Multimodal Contrastive learning with various data domains

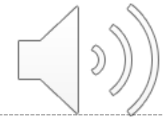
2023.11.17

Data **M**ining and **Q**uality **A**nalytics Lab

박진혁

Data Mining
Quality Analytics

Contents



Introduction

❖ 발표자 소개

- 박진혁
- Data Mining & Quality Analytics Lab(김성범 교수님)
- 석·박사 통합과정 8학기 재학 중(2019.8 ~)

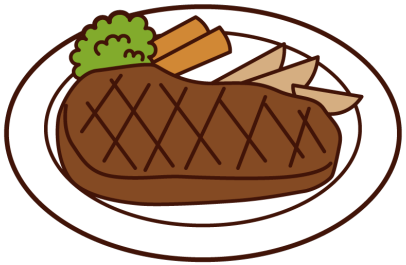
✓ 관심 분야

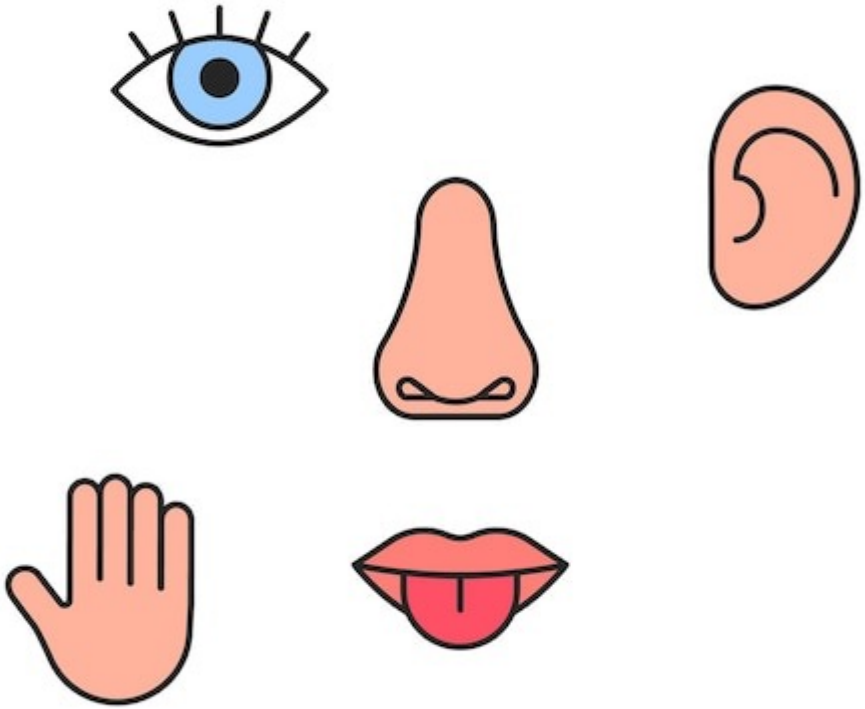
- Computer Vison
- Image Captioning
- Object Detection
- Multimodal Contrastive learning

What is Multi-modal?

❖ Multi-modal learning

- 하나의 데이터 타입이 아닌 서로 다른 데이터 타입을 같이 활용하는 학습방식
  - ➢ 이미지 + 텍스트, 텍스트 + 오디오, 이미지 + 오디오

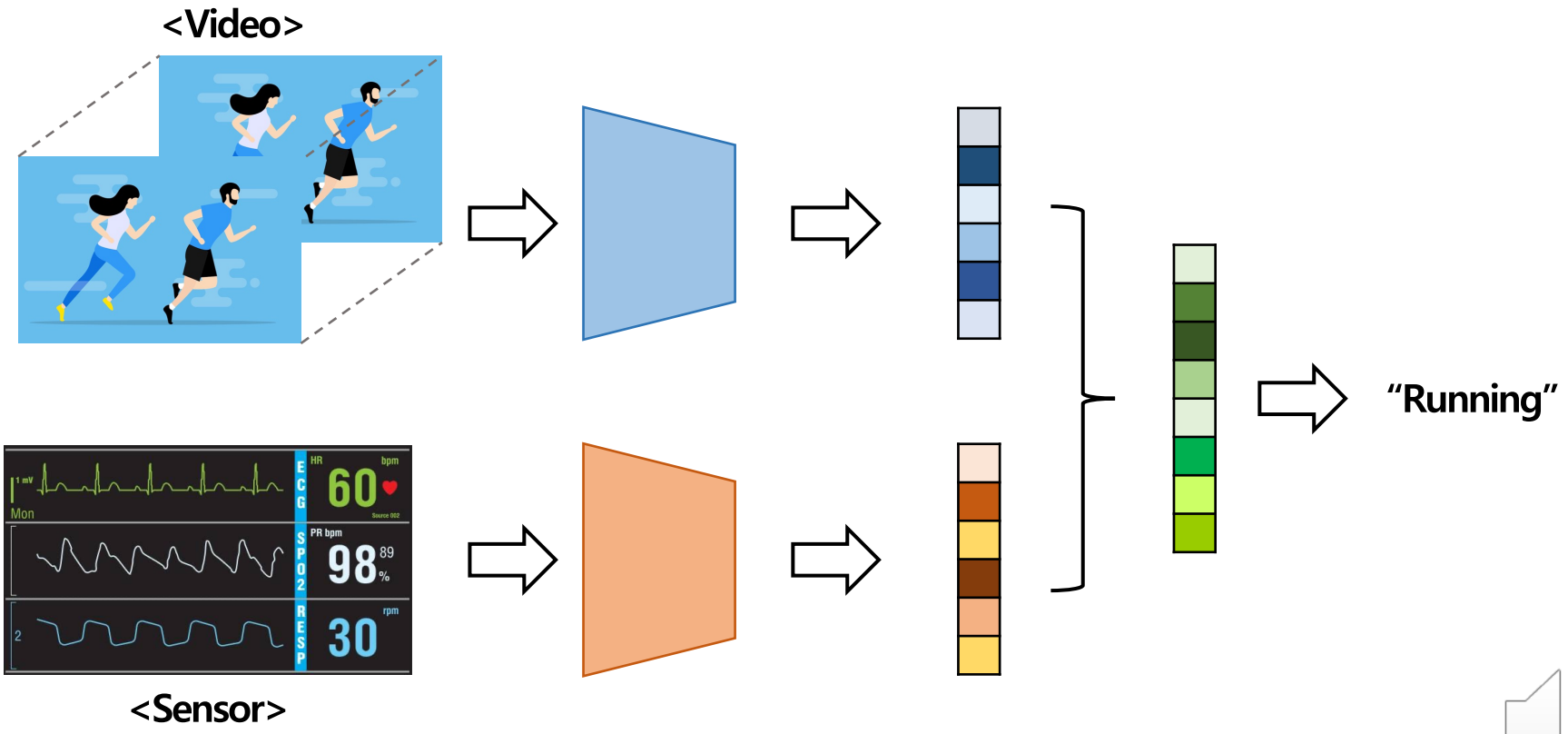




What is Multi-modal?

❖ Multi-modal learning

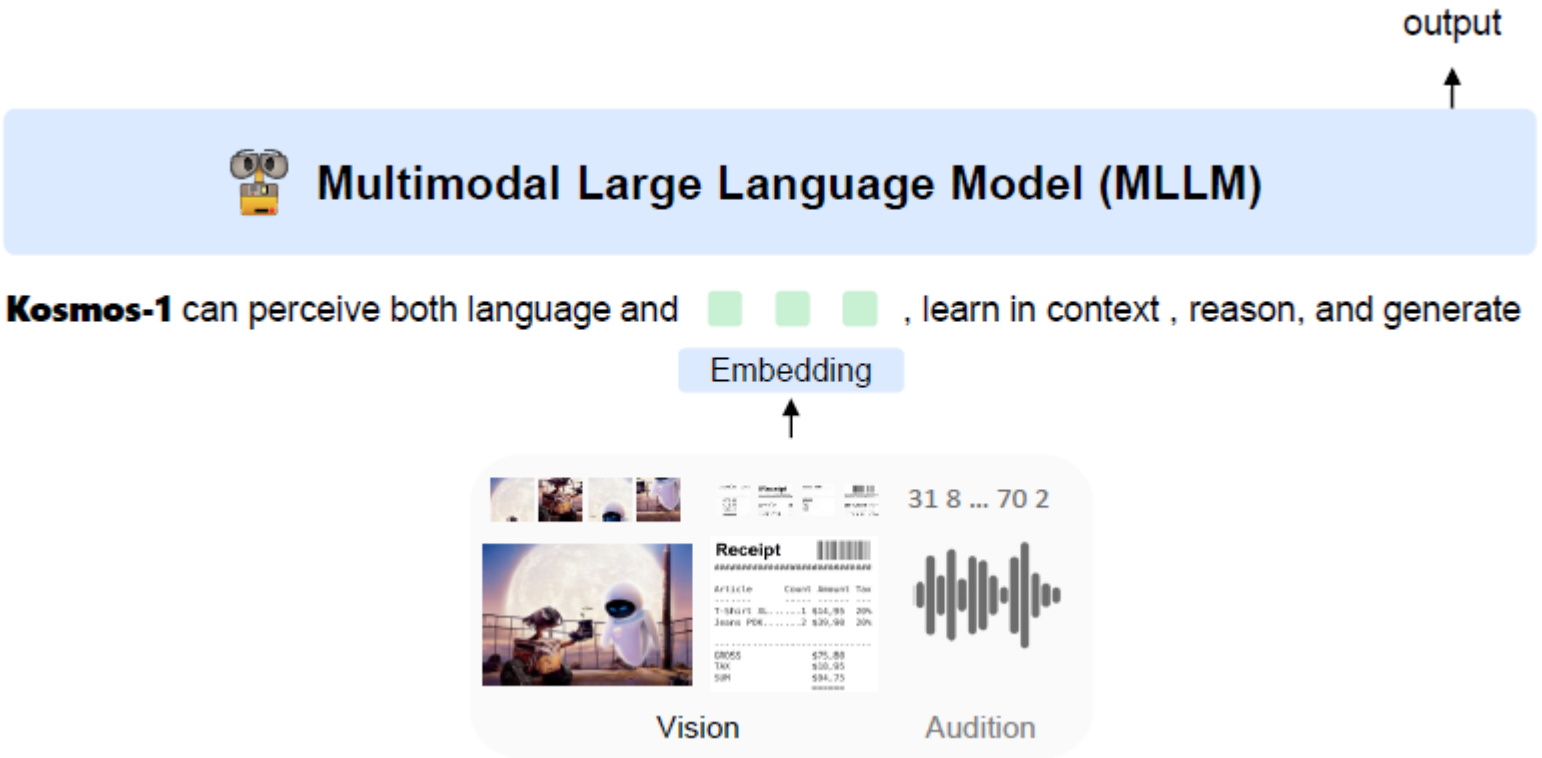
- 하나의 데이터 타입이 아닌 서로 다른 데이터 타입을 같이 활용하는 학습방식
  - ➢ 이미지 + 텍스트, 텍스트 + 오디오, 이미지 + 오디오...



What is Multi-modal?

❖ Language Is Not All You Need: Aligning Perception with Language Models(MS, 2023)

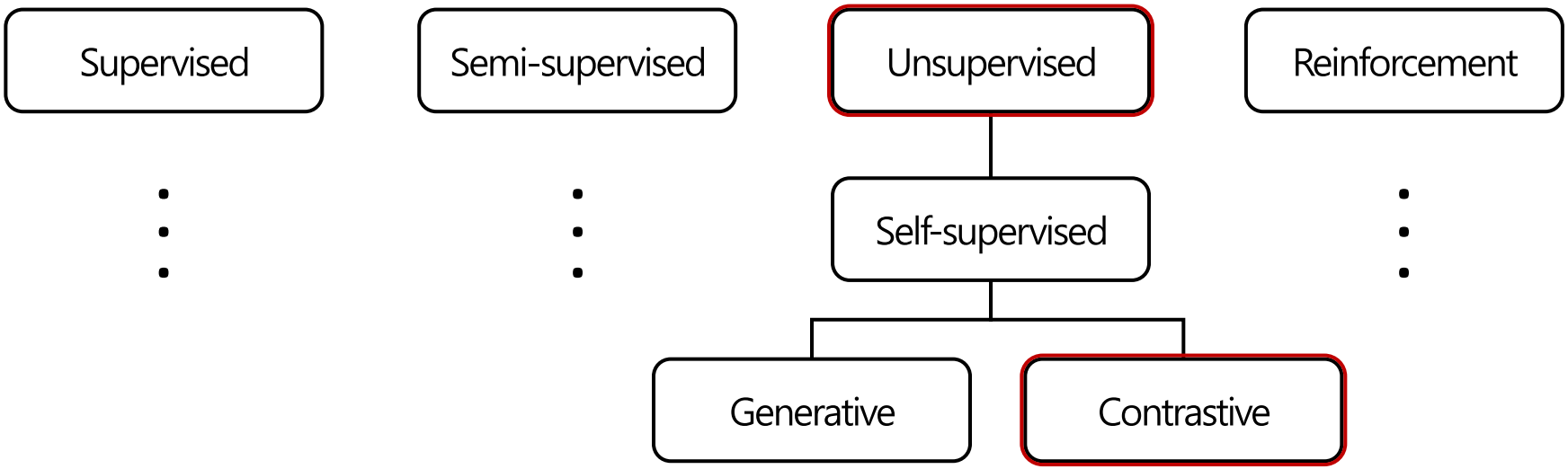
- Multimodal을 적용하여 기존의 task뿐만 아니라 nonverbal reasoning task도 수행할 수 있는 방법론
- Kosmos-1이후 Kosmos-2, Kosmos-G 등 다양한 MLLM 방법론이 나오고 있음



What is Contrastive learning?

❖ Contrastive learning

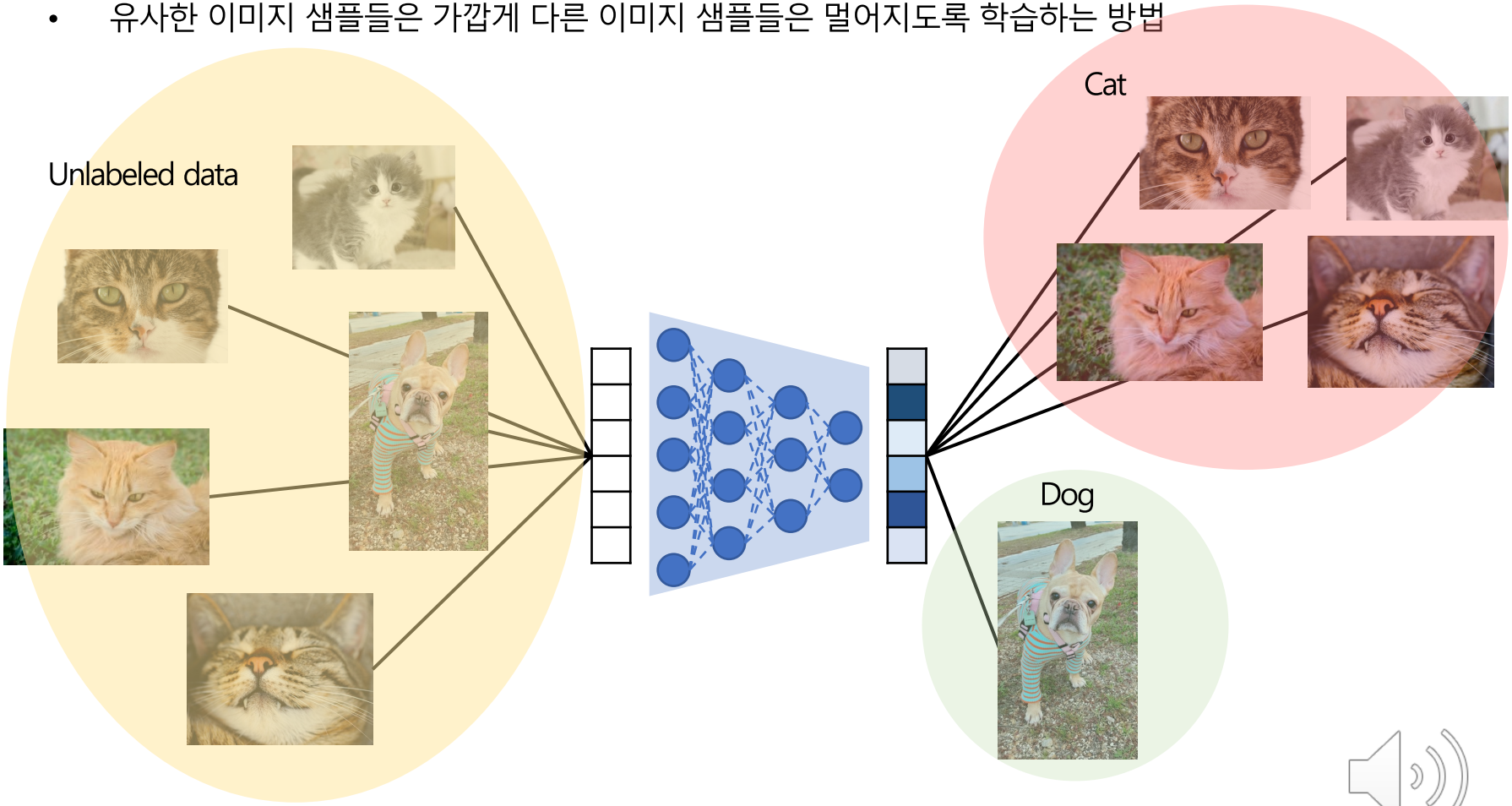
- Self-supervised learning의 방법론으로써 unlabelled dataset으로 부터 좋은 representation 얻고자 하는 방법
- 유사한 이미지 샘플들은 가깝게 다른 이미지 샘플들은 멀어지도록 학습하는 방법



What is Contrastive learning?

❖ Contrastive learning

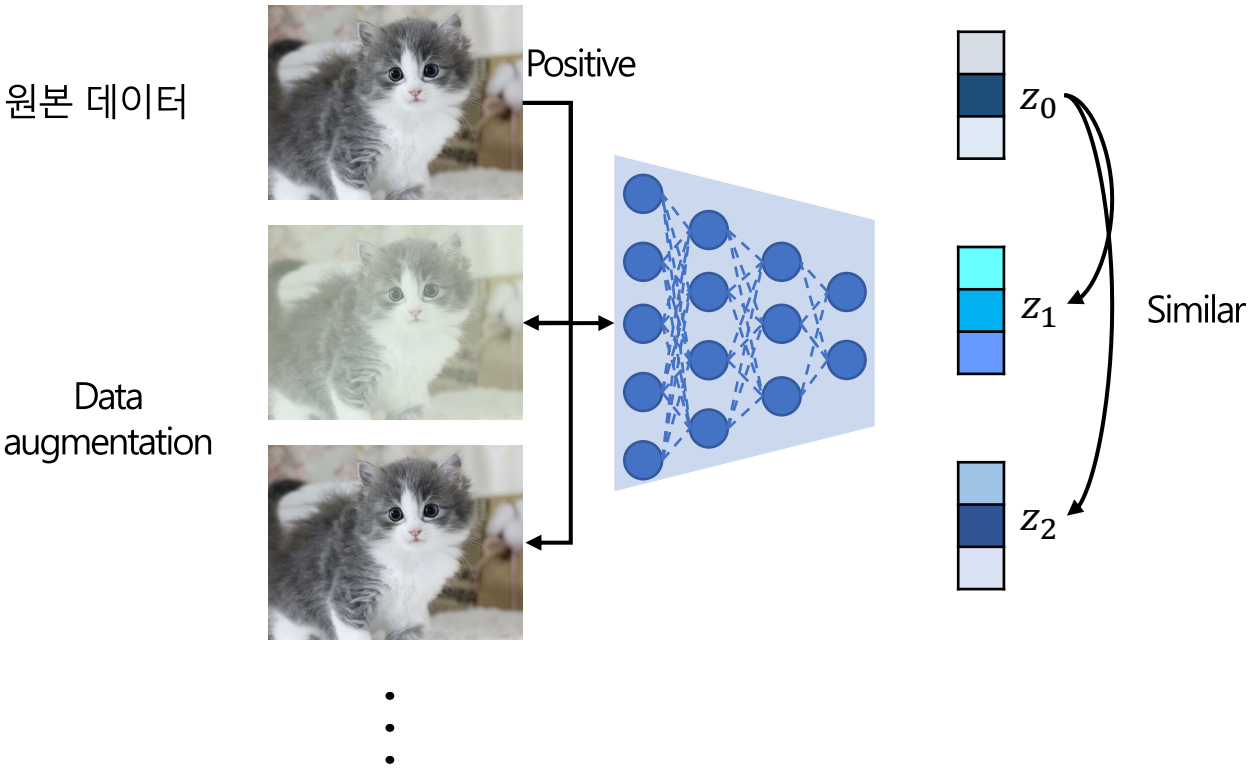
- Self-supervised learning의 방법론으로써 unlabelled dataset으로 부터 좋은 representation 얻고자 하는 방법
- 유사한 이미지 샘플들은 가깝게 다른 이미지 샘플들은 멀어지도록 학습하는 방법



What is Contrastive learning?

❖ Contrastive learning의 기본 학습방식

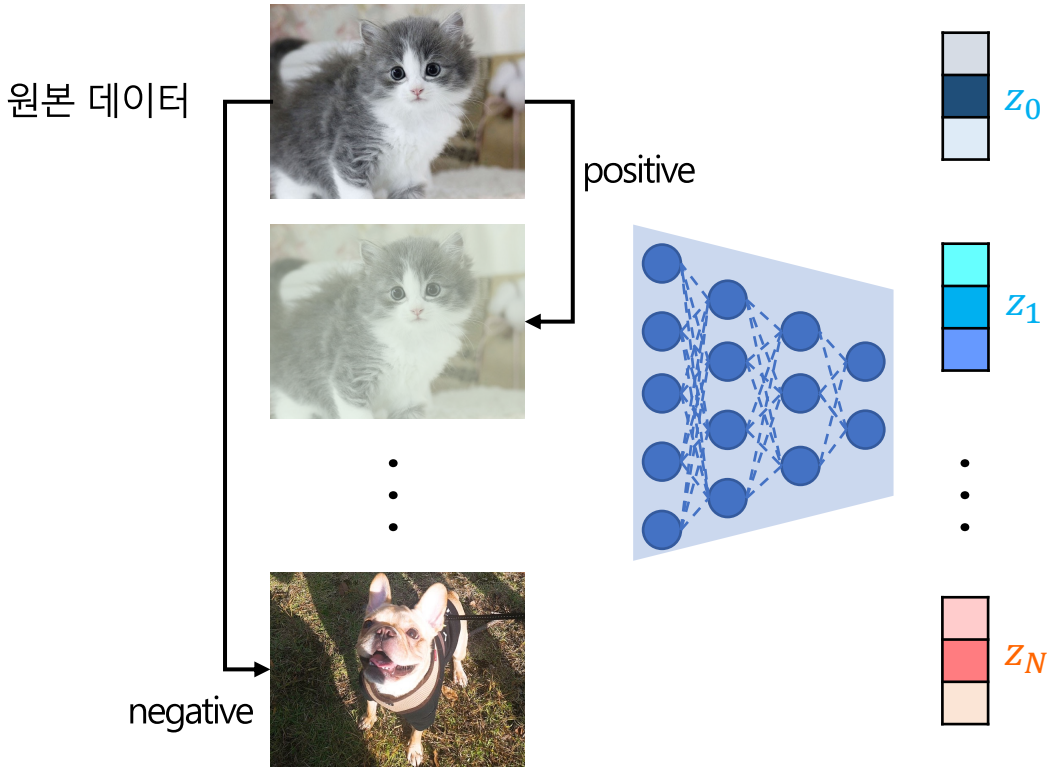
- Label이 없는 데이터를 사용하여 contrastive loss를 최소화하는 인코더를 학습
- Augmentation데이터는 원본데이터와 본질은 같음으로 모델 또한 서로 유사한 벡터를 생성하도록 학습



What is Contrastive learning?

❖ InfoNCE & NT-Xent Loss

- Contrastive learning에 사용되는 손실함수로써 cosine similarity를 사용
- 대표적으로 InfoNCE Loss, NT-Xent Loss 사용



$$sim(z_i, z_j) = \frac{z_i \cdot z_j}{\|z_i\| \cdot \|z_j\|}$$

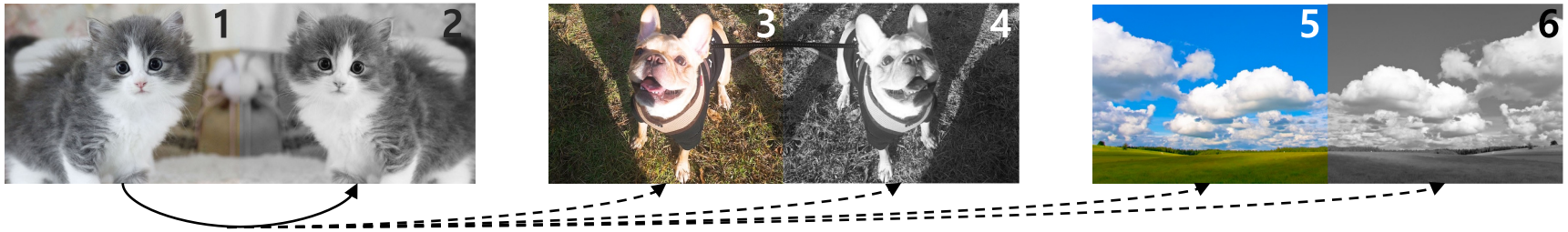
$$sim(z_0, z_1) = \frac{z_0 \cdot z_1}{\|z_0\| \cdot \|z_1\|} \approx 0.88$$

$$sim(z_0, z_N) = \frac{z_0 \cdot z_N}{\|z_0\| \cdot \|z_N\|} \approx 0.23$$

What is Contrastive learning?

$$sim(z_i, z_j) = \frac{z_i \cdot z_j}{\|z_i\| \cdot \|z_j\|}$$

❖ NT-Xent Loss



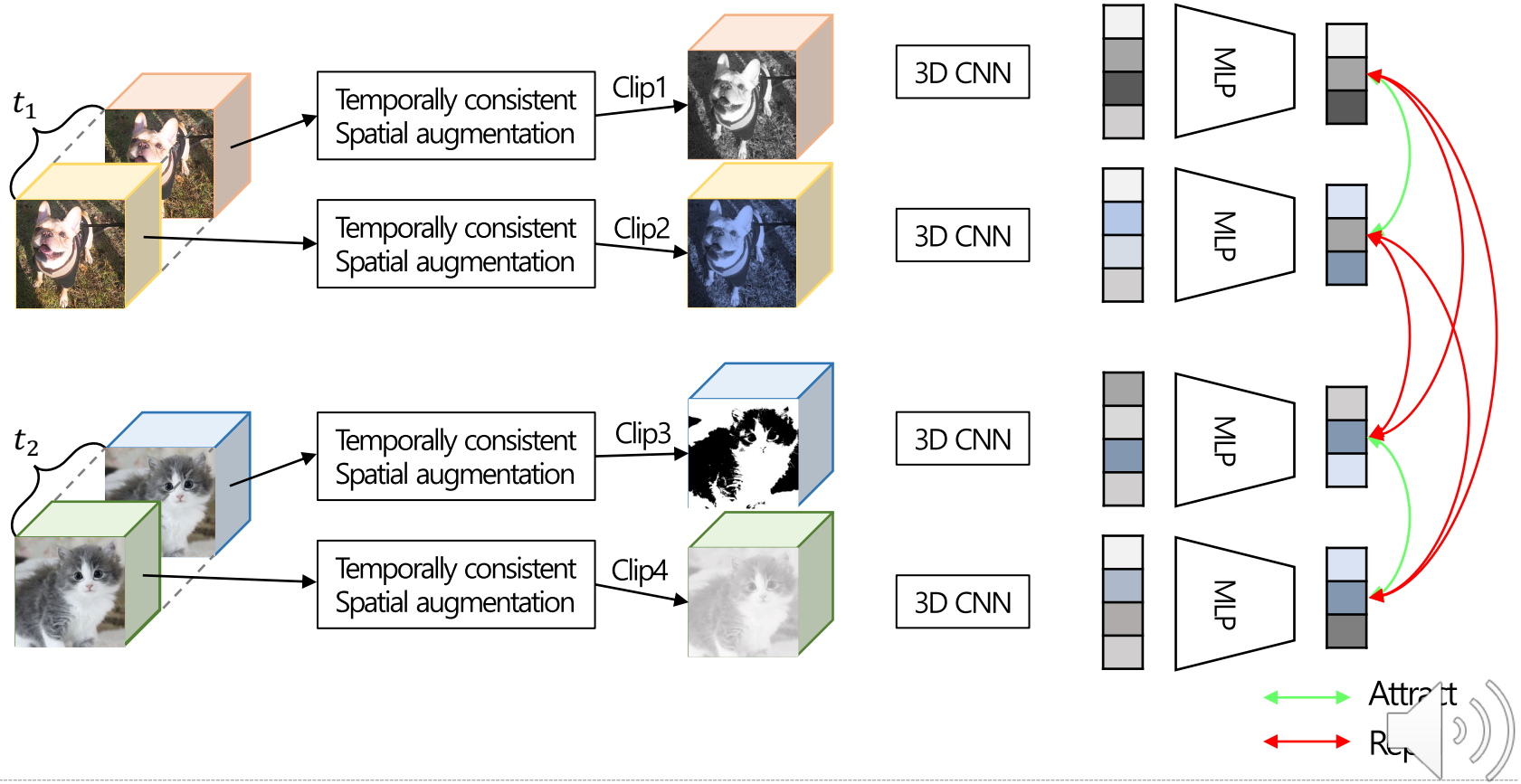
$$l(i,j) = -log \frac{\exp(s_{i,j}/\mathcal{T})}{\sum_{k=1}^{2N} l_{[k \neq i]} \exp(s_{i,k}/\mathcal{T})} \Rightarrow -log \frac{\exp(1,2)}{\exp(1,2) + \exp(1,3) + \exp(1,4) + \exp(1,5) + \exp(1,6)}$$

$$L = \frac{1}{2N} \sum_{k}^{N} [l(2k-1, 2k) + l(2k, 2k-1)]$$

Contrastive learning

❖ Contrastive learning in video data

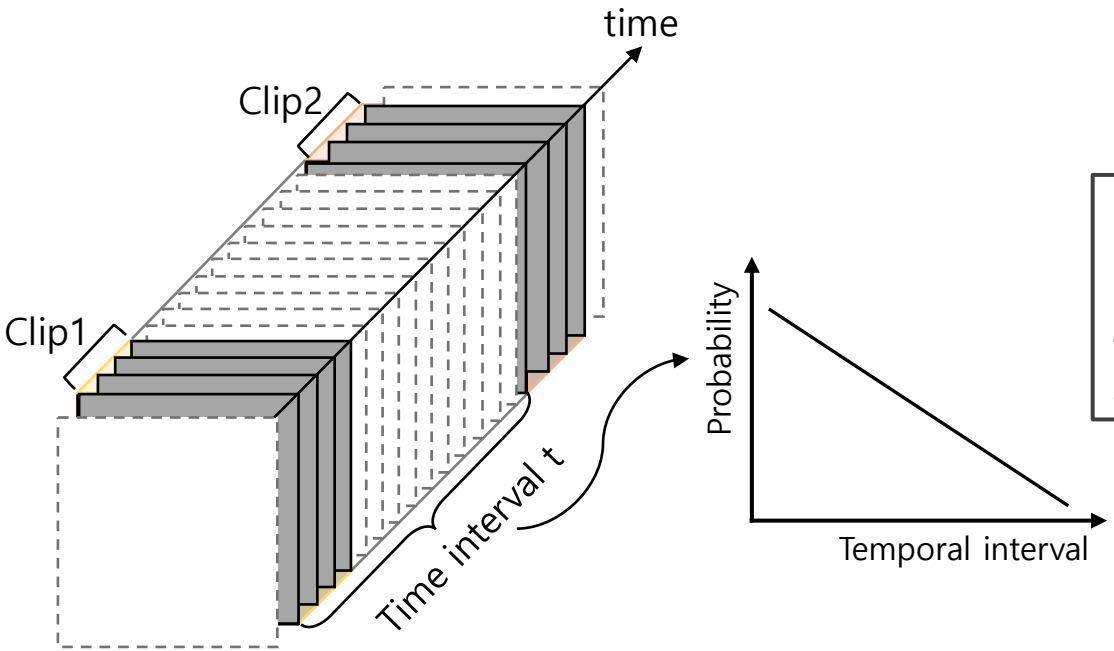
- Spatiotemporal Contrastive Video Representation Learning(CVPR, 2021)
- 레이블이 없는 비디오 데이터에 contrastive learning을 적용한 논문



CVRL

❖ Temporal Augmentation: a sampling perspective

- Contrastive learning에 필요한 positive pair를 생성하기 위해 video마다 2개의 clip을 생성
- 기존에는 temporal augmentation을 위해 clip을 섞거나 재생속도를 변화시킴 → temporal feature 손상
- 생성된 두개의 clip간 거리가 멀다면 다른 특성을 가진 clip이 positive로 생성될 수 있음 → Sampling strategy



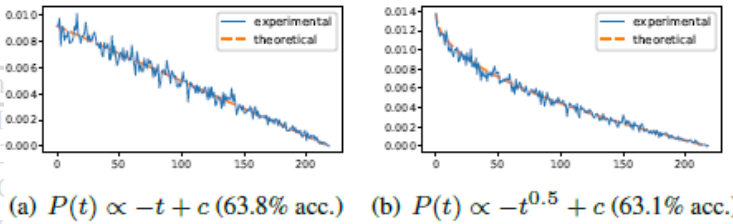
1. Video길이가 T라면 [0,T]를 가지는 분포에서 time interval t를 선택
2. [0,T-t]구간에서 Clip1을 sampling
3. Clip1으로 부터 +t시점에서 Clip2를 sampling

https://hellopotatoworld.tistory.com/23

CVRL

❖ Experiment & Results

- Sampling distribution이 일정하게 감소하는 경우 가장 좋은 성능을 보여줌
- Temporal, spatial augmentation 모두 적용했을 경우 가장 좋은 성능을 보여줌



(a) $P(t) \propto -t + c$ (63.8% acc.) (b) $P(t) \propto -t^{0.5} + c$ (63.1% acc.)

(c) $P(t) \propto -t^2 + c$ (62.9% acc.) (d) $P(t) \propto c$ (62.7% acc.)

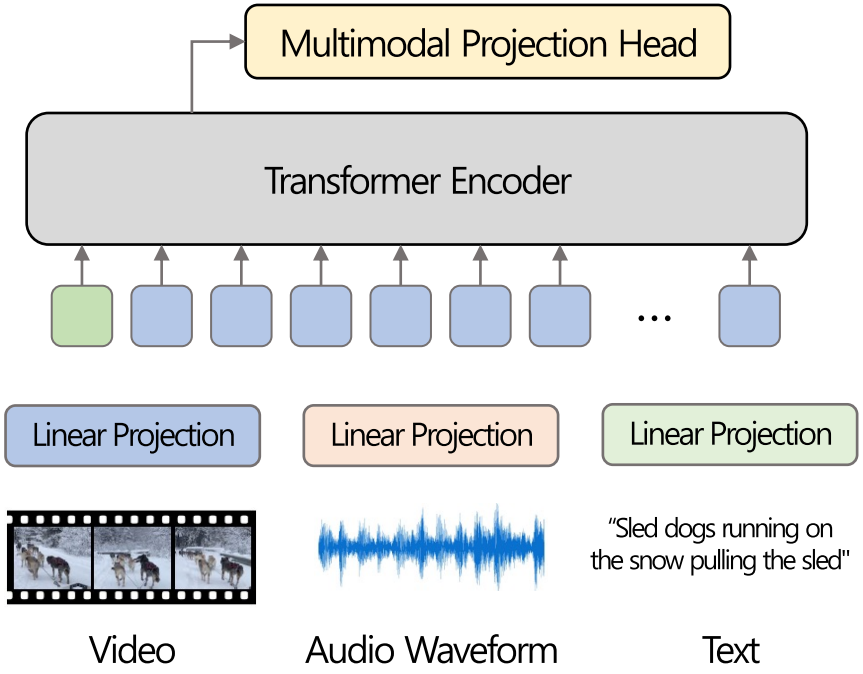
(e) $P(t) \propto t + c$ (62.4% acc.) (f) $P(t) \propto t^2 + c$ (61.9% acc.)

Temporal augmentation	Spatial augmentation	Temporal consistency	Accuracy (%) top-1	top-5
✓			33.0	57.3
	✓		40.9	66.6
✓	✓		52.3	76.0
✓	✓	✓	**63.8**	**85.2**

Multi-modal Contrastive learning

❖ VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text (NeurIPS, 2021)

- Transformer기반의 self-supervised learning
- 3(Video, Audio, Text) domain dataset을 self-supervised에 적용한 연구



VATT

❖ Tokenization and Positional Encoding

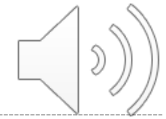
- Video: T×H×W×3으로 구성된 비디오를 [T/t]·[H/h]·[W/w]에 해당하는 패치로 나누어 줌

  이후 t×h×w×3의 패치를 d차원의 vector representation으로 생성

$$e_{i,j,k} = e_{\text{Temporal}_i} + e_{\text{Horizontal}_j} + e_{\text{Vertical}_k},$$

$$E_{\text{Temporal}} \in \mathbb{R}^{[T/t]\times d}, E_{\text{Horizontal}} \in \mathbb{R}^{[H/h]\times d}, E_{\text{Vertical}} \in \mathbb{R}^{[W/w]\times d}$$

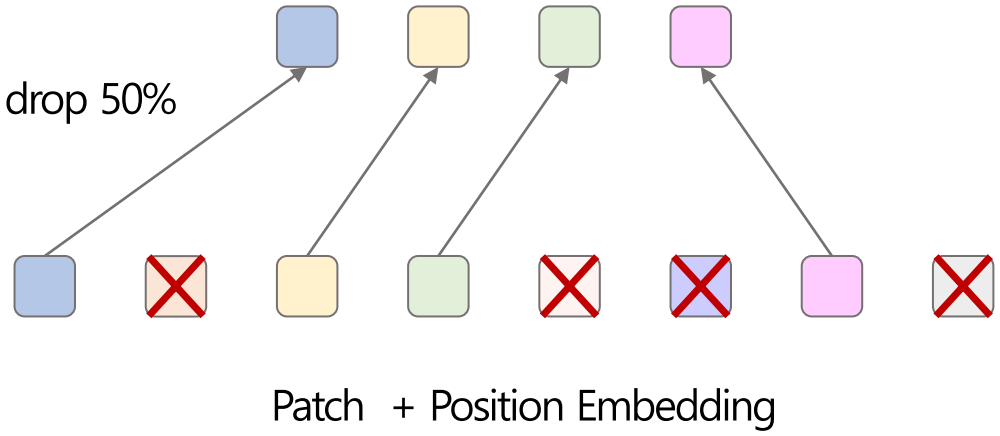
- Audio: T'길이의 1차원 데이터를 t'길이의 패치로 나눈 뒤 d차원의 vector representation으로 생성
- Text: 학습 데이터셋에 대해 v차원의 vocabulary를 정의한 뒤, v차원의 one-hot-vector로 매핑



VATT

❖ Drop token

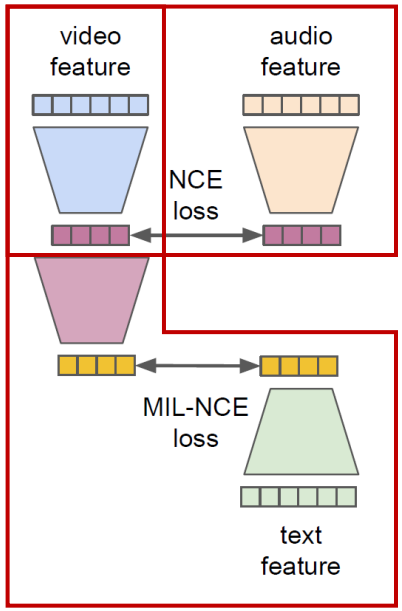
- Transformer기반 모델의 복잡도를 줄이기 위한 방법
- 비디오, 오디오에 대한 토큰 시퀀스를 입력으로 받을 경우 모든 토큰을 사용하지 않고 토큰 일부만을 입력으로 사용함
- Drop token을 적용 하여 원본 데이터의 해상도나 차원을 줄이지 않고 높은 정확도를 유지함



VATT

❖ Multimodal Contrastive learning loss

- (Video-audio), (Video-text) 쌍으로 학습진행
- NCE loss: Video-audio
- MIL-NCE loss: Video-text



$$\text{NCE}(\boldsymbol{z}_{v,va}, \boldsymbol{z}_{a,va}) = -\log\left(\frac{\exp(\boldsymbol{z}_{v,va}^{\top}\boldsymbol{z}_{a,va}/\tau)}{\exp(\boldsymbol{z}_{v,va}^{\top}\boldsymbol{z}_{a,va}/\tau) + \sum_{z'\in\mathcal{N}}\exp(\boldsymbol{z'}_{v,va}^{\top}\boldsymbol{z'}_{a,va}/\tau)}\right)$$

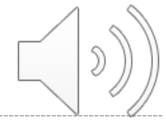
$$\text{MIL-NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\}) = -\log\left(\frac{\sum_{\boldsymbol{z}_{t,vt}\in\mathcal{P}}\exp(\boldsymbol{z}_{v,vt}^{\top}\boldsymbol{z}_{t,vt}/\tau)}{\sum_{\boldsymbol{z}_{t,vt}\in\mathcal{P}}\exp(\boldsymbol{z}_{v,vt}^{\top}\boldsymbol{z}_{t,vt}/\tau) + \sum_{z'\in\mathcal{N}}\exp(\boldsymbol{z'}_{v,vt}^{\top}\boldsymbol{z'}_{t,vt}/\tau)}\right)$$

$$\mathcal{L} = \text{NCE}(\boldsymbol{z}_{v,va}, \boldsymbol{z}_{a,va}) + \lambda\text{MIL-NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\})$$

VATT

❖ Experiments

- 사전학습
  - ➢ HowTo100M: 100만개의 영상 1억개 미만의 클립
  - ➢ AudioSet: 200만개의 영상에서 10초의 오디오 클립
  - ➢ AudioSet은 텍스트가 존재하지 않음으로 텍스트는 0으로 두고 MIL-NCE loss사용하지 않음
- Downstream Task and Dataset
  - ➢ Video action recognition
    - ✓ UCF101, HMDB51, Kinetics-400, Kinetics-600, Moments in Time
  - ➢ Audio event classification
    - ✓ ESC50, AudioSet
  - ➢ Zero-shot video retrieval
    - ✓ YouCook2, MSR-VTT
  - ➢ Image classification
    - ✓ ImageNet

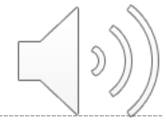


VATT

❖ Result

- Video action recognition

Method	Kinetics-400 Top-1	Kinetics-400 Top-5	Kinetics-600 Top-1	Kinetics-600 Top-5	Moments in Time Top-1	Moments in Time Top-5	TFLOPs
I3D [13]	71.1	89.3	71.9	90.1	29.5	56.1	-
R(2+1)D [26]	72.0	90.0	-	-	-	-	17.5
bLVNet [27]	73.5	91.2	-	-	31.4	59.3	0.84
S3D-G [96]	74.7	93.4	-	-	-	-	-
Oct-I3D+NL [20]	75.7	-	76.0	-	-	-	0.84
D3D [83]	75.9	-	77.9	-	-	-	-
I3D+NL [93]	77.7	93.3	-	-	-	-	10.8
ip-CSN-152 [87]	77.8	92.8	-	-	-	-	3.3
AttentionNAS [92]	-	-	79.8	94.4	32.5	60.3	1.0
AssembleNet-101 [77]	-	-	-	-	34.3	62.7	-
MoViNet-A5 [47]	78.2	-	82.7	-	39.1	-	0.29
LGD-3D-101 [69]	79.4	94.4	81.5	95.6	-	-	-
SlowFast-R101-NL [30]	79.8	93.9	81.8	95.1	-	-	7.0
X3D-XL [29]	79.1	93.9	81.9	95.5	-	-	1.5
X3D-XXL [29]	80.4	94.6	-	-	-	-	5.8
TimeSFormer-L [9]	80.7	94.7	82.2	95.6	-	-	7.14
VATT-Base	79.6	94.9	80.5	95.5	38.7	67.5	9.09
VATT-Medium	81.1	**95.6**	82.4	96.1	39.5	**68.2**	15.02
VATT-Large	**82.1**	95.5	**83.6**	**96.6**	**41.1**	67.7	29.80
VATT-MA-Medium	79.9	94.9	80.8	95.5	37.8	65.9	15.02



VATT

❖ Result

- Audio event classification & ImageNet classification & Zero-shot text-to-video retrieval

METHOD	mAP	AUC	d-prime
DaiNet [21]	29.5	95.8	2.437
LeeNet11 [55]	26.6	95.3	2.371
LeeNet24 [55]	33.6	96.3	2.525
Res1dNet31 [49]	36.5	95.8	2.444
Res1dNet51 [49]	35.5	94.8	2.295
Wavegram-CNN [49]	38.9	96.8	2.612
VATT-Base	**39.4**	**97.1**	**2.895**
VATT-MA-Medium	39.3	97.0	2.884

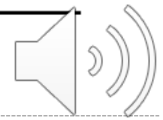
Audio event classification

METHOD	PRE-TRAINING DATA	TOP-1	TOP-5
iGPT-L [16]	ImageNet	72.6	-
ViT-Base [25]	JFT	**79.9**	-
VATT-Base	-	64.7	83.9
VATT-Base	HowTo100M	78.7	93.9

ImageNet classification

METHOD	BATCH	EPOCH	YouCook2 R@10	YouCook2 MedR	MSR-VTT R@10	MSR-VTT MedR
MIL-NCE [59]	8192	27	**51.2**	**10**	**32.4**	**30**
MMV [1]	4096	8	45.4	13	31.1	38
VATT-MBS	2048	4	45.5	13	29.7	49
VATT-MA-Medium	2048	4	40.6	17	23.6	67

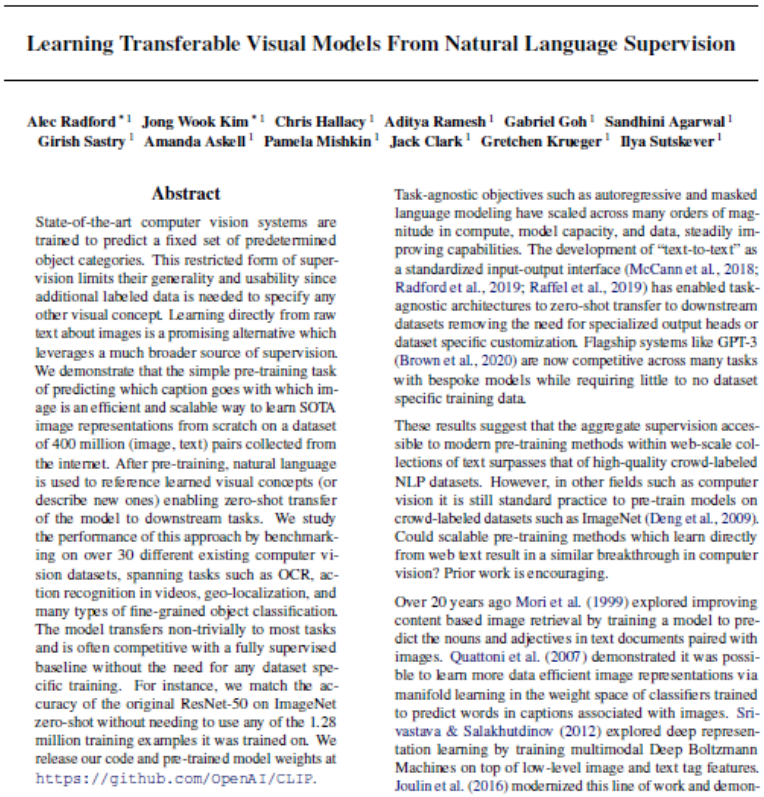
Zero-shot text-to-video retrieval



Multi-modal Contrastive learning

❖ Learning Transferable Visual Models From Natural Language Supervision(ICML, 2021)

- 대규모 데이터셋을 통해 강건한 일반화 성능을 가지는 pre-trained모델 생성
- Web-based image-text pair를 기반으로 visual representation을 사전학습 하는 방법론



Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks. We study the performance of this approach by benchmarking on over 30 different existing computer vision datasets, spanning tasks such as OCR, action recognition in videos, geo-localization, and many types of fine-grained object classification. The model transfers non-trivially to most tasks and is often competitive with a fully supervised baseline without the need for any dataset specific training. For instance, we match the accuracy of the original ResNet-50 on ImageNet zero-shot without needing to use any of the 1.28 million training examples it was trained on. We release our code and pre-trained model weights at https://github.com/OpenAI/CLIP.

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets removing the need for specialized output heads or dataset specific customization. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on crowd-labeled datasets such as ImageNet (Deng et al., 2009). Could scalable pre-training methods which learn directly from web text result in a similar breakthrough in computer vision? Prior work is encouraging.

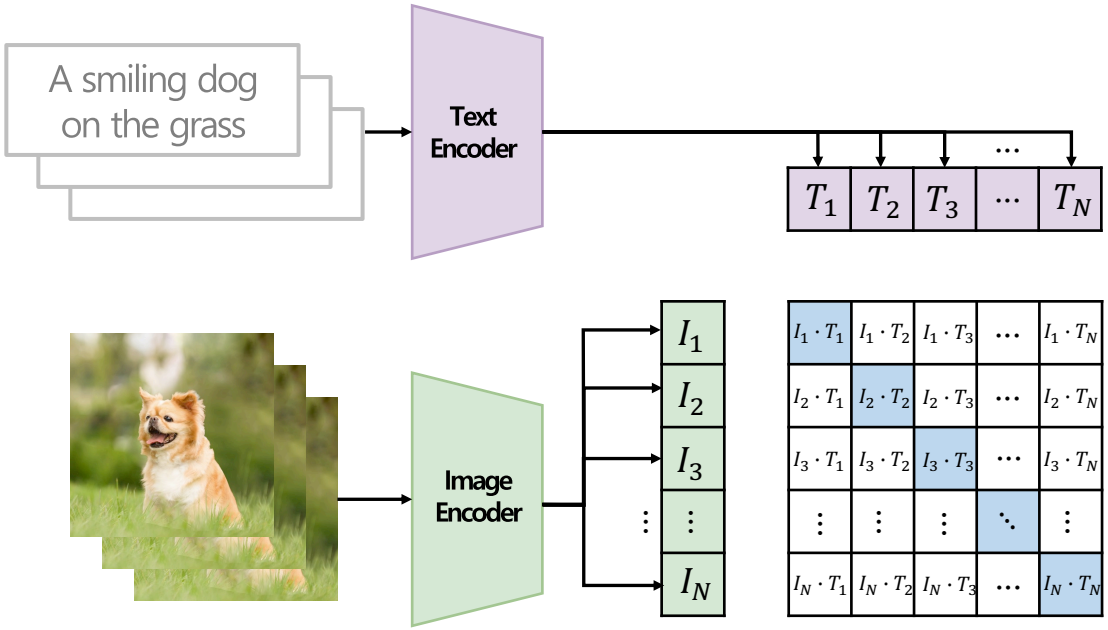
Over 20 years ago Mori et al. (1999) explored improving content based image retrieval by training a model to predict the nouns and adjectives in text documents paired with images. Quattoni et al. (2007) demonstrated it was possible to learn more data efficient image representations via manifold learning in the weight space of classifiers trained to predict words in captions associated with images. Srivastava & Salakhutdinov (2012) explored deep representation learning by training multimodal Deep Boltzmann Machines on top of low-level image and text tag features. Joulin et al. (2016) modernized this line of work and demon-

CLIP

❖ Experiments

- Image Encoder: ResNet-D
- Text Encoder: Transformer

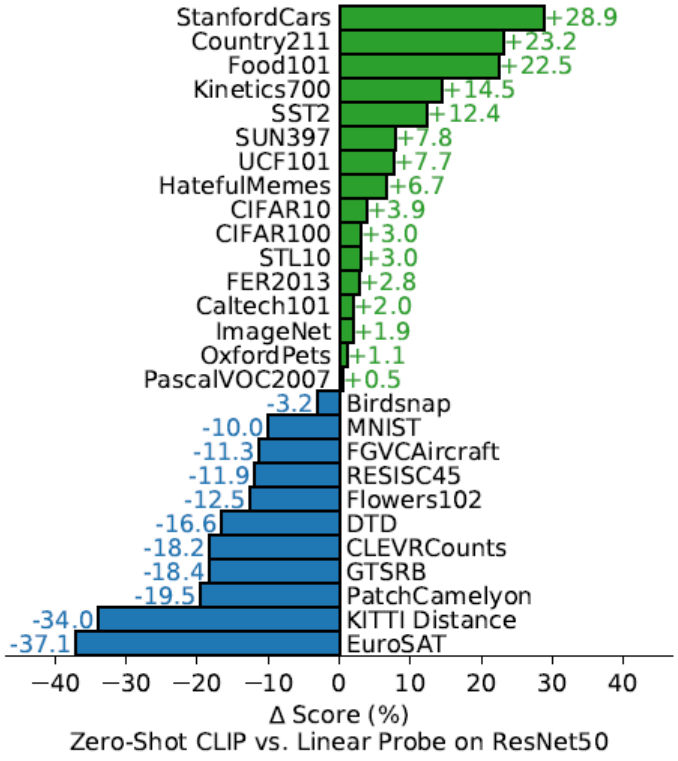
```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
```

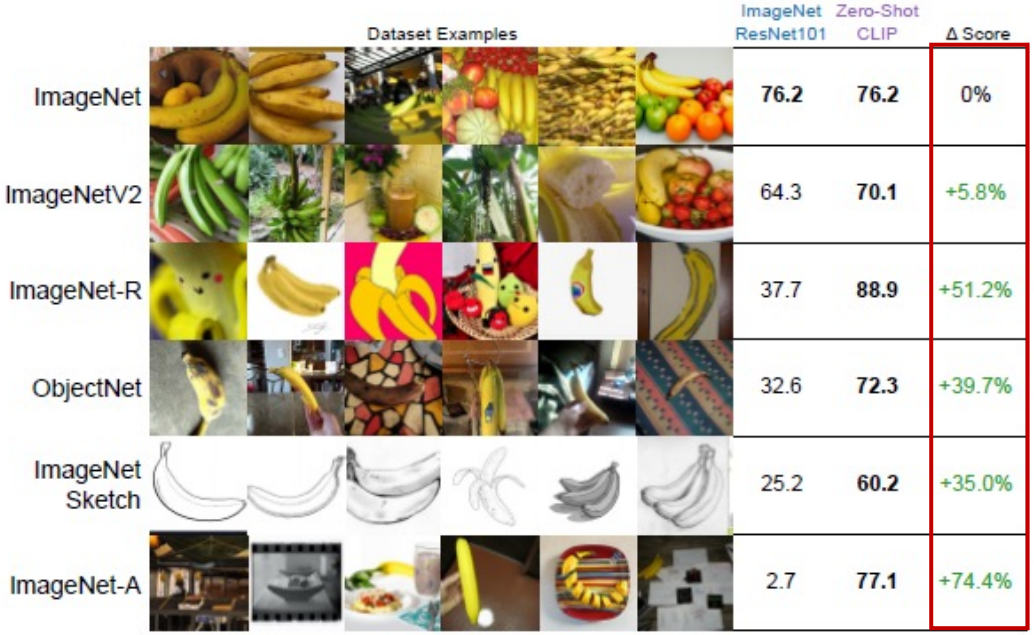


CLIP

❖ Results

- **Zero-shot CLIP vs Fully supervised ResNet50**
- Robust on natural distribution shift

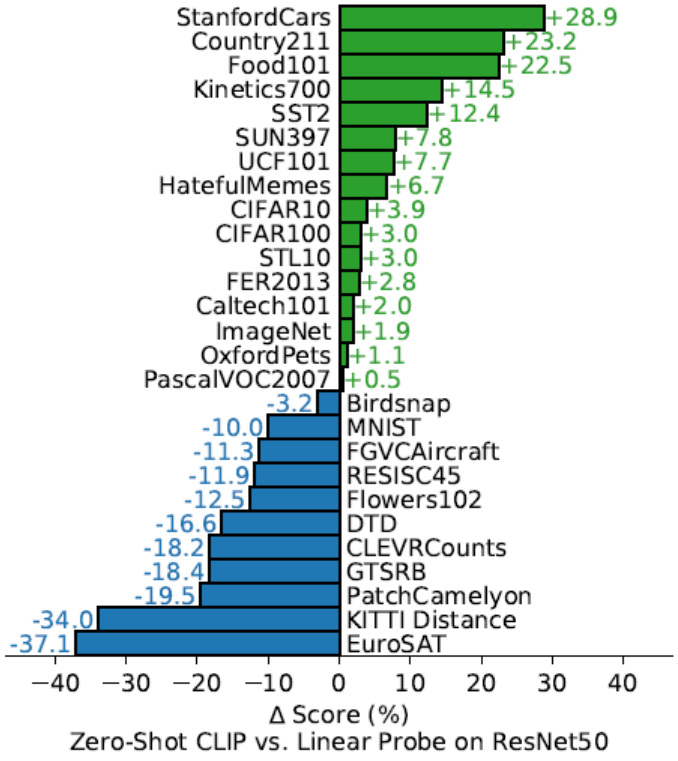


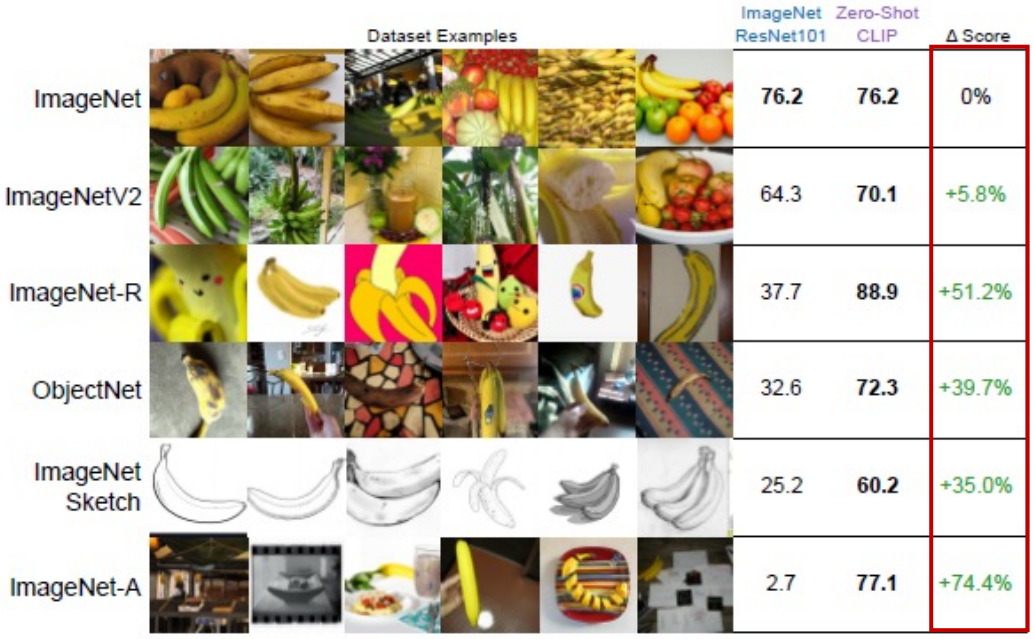


CLIP

❖ Results

- Zero-shot CLIP vs Fully supervised ResNet50
- **Robust on natural distribution shift**





Multimodal Contrastive learning

❖ Revisiting Multimodal Representation in Contrastive Learning: From Patch and Token Embeddings to Finite Discrete Tokens(CVPR, 2023)

- 기존 CLIP의 문제를 해결하기 위해 나온 방법론
- Finite Discrete Tokens(FDT)를 통해 이러한 문제를 해결

Revisiting Multimodal Representation in Contrastive Learning: From Patch and Token Embeddings to Finite Discrete Tokens

Yuxiao Chen[1*], Jianbo Yuan[2], Yu Tian[2], Shijie Geng[1,2], Xinyu Li[2],
Ding Zhou[2], Dimitris N. Metaxas[1†], Hongxia Yang[3]
[1]Rutgers University [2]ByteDance Inc. [3]Zhejiang University
{yc984, sg1309, dnm}@rutgers.edu,
{jianbo.yuan, yutian.yt, lixinyu.arthur, ding.zhou}@bytedance.com
hongxia.yang1@gmail.com

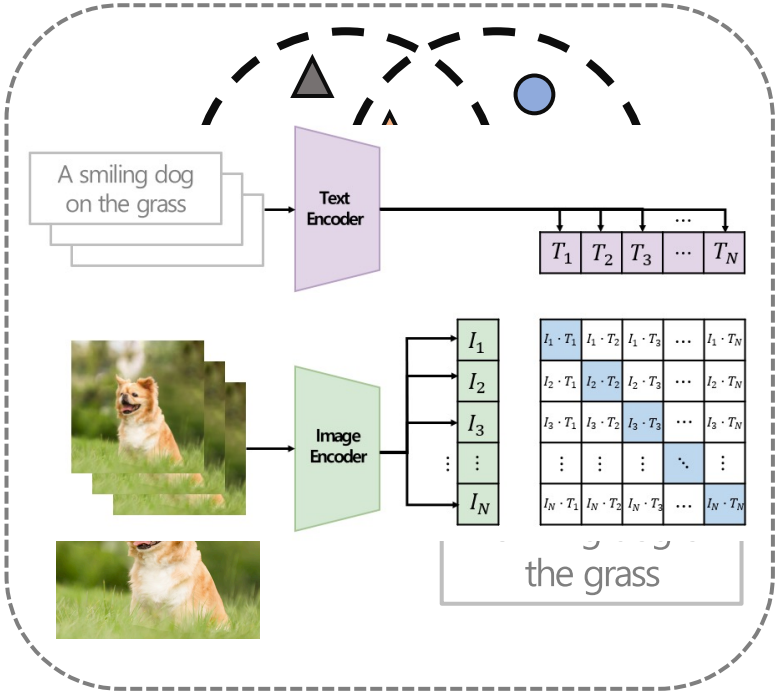
Abstract

Contrastive learning-based vision-language pre-training approaches, such as CLIP, have demonstrated great success in many vision-language tasks. These methods achieve cross-modal alignment by encoding a matched image-text pair with similar feature embeddings, which are generated by aggregating information from visual patches and language tokens. However, direct aligning cross-modal information using such representations is challenging, as visual patches and text tokens differ in semantic levels and granularities. To alleviate this issue, we propose a Finite Discrete Tokens (FDT) based multimodal representation. FDT is a set of learnable tokens representing certain visual-semantic concepts. Both images and texts are embedded using shared FDT by first grounding multimodal inputs to FDT space and then aggregating the activated FDT representations. The matched visual and semantic concepts are enforced to be represented by the same set of discrete tokens by a sparse activation constraint. As a result, the granularity gap between the two modalities is reduced. Through both quantitative and qualitative analyses, we demonstrate that using FDT representations in CLIP-style models improves cross-modal alignment and performance in visual recognition and vision-language downstream tasks. Furthermore, we show that our method can learn more comprehensive representations, and the learned FDT capture meaningful cross-modal correspondence, ranging from objects to actions and attributes.[1]

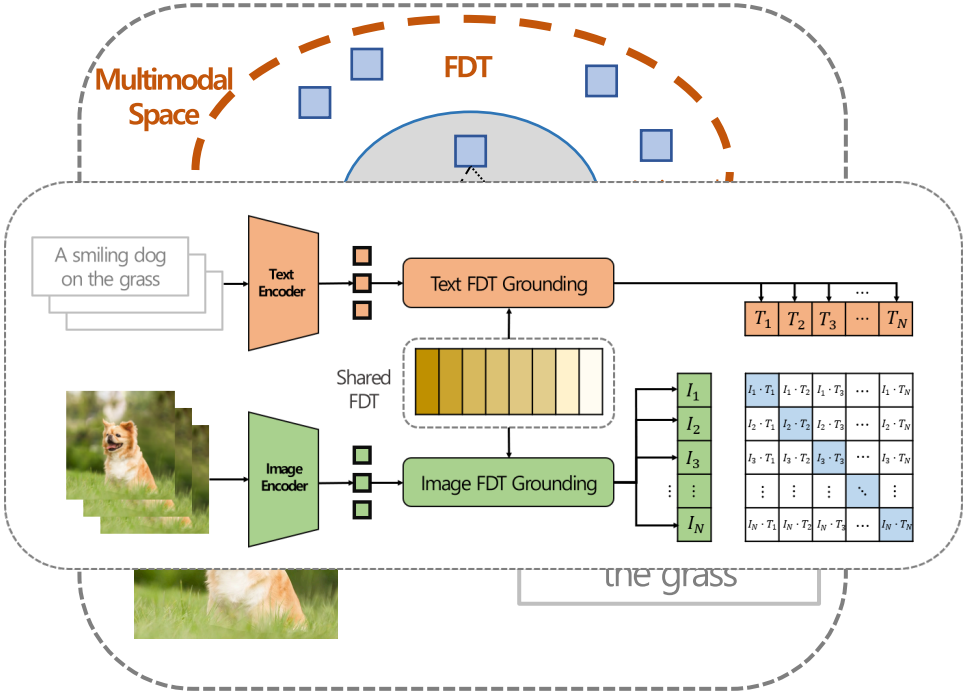
Multimodal Contrastive learning

❖ Motivation

- CLIP은 image와 text를 단순하게 align하여 학습을 진행하므로 성능저하와 특정 semantic concepts을 간과할 수 있음
- Learnable tokens(FDT)를 적용하여 modal간 semantic concepts를 공유할 수 있음 → 성능 향상

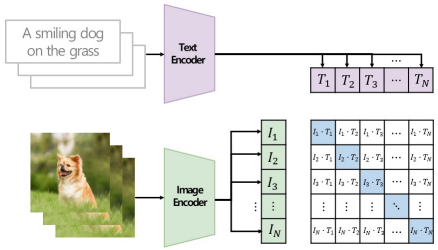


<CLIP>



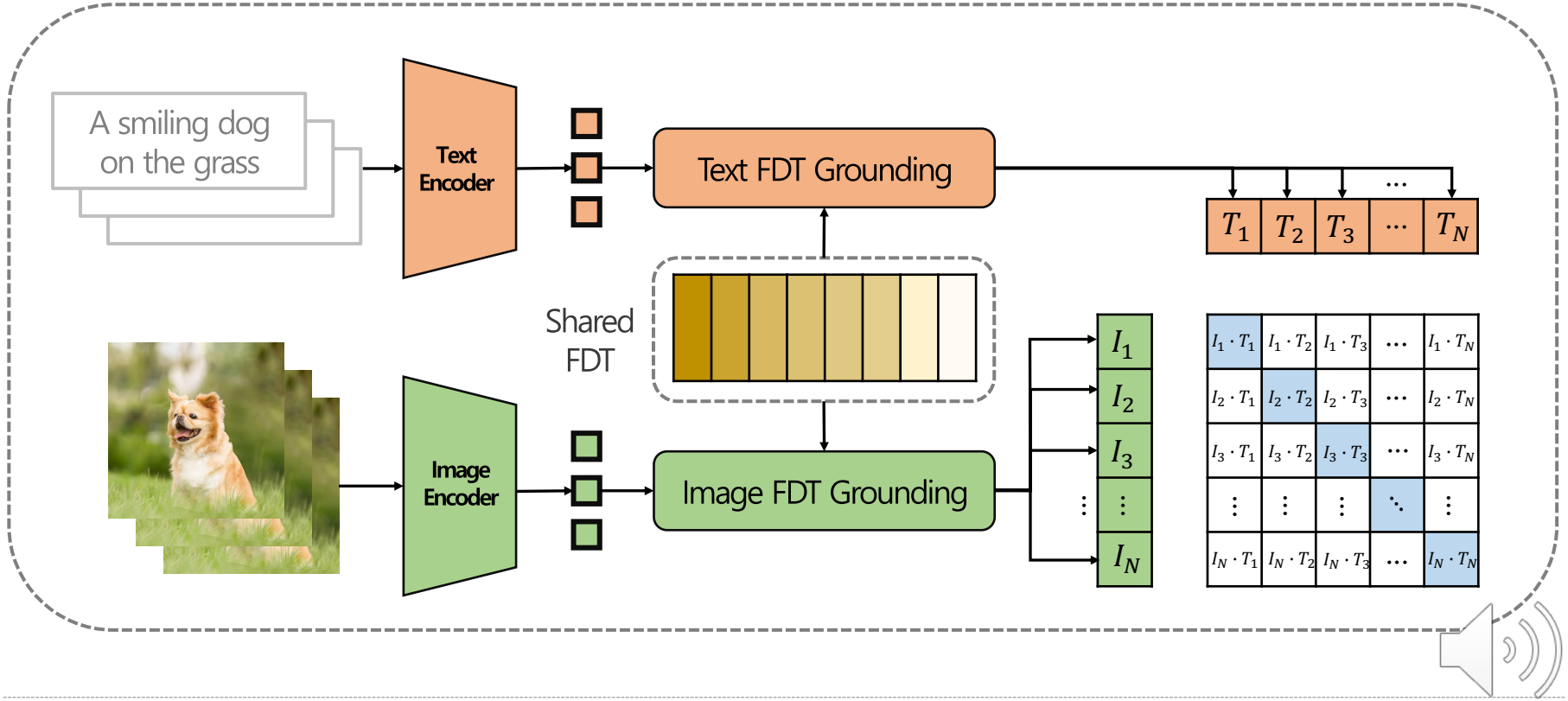
<FDT>

FDT

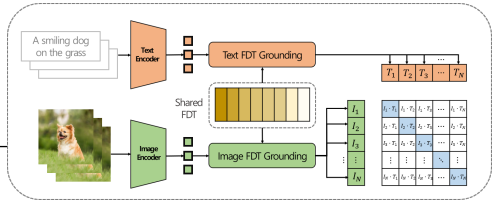


❖ Motivation

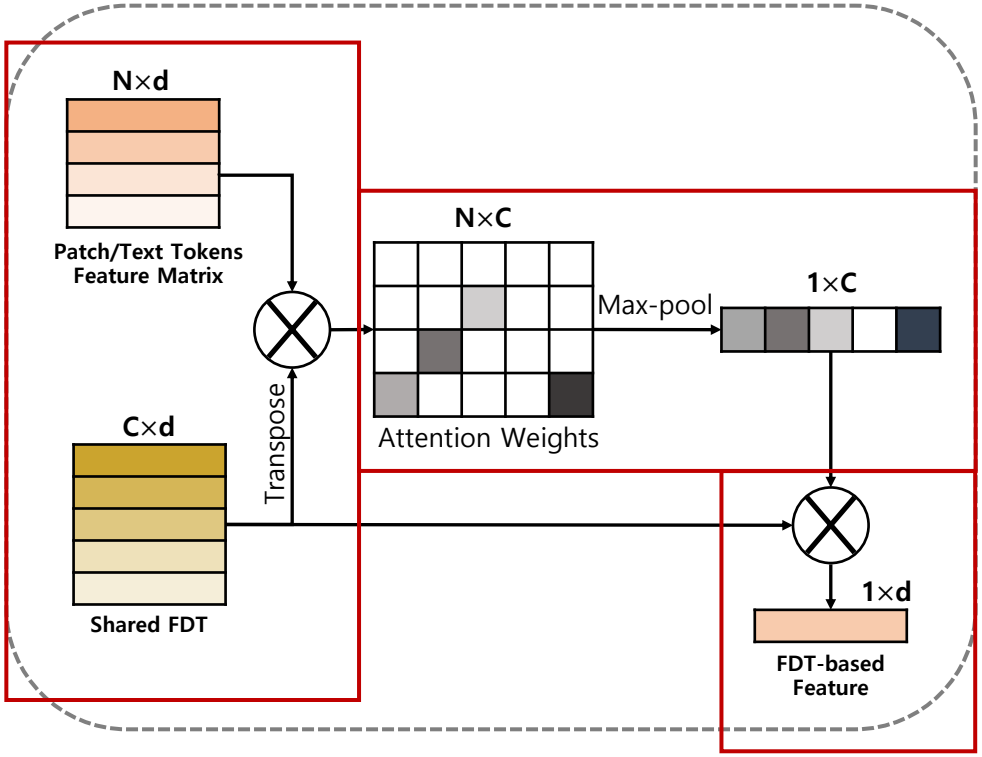
- CLIP은 image와 text를 단순하게 align하여 학습을 진행하므로 성능저하와 특정 semantic concepts을 간과할 수 있음
- Learnable tokens(FDT)를 적용하여 modal간 semantic concepts를 공유할 수 있음 → 성능 향상



FDT



❖ FDT(Finite Discrete Tokens)



$$r_i^t = \max_j < f_{t_j}, c_i >$$

$$w_i^t = \frac{e^{r_i^t}}{\sum_j^C e^{r_j^t}}$$

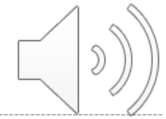
$$f_t^{\mathrm{FDT}} = \sum_i^C w_i^t \cdot c_i$$

$$\mathcal{L} = -\frac{1}{N}\sum_i^N \log \frac{\exp\left(\mathrm{sim}\left(f_{v_i}^{\mathrm{FDT}}, f_{t_i}^{\mathrm{FDT}}\right)/\tau\right)}{\sum_{j=1}^N \exp\left(\mathrm{sim}\left(f_{v_i}^{\mathrm{FDT}}, f_{t_j}^{\mathrm{FDT}}\right)/\tau\right)}$$
$$-\frac{1}{N}\sum_i^N \log \frac{\exp\left(\mathrm{sim}\left(f_{t_i}^{\mathrm{FDT}}, f_{v_i}^{\mathrm{FDT}}\right)/\tau\right)}{\sum_{j=1}^N \exp\left(\mathrm{sim}\left(f_{t_i}^{\mathrm{FDT}}, f_{v_j}^{\mathrm{FDT}}\right)/\tau\right)},$$

FDT

❖ Experiments

- Pre-training settings: 15M, 30M, and 145M
- Downstream Task
  - ➢ Zero-shot image classification
  - ➢ Linear probe image classification
  - ➢ Zero-shot image-text retrieval
  - ➢ VQA



FDT

❖ Experiments(**zero-shot image classification**)

- Pre-training settings: 15M
- CLIP, DeCLIP에 FDT 방법론을 적용한 경우 가장 우수한 성능을 보임

	C10	C100	F101	PETS	FLOW	SUN	DTD	CAL	IN	AVG
SLIP [36]	50.7	25.5	33.3	23.5	49.0	34.7	14.4	59.9	34.3	36.1
MS-CLIP-S [52]	-	-	-	-	-	-	-	-	36.7	-
CLIP [40]	60.4	33.5	39.6	23.1	54.0	42.0	17.0	65.5	37.0	41.3
FILIP [51]	65.1	34.2	43.2	24.1	52.8	50.8	24	68.9	39.5	44.7
DeCLIP [28]	72.8	40.3	49.9	36.2	60.1	48.8	26.4	72.7	43.2	50.0
CLIP+FDT (Ours)	67.7	39.9	42.9	25.8	55.5	45.5	26.5	69.6	39.3	45.9
DeCLIP+FDT (Ours)	**75.7**	**45.2**	**52.9**	**40.7**	**64.6**	**52.0**	**30.7**	**76.2**	**45.8**	**53.8**

Table 2. Zero-shot image classification accuracy (%) under the 15M setting. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. IN is ImageNet-1K. "AVG" is the average accuracy over all datasets.

	C10	C100	F101	PETS	FLOW	SUN	CARS	DTD	CAL	AIR	AVG
SLIP [36]	87.4	69.5	71.3	70.5	91.9	66.9	27.5	65.6	86.2	27.7	66.5
MS-CLIP-S [52]	87.2	66.7	76.0	62.1	93.8	71.7	27.5	69.4	81.6	**32.9**	66.9
CLIP [40]	88.3	68.6	72.1	72.5	92.6	69.5	29.8	67.8	86.2	27.7	67.5
FILIP [51]	86.5	66.6	71.7	69.2	93	69.6	30.0	66.4	85.7	27.0	66.6
DeCLIP [28]	89.4	69.6	75.9	71.4	95.7	71.6	30.1	66.9	89.0	26.7	68.6
CLIP+FDT (Ours)	89.1	**71.2**	74.4	73.0	93.4	70.8	31.4	69.4	87.7	27.9	68.8
DeCLIP+FDT (Ours)	**89.8**	**71.2**	**77.7**	**73.9**	**95.7**	**72.9**	**33.7**	**69.6**	**89.4**	26.9	**70.1**

Table 3. Linear probing image classification accuracy (%) under the 15M setting. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. Air is Aircraft. "AVG" is the average accuracy over all datasets.

FDT

❖ Experiments(**linear probing image classification**)

- Pre-training settings: 15M
- CLIP, DeCLIP에 FDT 방법론을 적용한 경우 가장 우수한 성능을 보임

	C10	C100	F101	PETS	FLOW	SUN	DTD	CAL	IN	AVG
SLIP [36]	50.7	25.5	33.3	23.5	49.0	34.7	14.4	59.9	34.3	36.1
MS-CLIP-S [52]	-	-	-	-	-	-	-	-	36.7	-
CLIP [40]	60.4	33.5	39.6	23.1	54.0	42.0	17.0	65.5	37.0	41.3
FILIP [51]	65.1	34.2	43.2	24.1	52.8	50.8	24	68.9	39.5	44.7
DeCLIP [28]	72.8	40.3	49.9	36.2	60.1	48.8	26.4	72.7	43.2	50.0
CLIP+FDT (Ours)	67.7	39.9	42.9	25.8	55.5	45.5	26.5	69.6	39.3	45.9
DeCLIP+FDT (Ours)	**75.7**	**45.2**	**52.9**	**40.7**	**64.6**	**52.0**	**30.7**	**76.2**	**45.8**	**53.8**

Table 2. Zero-shot image classification accuracy (%) under the 15M setting. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. IN is ImageNet-1K. "AVG" is the average accuracy over all datasets.

	C10	C100	F101	PETS	FLOW	SUN	CARS	DTD	CAL	AIR	AVG
SLIP [36]	87.4	69.5	71.3	70.5	91.9	66.9	27.5	65.6	86.2	27.7	66.5
MS-CLIP-S [52]	87.2	66.7	76.0	62.1	93.8	71.7	27.5	69.4	81.6	**32.9**	66.9
CLIP [40]	88.3	68.6	72.1	72.5	92.6	69.5	29.8	67.8	86.2	27.7	67.5
FILIP [51]	86.5	66.6	71.7	69.2	93	69.6	30.0	66.4	85.7	27.0	66.6
DeCLIP [28]	89.4	69.6	75.9	71.4	95.7	71.6	30.1	66.9	89.0	26.7	68.6
CLIP+FDT (Ours)	89.1	**71.2**	74.4	73.0	93.4	70.8	31.4	69.4	87.7	27.9	68.8
DeCLIP+FDT (Ours)	**89.8**	**71.2**	**77.7**	**73.9**	**95.7**	**72.9**	**33.7**	**69.6**	**89.4**	26.9	**70.1**

Table 3. Linear probing image classification accuracy (%) under the 15M setting. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. Air is Aircraft. "AVG" is the average accuracy over all datasets.

FDT

❖ Experiments(**different scales of taring data & image encoder architectures**)

- 145M개의 데이터로 학습한 모델이 가장 성능이 우수함
- Image encoder는 VIT-B/16인 경우 가장 성능이 우수함

	Setting	ZS CLS AVG Acc	LP CLS AVG Acc	ZS-Flickr30K IR R@1	ZS-Flickr30K TR R@1	ZS-Flickr30K rsum	ZS-MSCOCO IR R@1	ZS-MSCOCO TR R@1	ZS-MSCOCO rsum	VQAv2 overall
CLIP	15M	41.3	67.5	27.6	42.8	343.1	15.9	24.8	236.8	47.5
CLIP+FDT	15M	45.9(↑4.6)	68.8(↑1.3)	32.6(↑5.0)	51.0(↑8.2)	376.5(↑33.4)	19.4(↑3.5)	29.6(↑4.8)	263.1(↑26.3)	50.6(↑3.1)
CLIP	30M	56.8	73.8	43.6	58.8	431.3	23.3	34.8	300.8	50.6
CLIP+FDT	30M	61.2(↑ 4.4)	75.6 (↑ 1.8)	52.5(↑8.9)	70.8(↑12.0)	474.2(↑42.9)	28.3(↑5.0)	43(↑8.2)	337.1 (↑36.3)	53.4(↑2.8)
CLIP	145M	64	82.1	52.6	67.9	469.8	29.3	42.1	335.2	53.1
CLIP+FDT	145M	69.0(↑ 5.0)	82.3 (↑ 0.2)	56.3(↑3.7)	75.9(↑8.0)	489.4(↑19.6)	31.0(↑1.7)	46.4(↑4.3)	353.0(↑17.8)	55.2(↑2.1)

Table 5. Ablation study results when using different scales of training data. “ZS” means zero-shot. “AVG” is average. “ACC” is accuracy. “LP” stands for linear prob. “CLS” represents classification. “IR” and “TR” are image retrieval and text retrieval, respectively.

	ZS CLS AVG Acc	LP CLS AVG Acc	ZS-Flickr30K IR R@1	ZS-Flickr30K TR R@1	ZS-Flickr30K rsum	ZS-MSCOCO IR R@1	ZS-MSCOCO TR R@1	ZS-MSCOCO rsum	VQAv2 Overall
CLIP-ViT-B/32	41.3	67.5	27.6	42.8	343.1	15.9	24.8	236.8	47.5
CLIP-ViT-B/32+FDT	45.9(↑4.6)	68.8(↑1.3)	32.6(↑5.0)	51.0(↑8.2)	376.5(↑33.4)	19.4(↑3.5)	29.6(↑4.8)	263.1(↑26.3)	50.6(↑3.1)
CLIP-ViT-B/16	45.2	68.8	35.3	50.5	387.8	19.3	29.7	263.6	49.2
CLIP-ViT-B/16+FDT	49.9(↑4.7)	71.3(↑2.5)	41.6(↑6.3)	60.8(↑10.3)	425.5(↑37.7)	23.4(↑4.1)	35.3(↑5.6)	295.4 (↑31.8)	54.3(↑5.1)
CLIP-Swin-B	39.6	68.5	30.5	48.5	368.1	17.7	26.0	247.6	46.5
CLIP-Swin-B+FDT	42.4(↑2.8)	70.7(↑2.2)	39.6(↑9.1)	57.9(↑9.4)	415.5(↑47.4)	22.3(↑4.6)	33.8(↑7.8)	288.3(↑40.7)	51.6(↑5.1)

Table 6. Ablation Study results when using different image encoder architectures. “ZS” means zero-shot. “AVG” is average. “ACC” is accuracy. “LP” stands for linear prob. “CLS” represents classification. “IR” and “TR” are image retrieval and text retrieval.

FDT

❖ Visualization of Learned FDT

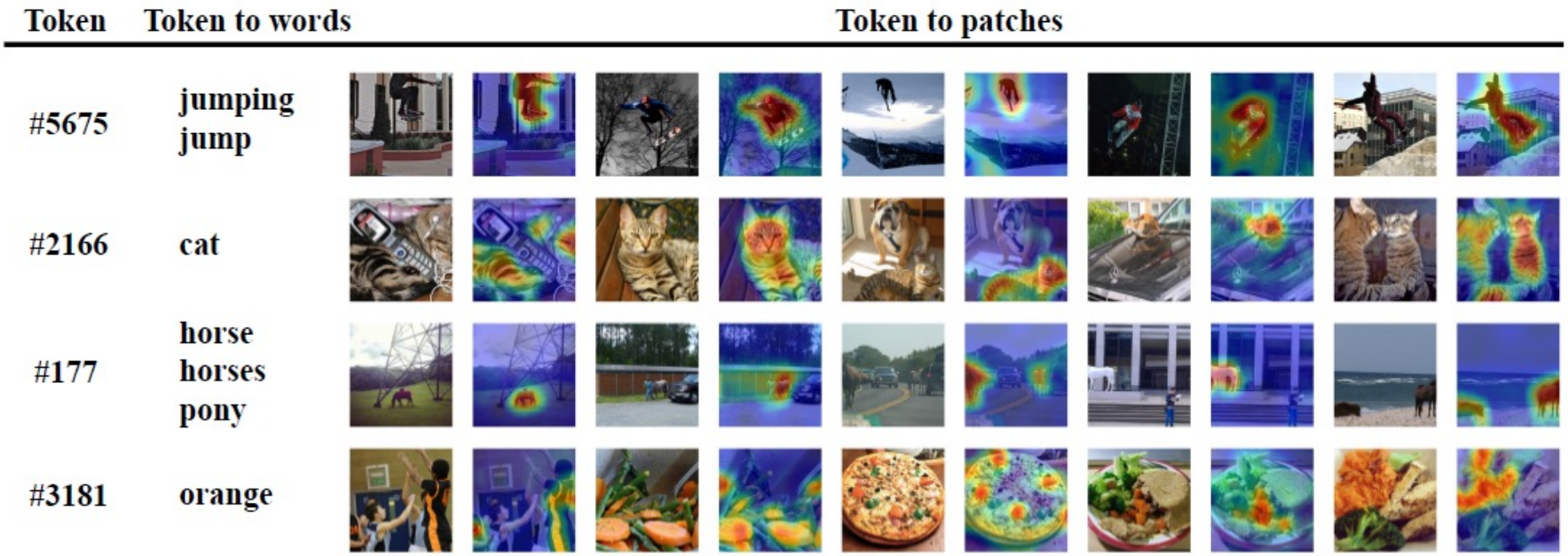


Figure 4. Example of the top-5 most relevant image patches and text tokens of four FDT tokens. Note that the redundant text tokens in the top-5 are removed. The color of the heatmap from blue to red denotes the relevance between patches and FDT from small to large.

Conclusion

❖ Multimodal learning

- 두개 이상의 데이터 타입을 학습에 동시에 사용하는 방법론
- 최근 세계적인 기업에서 multimodal을 활용한 MLLM방법론이 활발히 연구 진행

❖ Contrastive learning

- 유사한 데이터는 가깝게 다른 데이터는 멀게 학습하여 label 없이도 좋은 성능을 내고자 하는 방법론

❖ Multimodal Contrastive learning

- Label이 없는 여러 타입의 데이터를 사용하여 좋은 representation을 생성하는 방법론
- Model architecture는 비슷하지만 서로 다른 데이터 특성을 같은 공간에서 보다 잘 representation 시키고자 하는 연구들이 진행 중