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

NeurIPS 2021

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare
Shafiq Joty, Caiming Xiong, Steven C.H. Hoi
Salesforce Research

발제자:
윤예준

https://arxiv.org/abs/2107.07651

01. 연구배경

기존 VLP framework

- Region-based image features를 추출하는데 사전 학습된 object detector에 의존하고, multimodal encoder를 사용하여 image features를 word token과 align함

- multimodal encoder는 MLM, ITM과 같이 이미지와 텍스트의 joint understanding require task를 해결하도록 훈련됨

VLP framework limitations

- image features와 word token embeddings이 같은 공간에 mapping되므로, 하나의 공간에서 multimodal encoder가 interaction을 학습하기 어려움.

- object detector는 annotation-expensive, compute-expensive 함.

- 웹에서 수집된 이미지-텍스트 데이터세트는 noisy가 많으며, 이를 통해 학습하게 되면 MLM과 같은 objective에서 noisy가 많은 텍스트에 과적합되어 모델의 일반화 성능을 저하시킬 수 있음

01. 연구배경

앞서 말한 한계를 해결하기위해 multimodal encoder 전에 align하는 ALBEF를 제안

- image-text feature를 align하여 multimodal encoder가 cross-modal learning을 더 쉽게 하도록 만들어 줌

- unimodal encoder가 image-text feature의 semantic meaning을 더 잘 이해할 수 있도록 함

- image-text를 embed하기 위해 low-dimensional space를 학습하는데, 이는 image-text matching objective를 통해 정보가 더 담겨져 있는 sample를 찾을 수 있도록 도와줌

또한 Momentum Distillation (MoD) 제안

- noisy supervision 조건에서 학습을 개선시키기 위한 간단한 방법이며, 모델이 larger uncurated web dataset를 활용할 수 있도록 함

- pseudo-target을 생성하여 추가적인 지도학습을 가능하게 함

- web annotation과 다르게 reasonable한 data의 경우 모델이 penalize하지 않음

02. 제안

ALBE 방법
F

모델 구조

- Image encoder
    - ImageNet-1K로 사전학습된 ViT-B/16
    - [CLS] token 존재

- Text encoder
    - BERT base first 6 layers

- Multimodal encoder
    - BERT base last 6 layers

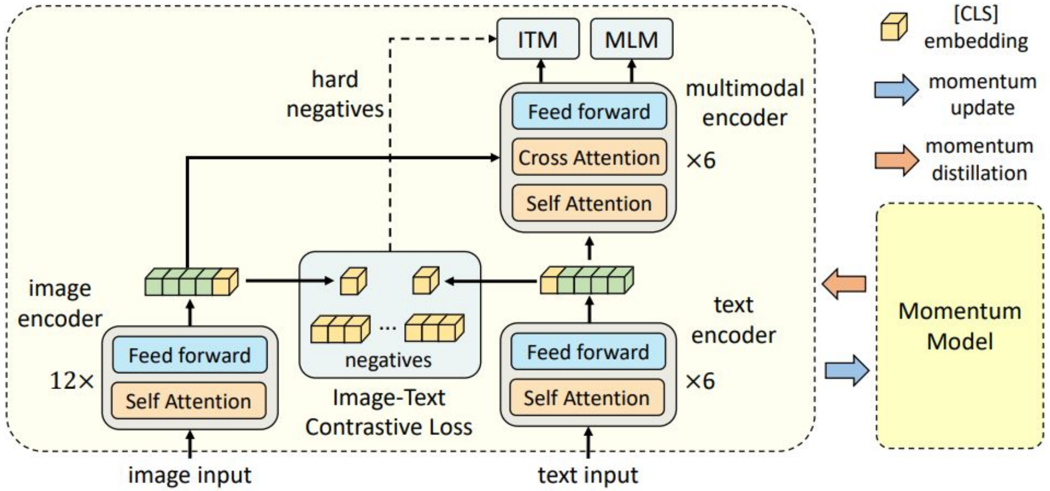


Figure 1: **Illustration of ALBEF.** It consists of an image encoder, a text encoder, and a multimodal encoder. We propose an image-text contrastive loss to align the unimodal representations of an image-text pair before fusion. An image-text matching loss (using in-batch hard negatives mined through 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.

02. 제안

ALBE방법
F

Pre-training Objectives

- Image-Text Contrastive Learning

$$s(I,T) = g_v(\boldsymbol{v}_{\text{cls}})^\top g'_w(\boldsymbol{w}'_{\text{cls}}) \text{ and } s(T,I) = g_w(\boldsymbol{w}_{\text{cls}})^\top g'_v(\boldsymbol{v}'_{\text{cls}})$$

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

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

- Masked Language Modeling

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

- Image-Text Matching
  - hard negative sample과 multimodal encoder의 [CLS] token 비교
    - ITC내에서 다항분포 확률값 기반 샘플링 I2T, T2I

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

Full pre-training objective of ALBEF is:

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

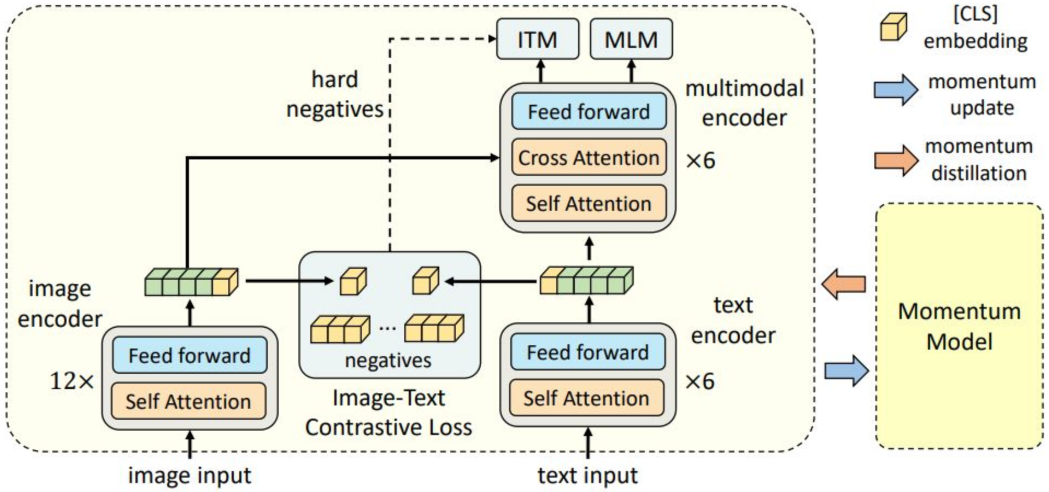


Figure 1: **Illustration of ALBEF.** It consists of an image encoder, a text encoder, and a multimodal encoder. We propose an image-text contrastive loss to align the unimodal representations of an image-text pair before fusion. An image-text matching loss (using in-batch hard negatives mined through 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.

02. 제안 방법

Momentum Distillation (MoD)

- Web image-text pair는 positive pairs임에도 weakly-correlated인 경우가 많아서 noisy함

- MLM과 ITC를 학습할 때 one-hot label을 사용함
- 그러나 그림2와 같이 이미지를 더 잘 설명하거나 의미상으로
  틀리지 않은 단어가 있을 수 있음(MLM)
- 또한, negative pair가 아님에도 negative로 여겨 similarity score가 낮아지게끔 학습할 수 있음 (ITC)

- 이를 해결하기 위해 momentum model로부터 pseudo-target을 만들고 one-hot label만으로 얻기 힘든 정보들을
  학습

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

- Image-Text Contrastive Learning

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

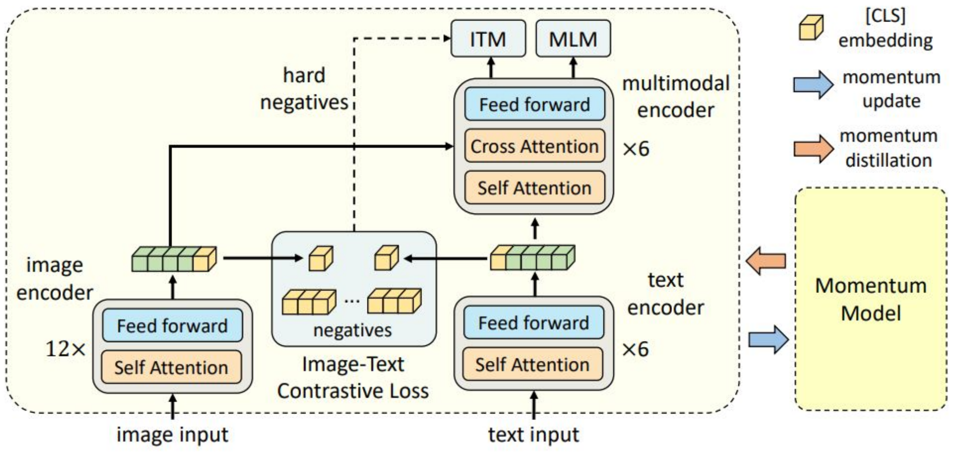


Figure 1: **Illustration of ALBEF.** It consists of an image encoder, a text encoder, and a multimodal encoder. We propose an image-text contrastive loss to align the unimodal representations of an image-text pair before fusion. An image-text matching loss (using in-batch hard negatives mined through 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.

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").

02. 제안 방법

Pre-training Datasets

- Following UNITER,
- Conceptual Captions, SBU captions
- COCO, Visual Genome
- noisier Conceptual 12M dataset

03. 실험 결과

Evaluation on the Proposed Methods

- 제안 방법 변형에 대한 downstream task 성능

#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 + $MoD_{Downstream}$)	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. $MoD_{Downstream}$: momentum distillation on downstream tasks.

03. 실험 결과

- Image-Text Retrieval
  - Flicker30K, COCO 벤치마크에 대해 평가 및 훈련 샘플 이용하여 fine-tune
  - Flicker30K 제로 샷 검색의 경우, COCO에서 fine-tune된 모델 사용
  - 미세 조정 시 ITC, ITM jointly optimize

Method	# Pre-train Images	Flickr30K (1K test set) TR			Flickr30K (1K test set) IR			MSCOCO (5K test set) TR			MSCOCO (5K test set) 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			Flickr30K (1K test set) 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.

03. 실험 결과

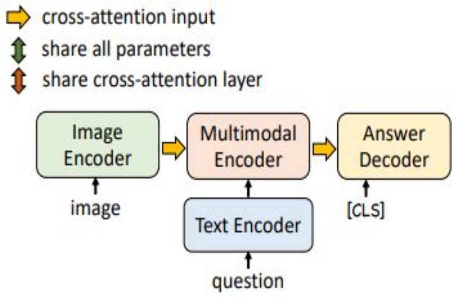
- Visual Entailment
    - UNITER에 따라 three-way classification 문제로 간주
    - MLP 사용하여 클래스 예측

- Visual Question Answering
    - 이미지와 질문이 주어졌을 때 모델이 답을 예측하는 것
    - 다중 분류 문제로 formulate하는 대신 답변 생성 문제로 간주

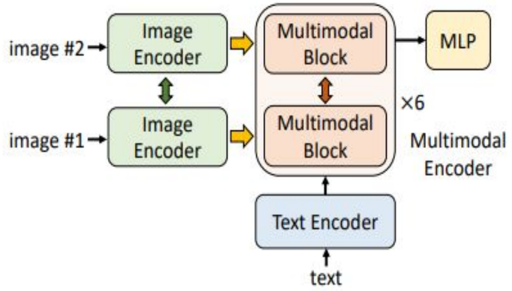
- Natural Language for Visual Reasoning
    - 모델이 텍스트가 이미지를 설명하는지 여부 예측
    - multimodal encoder의 [CLS] representation에 MLP 분류기 추가

Method	VQA test-dev	VQA test-std	NLVR² dev	NLVR² test-P	SNLI-VE val	SNLI-VE 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.



(a) VQA



(b) NLVR²

03. 실험 결과

- Visual Grounding
  - 특정 텍스트 설명에 해당하는 이미지 영역을 찾는 것을 목표
  - 바운딩 박스 annotation을 사용할 수 없는 weakly-supervised setting

Method	Val	TestA	TestB
ARN [57]	32.78	34.35	32.13
CCL [58]	34.29	36.91	33.56
$\text{ALBEF}_{\text{itc}}$	51.58	60.09	40.19
$\text{ALBEF}_{\text{itm}}$	**58.46**	**65.89**	**46.25**

Table 5: Weakly-supervised visual grounding on RefCOCO+ [56] dataset.

Figure 4: Grad-CAM visualization on the cross-attention maps in the 3rd layer of the multimodal encoder.

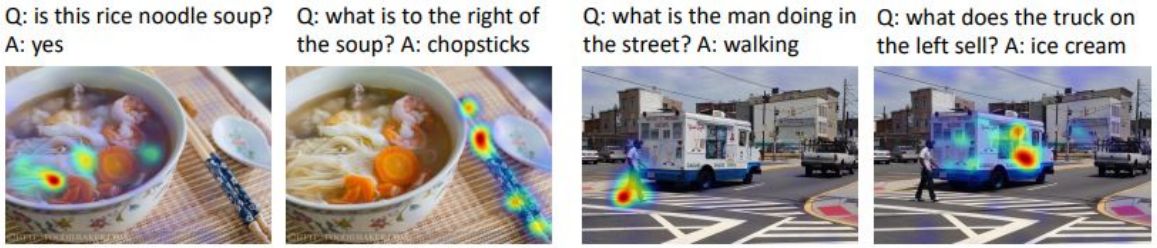


Figure 5: Grad-CAM visualizations on the cross-attention maps of the multimodal encoder for the VQA model.

Figure 6: Grad-CAM visualizations on the cross-attention maps corresponding to individual words.

04. 결론

- 시각-언어 표현 학습을 위한 새로운 프레임워크인 ALBEF 제안
- 다양한 평가를 통해 제안한 프레임워크의 효과 검증

감사합니
다.