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

AAI Lab. 서원희, 최창수

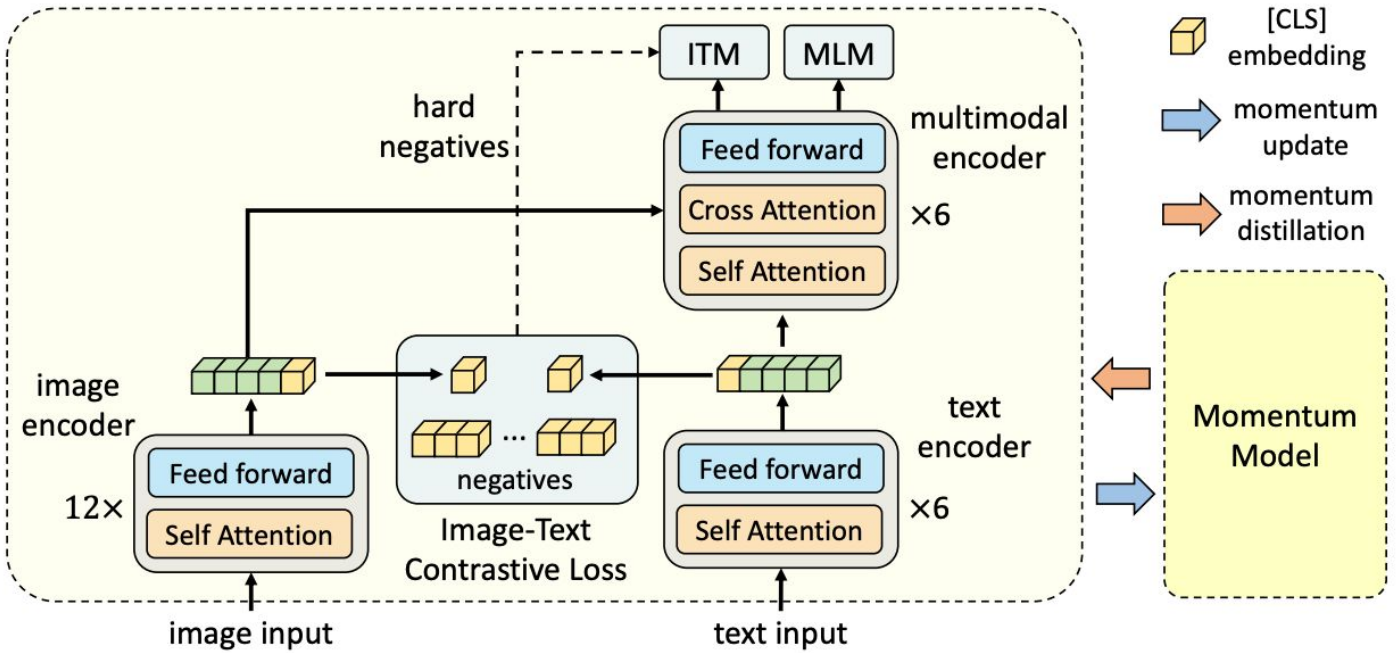
Index

1. Introduction

Introduction

논문소개

- Align before Fuse: Vision and Language Representation Learning with Momentum Distillation (NeurIPS 2021)
- Vision Language Pre-training(VLP): 이미지-텍스트쌍으로부터 멀티모달 표현을 학습하고 후속 연구를 위한 fine-tuning 양식을 제공

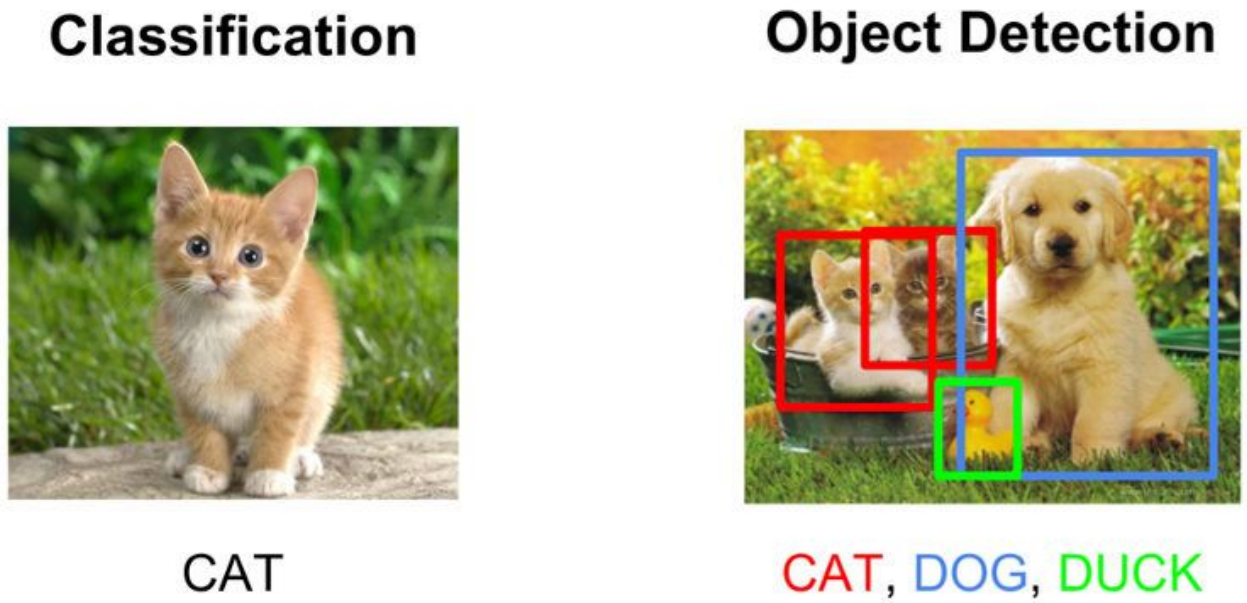


2. Background

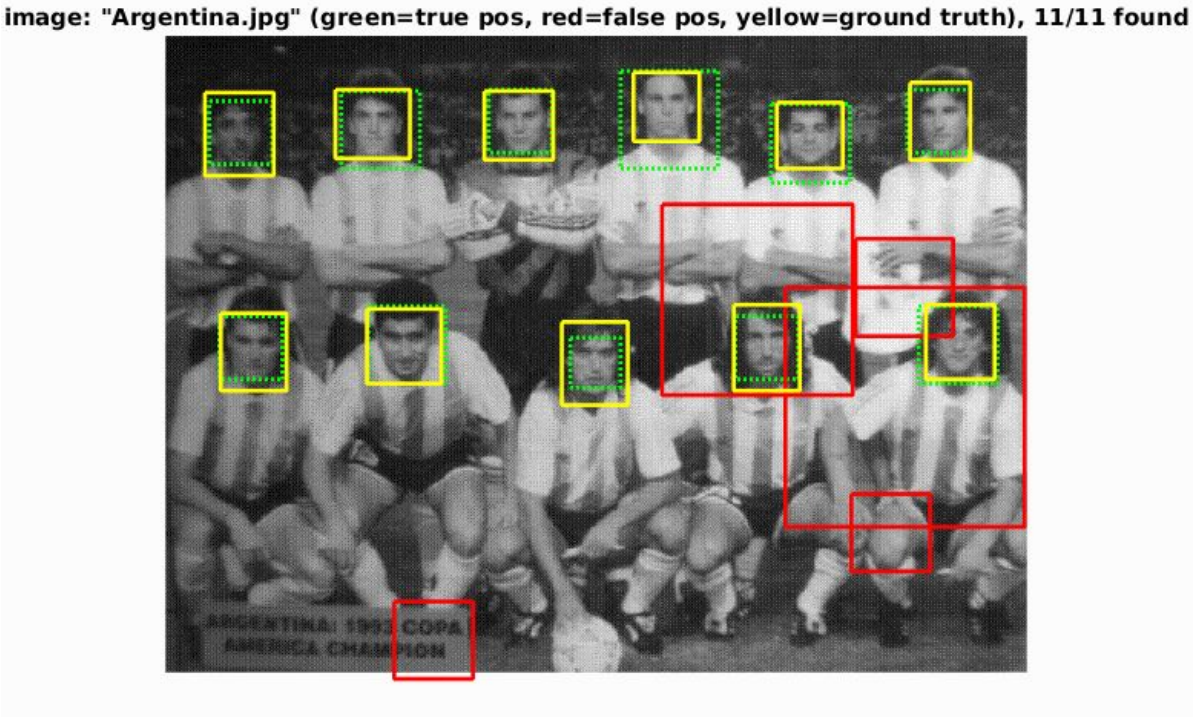
주요개념

[1] Object Detection

- 이미지를 classification + localization
  - localization: 객체라고 판단되는 곳에 직사각형(bounding box)을 그려주는 것
- 즉, 물체를 분류하고 객체에 bounding box를 그려줌



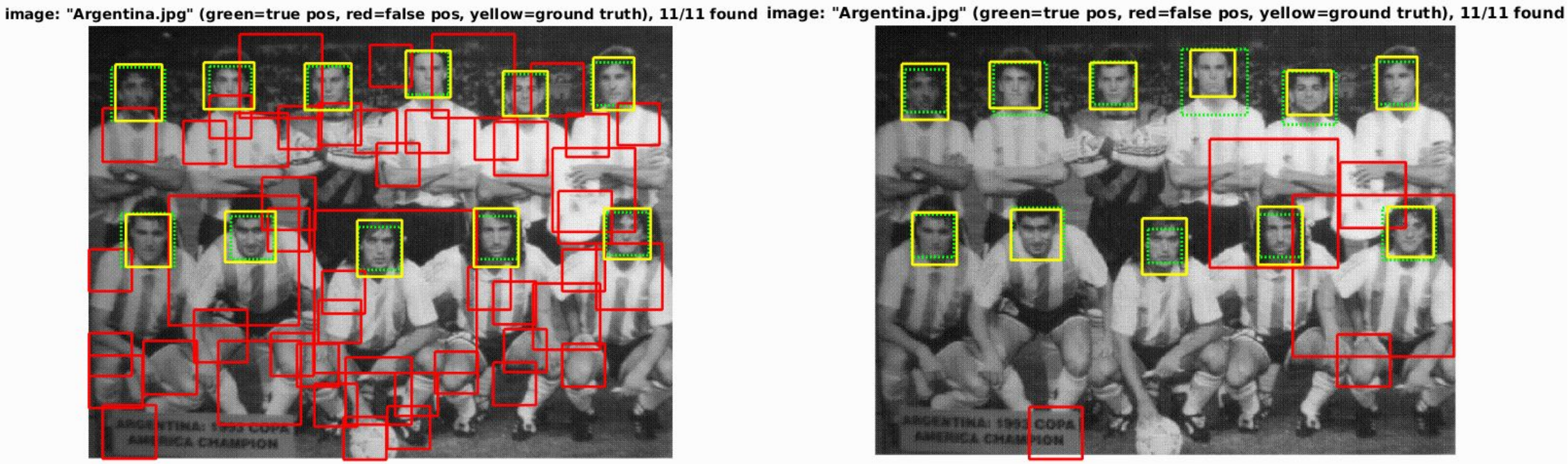
[2] Contractive hard negative mining



- 사람을 positive, 그 외의 것을 negative라고 할 때
  노란선이 정답, 초록 점선은 Ture positive, 빨간 선은 False Positive
- hard negative: 실제로는 negative인데 positive라고 예측하기 쉬운 데이터

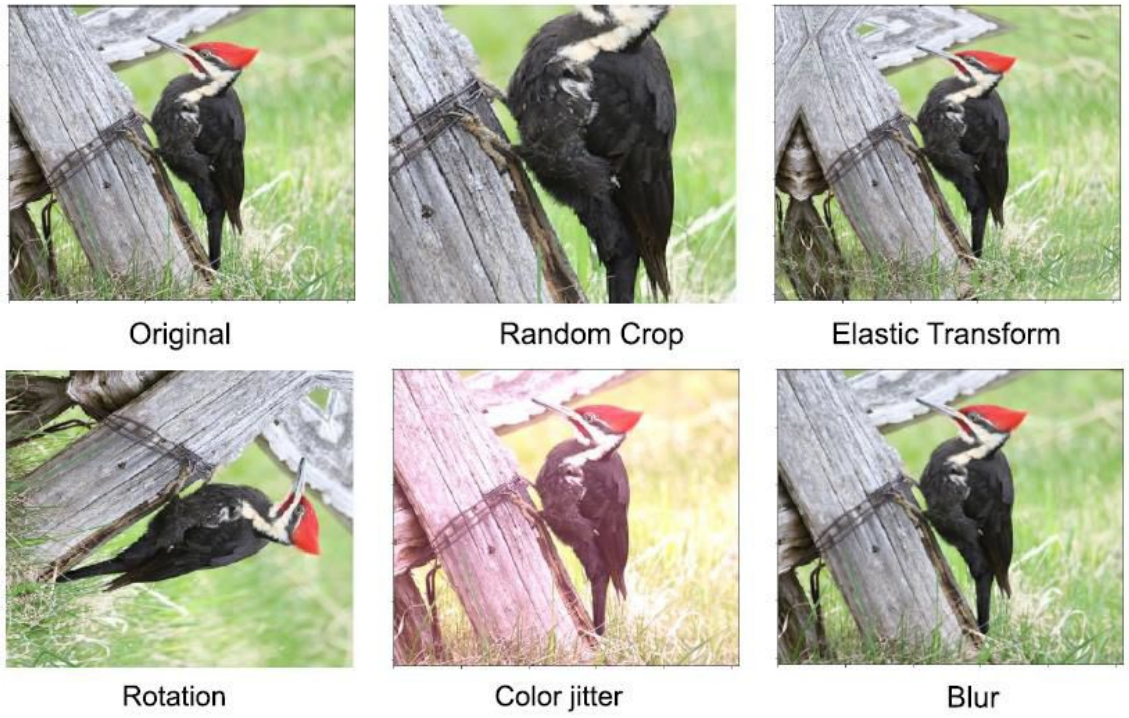
[2] Contractive hard negative mining

- hard negative mining: hard negative 데이터를 모으는 것
- hard negative mining으로 얻은 데이터를 원래의 데이터에 추가해서 재학습하면 false positive 오류에 강해진다.



[3] Contrastive Learning

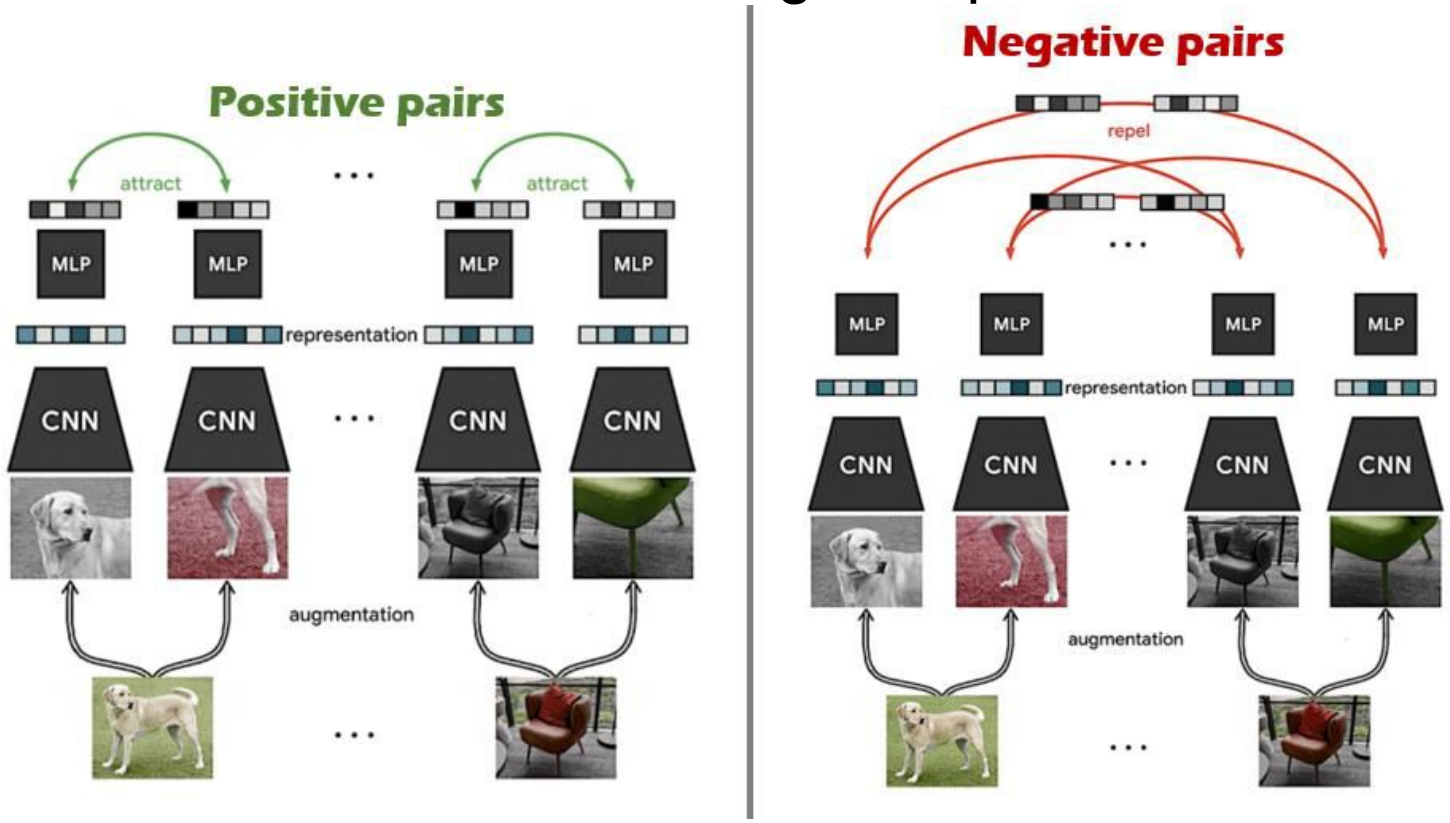
- 입력 샘플간의 비교를 통해 학습
- positive pair간의 유사도는 높이고 negative pair간의 유사도는 낮추는 방향



원본 이미지와 augmentation이 적용된 이미지처럼 서로 유사한 이미지(=positive pair)

[3] Contrastive Learning

- 같은 이미지에서 나온 이미지 패치는 positive pair,
  다른 이미지에서 나온 이미지 패치는 negative pair



[4] Low-Dimensional Representation

- 고차원 데이터에 대한 차원 축소 프로세스의 결과를 의미
- 데이터의 저차원 표현은 고차원 데이터로부터 가능한 많은 정보를 보유할 것으로 예상
- 일반적으로, 얼마나 차원을 줄일 수 있는가와 얼마나 많은 정보를 유지할 수 [illegible]

Why Is a Low-Dimensional Representation Important?

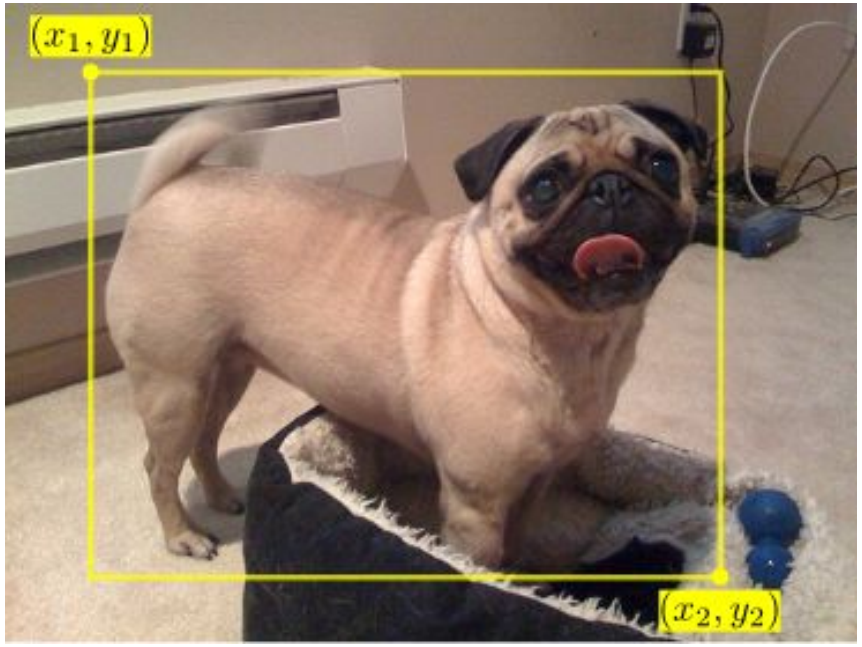
- noise를 제거하는 데에 사용
- 데이터의 feature를 추출할 때에 사용
- 일반적으로, 저차원 정보를 사용 할 때 원본 데이터의 대부분의 정보가 유지
- 저차원 표현을 사용하면 모델 성능을 크게 희생하지 않고도 학습 작업을 쉽게할 수 있음

3. Paper Review

선행 연구 동향 및 한계점

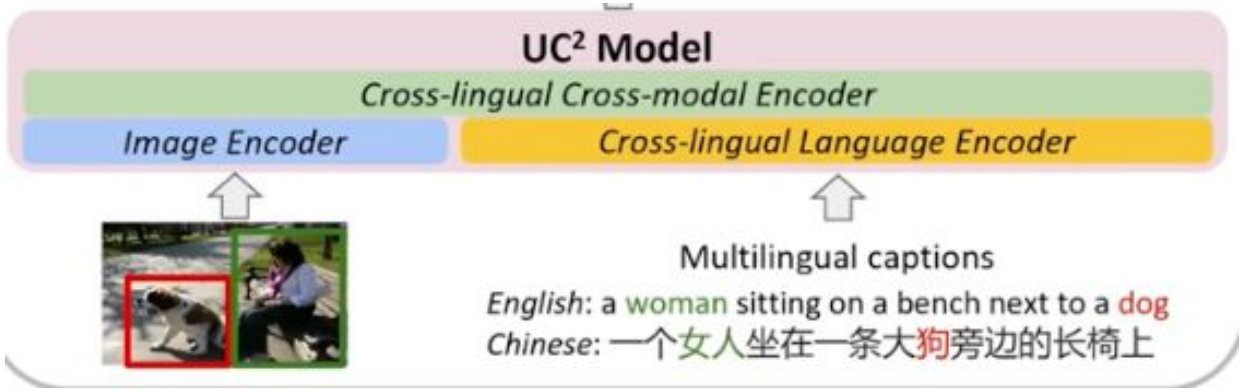
1. Vision-and-Language Pre-training(VLP)

- 대부분의 VLP 방법은 region-based image features를 추출하기 위해 object detector에 의존
- image와 text feature를 융합하기 위해 멀티모달 인코더 사용



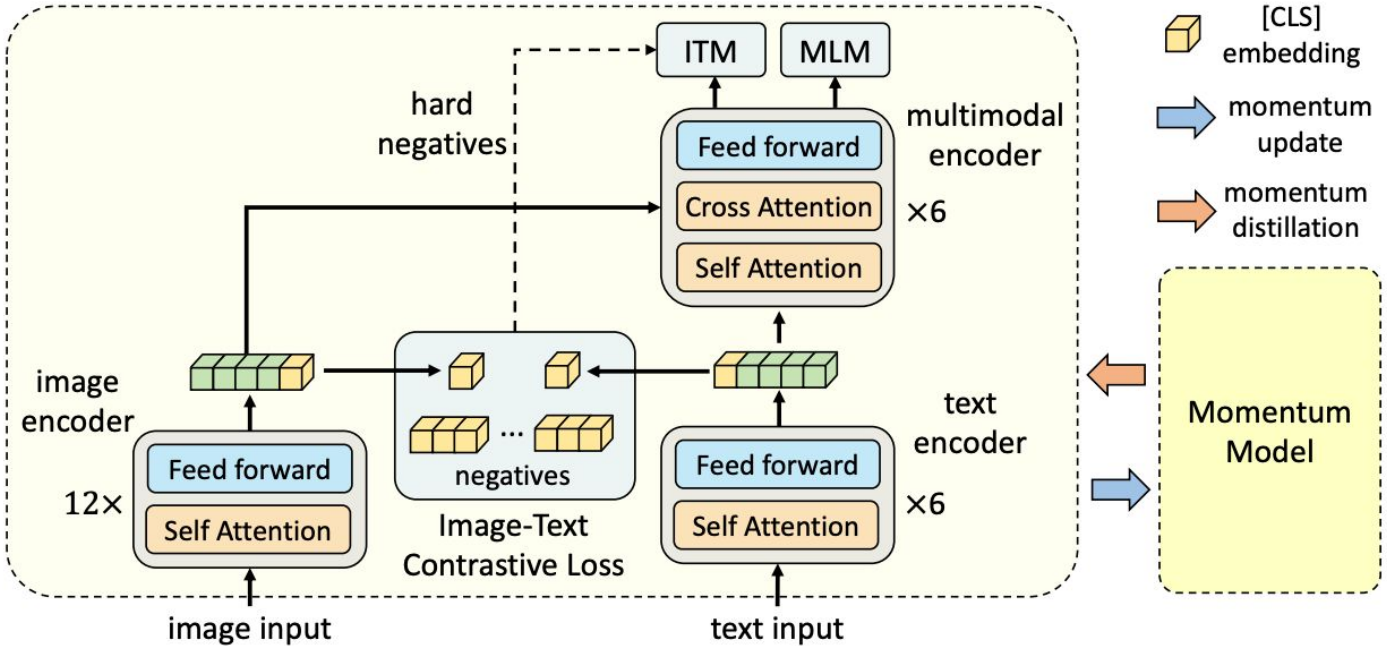
[한계점]

- image feature와 word token embeddings를 그들 자신의 space에 위치시켜서, 상호작용을 학습하기 어렵다. (image와 text의 연관성이 낮음)
- object detector는 사전학습시에 bounding box annotations와 고해상도 이미지를 요구하기 때문에 annotation-expensive하고 compute-expensive하다.
  - annotations: 인공지능이 데이터의 내용을 이해할 수 있도록 주석을 달아주는 작업
- 널리 사용되는 image-text dataset은 web에서 가져온 것이라 noisy에 overfit하여 모델의 일반화 성능을 떨어트릴 수 있다.



제안 방법론 및 Contribution

ALign BEfore Fuse(ALBEF) 구조



- Image encoder, Text encoder, Multimodal encoder로 구성됨
  - 입력 이미지 I 는 $\{\overrightarrow{v_{cls}}, \overrightarrow{v_1}, ..., \overrightarrow{v_N}\}$, 입력 텍스트 T 는 $\{\overrightarrow{w_{cls}}, \overrightarrow{w_1}, ..., \overrightarrow{w_N}\}$ 임베딩으로 인코딩 됨
  - Image encoder: visual transformer ViT-B/16의 12-layer 사용
  - Text encoder: BERTbase 모델의 first 6-layer 사용
  - Multimodal encoder: BERTbase 모델의 last 6-layer사용
- Multimodal encoder의 각 layer에서 cross attention을 통해 Image features와 Text feature를 Fusion

Pre-training Objectives

Image-Text Contrastive Learning(ITC)

- 멀티모달 인코더에서 image, text feature를 fusion하기 전 unimodal encoder를 학습 하는 것이 목적
- **같은 image-text pair(positive)로 부터 얻은 feature는 similarity가 높아지도록, 다른 image-text pair(negative)로 부터 얻은 features는 similarity가 낮아지도록 학습**

Masked Language Modeling(MLM)

- **이미지와 mask를 씌우지 않은 text를 활용해 mask 씌운 단어를 맞추는 적을 목적**
- BERT와 마찬가지로 input tokens의 15%를 랜덤하게 마스크함

Image-Text Matching(ITM)

- **image-text pair가 positive(matched)인지 negative(not matched)인지 예측하는 것이 목적**
- 멀티모달 인코더의 [CLS] embedding을 fully-connected layer를 거친후 softmax로 matching 여부를 예측

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

ITC and MLM

"polar bear in the [MASK]"

GT: wild
Top-5 pseudo-targets:
1. zoo
2. pool
3. water
4. pond
5. wild

"a man [MASK] along a road in front of nature in summer"

GT: standing
Top-5 pseudo-targets:
1. walks
2. walking
3. runs
4. running
5. goes

"a [MASK] waterfall in the deep woods"

GT: remote
Top-5 pseudo-targets:
1. small
2. beautiful
3. little
4. secret
5. secluded

GT: breakdown of the car on the road
Top-5 pseudo-targets:
1. young woman get out of the car near the road
2. a woman inspects her damaged car under a tree
3. a woman looking into a car after locking her keys inside
4. young woman with a broken car calling for help
5. breakdown of the car on the road

GT: the harbor a small village
Top-5 pseudo-targets:
1. the harbour with boats and houses
2. replica of the sailing ship in the harbour
3. ships in the harbor of the town
4. the harbor a small village
5. boats lined up alongside the geographical feature category in the village

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").

Image-Text Contrastive Learning(ITC)

- 멀티모달 인코더에서 image, text feature를 fusion하기 전 unimodal encoder를 학습 하는 것이 objective
- **같은 image-text pair(positive)로 부터 얻은 feature는 similarity가 높아지도록, 다른 image-text pair(negative)로 부터 얻은 features는 similarity가 낮아지도록 학습** $s(I,T) = g_v(v_{cls})^\top g_w(w_{cls})$
- g_v와 g_w는 [CLS]임베딩을 정규화된 lower-dimensional(256-d) representations으로 매핑하는 linear transformer
  - 즉, image input과 text input의 [CLS] token에 대한 embedding feature만으로 loss를 계산

각 이미지와 텍스트에 대해 softmax-normalized된 image-to-text및 text-to-image similiarity

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

τ : learnable temperature parameter, momentum을 적용할 때 쓰임, s가 p에 어느 정도 영향을 미치는지 조절

$y^{i2t}(I)$, $y^{t2i}(T)$: ground-truth one-hot similarity를 나타낸다고 하면 negative pair은 0의 확률값을 가지며 positive pair은 1의 확률 값을 가짐

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

image-text contrastive loss은 p와 y 사이의 cross-entropy H 로 정의 됨

Masked Language Modeling(MLM)

- **이미지와 mask를 씌우지 않은 contextual text를 활용해 mask 씌운 단어를 맞추는 것이 목표**
- BERT와 마찬가지로 input tokens의 15%를 랜덤하게 마스크 함
- $\hat{T}$ 를 masked text라 하고, $p^{msk}(I, \hat{T})$ 를 masked token에 대한 모델이 예측확률을 나타낸다면, MLM은 cross-entropy loss를 minimize해야 함.
- y^{msk}는 ground-truth token이 확률값 1을 갖는 one-hot vocabulary distribution

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

Image-Text Matching(ITM)

- **image-text pair가 positive(matched)인지 negative(not matched)인지 예측하는 objective**
- 멀티모달 인코더의 [CLS] embedding(joint representation of the image-text pair)을 fully-connected layer를 거친후 softmax로 matching 여부를 예측

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

  - 이때 batch내에서 negative sample을 선택 할때 random으로 고르는 것이 아니라 image 또는 text와 유사한 semantic을 가지는 hard negative sample로 ITM을 학습~~>~~ but how?
    - ITC 학습 과정에서 image-text similarity를 계산하는데, hard negative sample은 이를 활용해 추출함
    - 예를 들어, 하나의 image sample에 대한 negative text sample을 고를 때, 원래 positive pair에 해당하는 text를 제외하고 batch내 나머지 text들 중 similarity가 가장 높은 sample을 hard negative sample로 선택함