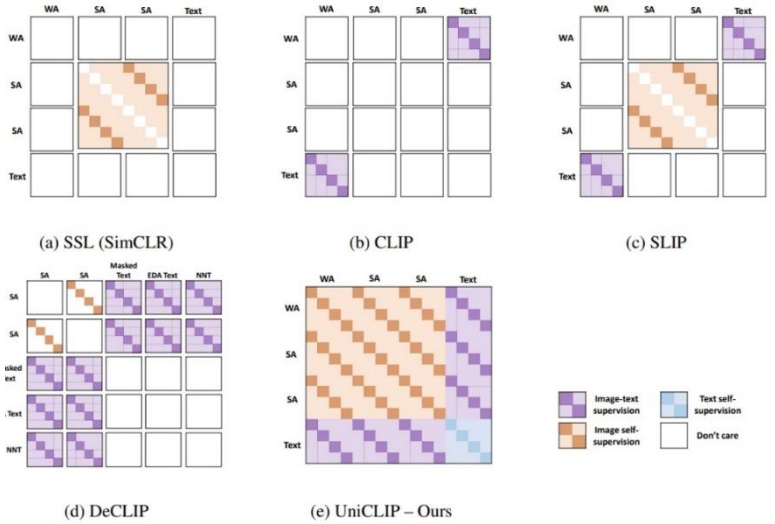
IAI
Industrial Artificial Intelligence Lab

UniCLIP: Unified Framework for Contrastive Language-Image Pre-training

Lee, J., Kim, J., Shon, H., Kim, B., Kim, S. H., Lee, H., & Kim, J. (2022). Uniclip: Unified framework for contrastive language-image pre-training. Advances in Neural Information Processing Systems, 35, 1008-1019.

Introduction

- 최근 이미지 비지도 학습 방법인 SimCLR[1]나 이미지-텍스트 쌍의 비지도 학습 방법인 CLIP[2]에 사용되는 contrastive learning은 representation learning에서 자주 사용
  - contrastive learning : positive pair와 negative pair를 정의하고, positive pair 간의 임베딩은 가까워지도록, negative pair 간의 임베딩은 멀어지도록 학습하는 방법
- 이후 SLIP[3], DeCLIP[4]과 같은 논문에서는 더욱 효율적인 학습을 위해 CLIP에 self-supervision loss를 추가
- 이미지-이미지 쌍과 같은 도메인 내(intra-domain) 쌍과 이미지-텍스트 쌍과 같은 도메인 간(inter-domain) 쌍에 대한 contrastive loss가 분리된 공간에서 독립적으로 정의된다는 한계 존재



(a) SSL (SimCLR) (b) CLIP (c) SLIP

(d) DeCLIP (e) UniCLIP – Ours

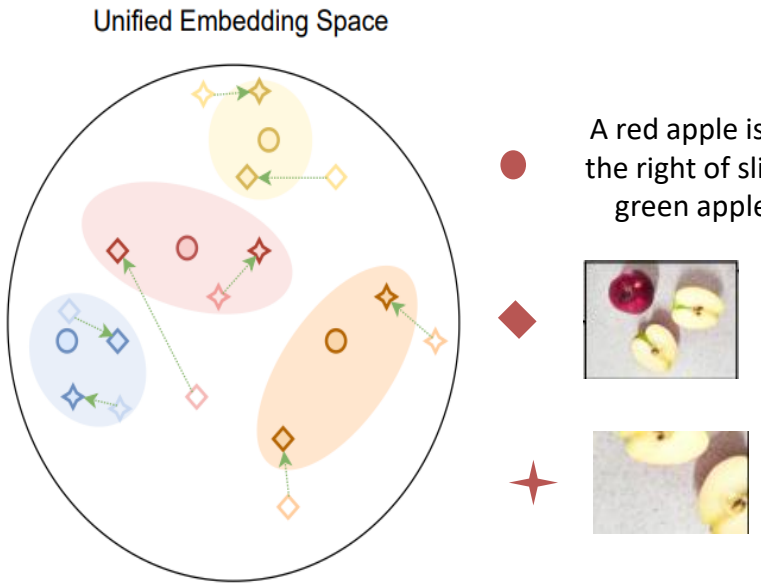
Introduction

- 목표 : 모든 가능한 intra-domain 및 inter-domain 간 대조 학습을 동일한 단일 통합 임베딩 공간에서 정의하는 이미지-텍스트 대조적 사전 학습 프레임워크를 구축
- 문제 : augmentation으로 인한 이미지-텍스트 misalignment 초래

  Ex1) Flip 적용 시, 오른쪽에 있는 것이 아니게 됨

  Ex2) grayscale 적용 시, 모델이 빨간색 사과인것을 알 수 없음

  Ex3) Crop 적용 시, 일부분만 보이기에 전체 그림에 대한 설명과 다르게 됨



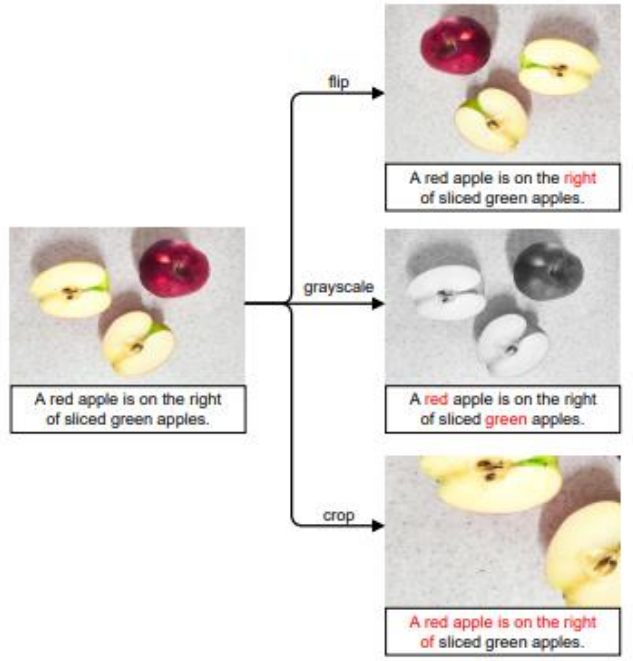
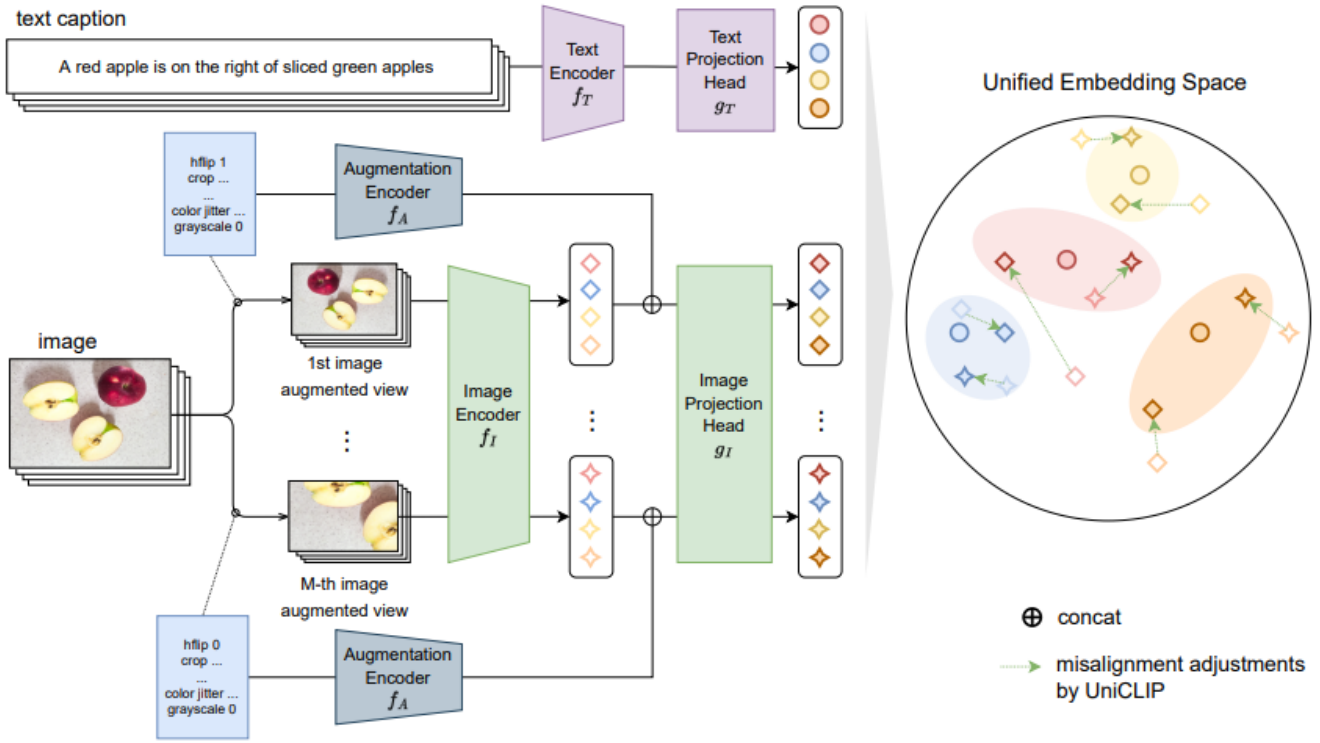


Figure 1: Image–text misalignments caused by data augmentations. The misaligned texts are highlighted in red (best viewed in color).

Framework Overview

- misalignment 문제를 해결하기 위해 UniCLIP에서는 augmentation의 정보를 임베딩에 반영할 수 있는 새로운 UniCLIP(Unified framework for Contrastive Language–Image Pretraining) 도입
- 구조는 크게 1) augmentation-aware feature embedding, 2) MP-NCE loss 3) domain dependent similarity measure 로 이루어짐



Augmentation-aware feature embedding

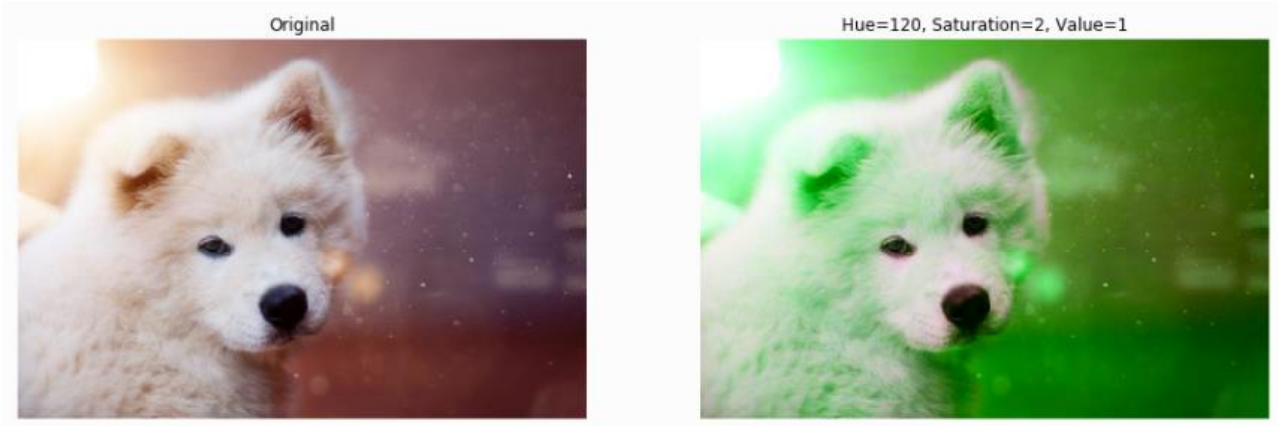
Augmentation Encoder

- 이미지에 어떤 종류의 augmentation이 얼마나 적용되었는지에 대한 정보를 제공하는 단계
- Augmentation 정보는 11차원의 벡터로 제공
  - ① RandomResizedCrop(x,y,w,h)

Augmentation-aware feature embedding

Augmentation Encoder

- 이미지에 어떤 종류의 augmentation이 얼마나 적용되었는지에 대한 정보를 제공하는 단계
- Augmentation 정보는 11차원의 벡터로 제공
  - ① RandomResizedCrop(x,y,w,h)
  - ② Color Jitter : (brightness, contrast, saturation)



color jitter은 이미지 data augmentation 기법의 하나로, 이미지의 Lightness, Hue 그리고 saturation 등을 임의로 변형

Augmentation-aware feature embedding

Augmentation Encoder

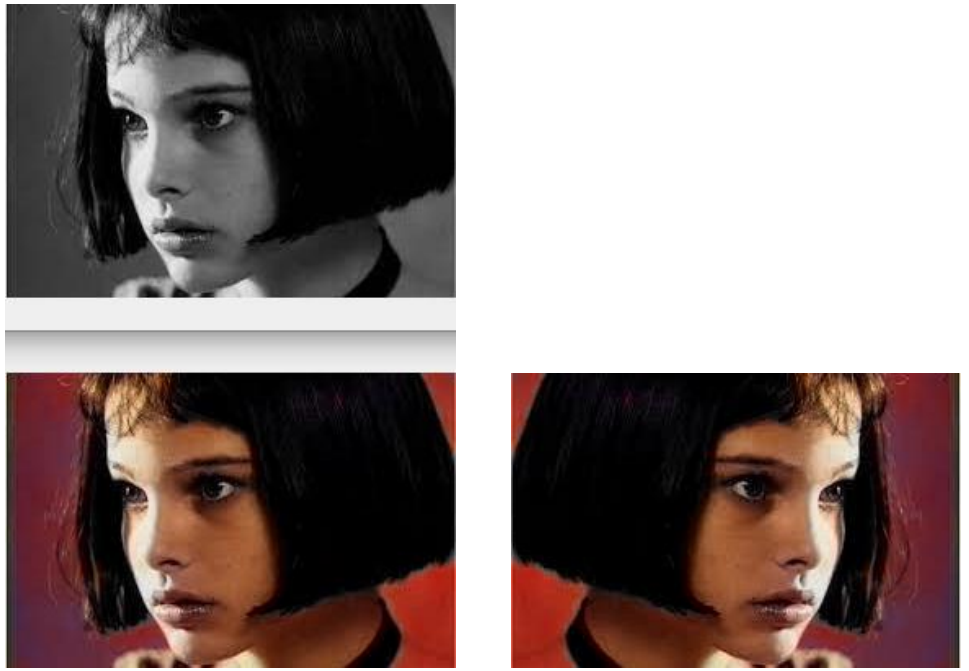
- 이미지에 어떤 종류의 augmentation이 얼마나 적용되었는지에 대한 정보를 제공하는 단계
- Augmentation 정보는 11차원의 벡터로 제공
  - ① RandomResizedCrop(x,y,w,h)
  - ② Color Jitter : (brightness, contrast, saturation)
  - ③ Gaussian Blur : (StDev)

가우시안 분포(Gaussian distribution) 함수를 근사하여 생성한 필터 마스크를 사용하는 필터링 기법

Augmentation-aware feature embedding

Augmentation Encoder

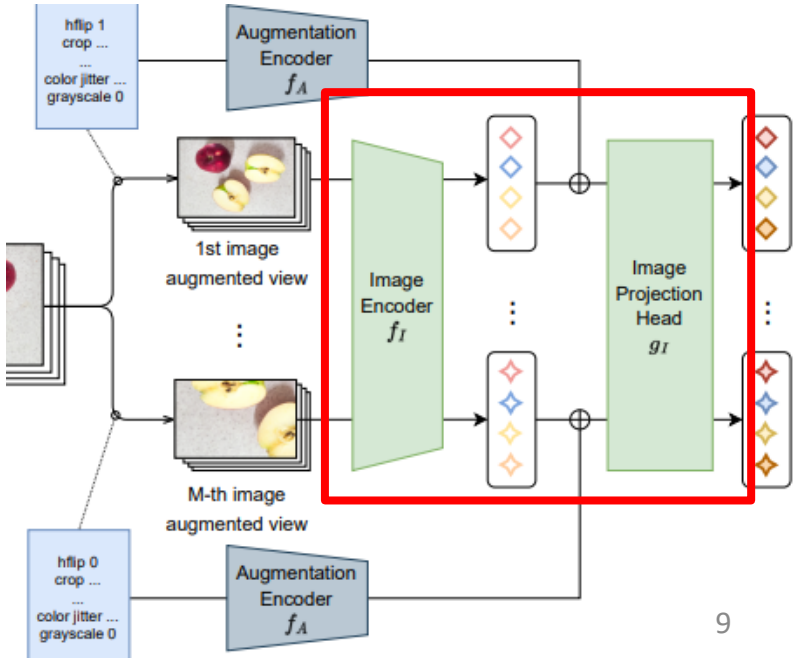
- 이미지에 어떤 종류의 augmentation이 얼마나 적용되었는지에 대한 정보를 제공하는 단계
- Augmentation 정보는 11차원의 벡터로 제공
  - ① RandomResizedCrop(x,y,w,h)
  - ② Color Jitter : (brightness, contrast, saturation)
  - ③ Gaussian Blur : (StDev)
  - ④ Horizontal Flip : (0 or 1)
  - ⑤ Grayscale Convert : (0 or 1)



Augmentation-aware feature embedding

Image Encoder & Image Projection Head

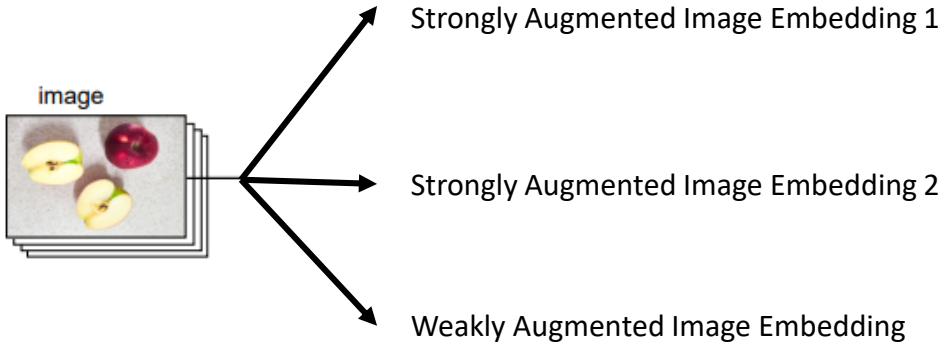
- 증강 정보 : [x,y,w,h,brightness,contrast,saturation,stdev,flip,grayscale]
- augmentation encoder를 통해 인코딩되고, 임베딩된 augmentation과 이미지 feature들을 concat해 projection head를 통과하여 최종 임베딩 feature 추출
- 원본 이미지와 augmented 이미지는 projection head g_I에만 positive pair라고 정보를 주기에, Image Encoder f_I는 robust한 generalization ability를 가지도록 학습이 됨

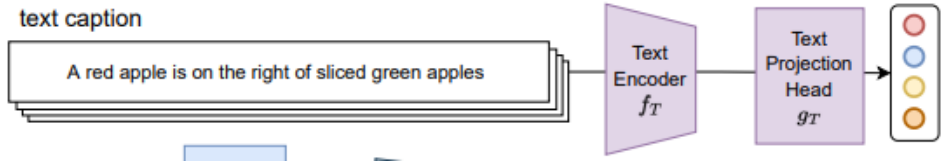


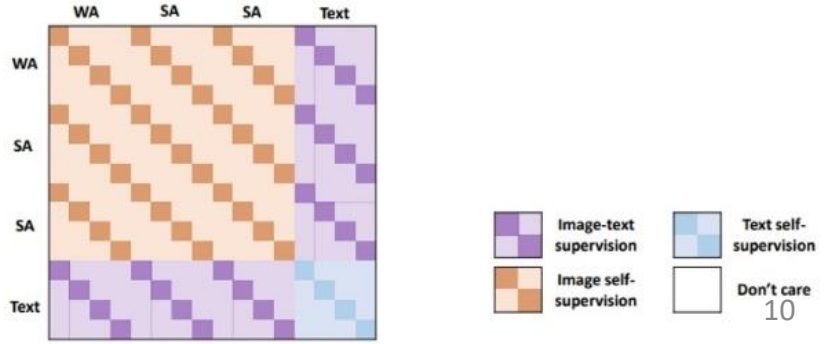
Augmentation-aware feature embedding

Text Encoder & Text Projection Head

- 논문에서는 1개의 image-text pair 인스턴스에 대해 총 3가지의 임의의 augmentation을 도입
- Prompt(text)를 augmentation 하지 않는 이유는 성능에 큰 영향을 끼치지 않음
- 모든 Image-Image 및 Image-text pair간의 동일한 공간에 mapping 하기 위해 이미지-이미지, 이미지-텍스트, 텍스트-텍스트 사이에서의 유사도를 학습







MP-NCE Loss

- 기존의 contrastive learning과 달리, 하나의 통합된 공간에서 여러 도메인의 임베딩을 비교하는 UniCLIP에서는 한 개보다 많은 positive 샘플이 존재
- 또한, 같은 모달(이미지-이미지, 텍스트-텍스) 사이에서의 학습이 멀티 모달(이미지-text) 간의 학습보다 쉬운 차이점 때문에, 학습에 난이도를 고려하여 loss식을 도입해야하는 필요성 존재
- 여러 개의 positive 샘플이 존재하는 상황에서도 InfoNCE loss를 적용할 수 있도록 확장한 MP-NCE (Multi-Positive NCE) loss를 제안

$$\mathcal{L}_i^{\text{MP-NCE}} = \mathbb{E}_{p \in P_i \cup \{i\}} \left[-w_{\mathcal{D}(i,p)} \log \frac{s_{i,p}}{s_{i,p} + \sum_{n \in N_i} s_{i,n}} \right]$$

$$P_i = \{j | (z_i, z_j) \text{ is a positive pair and } j \neq i\}$$
$$N_i = \{j | (z_i, z_j) \text{ is a negative pair}\}$$

z_i : i번째 임베딩
$S_{i,p}$: z_i 의 positive pair 사이의 거리

I번째 임베딩과 해당 임베딩과 positive pair 사이의 유사도가 높아지도록

MP-NCE Loss

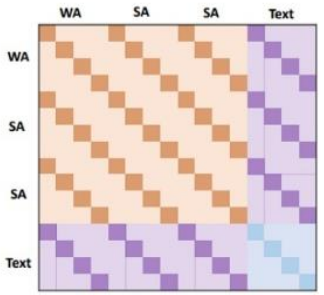
- 배치 내의 각 positive pair에 대한 InfoNCE loss의 평균을 취한 형태
- Inter 및 Intra domain 간 loss의 밸런스를 맞춰줄 수 있는 hyperparameter인 w 도입
- 예를 들어 배치 내 4개의 인스턴스에 대해 weak 및 strong augmentation에 대한 Image-Image pair의 loss는 1/9, image-text pair loss는 1/6, text-text pair 간 loss는 1/1로 weight를 regularize

도메인 별 균형 하이퍼파라미터

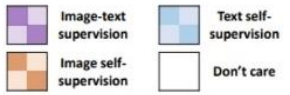
$$\mathcal{L}_i^{\text{MP-NCE}} = \mathbb{E}_{p \in P_i \cup \{i\}} \left[-w_{\mathcal{D}(i,p)} \log \frac{s_{i,p}}{s_{i,p} + \sum_{n \in N_i} s_{i,n}} \right]$$

$$P_i = \{j | (z_i, z_j) \text{ is a positive pair and } j \neq i\}$$
$$N_i = \{j | (z_i, z_j) \text{ is a negative pair}\}$$

z_i : i번째 임베딩
$S_{i,p}$: z_i 의 positive pair 사이의 거리



$w_D(i,p)$



Domain dependent similarity measure

- 같은 모달(이미지-이미지, 텍스트-텍스트) 사이에서의 학습이 멀티 모달(이미지-text) 인지에 따라 embedding이 다르기에, 서로 다른 similarity를 도입해야 함
- 본 논문은 도메인 별로 다르게 적용할 temperature T_D 와 offset b_D를 도입
- τ_D ($\tau_{\text{image-image}}$, $\tau_{\text{image-test}}$, $\tau_{\text{text-text}}$)
  - Cosine similarity는 -1~1 사이 이기에, score가 가질 수 있는 값의 범위를 넓혀 주기 위해 1보다 작은 temperature τ가 존재
  - 각 도메인간 특성이 달라 적절한 τ가 서로 다르기에 각 도메인 별 적절한 τ 각각 3개 존재
- offset_D ($\text{offset}_{\text{image-image}}$, $\text{offset}_{\text{image-test}}$, $\text{offset}_{\text{text-text}}$)
  - Contrastive learning 에서 positive pair인지 negative pair 인지를 유사도에 따라 결정하기 위한 임계값으로 offset 사용
  - 같은 도메인 상에 positive pair를 찾는 것이 다른 도메인 상에 positive pari를 찾는 것보다 쉽기에 서로 다른 3개의 offset 존재
- UniCLIP은 각 도메인 D마다 서로 다른 적정 temperature와 offset을 학습

$$s_{i,j} = \exp\left(\frac{1}{\tau} \cdot \frac{z_i^\top z_j}{\|z_i\|\|z_j\|}\right)$$

기존 contrastive learning에서의 similarity score

$$s_{i,j} = \exp\left(\frac{1}{\tau_{\mathcal{D}(i,j)}}\left(\frac{z_i^\top z_j}{\|z_i\|\|z_j\|} - b_{\mathcal{D}(i,j)}\right)\right)$$

UniCLIP에서 제안된 domain-dependent similarity score

UniCLIP Algorithm

- negative pair : 배치 내 서로 다른 인스턴스들의 조합
- Positive pair : 같은 원본 데이터에서 생성된 데이터
- $D_{(i,j)}$: 도메인에 따라 매핑(image-image : 1, image-text : 2, text-text : 3)

※ update networks, temperature, offset to minimize L 이라 적혀 있는데 temperature 및 offset 은 스칼라 값의 하이퍼파라미터인데 어떻게 자동으로 업데이트가 되는지 설명이 안되어있음

Algorithm A UniCLIP

Input: image encoder f_I, text encoder f_T, image projection head g_I, text projection head g_T, augmentation encoder f_A, batch size N, temperature $\tau \in \mathbb{R}^3$, offset $b \in \mathbb{R}^3$, weak augmentation distribution p_{wa}, strong augmentation distribution p_{sa}

for sampled mini-batch $\{(x_k^I, x_k^T)\}_{k=1}^N$ **do**
  for all $k \in \{1, \dots, N\}$ **do**
    draw augmentation instructions $\mathcal{A}_1 \sim p_{wa}, \mathcal{A}_2 \sim p_{sa}, \mathcal{A}_3 \sim p_{sa}$
    $z_k = g_I(f_I(\mathcal{A}_1(x_k^I)), f_A(\mathcal{A}_1))$
    $z_{k+N} = g_I(f_I(\mathcal{A}_2(x_k^I)), f_A(\mathcal{A}_2))$
    $z_{k+2N} = g_I(f_I(\mathcal{A}_3(x_k^I)), f_A(\mathcal{A}_3))$
    $z_{k+3N} = g_T(f_T(x_k^T))$
  end for
  for all $i \in \{1, \dots, 4N\}$ **do**
    for all $j \in \{1, \dots, 4N\}$ **do**
      $$\mathcal{D}(i,j) = \begin{cases} 1, & \text{if } i \le 3N \text{ and } j \le 3N \\ 3, & \text{if } i > 3N \text{ and } j > 3N \\ 2, & \text{otherwise} \end{cases}$$
      $$s_{i,j} = \exp\left(\frac{1}{\tau_{\mathcal{D}(i,j)}}\left(\frac{z_i^\top z_j}{\|z_i\|\|z_j\|} - b_{\mathcal{D}(i,j)}\right)\right)$$
    end for
    $P_i = \{j \in \{1, \dots, 4N\} \setminus \{i\} | (j-i)/N \in \mathbb{Z}\}$
    $N_i = \{1, \dots, 4N\} \setminus P_i \setminus \{i\}$
    $w = (1/9, 1/6, 1)$
    $\mathcal{L}_i = \mathbb{E}_{p \in P_i \cup \{i\}}\left[-w_{\mathcal{D}(i,p)} \log \frac{s_{i,p}}{s_{i,p} + \sum_{n \in N_i} s_{i,n}}\right]$
  end for
  $\mathcal{L} = \frac{1}{4N} \sum_{i=1}^{4N} \mathcal{L}_i$
  update networks, temperature, offset to minimize $\mathcal{L}$
end for

Experiment

- 다른 베이스라인에 비해 제안하는 방법이 zero-shot 및 linear proving에 대해 대부분 좋은 성능을 보임
- 전체 네트워크에 gradient가 업데이트 되게끔 하는 fine-tunning 진행시에도 가장 좋은 성능을 보임

Method	Pre-train dataset	Pets	CIFAR-10	CIFAR-100	SUN397	Food-101	Flowers	Cars	Caltech-101	Aircraft	DTD	ImageNet	**Average**
Zero-shot classification:													
CLIP-ViT-B/32	YFCC15M	19.4	62.3	33.6	40.2	33.7	6.3	2.1	55.4	1.4	16.9	31.3	27.5
SLIP-ViT-B/32	YFCC15M	28.3	72.2	45.3	45.1	44.7	6.8	2.9	65.9	1.9	21.8	38.3	33.9
DeCLIP-ViT-B/32	YFCC15M	30.2	72.1	39.7	**51.6**	46.9	7.1	**3.9**	70.1	2.5	**24.2**	41.2	35.4
UniCLIP-ViT-B/32	YFCC15M	**32.5**	**78.6**	**47.2**	50.4	**48.7**	**8.1**	3.4	**73.0**	**2.8**	23.3	**42.8**	**37.3**
DeCLIP-ResNet50† [21]	Open30M	-	-	-	-	-	-	-	-	-	-	49.3	-
UniCLIP-ViT-B/32	Open30M	69.2	87.8	56.5	61.1	64.6	8.0	19.5	84.0	4.7	36.6	**54.2**	49.7
Linear probing:													
CLIP-ViT-B/32	YFCC15M	71.2	89.2	72.1	70.1	71.4	93.2	34.9	84.3	29.7	60.9	61.1	67.1
SLIP-ViT-B/32	YFCC15M	75.4	90.5	75.3	73.5	77.1	96.1	43.0	87.2	34.1	71.1	68.1	71.9
DeCLIP-ViT-B/32	YFCC15M	76.5	88.6	71.6	75.9	79.3	96.7	42.6	88.0	32.6	69.1	69.2	71.8
UniCLIP-ViT-B/32	YFCC15M	**83.1**	**92.5**	**78.2**	**77.0**	**81.3**	**97.1**	**49.8**	**88.9**	**36.2**	**72.8**	**70.8**	**75.2**
UniCLIP-ViT-B/32	Open30M	85.4	95.1	81.5	79.2	84.4	97.3	67.3	91.1	39.0	77.2	74.0	79.1

Table 2: ImageNet-1k fine-tuning accuracy for the models pre-trained on YFCC15M.

Method	Accuracy
CLIP-ViT-B/32	72.27
SLIP-ViT-B/32	75.64
DeCLIP-ViT-B/32	74.34
UniCLIP-ViT-B/32	**76.54**

Experiment

- ImageNet-1k에 대한 제로샷 성능은 다음과 같음

Table F: ImageNet-1k zero-shot accuracy with varying the number of image views and text views.

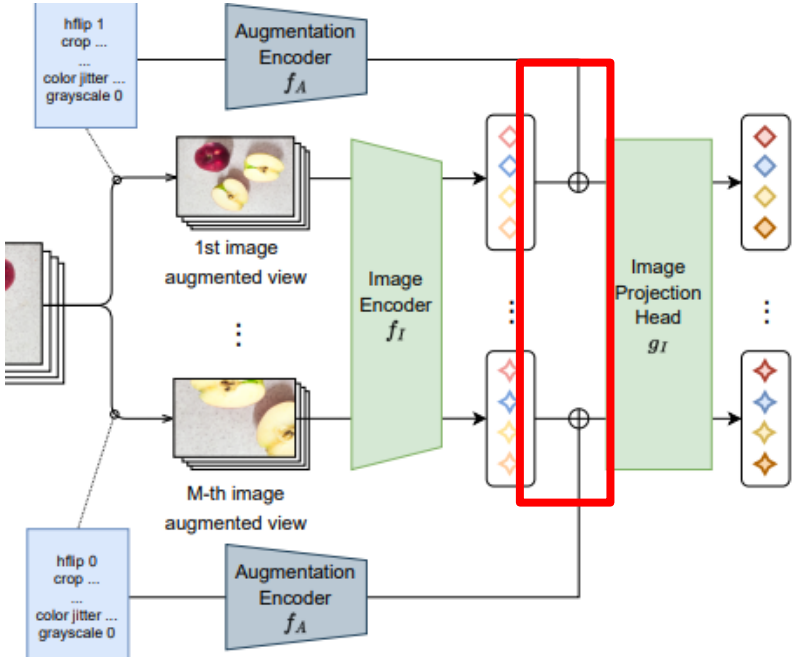
# of image views	# of text views	# of original pairs	Accuracy
1	1	192	21.80
2	1	128	25.54
2	2	96	24.60
3	1	96	**27.67**
3	2	72	24.57
4	1	72	**28.25**

Ablation Study

- Projection head의 모든 경우에서 projection head에 augmentation 정보를 알려주는 것이 더 좋은 정확도를 보임
- Projection head에서 MLP는 overfitting 현상이 발생하였으며, Resblock으로 head를 구성하는 것이 좋게 작용

(a) **Image projection head types.** One weak and two strong image augmentations are used.

Augmentation embedding	Head type	Accuracy
✗	MLP 3 layers	24.01
	MLP 6 layers	23.62
	1 ResBlock	**24.76**
	3 ResBlocks	24.46
✓	Linear layer	24.68
	MLP 3 layers	24.54
	MLP 6 layers	24.15
	1 ResBlock	27.67
	3 ResBlocks	**27.84**



Ablation Study

- 실험 결과 strong augmentation만 사용했을 때, 이미지가 원본에 비해 크게 바뀔 확률이 높기에 원래 text와 misalignment가 크게 작용
- 1개의 weak, 2개의 strong을 사용하는 것이 가장 효과적

(b) **Augmentation configurations.** 1-ResBlock head is used for no augmentation embedding config and 3-ResBlock head is used with augmentation embedding.

Augmentation embedding	Augmentation	Accuracy
✗	3 weak	24.49
	1 weak, 2 strong	**24.76**
	3 strong	22.60
✓	3 weak	23.40
	1 weak, 2 strong	**27.84**
	3 strong	26.43

Conclusion

- UniCLIP은 기존 SimCLR (이미지-이미지)와 CLIP (이미지-텍스트)에서 독립적으로 사용되던 contrastive loss를 단일 임베딩 공간으로 통합한 contrastive learning 프레임워크
- 하나의 통합된 임베딩 공간에서의 contrastive learning을 위해 UniCLIP은 architecture, contrastive loss, similarity score의 세 부분에서 새로운 문제를 해결하고 기존 contrastive learning을 확장
- 제안하는 방법은 computational overhead가 크지 않고 적용하기도 쉬워 그 활용성이 큼
- 또한, augmentation misalignment 문제를 완화하기에 데이터가 적어 강한 augmentation이 필요한 상황이거나 augmentation misalignment 문제에 취약한 데이터 세트의 학습 등에서는 UniCLIP의 활용이 더욱 중요

- 추후 연구로 조금 더 다양한 multi-modal dataset으로 확장하여 제안하는 방법의 성능 증명

Reference

1. Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.
2. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International Conference on Machine Learning. PMLR, 2021.
3. Mu, N., Kirillov, A., Wagner, D., & Xie, S. (2022, October). Slip: Self-supervision meets language-image pre-training. In European Conference on Computer Vision (pp. 529-544). Cham: Springer Nature Switzerland.
4. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., ... & Yan, J. (2021). Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208.
5. https://m.post.naver.com/viewer/postView.naver?volumeNo=34834542&memberNo=52249799