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

https://arxiv.org/abs/2107.07651

0. Introduction

- 비전과 언어 멀티모달 표현 학습 문제 다룸
- 기존엔 두 모달리티를 단순히 결합(fuse)하는 데 집중함
- 본 논문은 먼저 각 모달리티 표현을 의미적으로 정렬(align)함
- 정렬 후 결합하는 새로운 학습 방식 제안함
- momentum distillation 기법 도입해 학습 안정성과 효율 높임
- 다양한 멀티모달 태스크에서 성능 개선 확인함

1. Overview

- 비전과 언어 표현 학습에서 단순 결합보다 의미적 정렬이 중요함
- 논문은 두 단계로 학습 프로세스 구성함
- 첫째, Align Before Fuse: 모달리티별 표현을 먼저 정렬함
- 둘째, Momentum Distillation: momentum 기반 지식 증류로 학습 안정화함
- 이 방식을 통해 멀티모달 표현 품질을 크게 높임
- 이미지-텍스트 검색, VQA 등 다양한 태스크에서 우수한 성능 달성함

2. Challenges

- 비전과 언어 표현 공간이 달라서 바로 결합하면 잡음이 많아짐
- 멀티모달 모델 학습이 불안정해지는 문제 있음
- 모달리티 간 의미적 불일치 현상이 자주 발생함
- 기존 퓨전 방법은 정렬 과정 없이 결합하는 경우가 많음
- 이런 방식은 표현 일관성 유지에 한계가 있음
- 효과적인 지식 증류 방법도 찾기 어려운 과제임

3. Method

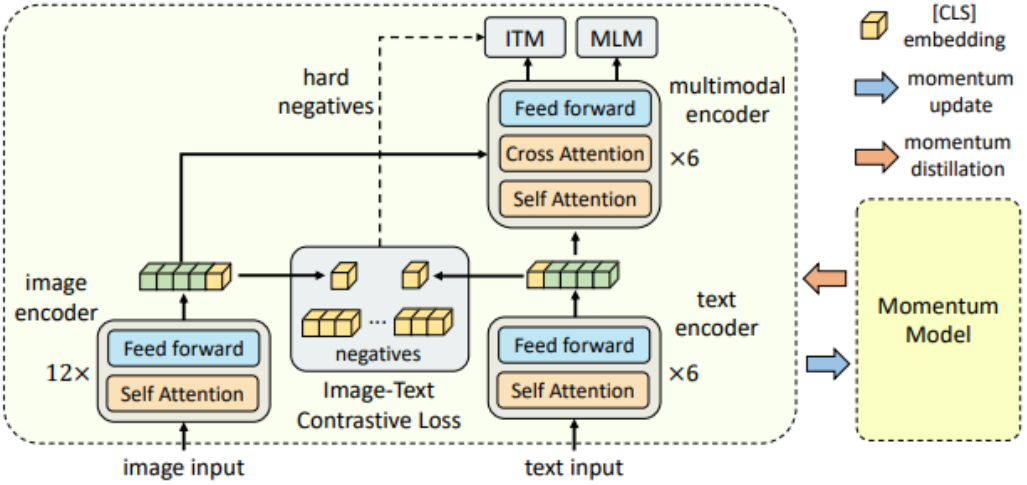

Figure 1: **Illustration of ALBEF.** It consists of an image encoder, a text encoder, and a multimodal encoder. We propose an image-text contrastive loss to align the unimodal representations of an image-text pair before fusion. An image-text matching loss (using in-batch hard negatives mined through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

- 각 모달리티에서 특징 추출 후 별도 네트워크로 표현 생성함
- 모달리티별 표현을 의미적으로 정렬하는 모듈 도입함
- 정렬 과정을 통해 두 모달리티 간 의미 간극을 좁힘
- momentum 기반 teacher 모델을 만들어 student 모델을 지도함
- momentum distillation으로 학습 안정성 높임
- 정렬과 결합을 단계별로 분리해 잡음 감소 효과 봄
- 이 과정에서 의미적 통일성과 표현 품질 크게 개선됨

4. Experiments

- MSCOCO, Flickr30k, Visual Genome 데이터셋 사용함
- 이미지-텍스트 검색, VQA, 이미지 캡셔닝 태스크로 평가함
- 기존 멀티모달 학습 방법들과 성능 비교함
- 동일 backbone 기반으로 ablation study 진행함
- 각 구성 요소별 효과를 자세히 분석함
- 학습 안정성과 표현력 향상 여부 집중 점검함

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

Method	VQA test-dev	VQA test-std	NLVR^2 dev	NLVR^2 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.

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

- Align Before Fuse가 단순 fusion보다 성능 더 좋음
- Momentum Distillation 도입 시 학습이 안정적임
- 표현의 일반화 능력도 함께 개선됨
- 이미지-텍스트 검색, VQA 등에서 SOTA 수준 성능 달성함

- Ablation 결과 각 구성 요소가 성능 향상에 기여함 확인됨
- 실험 통해 본 방법의 효과와 우수성 입증됨

6. Insight

- 멀티모달 학습에서 의미적 정렬이 결합보다 더 중요함
- momentum 기반 지식 증류가 학습 안정성에 큰 도움 됨
- ‘align before fuse’ 패러다임이 멀티모달 AI 발전에 기여 가능함
- 표현의 의미 일관성과 안정성에 집중한 학습 전략 필요함
- 본 연구는 향후 다양한 멀티모달 응용에 확장 적용 가능성 보여줌
- 하지만 네트워크 구조와 추가 모듈로 계산 비용이 늘어남
- momentum distillation 최적화가 까다로워 실무 적용에 어려움 있을 수 있음
- 데이터셋 편향 문제나 일반화에 대한 언급이 다소 부족함
- 최신 멀티모달 방법과 비교가 제한적임
- 실시간 처리나 경량화 측면에서는 추가 연구가 필요함