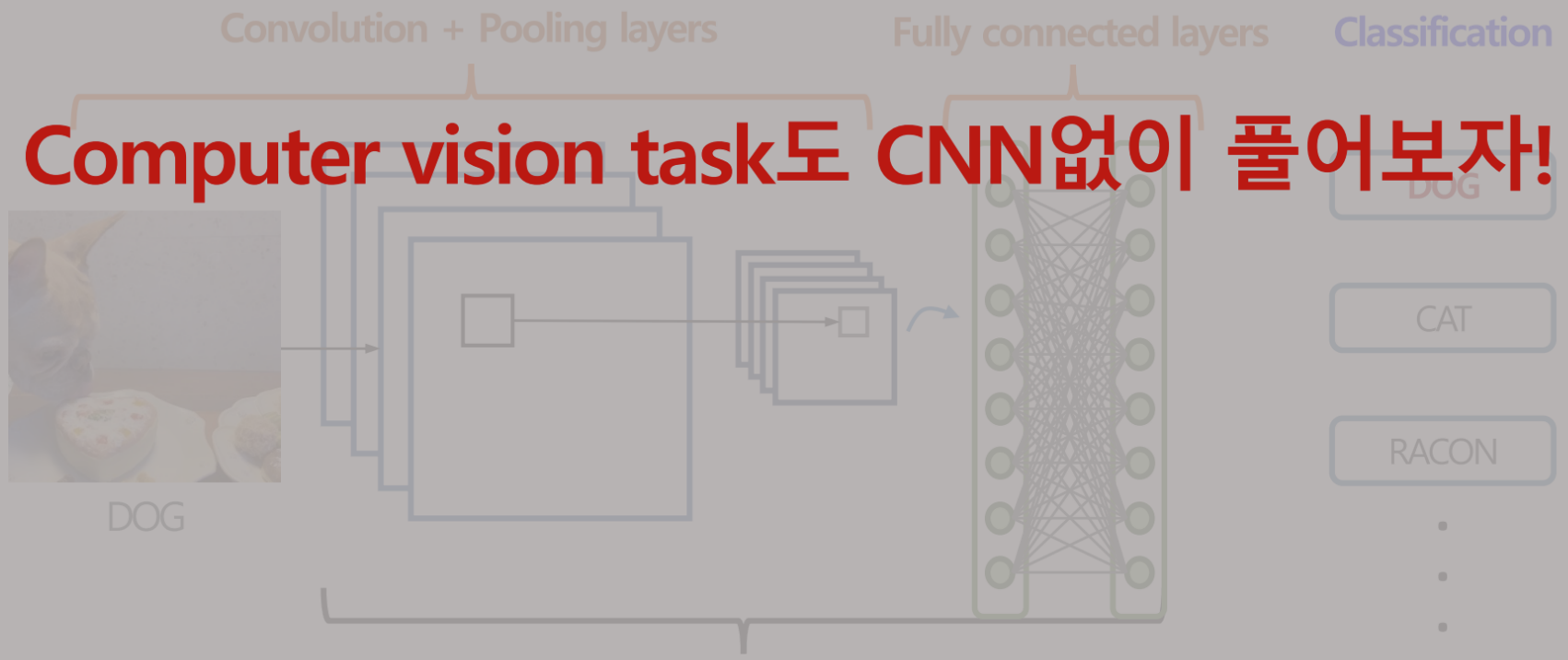
❖ From CNN to ViT

- ➢ Convolutional Neural Network
  - ➢ Computer vision분야에서 가장 많이 사용되는 architecture
  - ➢ 이미지를 입력 받아 이미지의 공간정보를 유지한 채 학습함



1. Introduction

- From CNN to ViT
  - Convolutional Neural Network
    - Computer vision분야에서 가장 많이 사용되는 architecture
    - 이미지를 입력 받아 이미지의 공간정보를 유지한 채 학습함

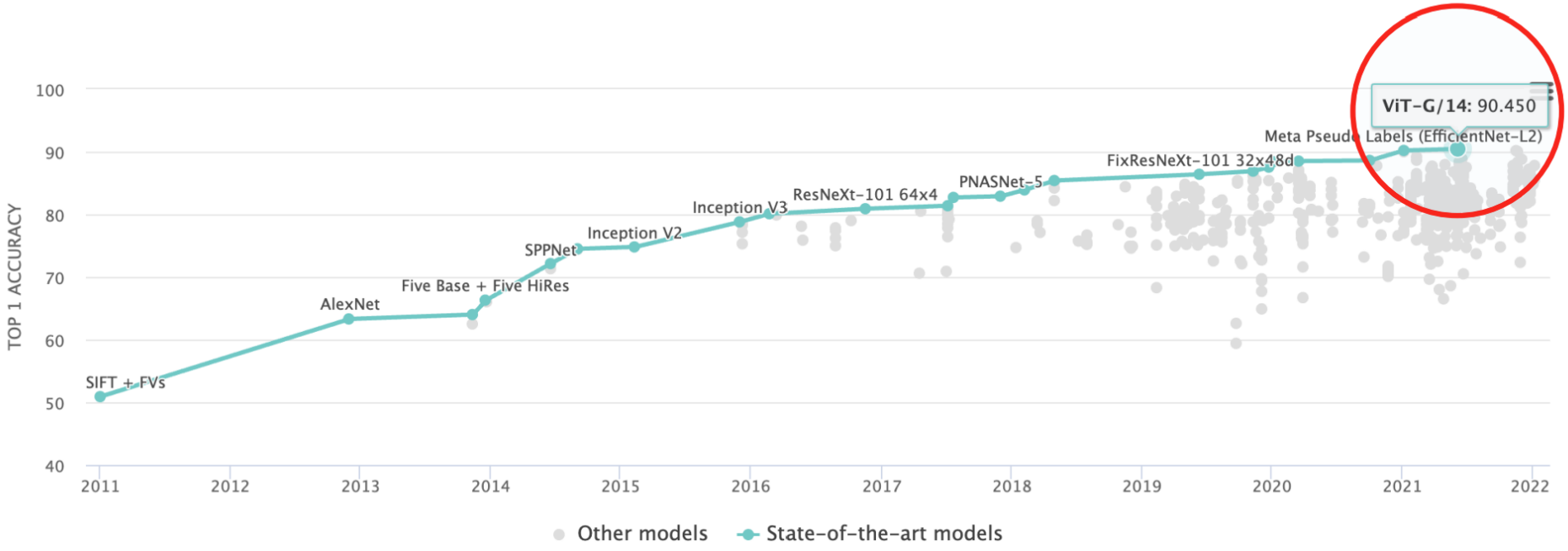


Computer vision task도 CNN없이 풀어보자!

1. Introduction

❖ From CNN to ViT

- ➢ Computer vision분야에서 합성곱 신경망이 아닌 Transformer를 적용
- ➢ Transformer architecture가 computer vision분야에서 수 많은 SOTA를 달성
- ➢ Computer vision분야에 NLP architecture의 적용 가능성을 보여준 논문



2. What is ViT?

- Vision Transformer(ViT)
  - 2023년 9월 12일 기준 20,997회 인용
  - Google Research에서 발표
  - Transformer architecture를 활용하여 image classification을 수행

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]

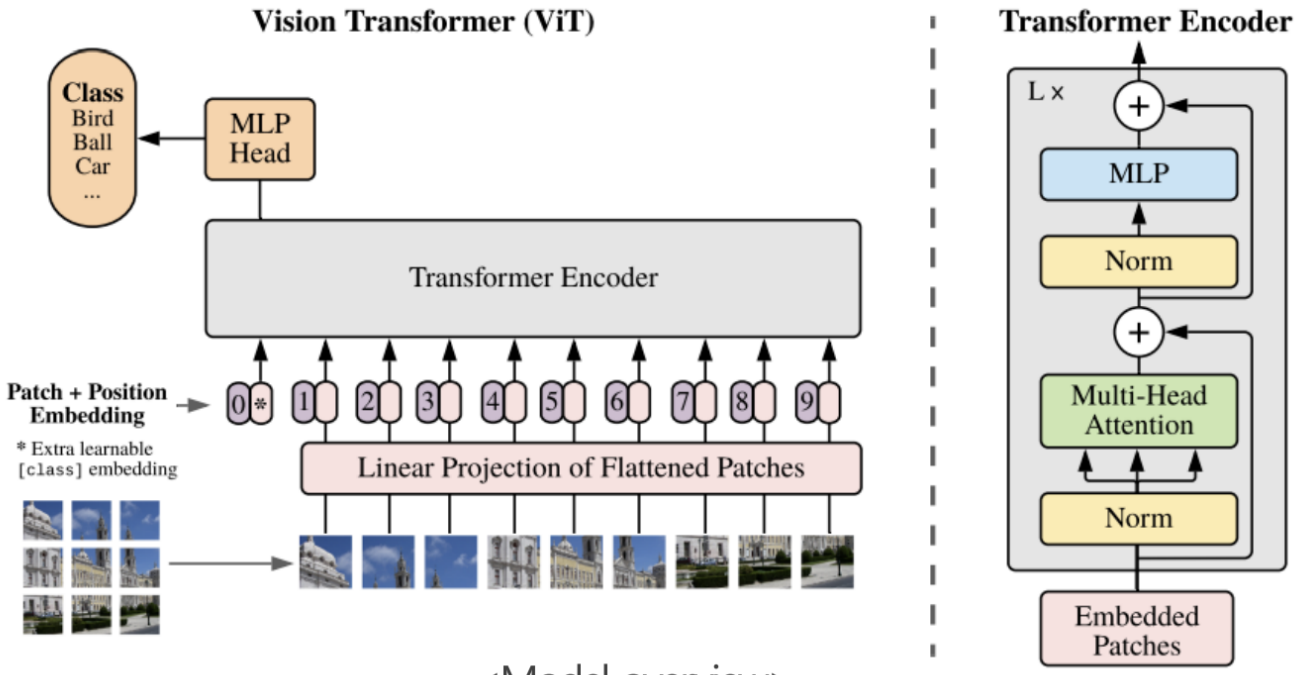
[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

2. What is ViT?

❖ Vision Transformer(ViT)

- ➢ 입력 이미지
- ➢ 모델 아키텍처
- ➢ 대용량 데이터의 사전학습

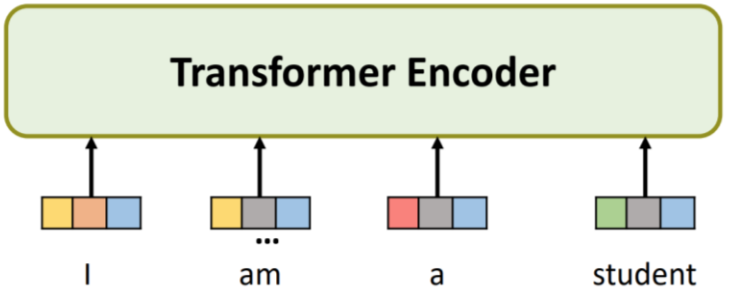


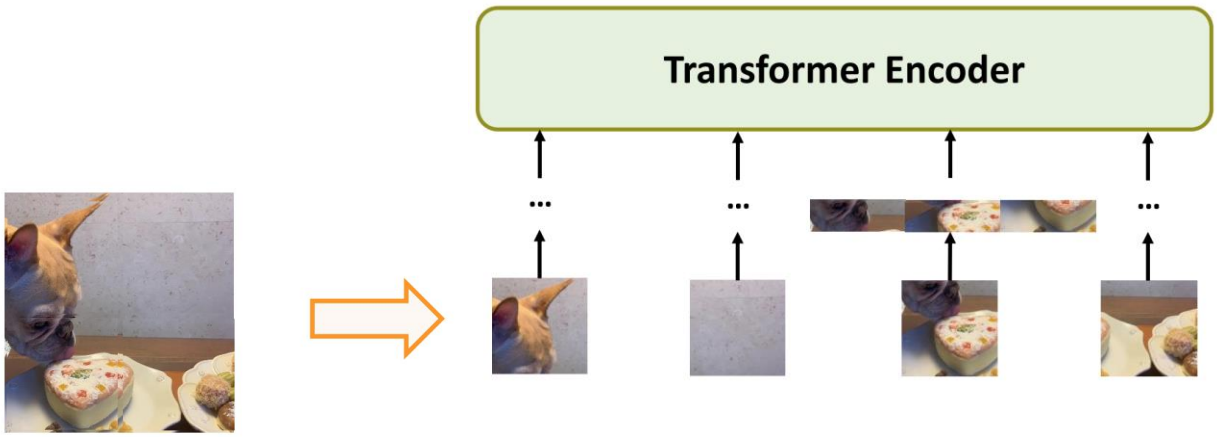
<Model overview>

2. What is ViT?

❖ Vision Transformer(ViT)

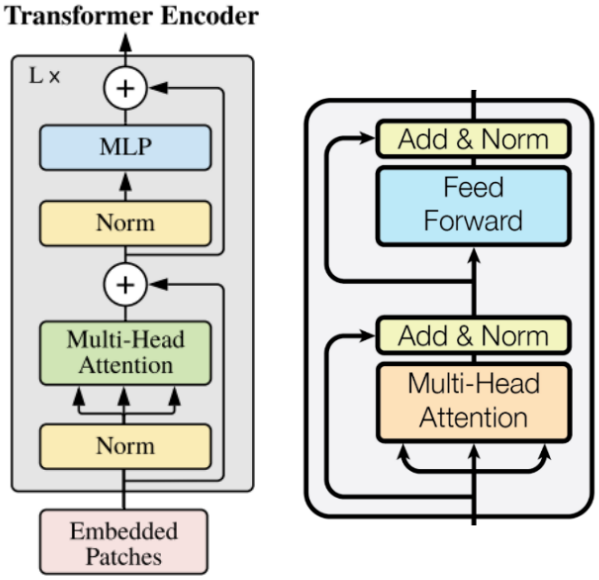
- ➢ 입력 이미지
  - ➢ 기존 Transformer에서는 token embedding을 입력
  - ➢ 이미지를 patch형태로 나눠서 입력
  - ➢ 모든 patch를 벡터와 flatten하게 생성한 뒤 입력





2. What is ViT?

- Vision Transformer(ViT)
  - 모델 아키텍처
    - 기존 Transformer의 encoder부분과 거의 유사함
    - Transformer와 차이점: Norm의 위치, GELU사용
  - 대용량 데이터의 사전학습
    - 대용량 데이터셋으로 사전학습을 한 뒤 downstream task로 fine-tuning을 진행



<ViT> <Transformer>

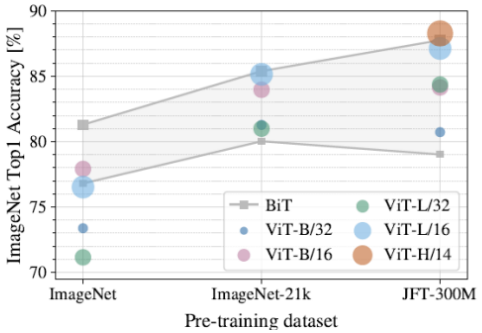
私は学生です。

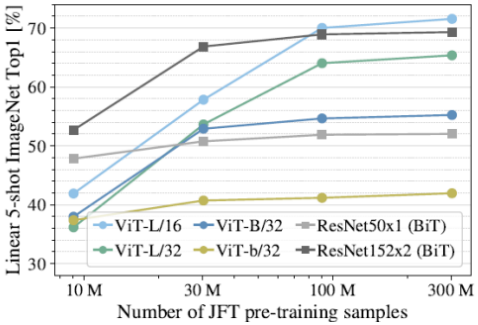
Je suis étudiant.

2. What is ViT?

- Vision Transformer(ViT)
  - Result
    - JFT로 사전학습을 진행한 뒤, 각 데이터 셋에 Transfer Learning진행
    - CNN기반의 SoTA모델과 비교해서 ViT가 SoTA를 갱신함
    - 사전학습을 진행할 경우 데이터 셋이 클수록 ViT성능이 향상
      - Convolutional Inductive Bias가 존재하지 않기 때문에 많은 데이터가 필요

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k





3. How to train your ViT?

❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

- 2023년 9월 12일 기준 309회 인용
- Google Research에서 발표
- ViT model을 효율적으로 학습시키는 방법론에 대한 비교분석 실험 진행

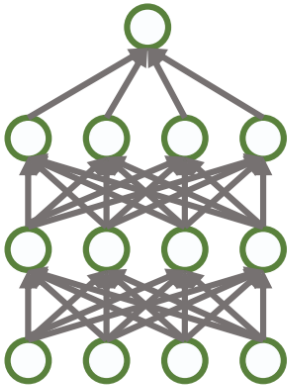
How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers

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

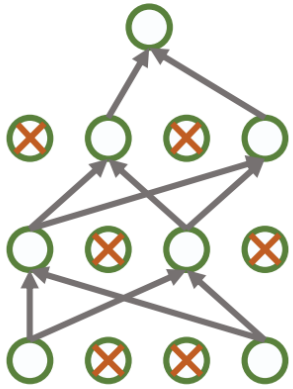
3. How to train your ViT?

❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

- **Regularization**
  - **Dropout to intermediate activations of ViT**
  - Stochastic depth regularization technique
- Data augmentations
  - Mixup
  - RandAugment



<Standard Neural Net>

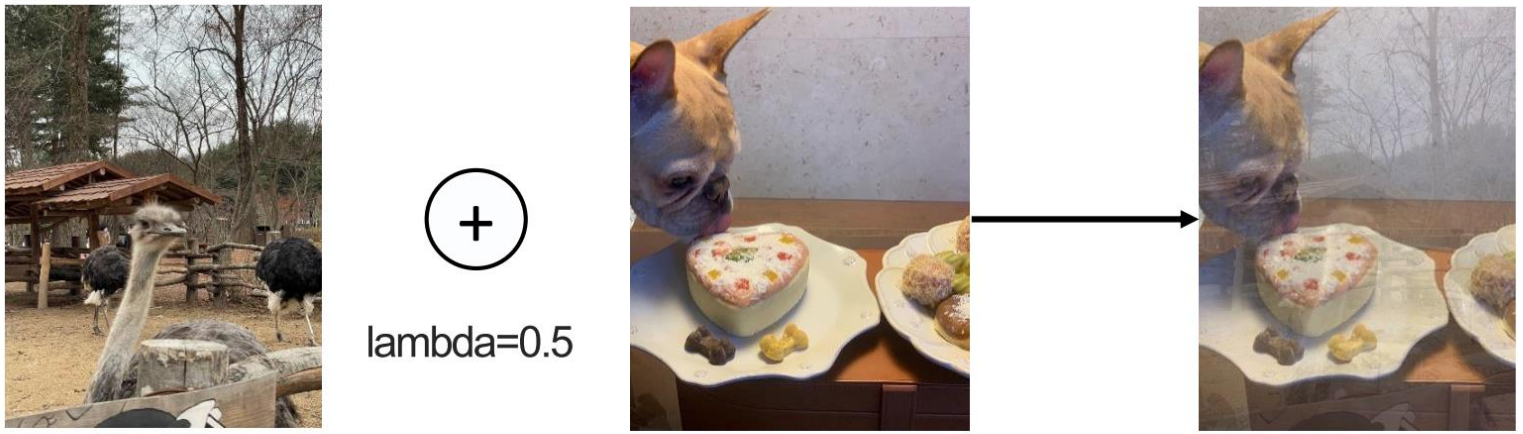


<After applying dropout>

3. How to train your ViT?

❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

- Regularization
  - Dropout to intermediate activations of ViT
  - Stochastic depth regularization technique
- **Data augmentations**
  - **Mixup**
  - RandAugment



3. How to train your ViT?

❖ 딥러닝 모델 과적합(Overfitting) 문제

- 데이터 수에 비해 많은 파라미터 수를 갖는 딥러닝 모델은 과적합 문제에 빠지기 쉬움
- 과적합된 모델의 경우, 학습에서 보지 못한 새로운 데이터에 대한 예측 성능이 상당히 낮음
- 이러한 문제를 해결하기 위해 **규제화(regularization)** / **데이터 증강(data augmentation)** 기법을 적용

regularization method

- ✓ weight decay
- ✓ dropout
- ✓ early stopping
- ✓ etc..

data augmentation method

- ✓ mix-up
- ✓ autoaugment
- ✓ flipping & cropping
- ✓ etc..

Mix-up Algorithm

❖ Mix-up 알고리즘

- 2023.09.12 기준 7,576회 인용
- 다양한 연구 분야에서 활발하게 사용되고 있는 데이터 증강기법

mixup: BEYOND EMPIRICAL RISK MINIMIZATION

Hongyi Zhang
MIT

Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz*
FAIR

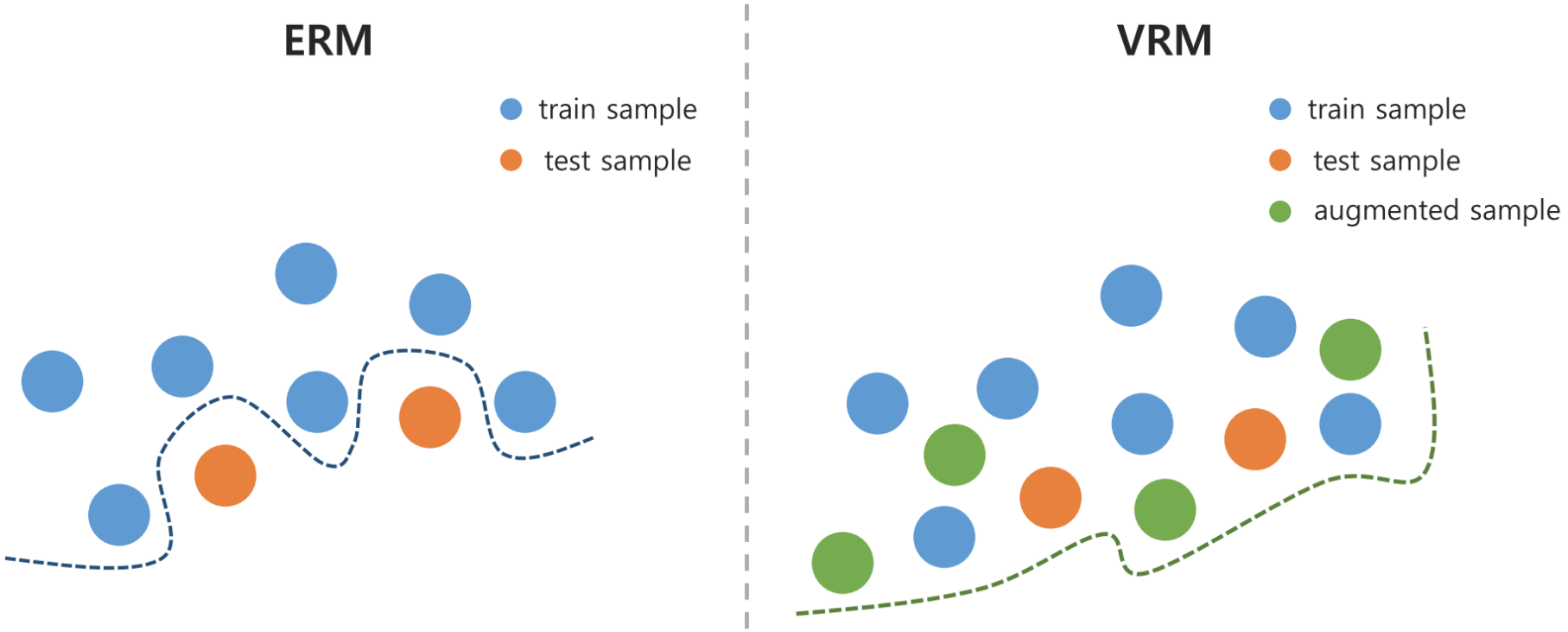
ABSTRACT

Large deep neural networks are powerful, but exhibit undesirable behaviors such as memorization and sensitivity to adversarial examples. In this work, we propose *mixup*, a simple learning principle to alleviate these issues. In essence, *mixup* trains a neural network on convex combinations of pairs of examples and their labels. By doing so, *mixup* regularizes the neural network to favor simple linear behavior in-between training examples. Our experiments on the ImageNet-2012, CIFAR-10, CIFAR-100, Google commands and UCI datasets show that *mixup* improves the generalization of state-of-the-art neural network architectures. We also find that *mixup* reduces the memorization of corrupt labels, increases the robustness to adversarial examples, and stabilizes the training of generative adversarial networks.

Mix-up Algorithm

❖ ERM(Empirical Risk Minimization) 과 VRM(Vicinal Risk Minimization)

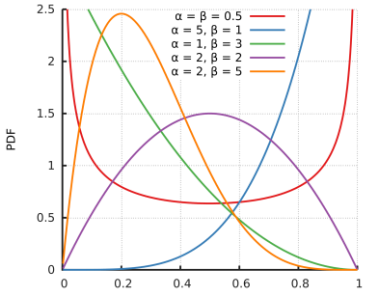
- 신경망 모델은 주어진 훈련 데이터셋에 대한 에러를 최소화하는 방향으로 학습을 진행
- ERM 기반 학습은 훈련 데이터에 과적합 되는 위험이 존재
- VRM 기반 학습은 주어진 훈련 데이터의 근방 분포를 활용하여 신경망 학습의 도움을 주고자 함



Mix-up Algorithm

❖ Mix-up 알고리즘 동작 원리

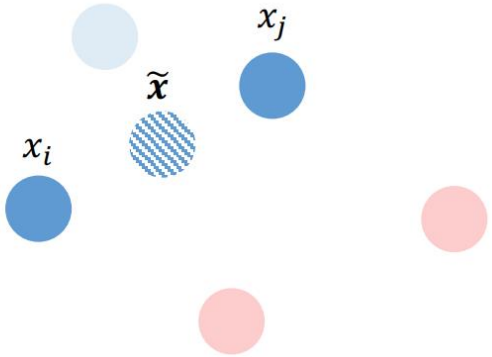
- 두 데이터 샘플로부터 선형 보간법을 통해 새로운 샘플을 생성하는 데이터 증강 기법
- 간단한 동작 원리에도 좋은 일반화 성능을 보장하여 다양한 딥러닝 연구 분야에서 사용되고 있는 기법

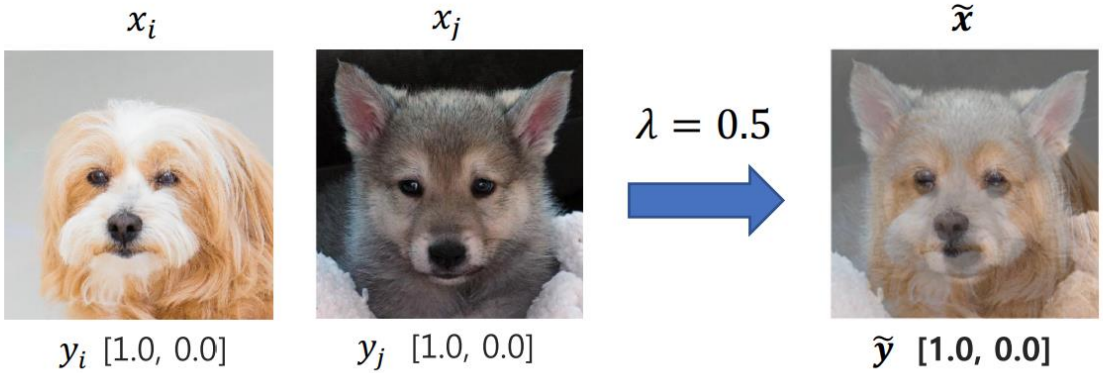


$\lambda \sim Beta(\alpha, \alpha)$ ※ 베타 분포

Mix-up formulation

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \quad \text{where } x_i, x_j \text{ are raw input vectors}$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j, \quad \text{where } y_i, y_j \text{ are one-hot label encodings}$$





Mix-up Algorithm

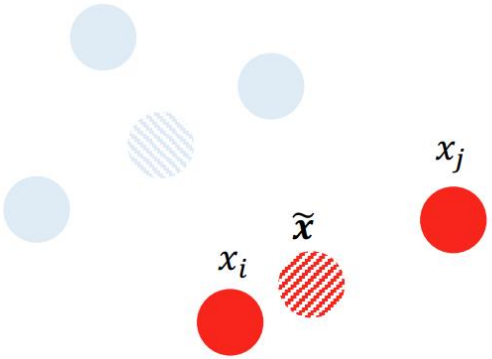
❖ Mix-up 알고리즘 동작 원리

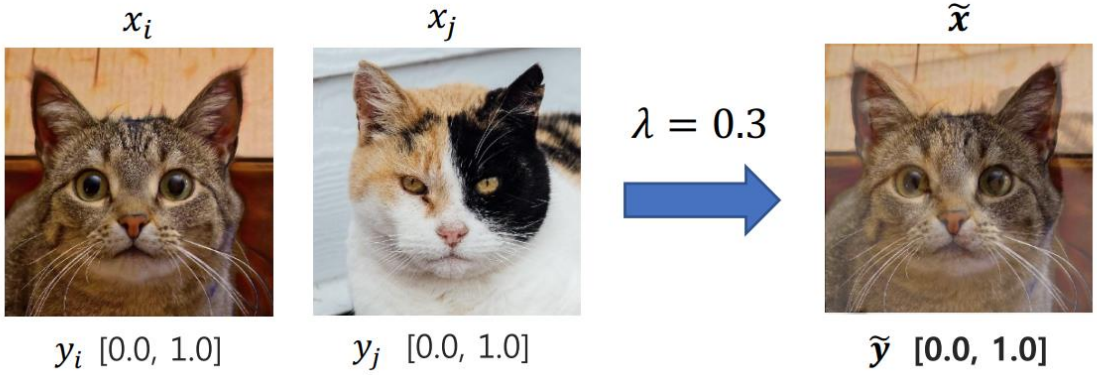
- 두 데이터 샘플로부터 선형 보간법을 통해 새로운 샘플을 생성하는 데이터 증강 기법
- 간단한 동작 원리에도 좋은 일반화 성능을 보장하여 다양한 딥러닝 연구 분야에서 사용되고 있는 기법

Mix-up formulation

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \quad \text{where } x_i, x_j \text{ are raw input vectors}$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j, \quad \text{where } y_i, y_j \text{ are one-hot label encodings}$$

$\lambda \sim Beta(\alpha, \alpha)$





Mix-up Algorithm

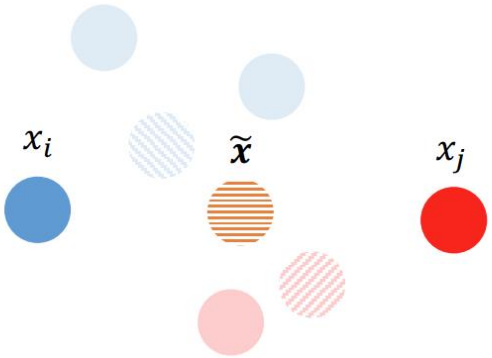
❖ Mix-up 알고리즘 동작 원리

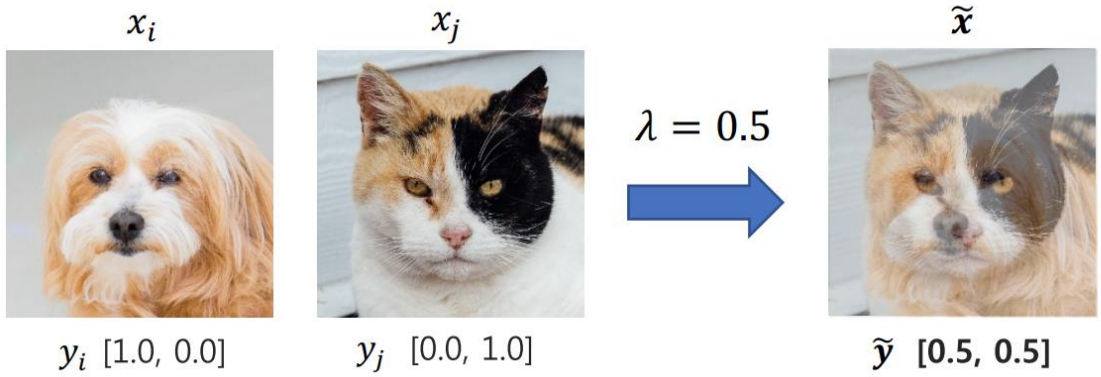
- 두 데이터 샘플로부터 선형 보간법을 통해 새로운 샘플을 생성하는 데이터 증강 기법
- 간단한 동작 원리에도 좋은 일반화 성능을 보장하여 다양한 딥러닝 연구 분야에서 사용되고 있는 기법

Mix-up formulation

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \quad \text{where } x_i, x_j \text{ are raw input vectors}$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j, \quad \text{where } y_i, y_j \text{ are one-hot label encodings}$$

$\lambda \sim Beta(\alpha, \alpha)$



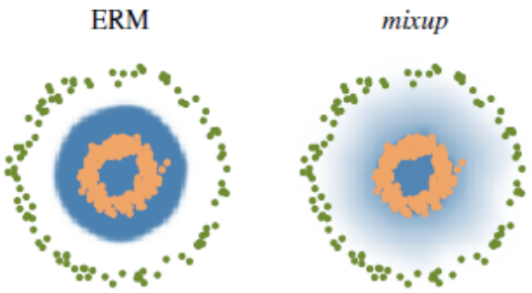


Mix-up Algorithm

❖ Mix-up 알고리즘 실험 결과

- ERM 학습 모델과 비교해 mixup을 적용한 모델의 경우, 부드러운 결정 경계선(Decision Boundary)를 생성
- CIFAR 이미지 데이터셋에 대해서 mixup을 적용했을 때, ERM 학습 모델보다 좋은 예측 성능을 보임

Decision Boundary Visualization

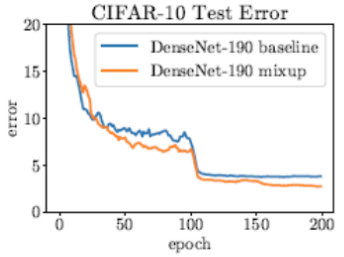


(b) Effect of *mixup* ($\alpha = 1$) on a toy problem. Green: Class 0. Orange: Class 1. Blue shading indicates $p(y = 1|x)$.

Errors in CIFAR Datasets

Dataset	Model	ERM	*mixup*
CIFAR-10	PreAct ResNet-18	5.6	**4.2**
	WideResNet-28-10	3.8	**2.7**
	DenseNet-BC-190	3.7	**2.7**
CIFAR-100	PreAct ResNet-18	25.6	**21.1**
	WideResNet-28-10	19.4	**17.5**
	DenseNet-BC-190	19.0	**16.8**

(a) Test errors for the CIFAR experiments.



(b) Test error evolution for the best ERM and *mixup* models.

Mix-up Algorithm

❖ Mix-up 알고리즘 실험 결과

- 이미지 데이터가 아닌 Speech / Tabular 데이터에 mixup을 적용해도 예측 성능 향상을 보여줌
- mixup을 통해 불안정한 학습을 하는 GAN(Generative Adversarial Networks)이 안정적으로 학습하도록 도와줌

Errors in Speech/Tabular Dataset

Model	Method	Validation set	Test set
LeNet	ERM	**9.8**	**10.3**
	mixup ($\alpha = 0.1$)	10.1	10.8
	mixup ($\alpha = 0.2$)	10.2	11.3
VGG-11	ERM	5.0	4.6
	mixup ($\alpha = 0.1$)	4.0	3.8
	mixup ($\alpha = 0.2$)	**3.9**	**3.4**

Figure 4: Classification errors of ERM and *mixup* on the Google commands dataset.

Dataset	ERM	*mixup*
Abalone	74.0	73.6
Arcene	57.6	**48.0**
Arrhythmia	56.6	**46.3**

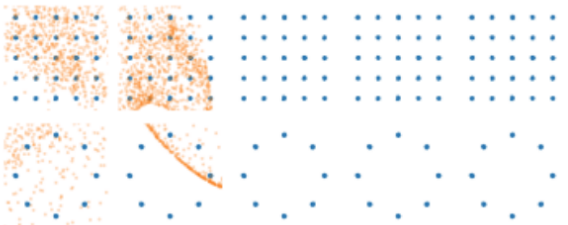
Dataset	ERM	*mixup*
Htru2	2.0	2.0
Iris	21.3	**17.3**
Phishing	16.3	15.2

Table 4: ERM and *mixup* classification errors on the UCI datasets.

Effect of mixup on GAN training

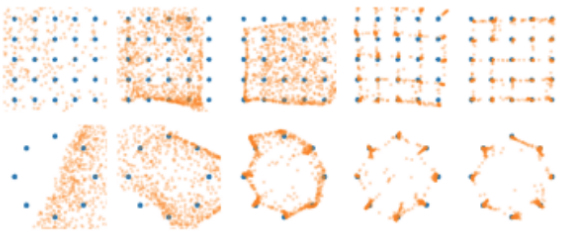
$$\max_g \min_d \mathop{\mathbb{E}}_{x,z} \ell(d(x), 1) + \ell(d(g(z)), 0)$$

ERM GAN



$$\max_g \min_d \mathop{\mathbb{E}}_{x,z,\lambda} \ell(d(\lambda x + (1-\lambda)g(z)), \lambda)$$

mixup GAN ($\alpha = 0.2$)



Advanced Mix-up Algorithm

❖ CutMix 알고리즘

- 특정 샘플의 일부 영역을 다른 샘플의 일부 영역으로 대체하여 새로운 샘플을 생성
  - ➢ 전체 영역 중 선택된 일부 영역의 비율이 λ 값으로 사용
- CutMix는 Classification, Localization, Detection task에서 좋은 성능을 보여준다고 제안

x_i

1	**1**	**1**	0
1	**1**	**1**	0
1	**1**	**1**	0
0	0	0	0

M

y_i [1.0, 0.0]

x_j

0	0	0	**1**
0	0	0	**1**
0	0	0	**1**
1	**1**	**1**	**1**

$(1 - M)$

y_j [0.0, 1.0]

$$\tilde{x} = M \odot x_i + (1 - M) \odot x_j$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \quad , \lambda = \frac{\text{선택된 일부 영역}}{\text{전체 영역}} = \frac{9}{16}$$

$\tilde{x}$

$\tilde{y}$ [0.56, 0.44]

Advanced Mix-up Algorithm

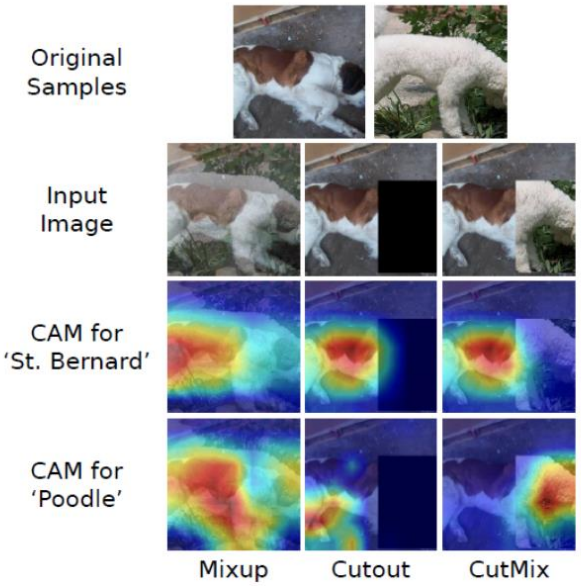
❖ CutMix 알고리즘 실험

- 다른 방법론들과 비교해 Classification, Localization, Detection task에서 좋은 성능을 보여줌
- CutMix를 통해 생성된 샘플로부터 각 클래스의 CAM이 적절하게 출력됨을 시각적으로 확인

results on each task

	ResNet-50	Mixup [47]	Cutout [3]	CutMix
Image				
Label	Dog 1.0	Dog 0.5 Cat 0.5	Dog 1.0	Dog 0.6 Cat 0.4
ImageNet Cls (%)	76.3 (+0.0)	77.4 (+1.1)	77.1 (+0.8)	**78.6** **(+2.3)**
ImageNet Loc (%)	46.3 (+0.0)	45.8 (-0.5)	46.7 (+0.4)	**47.3** **(+1.0)**
Pascal VOC Det (mAP)	75.6 (+0.0)	73.9 (-1.7)	75.1 (-0.5)	**76.7** **(+1.1)**

CAM visualization



Advanced Mix-up Algorithm

❖ CutMix 알고리즘 한계점

- 다른 샘플의 일부 영역으로 대체하는 과정에서 불필요한 정보를 포함하는 영역을 사용할 수 있음
- 이러한 한계점을 개선하기 위한 연구가 최근 활발하게 진행되고 있음

x_i

1	1	0	0
1	1	0	0
0	0	0	0
0	0	0	0

y_i [1.0, 0.0]

x_j

0	0	1	1
0	0	1	1
1	1	1	1
1	1	1	1

y_j [0.0, 1.0]

$\tilde{x}$

$\tilde{y}$ [0.25, 0.75]

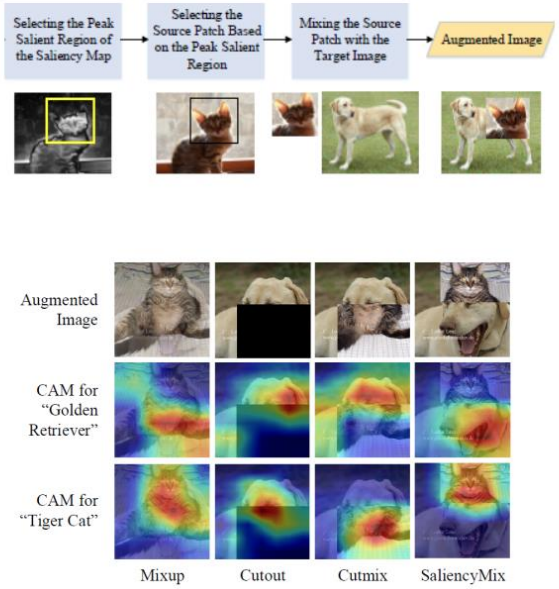
치타?

Advanced Mix-up Algorithm

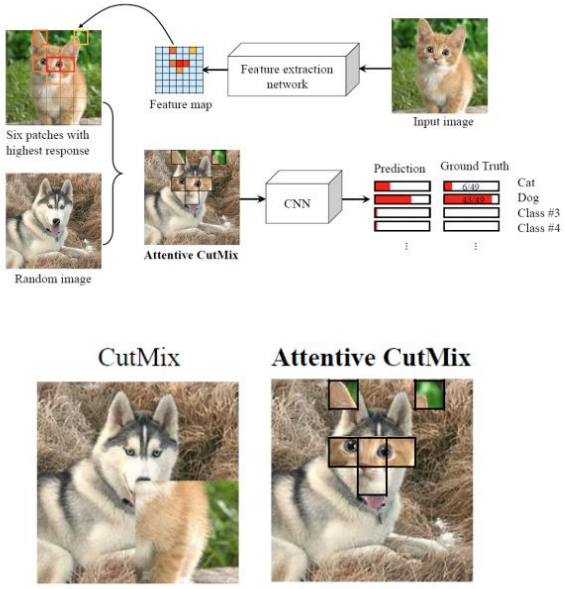
❖ 중요 영역을 고려한 개선된 Mix-up 알고리즘

- 최근 다양한 방법론에서 샘플 내 중요 영역 정보를 이용하여 두 샘플을 섞는 방안을 제안
- 각 방법론 마다 중요 영역을 찾는 방법과 섞는 방식이 다름

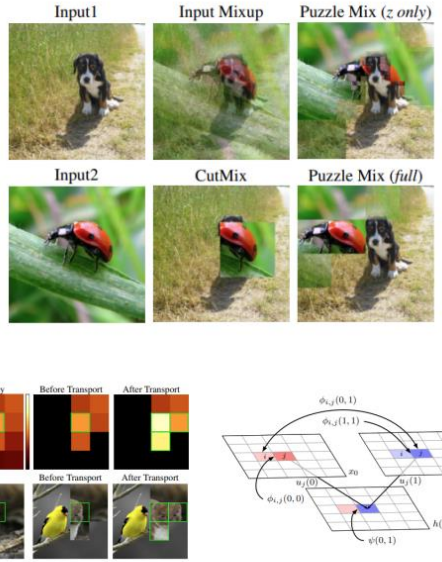
Saliency Mix



Attentive Mix



Puzzle Mix



Conclusion

❖ 결론

- 모델의 일반화 성능 향상을 위한 데이터 증강기법 중 Mix-up 알고리즘에 대해 설명
- 다양한 실험 조건에서 좋은 성능을 보임을 확인
- 이후 Mix-up 알고리즘을 적용 및 개선한 여러 연구들이 진행되었음
  - ➢ 모델링 고도화 / 적용 데이터셋 확장 / 다른 연구 분야에 적용
  - ➢ Manifold mixup / CutMix 등 여러 개선된 Mix-up 알고리즘이 연구됨
- 최근에는 중요 영역 정보를 이용한 개선된 Mix-up 연구가 진행되고 있음
- 본 세미나를 통해 Mix-up에 대한 전반적인 이해를 돕고, 다양한 연구 분야에 활용되기를 기대