CVPR 2022 Challenge

- Align before Fuse: Vision and Language Representation Learning with Momentum Distillation
- Solving imageNet: a Unified Scheme for Training any Backbone to Top Results

22.06.01

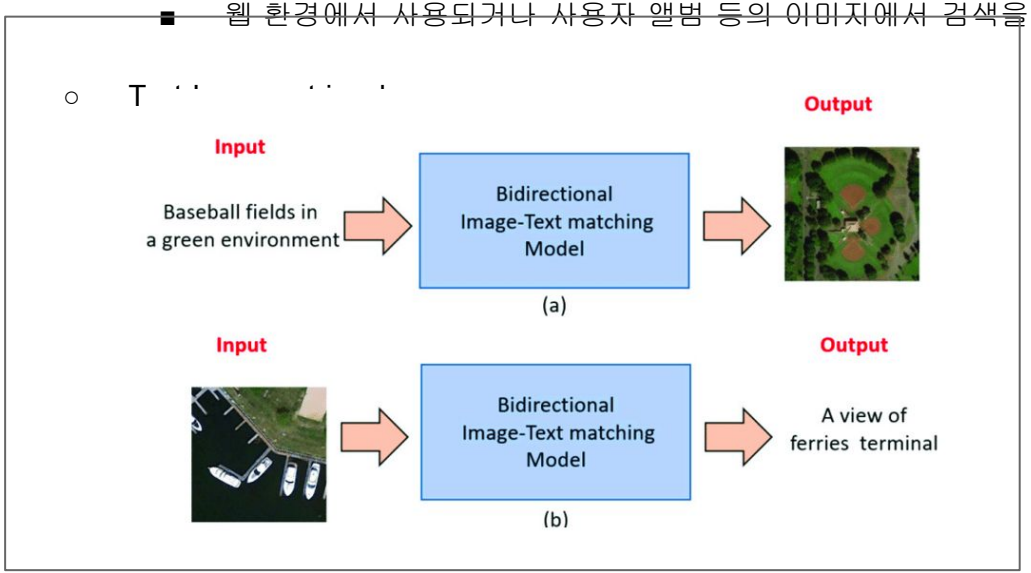
발표자 이지현

CVPR 2022 Challenge

- 참여 챌린지 정보
  - 링크 : https://retailvisionworkshop.github.io/#challenge
  - AliProducts Challenge: Large-scale Cross-Modal Product Retrieval
  - Timeline
    - March 25, 2022 : Registration opens
    - March 31, 2022 : Entire training data released
    - May 31, 2022 : Registration deadline and Test data released
    - **June 9, 2022 : Submission ends**
    - **June 10, 2022 : Deadline for submitting technical reports**
    - June 12, 2022 : Challenge award

CVPR 2022 Challenge

- Retrieval
  - Instance-level image retrieval
    - 간단하게 생각하면, 이미지를 넣어 유사한 이미지를 반환해주는 검색
    - 좀 더 구체적으로는 이미지 내 오브젝트와 가장 유사한 오브젝트를 찾아서 반환해주는 시스템
    - 단순하게 동일한 이미지를 찾는 문제 수준에서 semantic 정보까지 활용하여 유사한 이미지를 찾는 것까지 고려할 수 있음
    - 보통 매우 큰 이미지 집합 내에서 검색
    - 웹 환경에서 사용되거나 사용자 앨범 등의 이미지에서 검색을 제공하는 등 여러 응용 예제가 있음
  - T



Challenge

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare
Shafiq Joty, Caiming Xiong, Steven C.H. Hoi
Salesforce Research
{junnan.li,rselvaraju,akhilesh.gotmare,sjoty,shoi}@salesforce.com

https://arxiv.org/pdf/2107.07651.pdf
https://github.com/salesforce/ALBEF

Transformer 기반 Visual - Text multimodal learning

- 기존 transfomer 기반의 multimodal encoder는 model이 visual, text tokens을 함께 학습하도록 한다.
  - visual word tokens이 unligned 되어있어 multimodal encoder 가 image-text interaction을 학습하는 것이 어렵다
- 본 논문에서는 image-text representation을 합치기 전 align하기 위한 contrastive loss를 소개. ⇒ 그래서 본 논문의 제목도 Align before Fuse.
  - BBOX annotation, high-resolution image가 필요없다는 장점 (기존 VLP 모델 region-based image features를 추출하는데, 사전학습된 object detector module을 사용하곤 함)
- 또한 Momentum distillation 으로 large noisy data set (web dataset) 에서 발생할수 있는 문제를 해결하고자 함 (image - text pair 가 다른수도 있고, 상당히 비슷할수도 있음)
  - momentum model에 의해 생성된 pseudo-targets으로 부터 self-training 하는 것
  - ALBEF에 대해 이론적인 분석 제공 ⇒ 실제 label distribution 과 teacher의 pseudo label distribution간이 Mutual information을 높게 학습하다는 것과 동치임을 설명
  - 결국 Momentum distill을 통해 image-text pair를 학습시킬때 pair에 대한 다양한 view를 생성하는 것으로 해석가능. (data augmentation ?)
- 2가지 noisy large dataset, 2가지 in-domain dataset 에서 5가지 task에 대한 실험
- SOTA 달성 및 github 공개
- 비교적 관련 VLP task의 논문성능 대비 여전히 좋은성능을 내고있음.

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Vision and Language Pre-training (VLP)

Vision-and-Pre training (VLP) aims to learn multimodal representations from large-scale image-text paris that can improve downstream Vision-and-Language (V+L) tasks performance

- **Image-text Retrieval**
- Visual Alignment
- Visual Questions Anserwering (VQA)
- Natural language for Visual Reasoning (NLVR)
- Visual Grounding

Context

Most existing VLP Methods rely on pre-training object detectors to extract region based image features and employ a multimodal encoder to fuse the image features with word tokens,

This VLP framework suffers from several key limitations

1. Challenging for the multimodal encoder to model image and text iterations
2. Object detector is both annotation-expensive and compute-expensive
3. Existing pre-training objectives such as MLM may overfit to the noisy image-text pairs and degrade the model's generalization performance

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

ALBEF Model Architecture

- image encoder - 12 layer ViT-B/16
- text encoder - first 6 layer of BERT_base
- multimodal encoder - last 6 layer of the BERT_base

The full pre-training objective of ALBEF is:

$$\mathcal{L} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itm}}$$

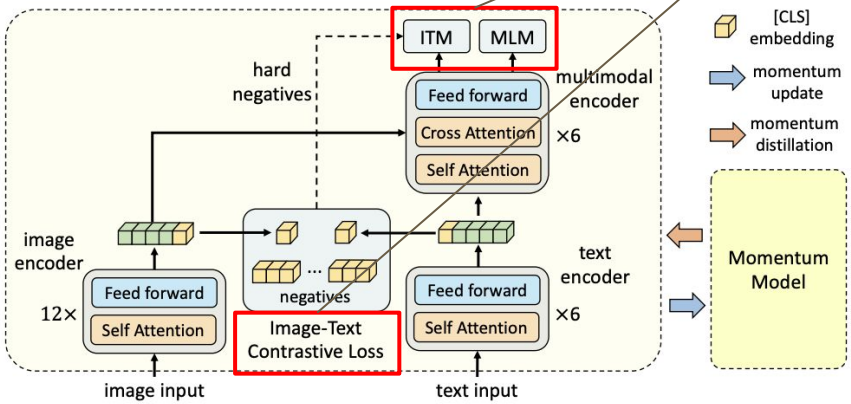


Figure 1: **Illustration of ALBEF.** It consists of an image encoder, a text encoder, and a multimodal encoder. We propose an image-text contrastive loss to align the unimodal representations of an image-text pair before fusion. An image-text matching loss (using in-batch hard negatives mined through contrastive similarity) and a masked-language-modeling loss are applied to learn multimodal interactions between image and text. In order to improve learning with noisy data, we generate pseudo-targets using the momentum model (a moving-average version of the base model) as additional supervision during training.

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

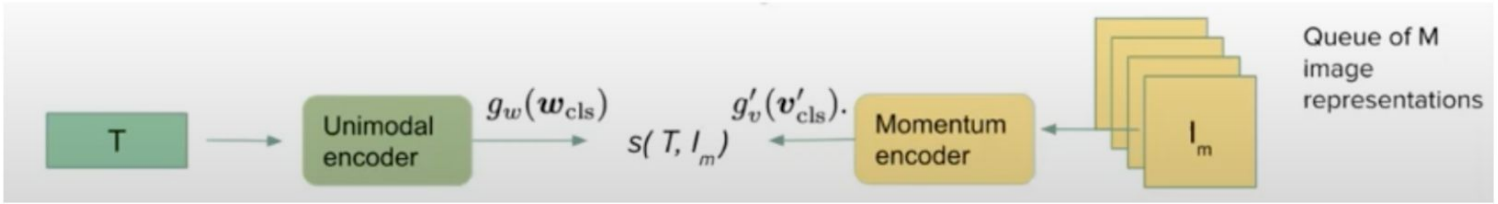
The full pre-training objective of ALBEF is:

$$\mathcal{L} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itm}}$$

ITC Loss (Image-Text Contrastive Learning)

- fusion 시키기 전에 unimodal representations 을 더 잘 학습하게 하기 위함.
  - Input Image is encoded into a sequence of embeddings $\{\boldsymbol{v}_{\text{cls}}, \boldsymbol{v}_1, ..., \boldsymbol{v}_N\}$,
  - Text is encoded into sequence of embeddings $\{\boldsymbol{w}_{\text{cls}}, \boldsymbol{w}_1, ..., \boldsymbol{w}_N\}$.
  - Similarity function (g 는 linear transformation) $s(I,T) = g_v(\boldsymbol{v}_{\text{cls}})^\top g'_w(\boldsymbol{w}'_{\text{cls}})$ and $s(T,I) = g_w(\boldsymbol{w}_{\text{cls}})^\top g'_v(\boldsymbol{v}'_{\text{cls}})$.
  - maintain two queues to store the most recent M image-text representations from the momentum unimodal encoders:

$g'_v(\boldsymbol{v}'_{\text{cls}})$ and $g'_w(\boldsymbol{w}'_{\text{cls}})$. : The normalized features from the momentum encoders



Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

The full pre-training objective of ALBEF is:

$$\mathcal{L} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itm}}$$

ITC Loss (Image-Text Contrastive Learning)

$$p_m^{\text{i2t}}(I) = \frac{\exp(s(I, T_m)/\tau)}{\sum_{m=1}^{M} \exp(s(I, T_m)/\tau)}, \qquad p_m^{\text{t2i}}(T) = \frac{\exp(s(T, I_m)/\tau)}{\sum_{m=1}^{M} \exp(s(T, I_m)/\tau)} \tag{1}$$

The image-text contrastive loss is defined as the cross-entropy H between $\boldsymbol{p}$ and $\boldsymbol{y}$:

$$\mathcal{L}_{\text{itc}} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D}\left[\text{H}(\boldsymbol{y}^{\text{i2t}}(I), \boldsymbol{p}^{\text{i2t}}(I)) + \text{H}(\boldsymbol{y}^{\text{t2i}}(T), \boldsymbol{p}^{\text{t2i}}(T))\right] \tag{2}$$

GT (one-hot encoding) GT (one-hot encoding)

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

The full pre-training objective of ALBEF is:

$$\mathcal{L} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itm}}$$

MLM Loss (Masked Language Modeling)

- Utilizes both image and contextual text to predict the masked workds
- Following BERT, we randomly mask out the input text tokens with 15% probability and replace them with a special tokan [MASK]
- MLM minimizes the cross-entropy loss:

n. MLM minimizes a cross-entropy loss:

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{(I,\hat{T})\sim D} \text{H}(\boldsymbol{y}^{\text{msk}}, \boldsymbol{p}^{\text{msk}}(I, \hat{T})) \quad (3)$$

GT (one-hot) vocabulary distribution

Predicted prob for a masked token

Masked text

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

The full pre-training objective of ALBEF is:

$$\mathcal{L} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itm}}$$

ITM Loss (Image-Text Matching)

- Predicts whether a pair of image and text is positive(Matched) or negative(UNmatched)
- Use the multimodal encoder's output embedding of the [CLS] token as the joint representation of the image-text pair
- Append a fully-connected layer followed by softmax to predict a two-class probability distribution p^itm
- ITM loss is:

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(I,T)\sim D} \mathrm{H}(\boldsymbol{y}^{\text{itm}}, \boldsymbol{p}^{\text{itm}}(I,T))$$

2-d on-hot vector for GT label

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Momentum Distillation

- A self-training method which learns from pseudo-targets produced by a momentum model, which is a continuously-evolving teacher which consists of exponential-moving-average (EMA) version of the unimodal and multimodal encoders.

$$\theta_{mo} \leftarrow m*\theta_{mo} + (1-m)*\theta$$

Momentum Distillation for ITC

$$\mathcal{L}_{\text{itc}}^{\text{mod}} = (1-\alpha)\mathcal{L}_{\text{itc}} + \frac{\alpha}{2}\mathbb{E}_{(I,T)\sim D}\left[\text{KL}(\boldsymbol{q}^{\text{i2t}}(I) \parallel \boldsymbol{p}^{\text{i2t}}(I)) + \text{KL}(\boldsymbol{q}^{\text{t2i}}(T) \parallel \boldsymbol{p}^{\text{t2i}}(T))\right] \quad (6)$$

from momentum unimodal encoder

from unimodal encoder

Momentum Distillation for MLM

$$\mathcal{L}_{\text{mlm}}^{\text{mod}} = (1-\alpha)\mathcal{L}_{\text{mlm}} + \alpha\mathbb{E}_{(I,\hat{T})\sim D}\text{KL}(\boldsymbol{q}^{\text{msk}}(I,\hat{T}) \parallel \boldsymbol{p}^{\text{msk}}(I,\hat{T})) \quad (7)$$

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Momentum Distillation

- A self-training method which learns from pseudo-targets produced by a momentum model, which is a continuously-evolving teacher which consists of exponential-moving-average (EMA) version of the unimodal and multimodal encoders.

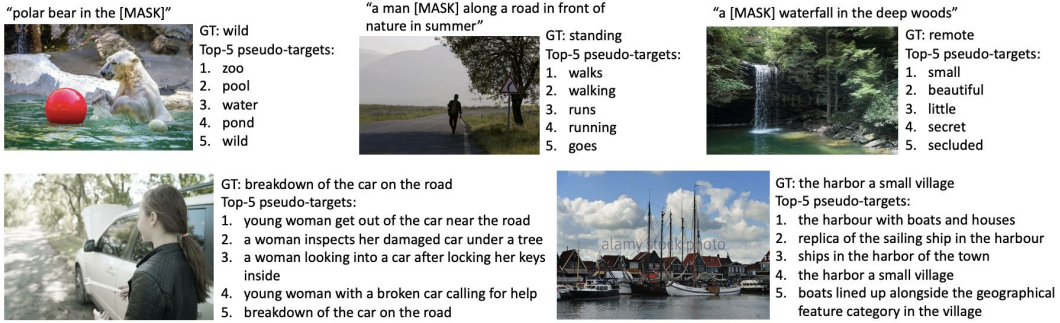


Figure 2: Examples of the pseudo-targets for MLM (1st row) and ITC (2nd row). The pseudo-targets can capture visual concepts that are not described by the ground-truth text (*e.g.* "beautiful waterfall", "young woman").

D Additional Examples of Pseudo-targets

Figure 11: Examples of the top-5 most similar texts selected by the momentum model for ITC.

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Pre training Data

- Two web datsets: Conceptual Captions, SBU Captions
- Two in-domain datasets: COCO and Visual Genome
- From the datasets above: total number of unique images is 4.0M. number of image-text pair is 5.1M
- To show the method is scalable, They also include the much noiser Conceptual 12M datasets. Increasing the total number of images to 14.1M

Vision and Language Pre-training (VLP)

Vision-and-Pre training (VLP) aims to learn multimodal representations from large-scale image-text paris that can improve downstream Vision-and-Language (V+L) tasks performance

- **Image-text Retrieval**
- Visual Alignment
- Visual Questions Anserwering (VQA)
- Natural language for Visual Reasoning (NLVR)
- Visual Grounding

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Results

#Pre-train Images	Training tasks	TR (flickr test)	IR (flickr test)	SNLI-VE (test)	NLVR^2 (test-P)	VQA (test-dev)
4M	MLM + ITM	93.96	88.55	77.06	77.51	71.40
	ITC + MLM + ITM	96.55	91.69	79.15	79.88	73.29
	ITC + MLM + ITM_{hard}	97.01	92.16	79.77	80.35	73.81
	ITC_{MoD} + MLM + ITM_{hard}	97.33	92.43	79.99	80.34	74.06
	Full (ITC_{MoD} + MLM_{MoD} + ITM_{hard})	97.47	92.58	80.12	80.44	74.42
	ALBEF (Full + $\text{MoD}_{\text{Downstream}}$)	97.83	92.65	80.30	80.50	74.54
14M	ALBEF	98.70	94.07	80.91	83.14	75.84

Table 1: Evaluation of the proposed methods on four downstream V+L tasks. For text-retrieval (TR) and image-retrieval (IR), we report the average of R@1, R@5 and R@10. ITC: image-text contrastive learning. MLM: masked language modeling. ITM_{hard}: image-text matching with contrastive hard negative mining. MoD: momentum distillation. $\text{MoD}_{\text{Downstream}}$: momentum distillation on downstream tasks.

Method	VQA		NLVR^2		SNLI-VE	
	test-dev	test-std	dev	test-P	val	test
VisualBERT [13]	70.80	71.00	67.40	67.00	-	-
VL-BERT [10]	71.16	-	-	-	-	-
LXMERT [1]	72.42	72.54	74.90	74.50	-	-
12-in-1 [12]	73.15	-	-	78.87	-	76.95
UNITER [2]	72.70	72.91	77.18	77.85	78.59	78.28
VL-BART/T5 [54]	-	71.3	-	73.6	-	-
ViLT [21]	70.94	-	75.24	76.21	-	-
OSCAR [3]	73.16	73.44	78.07	78.36	-	-
VILLA [8]	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF (4M)	74.54	74.70	80.24	80.50	80.14	80.30
ALBEF (14M)	**75.84**	**76.04**	**82.55**	**83.14**	**80.80**	**80.91**

Table 4: Comparison with state-of-the-art methods on downstream vision-language tasks.

Method	# Pre-train Images	Flickr30K (1K test set)						MSCOCO (5K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER	4M	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
VILLA	4M	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
OSCAR	4M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
ALIGN	1.2B	95.3	99.8	100.0	84.9	97.4	98.6	77.0	93.5	96.9	59.9	83.3	89.8
ALBEF	4M	94.3	99.4	99.8	82.8	96.7	98.4	73.1	91.4	96.0	56.8	81.5	89.2
ALBEF	14M	**95.9**	**99.8**	**100.0**	**85.6**	**97.5**	**98.9**	**77.6**	**94.3**	**97.2**	**60.7**	**84.3**	**90.5**

Table 2: Fine-tuned image-text retrieval results on Flickr30K and COCO datasets.

Method	# Pre-train Images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER [2]	4M	83.6	95.7	97.7	68.7	89.2	93.9
CLIP [6]	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [7]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	4M	90.5	98.8	99.7	76.8	93.7	96.7
ALBEF	14M	**94.1**	**99.5**	**99.7**	**82.8**	**96.3**	**98.1**

Table 3: Zero-shot image-text retrieval results on Flickr30K.

Solving ImageNet: a Unified Scheme for Training any Backbone to Top Results

Tal Ridnik, Hussam Lawen, Emanuel Ben-Baruch, Asaf Noy
DAMO Academy, Alibaba Group
tal.ridnik@alibaba-inc.com

https://arxiv.org/pdf/2204.03475.pdf

어떠한 아키텍처던 관계 없이 **하나의 unified 된 training scheme 을 가지고**, CNN, Transformer, NLP 등에 좋은 성능을 보이고자 함.

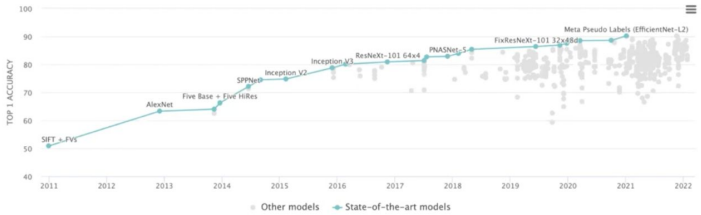
Solving imageNet: a Unified Scheme for Training any Backbone to Top Results

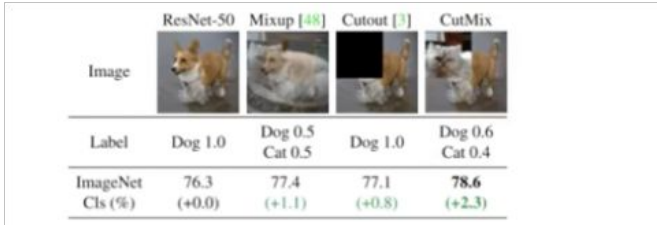
Previous works

- **Regularizations**
  - Stronger augmentations: AutoAugment, RandAugment (다양한 augmentation 방법)
  - **Image-based regularizations Cutout, Cutmix and Mixup**
    - **이미지에서 적용할 수 있는 새로운 augmentation**
  - Architecture regularizations like drop-path, drop-block
    - 아키텍처 특정한 부분은 weight 업데이트 안함.
  - Label-smoothing
  - Progressive image resizing during training
  - **Different train-test resolutions**
    - **train, test resolutions 을 다르게 보는 것.**

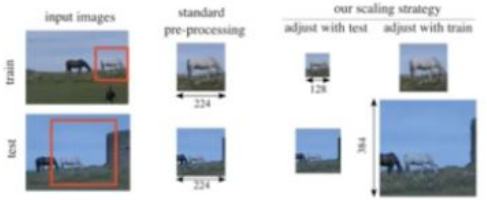
- **Training configuration**
  - More training epochs
  - Dedicated optimizer for large batch size (LAMB Optimizer), Scaling learning rate with batch size
  - Exponential-moving average (EMA) of model weights
  - Improved weights initializations
  - Decoupled weight decay (AdamW)

- **Architecture**





Yun, S. et al., CutMix: Regularization strategy to train strong classifiers with localizable features. ICCV 2019



Fixing the train-test resolution discrepancy. NeurIPS 2019

Solving imageNet: a Unified Scheme for Training any Backbone to Top Results

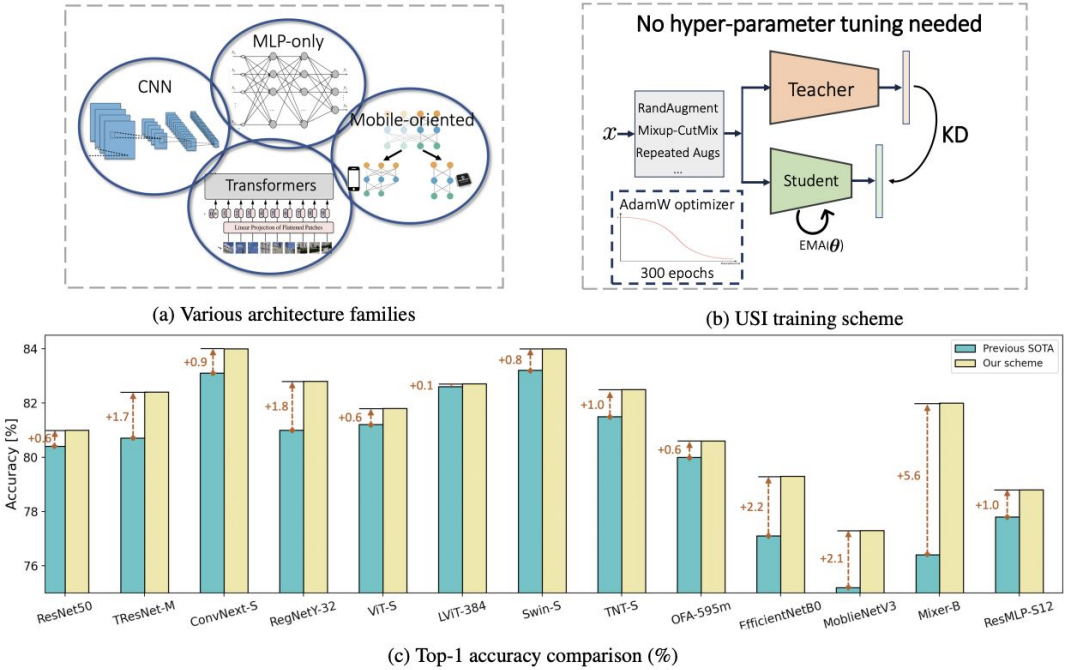
Motivation - Architecture 와 관계없이 잘 작동하는 training scheme 제안 필요

- **Architecture 마다 맞춤형 training scheme 이 적용됨**
  - ResNet 계열 (TResNet, SEResNet, ResNet-D, …)
    - 일반적으로 다양한 training scheme 에 잘 작동함
    - (Ross Wightman et al., 2021) 에서 제안한 방법이 ResNet 계열을 학습시키는데 standard 가 되었다고 함.
  - Mobile-oriented models
    - Depth-wise convolutions 에 많이 의존
    - RMSProp optimizer, waterfall learning rate scheduling and EMA
  - Transformer- based, NLP-only models
    - Inductive bias 가 없어 훈련하기 어려움 -> longer training (1000 epochs), strong cutmix-mixup and drop-path regularizations, large weight-decay and repeated augmentations

- **어떤 한 모델에 대한 맞춤형 training scheme 은 다른 모델에 적용하면 성능이 낮아짐**
  - ResNet50을 위한 training scheme 을 EfficientNet v2 model 에 적용했을 때, 맞춤형 training scheme 을 적용했을 때 보다 3.3%의 성능 하락을 보임 (Mingxing Tan et al., PMLR, 2021)

Solving imageNet: a Unified Scheme for Training any Backbone to Top Results

Unified training scheme for ImageNet without any hyper-parameter tuning or tailor-made tricks per model



(a) Various architecture families

(b) USI training scheme

(c) Top-1 accuracy comparison (%)

Figure 1: **Our unified training scheme for ImageNet, USI**. With USI, we can train any backbone to top results on ImageNet, without any hyper-parameter tuning or adjustments per architecture.

Solving imageNet: a Unified Scheme for Training any Backbone to Top Results

Methods

- **Knowledge Distillation (KD) 적용**
  - ResNet50 의 image classification, DeIT, NAS 등 소개하는 previous work 에서 KD 가 성능 향상에 중요한 역할을 함.
  - **But, KD is not a common practice for ImageNet training.**
  - 그럼에도, **KD** 를 사용해야 하는 이유.
    - **(b) Wing, airplane : Teacher network 는 image 가 완전히 mutually-exclusive 하지 않은 case 를 보완한다.**
    - **(c) Hem 55.5% 사람이 봐도 애매한데, 그 애매함을 teacher 의 classification 결과가 반영한다.**
    - **(d) Task 로 보면 틀린 답이지만, English setter 가 이미지에서 main object 라고 볼 수 있다.**

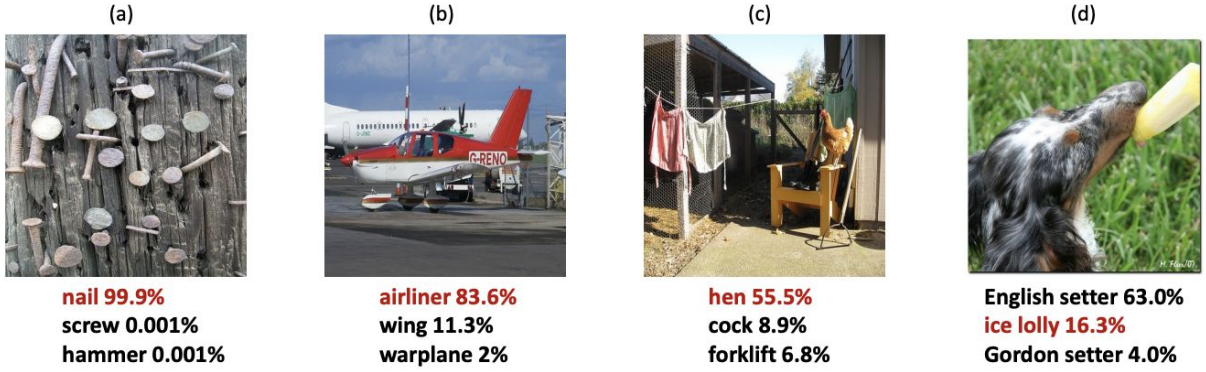
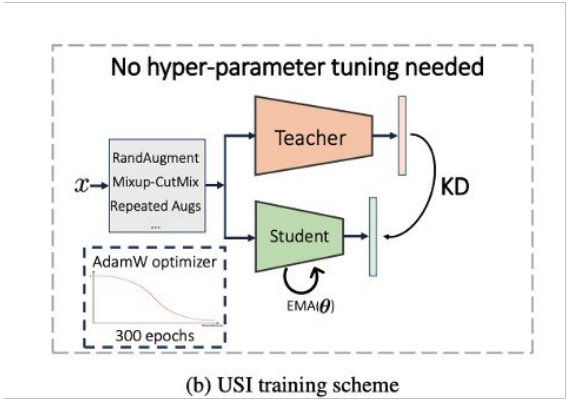


Figure 2: **Examples for teacher predictions**. ImageNet ground-truth labels are highlighted in red. Unlike the ground-truth, the teacher predictions account for similarities and correlations between classes, objects' saliency, pictures with several objects, and more. The teacher predictions would also better represent the content of images under strong augmentations.

Solving imageNet: a Unified Scheme for Training any Backbone to Top Results

Methods

- **Knowledge Distillation (KD) 적용**
  - **그럼에도, KD 를 사용해야 하는 이유.**
    - (b) Wing, airplane : Teacher network 는 image 가 완전히 mutually-exclusive 하지 않은 case 를 보완한다.
    - (c) Hem 55.5% 사람이 봐도 애매한데, 그 애매함을 teacher 의 classification 결과가 반영한다.
    - (d) Task 로 보면 틀린 답이지만, English setter 가 이미지에서 main object 라고 볼 수 있다.
  - **즉, Teacher label 에 GT label 보다 더 많은 정보가 포함되어 있음 (class 간의 유사성과 상관관계)**
  - **Label error 을 보정할 수 있음. Label smoothing 을 따로 할 필요가 없음.**
  - **Lead to a more effective and robust optimization process, compared to training with hard-labels only.**
    - **hard label 만을 사용해서 process 하는 것보다, 좀 더 optimizational 한 결과 값을 포함할 수 있다.**
  - **KD 를 활용해 아키텍처가 달라고 동일한 training configuration 을 적용할 수 있도록 제안.**



(b) USI training scheme

Procedure	Value
Train resolution	224
Test resolution	224
Epochs	300
Optimizer	AdamW
Weight decay	2e-2
Learning rate	2e-3
LR decay	One-cycle policy
Mixup alpha	0.8
Cutmix alpha	1.0
Augmentations	Rand-augment (7/0.5)
Test crop ratio	0.95
Repeated Augs	3
Base loss	Cross entropy
KD loss	KL-divergence
KD temperature	1
α_{kd}	5
Teacher	TResNet-L
Batch size	512 to 3456

Table 1: **USI training configuration**. With USI, exactly the same training recipe is applied to any backbone, and no hyper-parameter tuning is needed.

Solving imageNet: a Unified Scheme for Training any Backbone to Top Results

Experiments

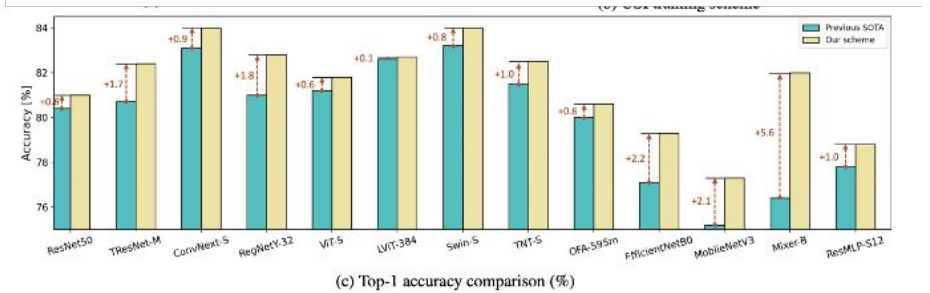
- USI 의 robustness 검증
- 제안한 training scheme (KD), loss function 이 잘 작동함을 확인
- 추가로 성능 향상할 수 있는 방법 제안
- Application: Speed-Accuracy comparison

USI 의 robustness 검증

- 모델들에 똑같이 USI 를 적용했을 때, tailor-made schemes 을 적용한 각 논문의 Top1 accuracy 보다 좋은 성능을 보임

Model Type	Model Name	USI Top1 Acc. [%]	Comparable Training Scheme			
			Top1 Acc. [%]	Epochs	KD	Additional Details
CNN	ResNet50	81.0	80.4 [42]	600	no	ResNet-strike-back, A1 config
			80.2 [47]	300	yes	Relabel-based KD
	TResNet-M	82.4	80.7 [30]	300	no	
	ConvNext-S	84.0	83.1 [25]	300	no	
	RegNetY-32	82.8	81.0 [28]	100	no	
Transformer	ViT-S	81.8	79.8 [8]	300	no	Original paper scheme
			81.2 [38]	1000	yes	DeiT scheme
	LeViT-384	82.7	82.6 [11]	1000	yes	DeiT scheme
	Swin-S	84.0	83.2 [24]	300	no	
	TNT-S	82.5	81.5 [12]	300	no	
Mobile-Oriented CNN	OFA-595m	80.6	80.0 [3]	255	yes	KD from super network
	EfficientnetB0	79.3	77.1 [34]	not stated	no	
	MobileNetV3	77.3	75.2 [17]	not stated	no	
MLP-Only	Mixer-B	82.0	76.4 [36]	255	no	
	ResMLP-S12	78.8	77.8 [37]	400	yes	

Table 2: Comparison of our proposed scheme, USI, to previous state-of-the-art results. Train and test resolution - 224.



(c) Top-1 accuracy comparison (%)

Solving imageNet: a Unified Scheme for Training any Backbone to Top Results

USI 의 robustness 검증

- 모델들에 똑같이 USI 를 적용했을 때, tailor-made schemes 을 적용한 각 논문의 Top1 accuracy 보다 좋은 성능을 보임

Batch Size	Top1 Acc. [%]	Training speed [img/sec]
512	82.3	1100
1024	82.5	1900
2048	82.3	3000
2752	82.4	4100
3456	82.4	4300
3456		4900 (no-KD reference)

Table 3: Accuracy and training speed, for different batch sizes. Model tested - TResNet-M.

Student	Student Type	Teacher	Teacher Type	Top1 Acc. [%]
ResNet50	CNN	TResNet-L	CNN	81.0
		Volo-d1	Transformer	80.9
LeViT384	Transformer	TResNet-L	CNN	82.7
		Volo-d1	Transformer	82.7

Table 4: Testing different students with different teachers.

In Table 9 we test whether adding drop-path to our scheme, when training a Transformer-based model, would improve results.

Drop-path	Top1 Acc. [%]
0	82.7
0.1	82.6
0.2	82.5

Table 9: Accuracy for different values of drop-path regularization. Model tested - LeViT-384

Solving imageNet: a Unified Scheme for Training any Backbone to Top Results

제안한 training scheme (KD), loss function 이 잘 작동함을 확인

- ImageNet Training 에서 KD 가 효과적임을 입증
- Vanilla softmax probabilities 를 사용하는 것이 좋음

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(\mathbf{p}^s(1), \mathbf{y}) + \alpha_{\text{kd}}\mathcal{L}_{\text{KL}}(\mathbf{p}^s(\tau), \mathbf{p}^t(\tau)), \quad (2)$$

KD relative weight, α_{kd}	Top1 Acc. [%]
0 (no KD loss)	76.2
1	80.8
5	82.7
10	82.6
20	82.7
∞ (no CE loss)	82.7

Table 6: Accuracy for different KD relative weights. Model tested - LeViT-384

3.6.2 KD Temperature

In Table 7 we investigate the impact of KD Temperature (τ in Eq. 2) on the accuracy.

KD Temperature, τ	Top1 Acc. [%]
0.1	79.3
1	82.7
2	82.7
5	81.7
10	81.4

Table 7: Accuracy for different KD temperatures. Model tested - LeViT-384

Solving imageNet: a Unified Scheme for Training any Backbone to Top Results

추가로 성능을 더 높일 수 있는 방법

- Epoch 에 관한 USI 의 default configuration 은 300 이지만, 더 긴 training epoch 으로 성능을 향상 시킬 수 있다.
- Augmentation 은 적용하는게 좋음.

Training epochs	Top1 Acc. [%]
100	80.0
200	81.9
300	82.7
600	83.0
1000	83.2

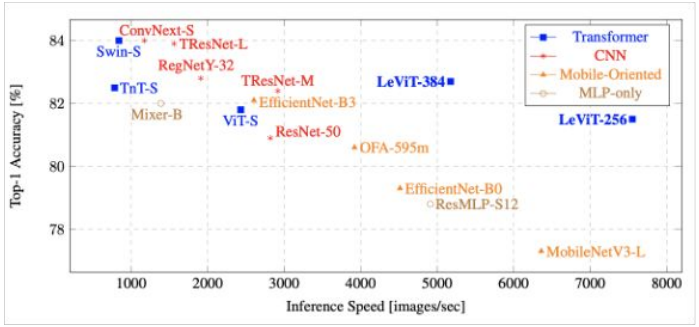
Table 5: Accuracy for different numbers of epochs. Model tested - LeViT-384.

Augmentation Type	Top1 Acc. [%]
None	82.0
Cutout	82.4
Mixup-Cutmix	82.7

Table 8: Accuracy for different augmentations. Model tested - LeViT-384

Speed-Accuracy comparison

- USI 를 활용해 모든 backbone 에 대해 동일한 하이퍼 파라미터를 적용했고, 이에 따라 재현성과 신뢰도가 높은 speed-accuracy trade-off 비교가 가능하다.



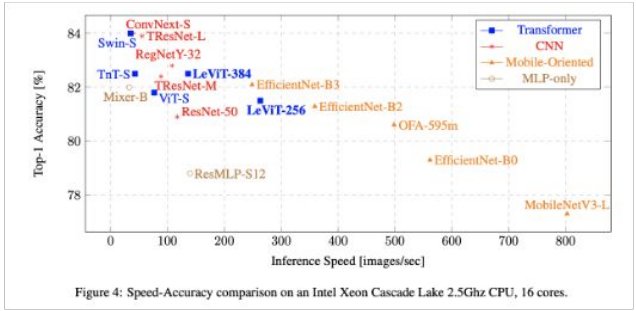


Figure 4: Speed-Accuracy comparison on an Intel Xeon Cascade Lake 2.5Ghz CPU, 16 cores.

Thank you :)