The Modality gap in Multi-modal Contrastive Learning

2026년 4월 3일

박현우

발표자 소개

- **박현우 (Hyunwoo Park)**
  - 고려대학교 산업경영공학과 석사과정 (2025.03 ~ Present)
  - Data Mining & Quality Analytics Labs. (김성범 교수님)
- **Research Interest**
  - Multimodal Learning
  - Out-of-distribution Detection
- **Contact**
  - phwnob20@korea.ac.kr

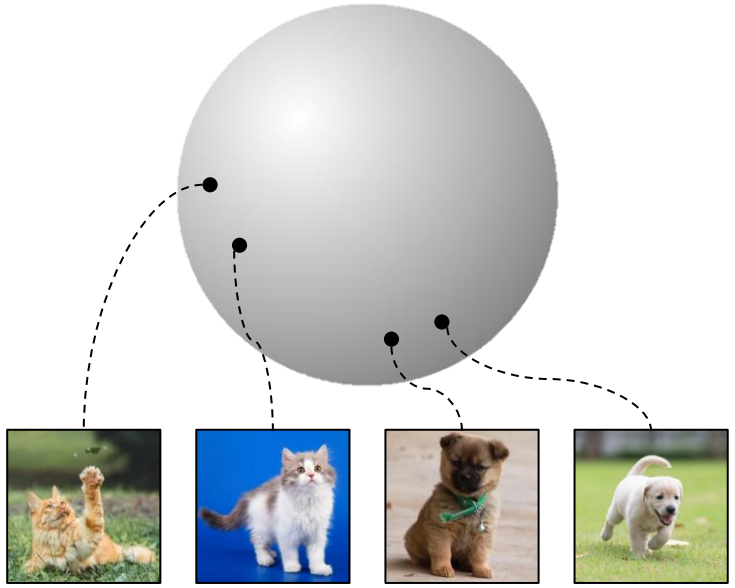
Seminar Outline

- Introduction
- Understanding Modality gap
- Conclusion

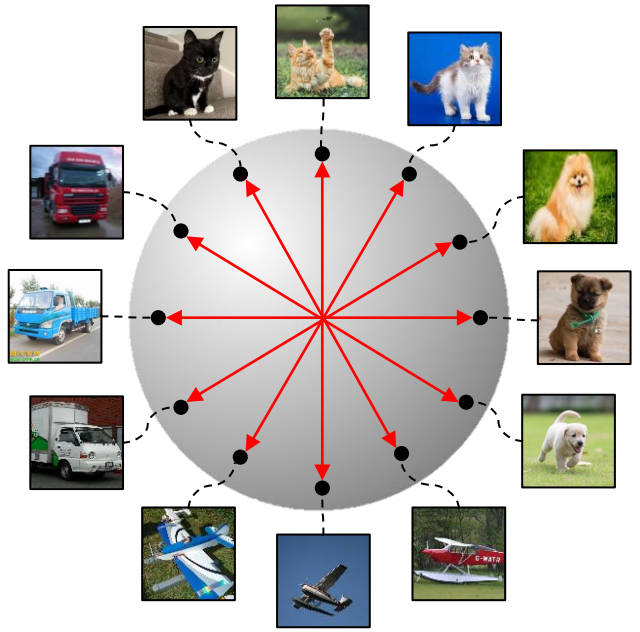
Introduction

Contrastive Learning

- Contrastive Learning은 Alignment와 Uniformity를 향상시키는 방향으로 훈련됨 [1]
- Alignment와 Uniformity가 균형을 잘 이룰수록, 모델이 좋은 representation을 갖게 됨



Alignment



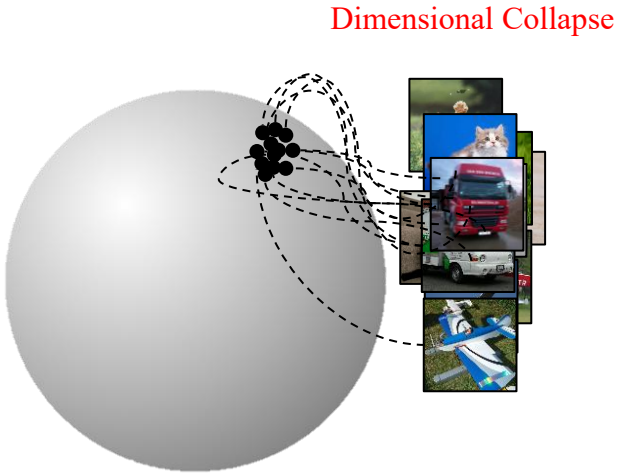
Uniformity

[1] Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. (2020 ICML)

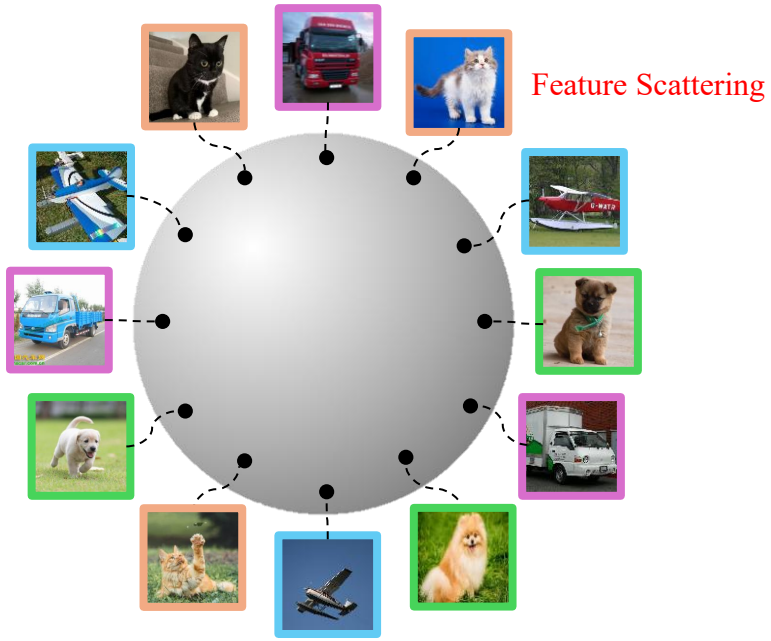
Introduction

Contrastive Learning

- Contrastive Learning은 Alignment와 Uniformity를 향상시키는 방향으로 훈련됨 [1]
- Alignment와 Uniformity가 균형을 잘 이룰수록, 모델이 좋은 representation을 갖게 됨



Alignment에 치우쳐져 학습된 경우



Uniformity에 치우쳐져 학습된 경우

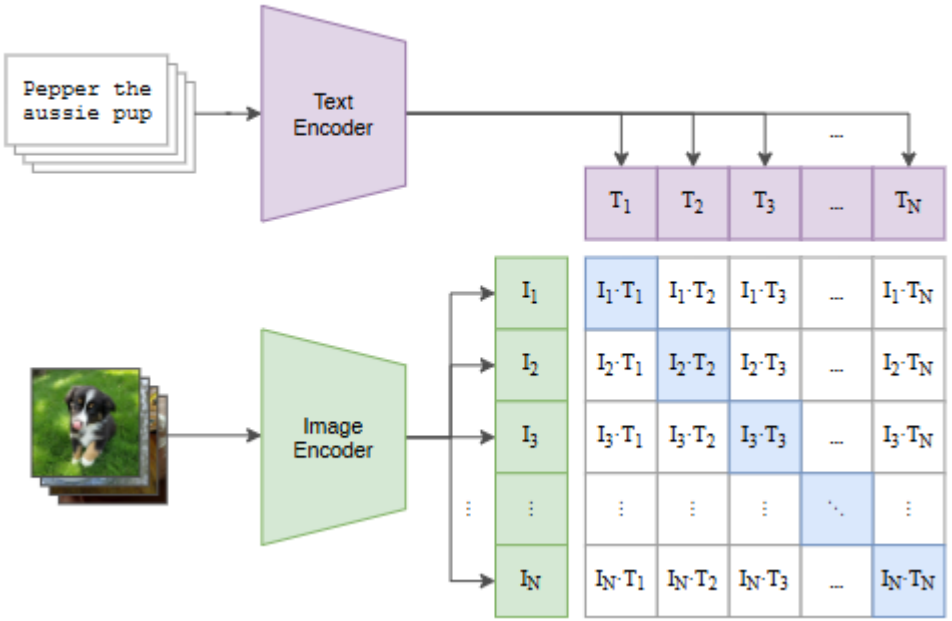
Alignment와 Uniformity가
균형을 이뤄 학습된 경우

[1] Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. (2020 ICML)

Introduction

CLIP (Contrastive Language-Image Pre-training)

- CLIP: Contrastive Language Image Pretraining [2]
  - Image-Text Pair에 대해 Contrastive Learning으로 학습한 Foundation Model
  - 대조학습이 멀티 모달리티의 피처 정렬에도 효과적임을 보여줌

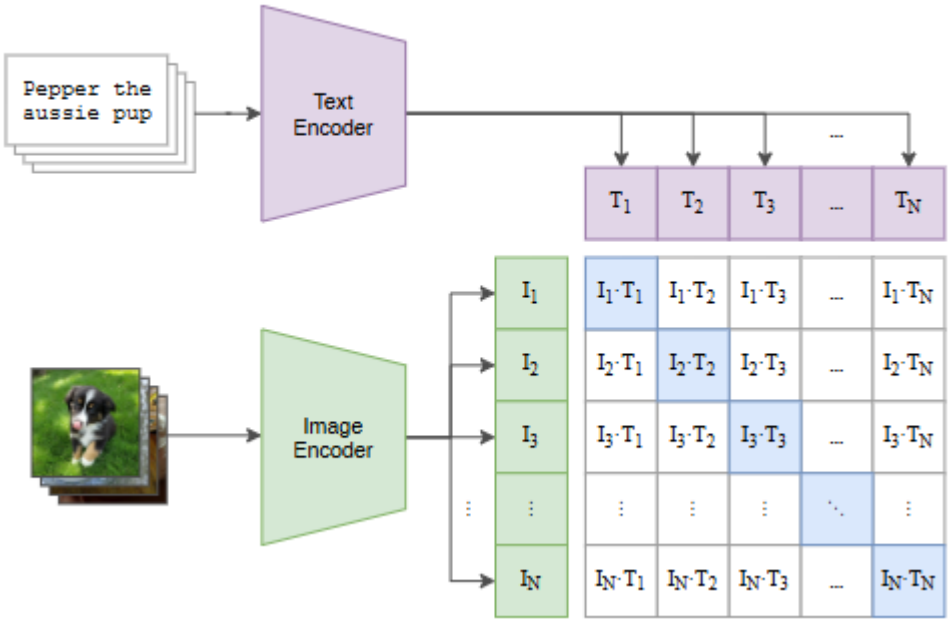


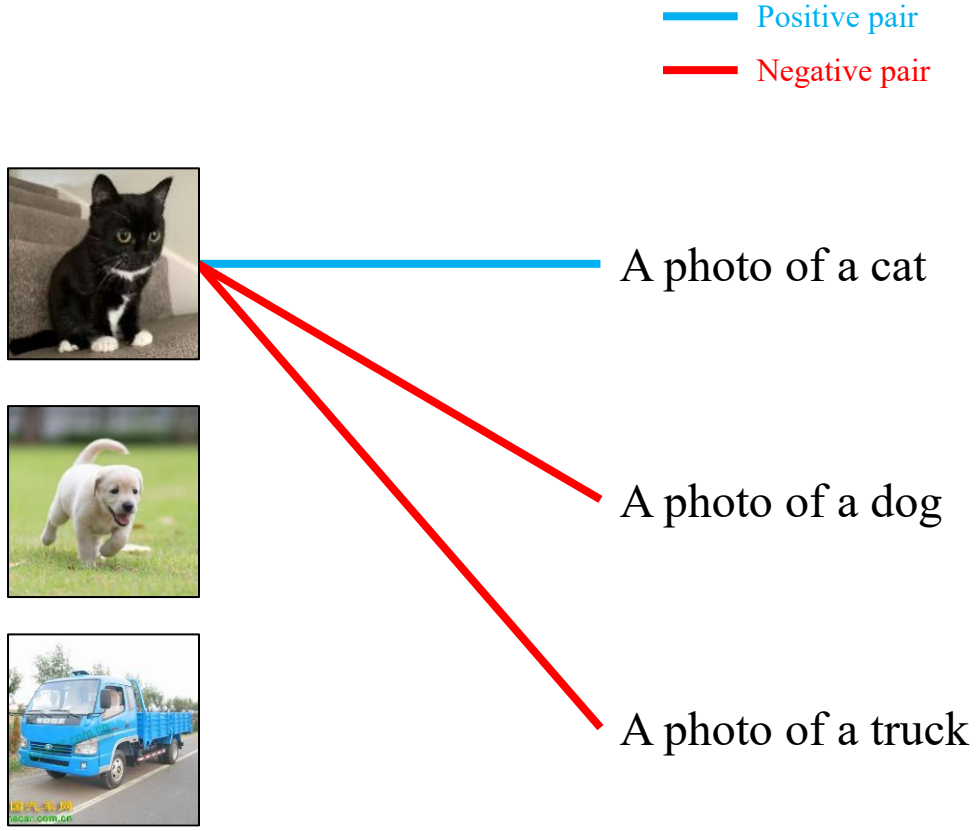
[2] Learning Transferable Visual Models From Natural Language Supervision. (2021 ICML)

Introduction

CLIP (Contrastive Language-Image Pre-training)

- CLIP: Contrastive Language Image Pretraining [2]
  - Image-Text Pair에 대해 Contrastive Learning으로 학습한 Foundation Model
  - 대조학습이 멀티 모달리티의 피처 정렬에도 효과적임을 보여줌



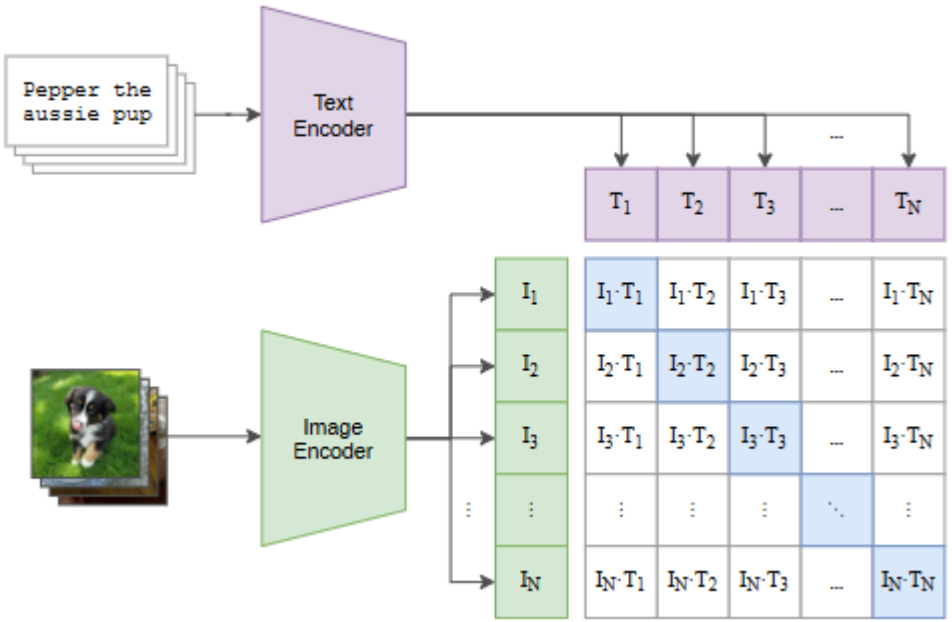


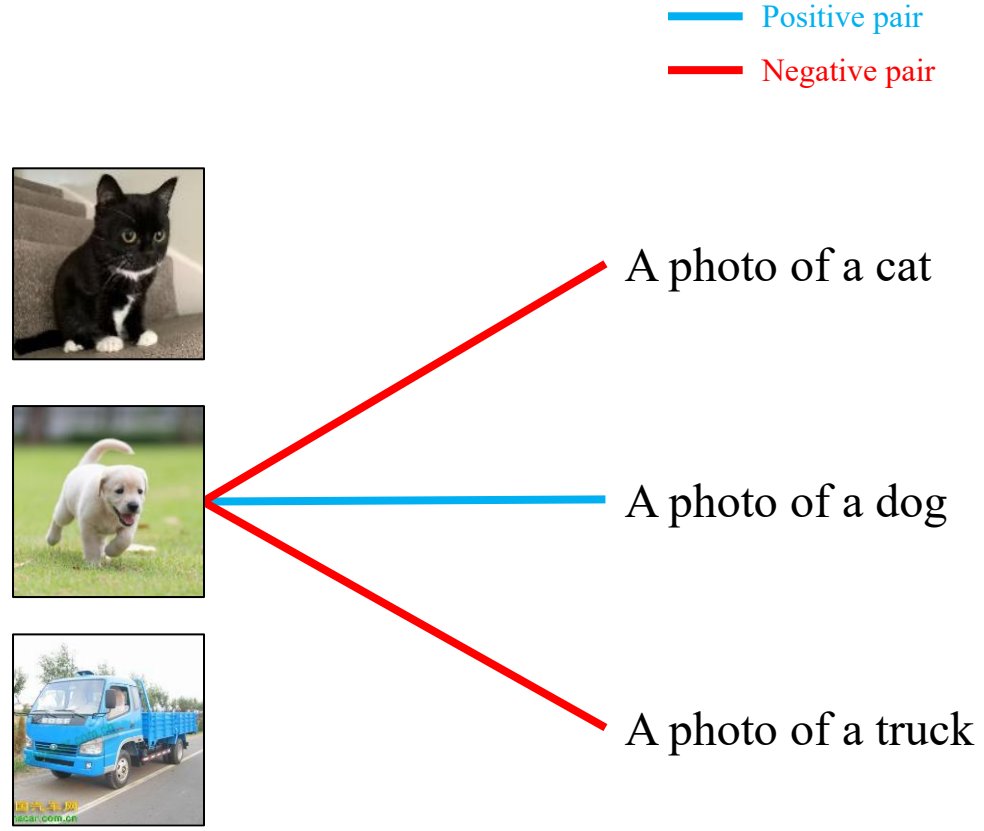
[2] Learning Transferable Visual Models From Natural Language Supervision. (2021 ICML)

Introduction

CLIP (Contrastive Language-Image Pre-training)

- CLIP: Contrastive Language Image Pretraining [2]
  - Image-Text Pair에 대해 Contrastive Learning으로 학습한 Foundation Model
  - 대조학습이 멀티 모달리티의 피처 정렬에도 효과적임을 보여줌



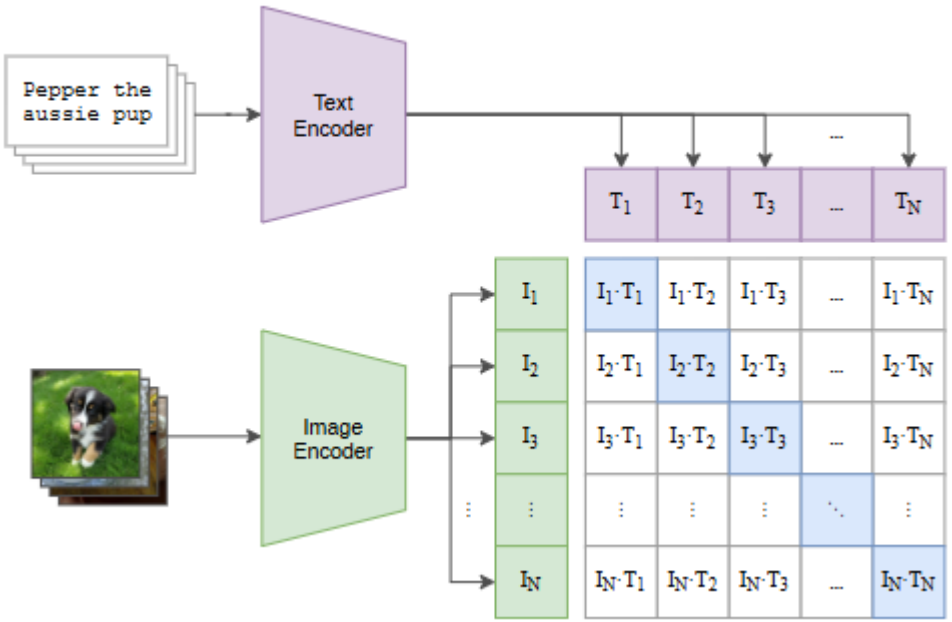


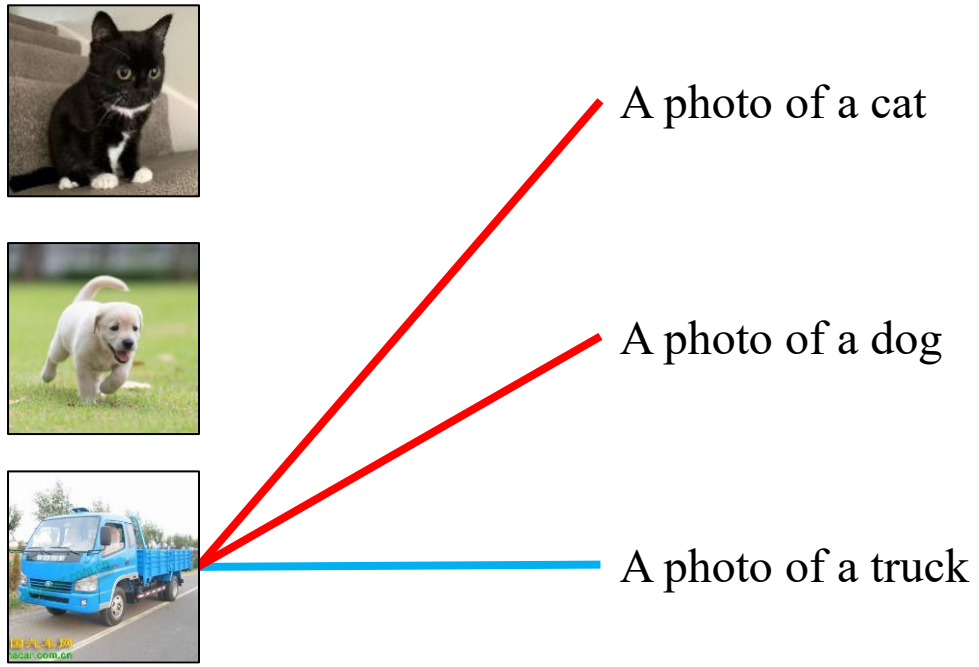
[2] Learning Transferable Visual Models From Natural Language Supervision. (2021 ICML)

Introduction

CLIP (Contrastive Language-Image Pre-training)

- CLIP: Contrastive Language Image Pretraining [2]
  - Image-Text Pair에 대해 Contrastive Learning으로 학습한 Foundation Model
  - 대조학습이 멀티 모달리티의 피처 정렬에도 효과적임을 보여줌



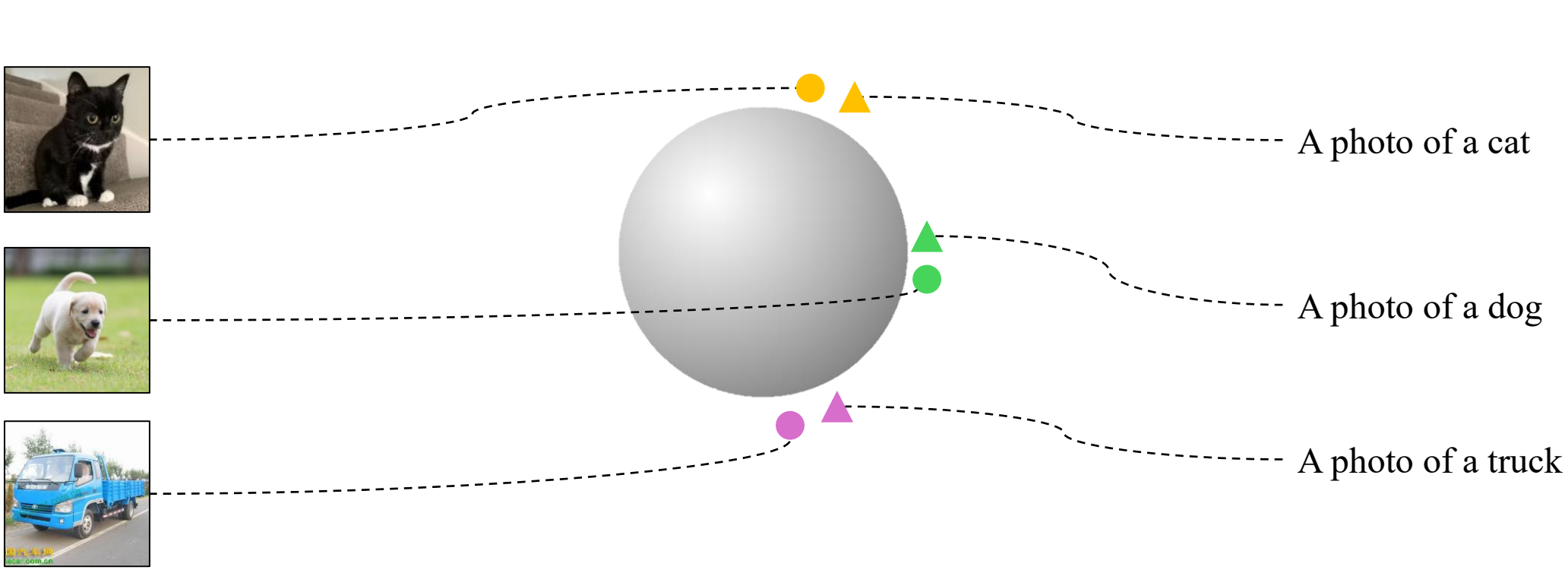


[2] Learning Transferable Visual Models From Natural Language Supervision. (2021 ICML)

Modality Gap

Expectation vs. Reality in Modality Alignment

- CLIP 모델에서 같은 의미를 갖는 Image-Text Pair에 대해 임베딩 공간에서 가까운 위치에 존재해야할 것임.
- 대조학습 기반의 멀티모달 모델에서 임베딩이 분리되어 존재하는 모달리티 갭 현상이 있음을 발견 [3]

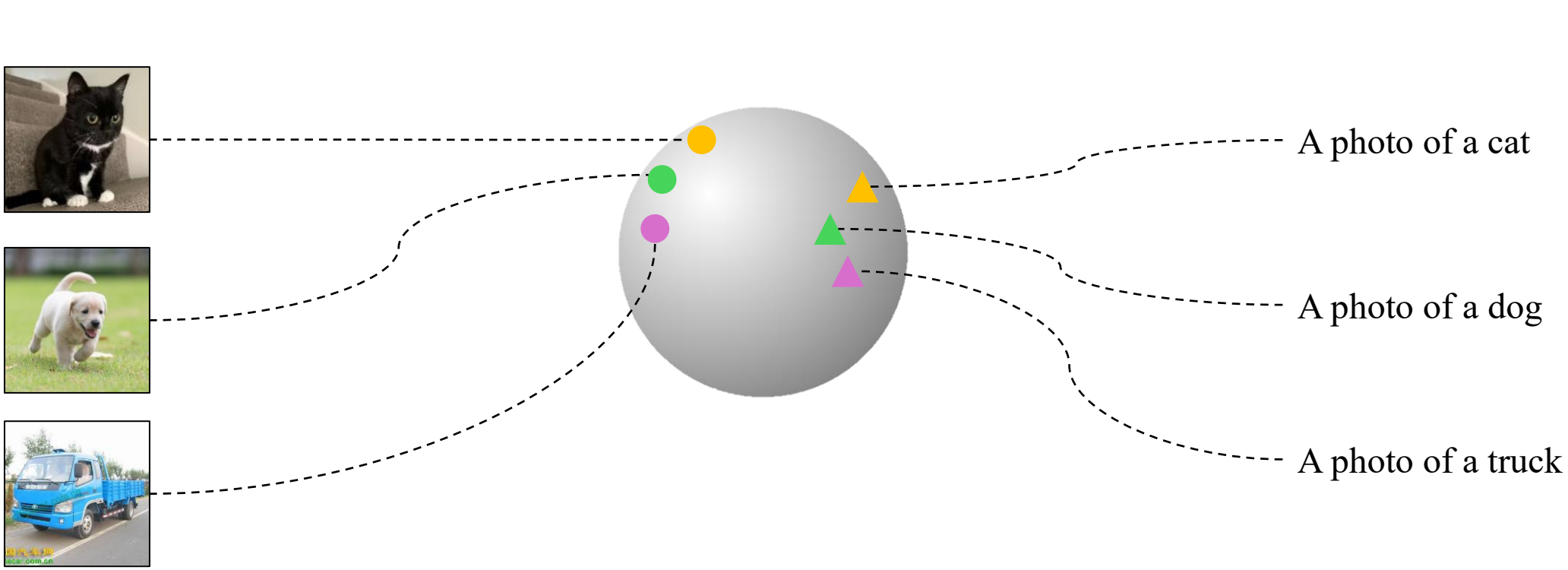


[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Modality Gap

Expectation vs. Reality in Modality Alignment

- CLIP 모델에서 같은 의미를 갖는 Image-Text Pair에 대해 임베딩 공간에서 가까운 위치에 존재해야할 것임.
- 대조학습 기반의 멀티모달 모델에서 임베딩이 분리되어 존재하는 모달리티 갭 현상이 있음을 발견 [3]



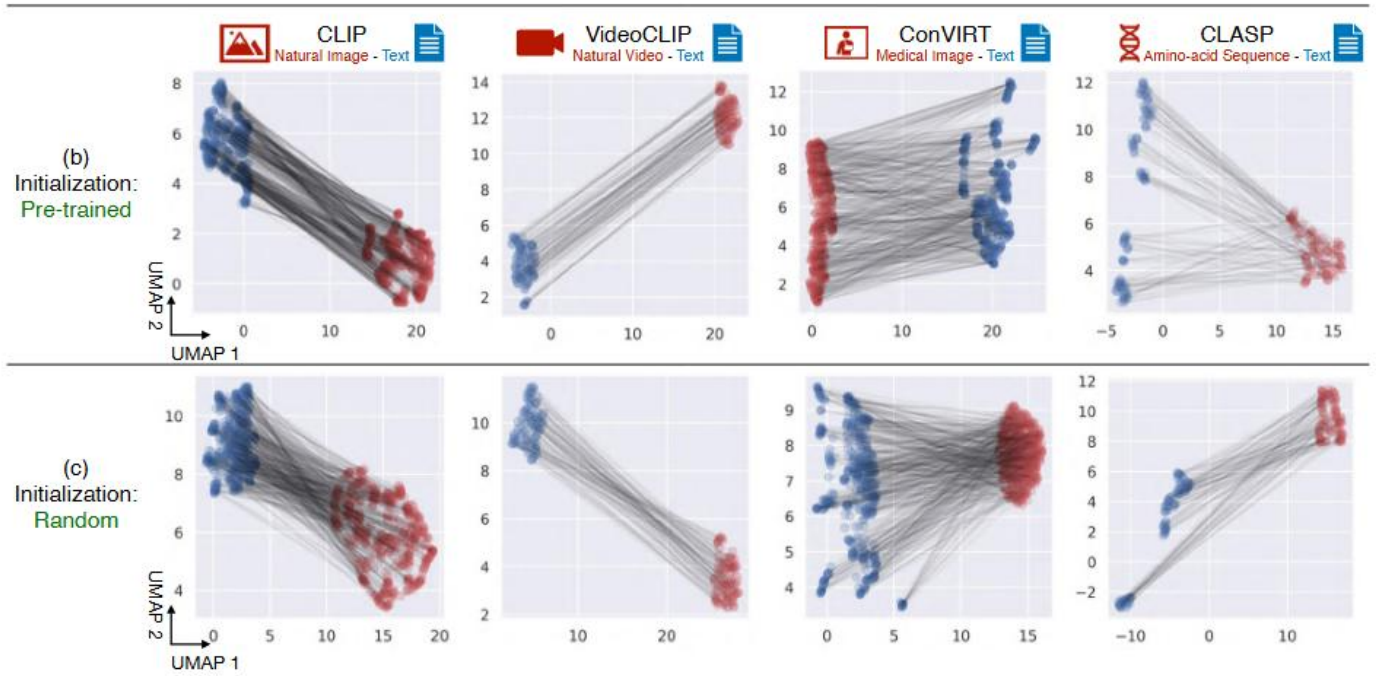
[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap

Modality Gap at random initialization

- 모달리티 갭은 초기화 단계와 최적화 과정의 특성이 결합되어 나타남 [3]
- 학습이 완료된 모델 뿐 아니라, 랜덤하게 초기화 된 모델에서도 갭은 존재함

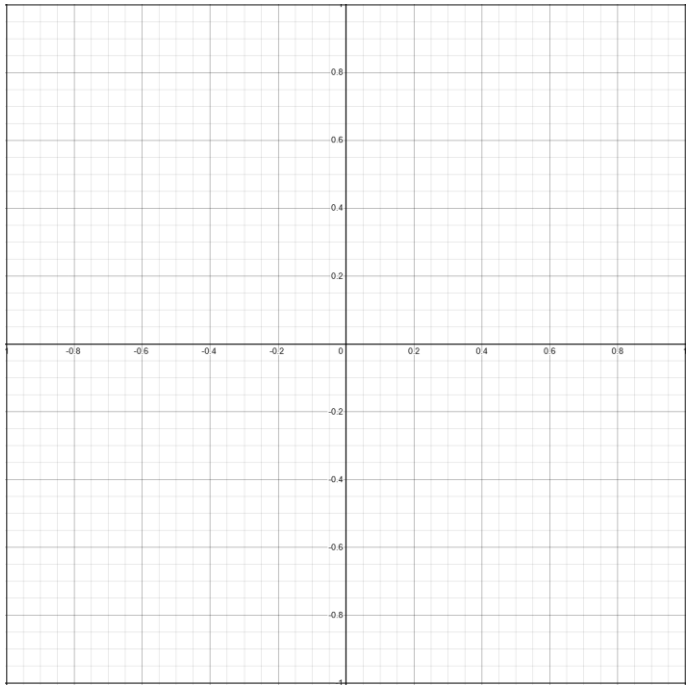


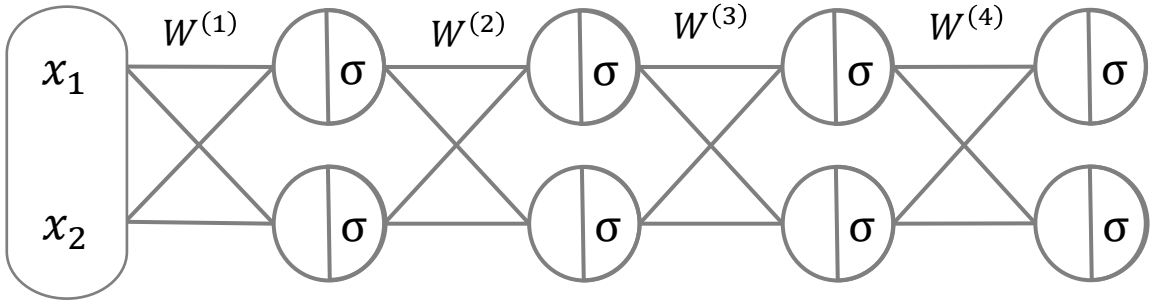
[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap

Cone Effect

- 랜덤 초기화된 모델에서 발생하는 갭은 딥러닝 모델 고유의 구조적 특성인 Cone Effect에 의해 발생함
- Cone Effect란, 데이터가 임베딩 공간 전체를 사용하지 않고 좁은 영역에만 집적으로 사용되는 현상을 의미함





$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

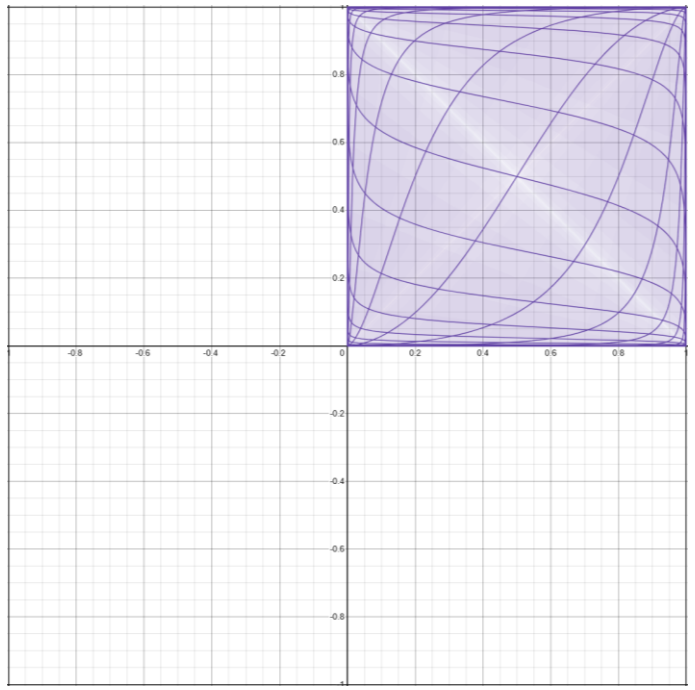
$$x = [x_1, x_2]^T \in \mathbb{R}^2$$

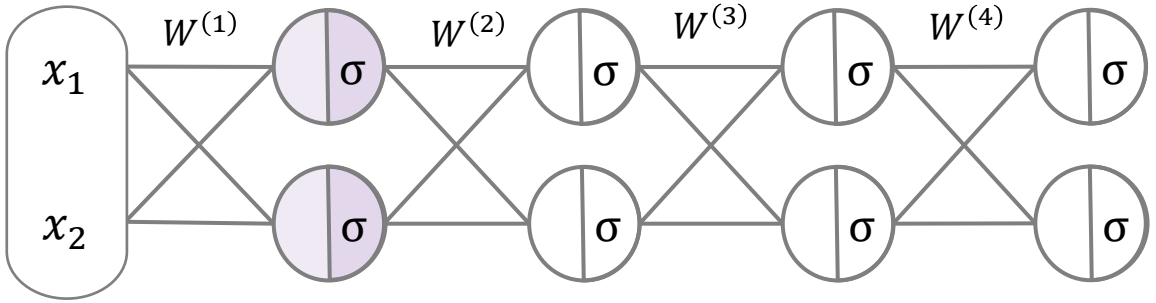
[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap

Cone Effect

- 랜덤 초기화된 모델에서 발생하는 갭은 딥러닝 모델 고유의 구조적 특성인 Cone Effect에 의해 발생함
- Cone Effect란, 데이터가 임베딩 공간 전체를 사용하지 않고 좁은 영역에만 집적으로 사용되는 현상을 의미함





$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\boldsymbol{h^{(1)} = \sigma(W^{(1)}x + b^{(1)})}$$

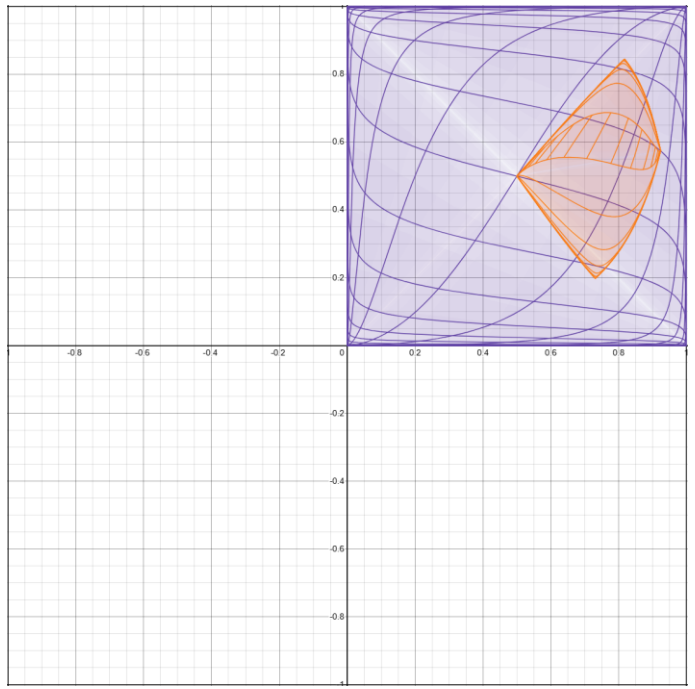
$$x = [x_1, x_2]^T \in \mathbb{R}^2$$

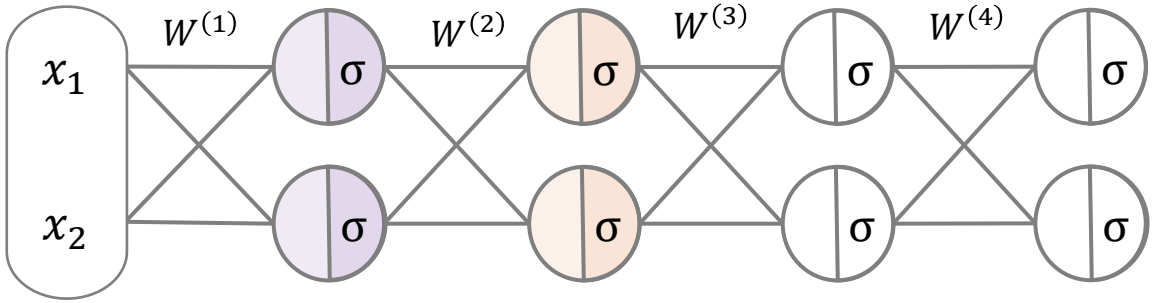
[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap

Cone Effect

- 랜덤 초기화된 모델에서 발생하는 갭은 딥러닝 모델 고유의 구조적 특성인 Cone Effect에 의해 발생함
- Cone Effect란, 데이터가 임베딩 공간 전체를 사용하지 않고 좁은 영역에만 집적으로 사용되는 현상을 의미함





$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\boldsymbol{h^{(2)} = \sigma(W^{(2)} h^{(1)} + b^{(2)})}$$

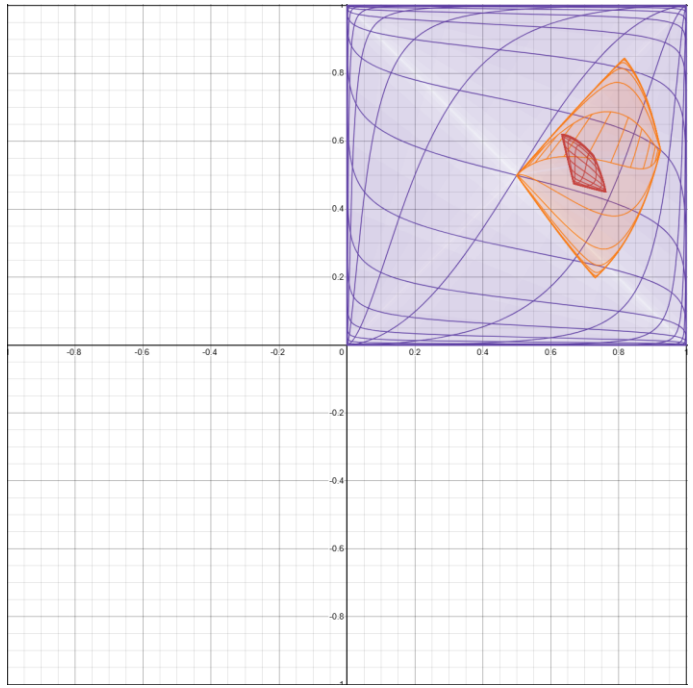
$$x = [x_1, x_2]^T \in \mathbb{R}^2$$

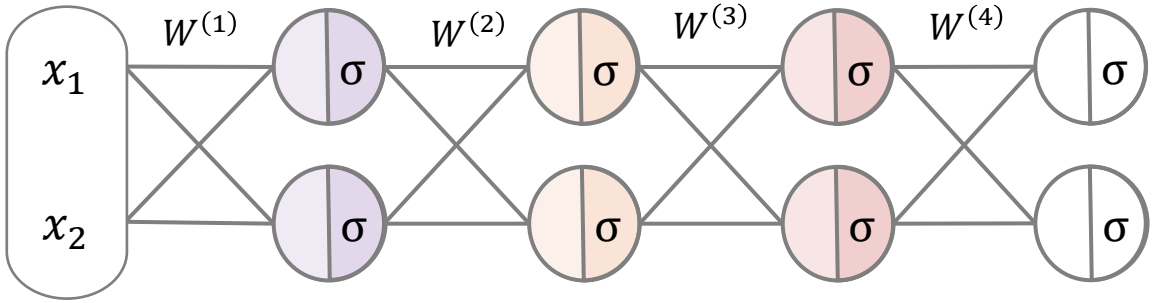
[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap

Cone Effect

- 랜덤 초기화된 모델에서 발생하는 갭은 딥러닝 모델 고유의 구조적 특성인 Cone Effect에 의해 발생함
- Cone Effect란, 데이터가 임베딩 공간 전체를 사용하지 않고 좁은 영역에만 집적으로 사용되는 현상을 의미함





$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\boldsymbol{h^{(3)} = \sigma(W^{(3)} h^{(2)} + b^{(3)})}$$

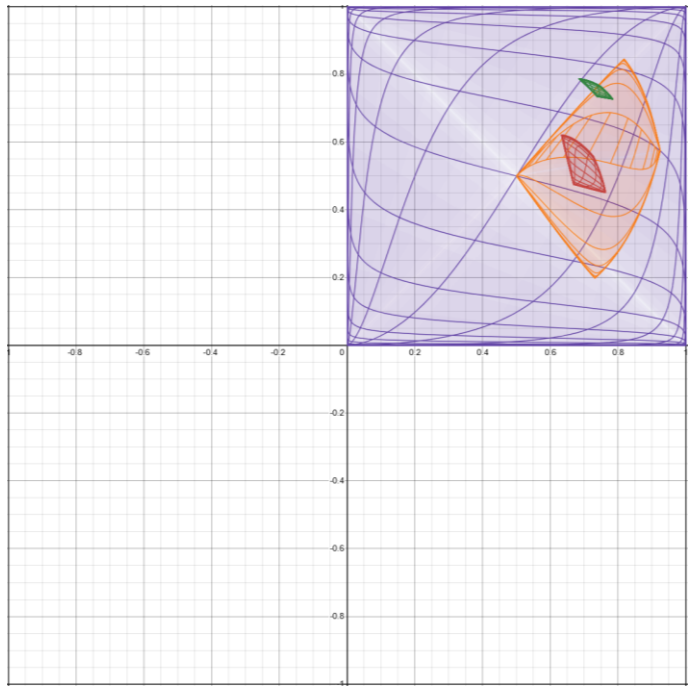
$$x = [x_1, x_2]^T \in \mathbb{R}^2$$

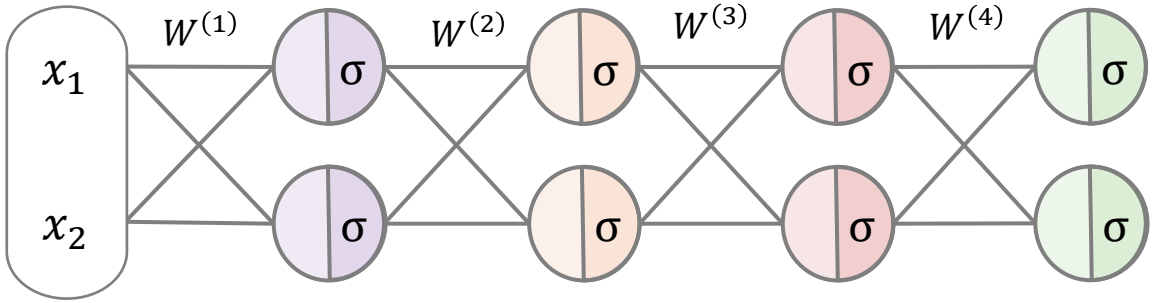
[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap

Cone Effect

- 랜덤 초기화된 모델에서 발생하는 갭은 딥러닝 모델 고유의 구조적 특성인 Cone Effect에 의해 발생함
- Cone Effect란, 데이터가 임베딩 공간 전체를 사용하지 않고 좁은 영역에만 집적으로 사용되는 현상을 의미함





$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\boldsymbol{h^{(4)} = \sigma(W^{(4)} h^{(3)} + b^{(4)})}$$

$$x = [x_1, x_2]^T \in \mathbb{R}^2$$

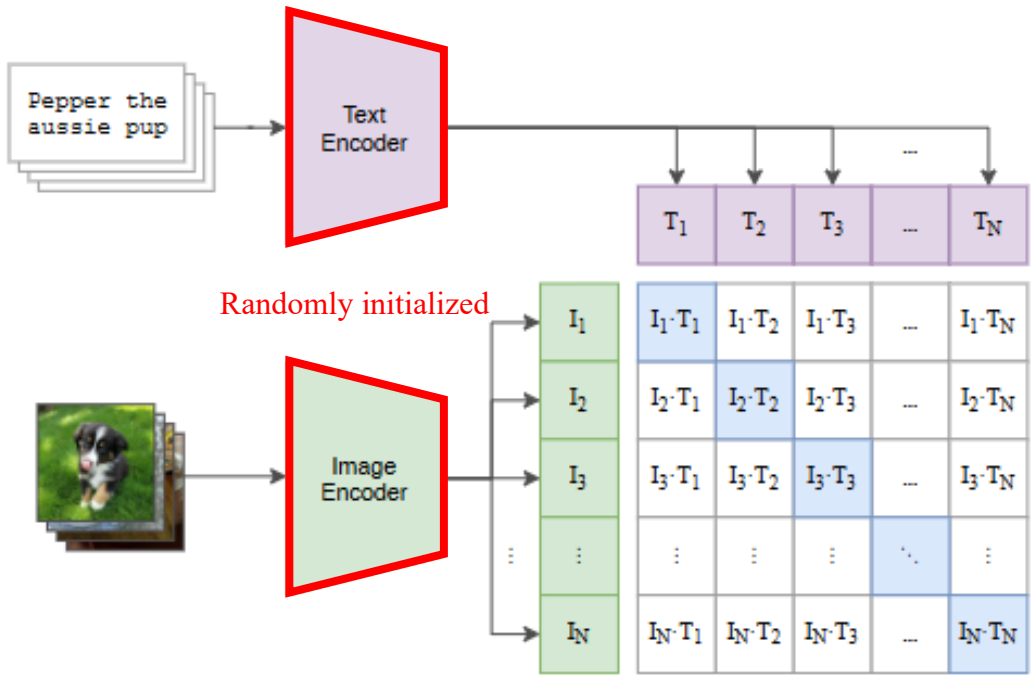
→ 레이어를 지날수록 임베딩 공간이 국소 영역으로 수축됨 (Cone effect)

[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap

Modality Gap at random initialization

- CLIP은 독립적인 두 개의 신경망을 사용하고, 이는 무작위 가중치로 초기화된 상태로 학습을 시작함
- 완전히 같은 구조의 모델이더라도, 무작위 가중치로 초기화된 모델들의 임베딩 공간은 다른 위치에 존재함

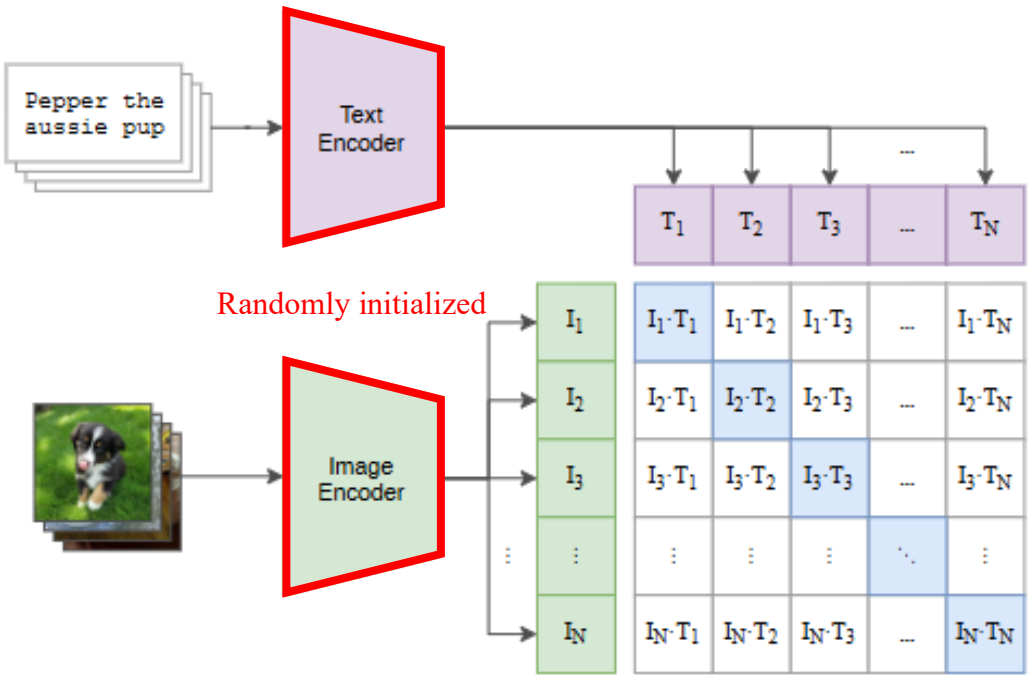


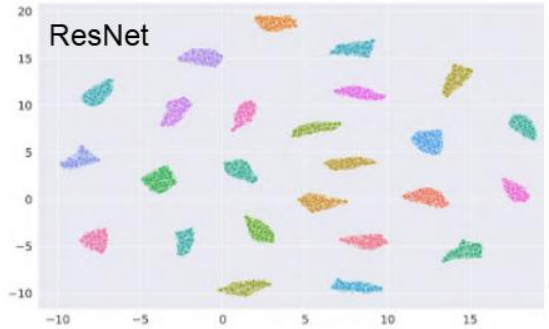
[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

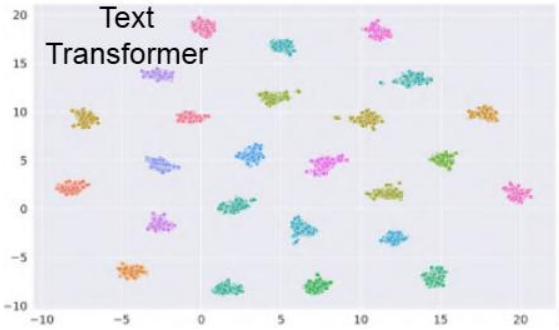
Mind the Gap

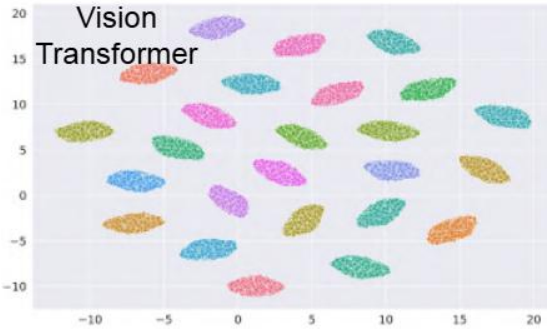
Modality Gap at random initialization

- CLIP은 독립적인 두 개의 신경망을 사용하고, 이는 무작위 가중치로 초기화된 상태로 학습을 시작함
- 완전히 같은 구조의 모델이더라도, 무작위 가중치로 초기화된 모델들의 임베딩 공간은 다른 위치에 존재함









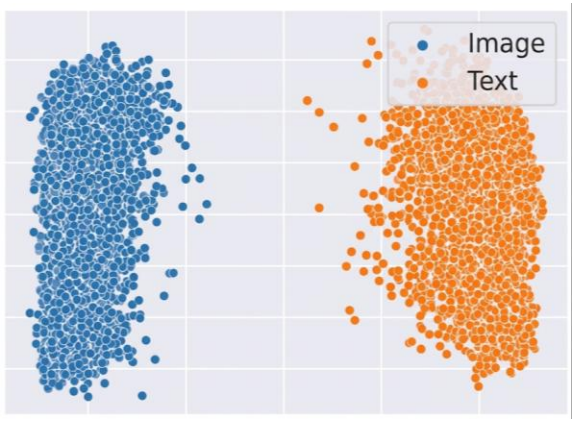
[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap

Modality Gap during model optimization

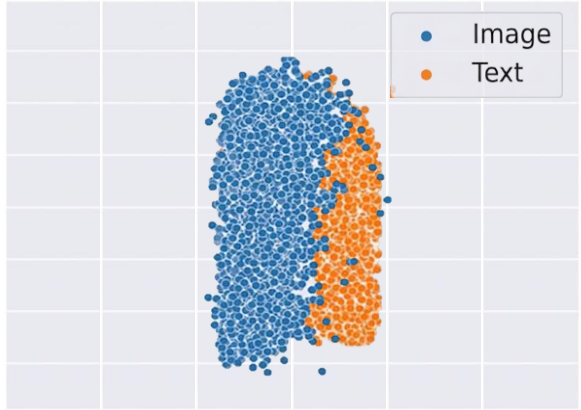
- 학습 과정에서 모달리티 갭은 왜 줄어들지 않는가?
- 학습이 완료된 CLIP 모델을 이용해 임베딩 공간을 강제로 보정하는 실험을 수행

$$\vec{gap} = \mu(IMG) - \mu(TXT)$$
$$x_{shift,i} = Norm(x_i - \alpha \cdot \vec{gap})$$
$$y_{shift,i} = Norm(y_i + \alpha \cdot \vec{gap})$$



$\alpha = 0$

$\alpha = 0.3$



$\alpha = 0.5$

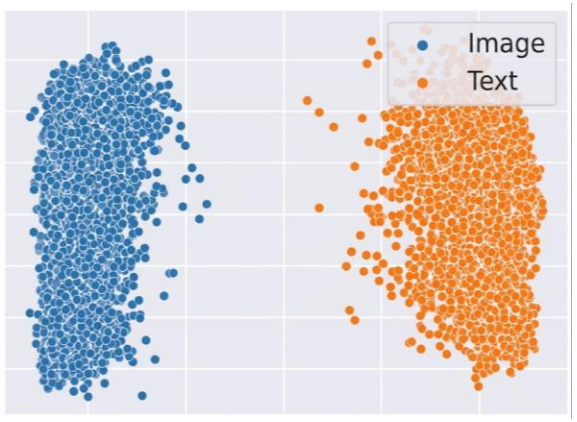
[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap

Modality Gap during model optimization

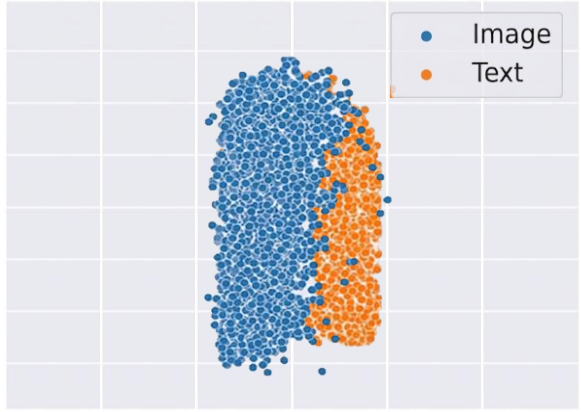
- 학습 과정에서 모달리티 갭은 왜 줄어들지 않는가?
- 학습이 완료된 CLIP 모델을 이용해 임베딩 공간을 강제로 보정하는 실험을 수행

$$\vec{gap} = \mu(IMG) - \mu(TXT)$$
$$x_{shift,i} = Norm(x_i - \alpha \cdot \vec{gap})$$
$$y_{shift,i} = Norm(y_i + \alpha \cdot \vec{gap})$$

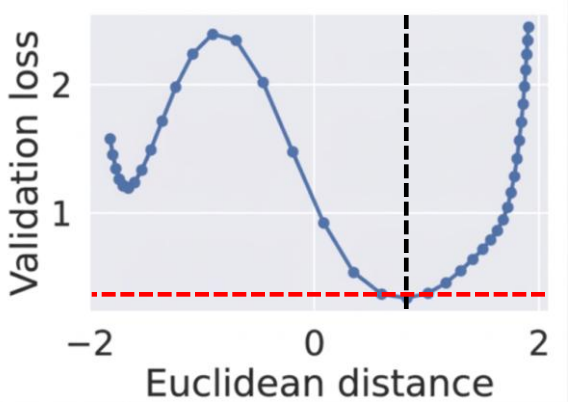


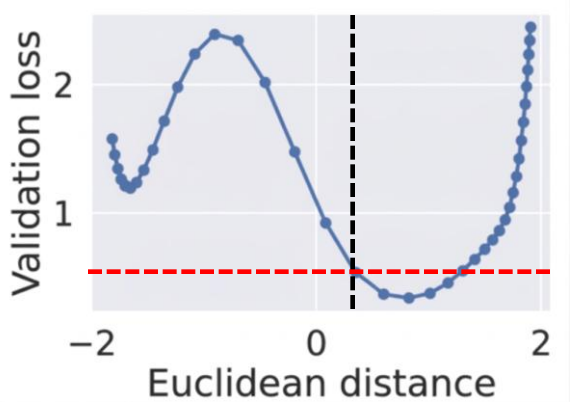
$\alpha = 0$

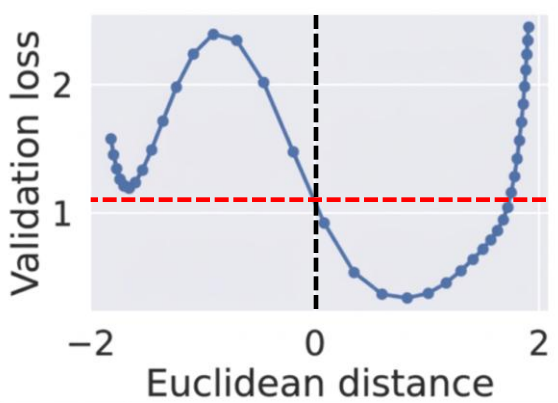
$\alpha = 0.3$



$\alpha = 0.5$







[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap

Modality Gap during model optimization

- CLIP은 훈련에 대조학습을 사용하고, 이는 Uniformity와 Alignment를 증가시키는 방향으로 학습된다고 했음.
- 거리가 가까워진다면 Alignment가 향상되는 것인데 왜 손실이 증가했을까?

Positive 샘플의 유사도 ↑

$$\mathcal{L}_{\mathcal{I}\to\mathcal{T}} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{exp(x_i \cdot y_i/\tau)}{\sum_{j=1}^{N} exp(x_i \cdot y_j/\tau)}$$

배치 내 샘플의 유사도의 합 ↓

N:	배치 내 샘플 수
x, y:	이미지, 텍스트 벡터
i, j:	Positive, Negative 샘플 인덱스
τ:	Temperature

[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Mind the Gap

Modality Gap during model optimization

- CLIP은 훈련에 대조학습을 사용하고, 이는 Uniformity와 Alignment를 증가시키는 방향으로 학습된다고 했음.
- 거리가 가까워진다면 Alignment가 향상되는 것인데 왜 손실이 증가했을까?

Positive 샘플의 유사도 ↑

$$\mathcal{L}_{\mathcal{I}\to\mathcal{T}} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{exp(x_i \cdot y_i/\tau)}{\sum_{j=1}^{N} exp(x_i \cdot y_j/\tau)}$$

배치 내 샘플의 유사도의 합 ↓

N: 배치 내 샘플 수
x, y: 이미지, 텍스트 벡터
i, j: Positive, Negative 샘플 인덱스
τ: Temperature

만약 τ에 0.01이 입력된다면?

$$= -\frac{1}{N}\sum_{i=1}^{N} log \frac{exp(100 \cdot x_i \cdot y_i)}{\sum_{j=1}^{N} exp(100 \cdot x_i \cdot y_j)}$$

$$= -\frac{1}{N}\sum_{i=1}^{N} log \frac{exp(100 \cdot x_i \cdot y_i)}{exp(100 \cdot x_i \cdot y_1) + exp(100 \cdot x_i \cdot y_2) + \cdots + exp(100 \cdot x_i \cdot y_N)}$$

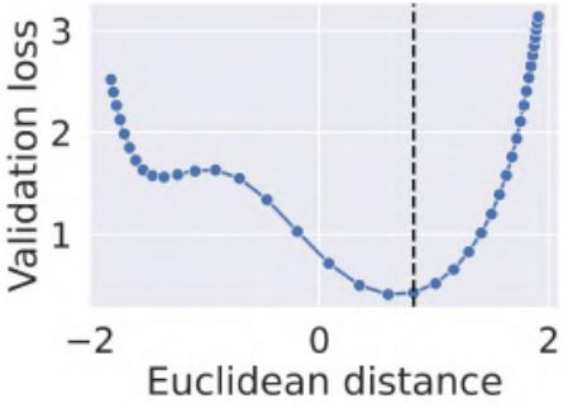
분자가 작아서 주는 패널티 < 분모가 커서 주는 패널티

Mind the Gap

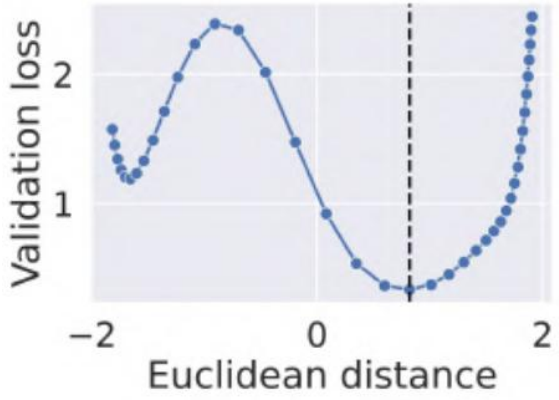
Modality Gap during model optimization

- 실제로 τ 가 1일 때, 모달리티 간의 거리가 최소인 지점이 손실이 가장 낮은 점임을 볼 수 있음.

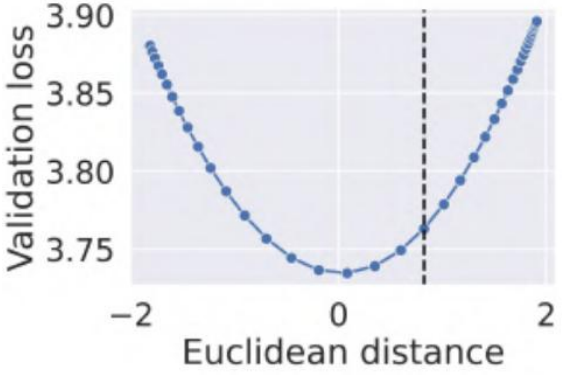
$\tau = 0.01$



$\tau = 0.2$



$\tau = 1$



[3] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. (2022 NeurIPS)

Two Effects, One Trigger:
On the Modality gap, Object bias and Information Imbalance in Contrastive Vision-Language Models (2025 ICLR)

Two Effects, One Trigger

Information Imbalance

- 이미지는 텍스트 데이터에 비해 더 많은 정보를 담고 있음 (질감, 색상, 구도, 배경 등)
- 정보 불균형은 한 모달리티가 다른 모달리티보다 더 많은 정보를 담고 있는 상태를 의미함

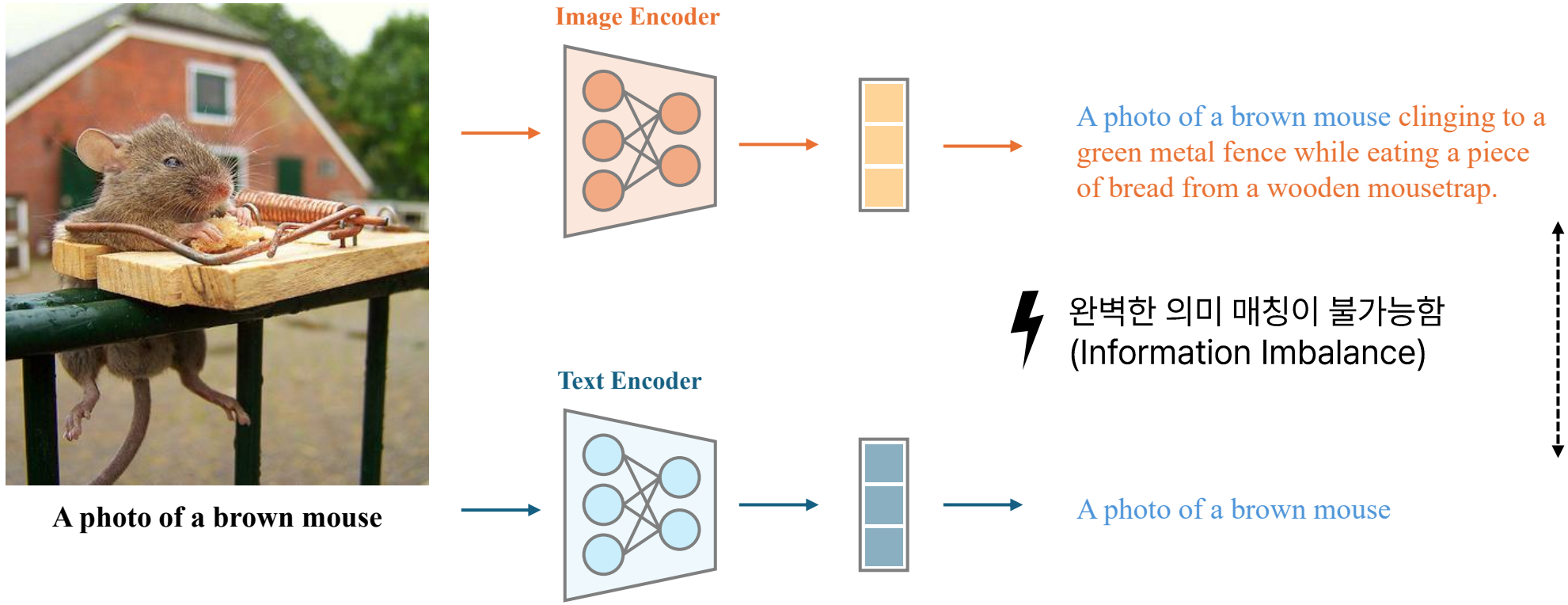
A photo of a brown mouse

[4] Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. (2025 ICLR)

Two Effects, One Trigger

Information Imbalance

- 이미지는 텍스트 데이터에 비해 더 많은 정보를 담고 있음 (질감, 색상, 구도, 배경 등)
- 정보 불균형은 한 모달리티가 다른 모달리티보다 더 많은 정보를 담고 있는 상태를 의미함

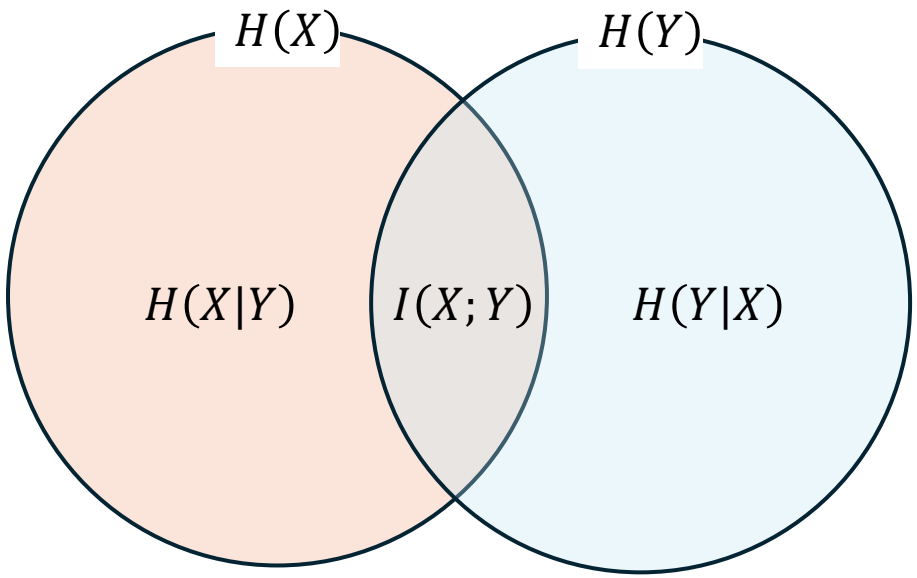


[4] Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. (2025 ICLR)

Two Effects, One Trigger

Insufficient mutual information

- 정보 이론에서 두 변수의 정렬 정도는 상호 정보량($I(X;Y)$)에 의해 결정됨
- 이미지는 정보가 풍부하지만, 텍스트는 정보가 적어 이미지와 텍스트가 완벽히 정렬하는 데에는 한계가 있음

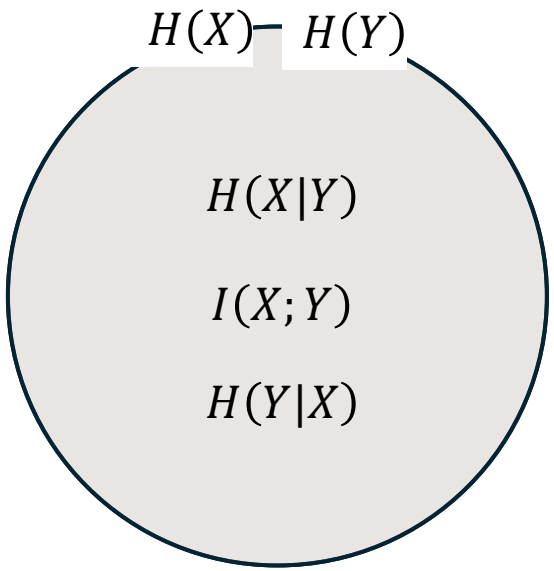


[4] Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. (2025 ICLR)

Two Effects, One Trigger

Insufficient mutual information

- 정보 이론에서 두 변수의 정렬 정도는 상호 정보량($I(X;Y)$)에 의해 결정됨
- 이미지는 정보가 풍부하지만, 텍스트는 정보가 적어 이미지와 텍스트가 완벽히 정렬하는 데에는 한계가 있음

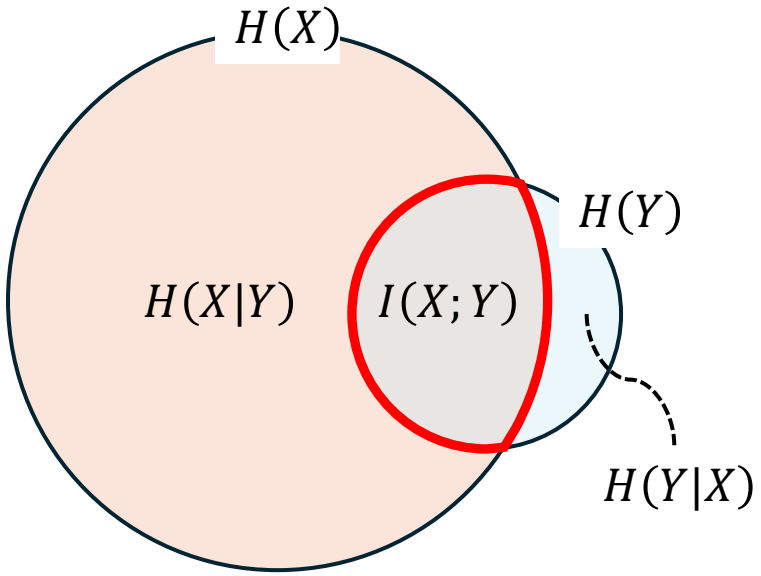


[4] Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. (2025 ICLR)

Two Effects, One Trigger

Insufficient mutual information

- 정보 이론에서 두 변수의 정렬 정도는 상호 정보량($I(X;Y)$)에 의해 결정됨
- 이미지는 정보가 풍부하지만, 텍스트는 정보가 적어 이미지와 텍스트가 완벽히 정렬하는 데에는 한계가 있음



$X =$ 

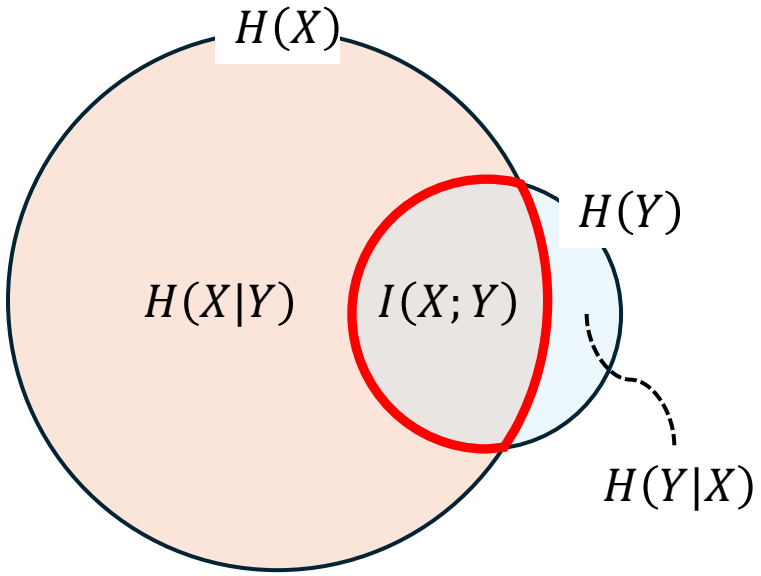
$Y =$ **A photo of a brown mouse**

[4] Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. (2025 ICLR)

Two Effects, One Trigger

Conditional entropy – Object Bias

- 모델은 이미지와 텍스트를 최대한 정렬하도록 학습됨.
- 정보 불균형 상황에서 이미지 인코더는 확실한 개념에 집중하는 방향으로 대응하게 됨 (Object Bias)

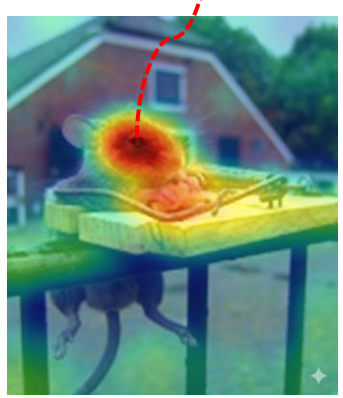


$X =$ 

$Y =$ **A photo of a brown mouse**

확실한 개념에
집중하도록 학습됨

Image Encoder

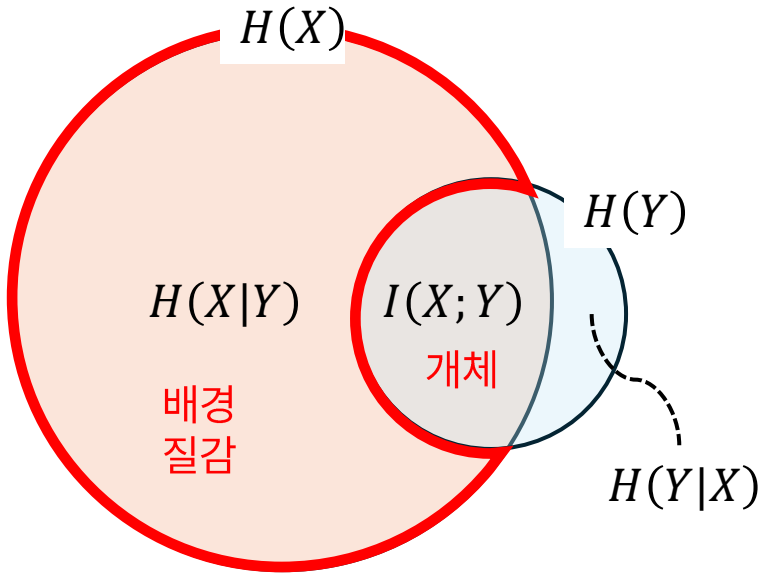


[4] Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. (2025 ICLR)

Two Effects, One Trigger

Conditional entropy – Modality Gap

- 남겨진 이미지 정보($H(X|Y)$) 중 어떤 것을 텍스트와 정렬해야 할지 불확실함
- 모델은 이런 불확실함을 반영하기 위해 모든 텍스트와 거리를 균일하게 방향으로 대응하게 됨 (Modality Gap)



$X =$ 

$Y =$ **A photo of a brown mouse**

데이터의 자체의 높은 불확실성

→ 출력에 엔트로피가 증가하도록 학습

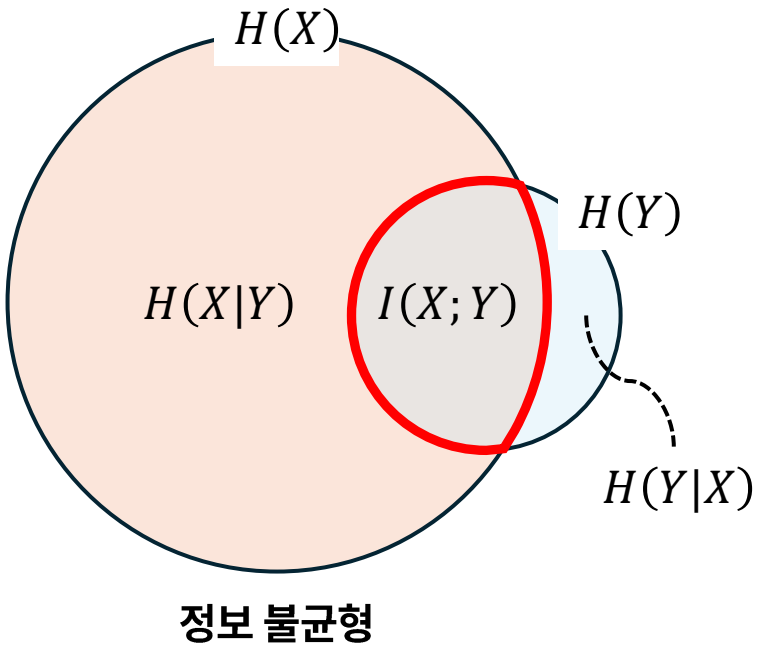
→ 모든 텍스트과 거리를 두는 Modality Gap 생성

[4] Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. (2025 ICLR)

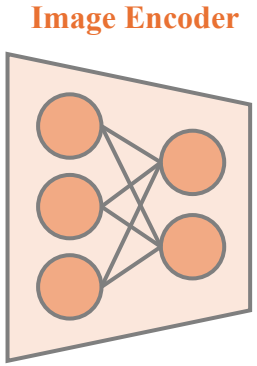
Two Effects, One Trigger

Conditional entropy – Modality Gap

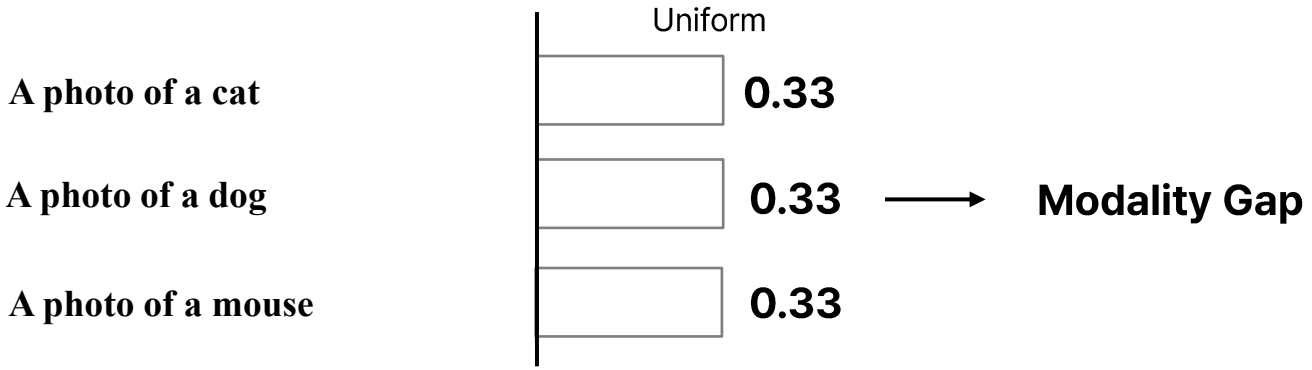
- 이미지와 텍스트는 서로 가지고 있는 정보량이 달라 정보가 불균형한 상황
- 모델이 확실한 정보만 집중하려는 힘이 작용해 **Object bias** 현상이 나타남
- 텍스트가 가진 정보 외의 이미지 정보를 불확실성으로 처리하기 위한 힘이 작용해 **Modality gap** 현상이 나타남



정보 불균형



Object Bias

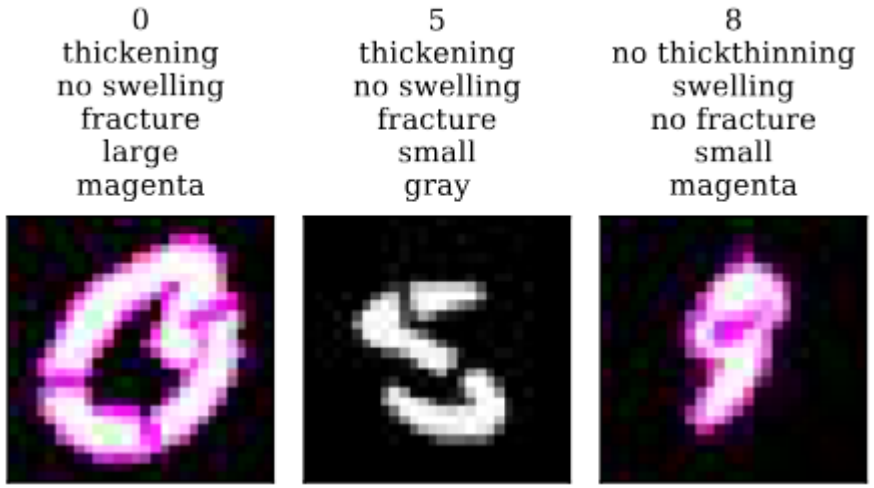


[4] Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. (2025 ICLR)

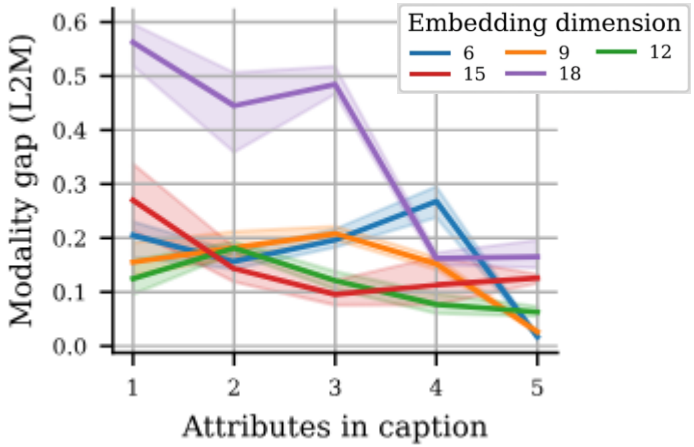
Two Effects, One Trigger

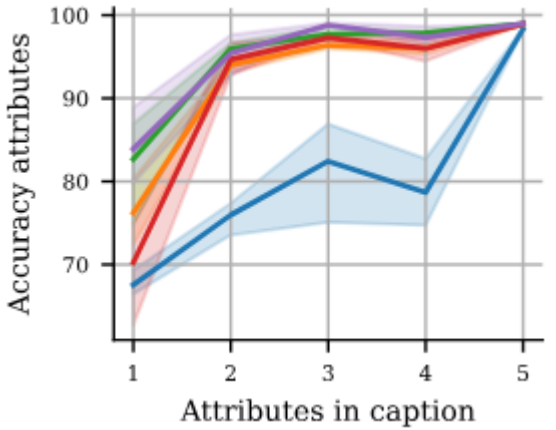
Experiment

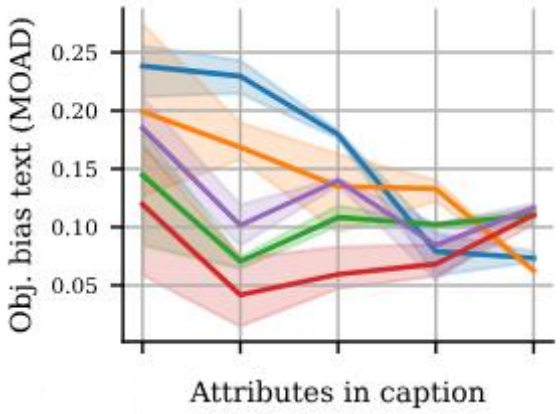
- Multimodal Attributes and Digits(MAD) 데이터셋을 이용해 모달리티 갭과 객체 편향이 일어나는지 확인함
- 이미지는 그대로 두고, 캡션에 포함되는 단어의 수를 조정하며 실험

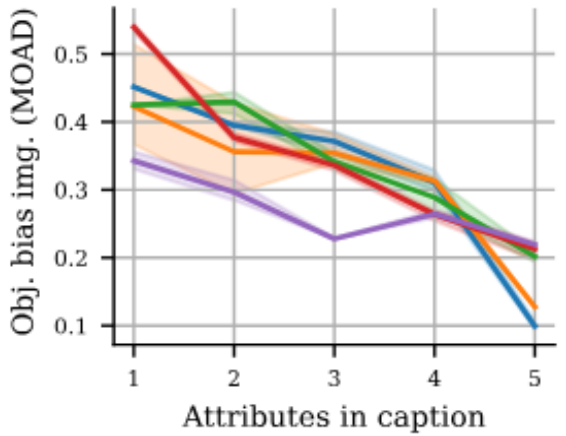


MAD Dataset









[4] Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. (2025 ICLR)

Conclusion

- **Modality Gap**
  - 학습된 이미지와 텍스트 **임베딩이 섞이지 않고 두 영역으로 분리**되는 현상

- **기하학 및 최적화적 관점**
  - 신경망 특성상 임베딩이 좁은 영역으로 수축하는 Cone Effect 현상 발생
  - 대조 손실 함수가 분모의 패널티를 줄이기 위해 **모달리티 간 거리를 벌리는 방향**으로 학습

- **정보 이론적 관점**
  - **이미지가 텍스트보다 많은 정보를 담고 있어** 완벽한 의미 정렬이 불가능
  - 불균형 속에서 모델이 확실한 "객체" 정보에만 과도하게 집중됨 (Object Bias)
  - 캡션이 설명하지 못하는 잉여 정보(불확실성)을 처리하기 위해 **모든 텍스트와 일정한 거리를 두도록 학습**됨 (Modality Gap)

- 모달리티 갭은 모델이 **정보 불균형과 불확실성을 조절하기 위해** 찾아낸 타협점이라고 할 수 있음.

고맙습니다