VLM 끄적끄적 2

Pretraining Task (1): Contrastive Learning

1. CLIP: Image-Text CL

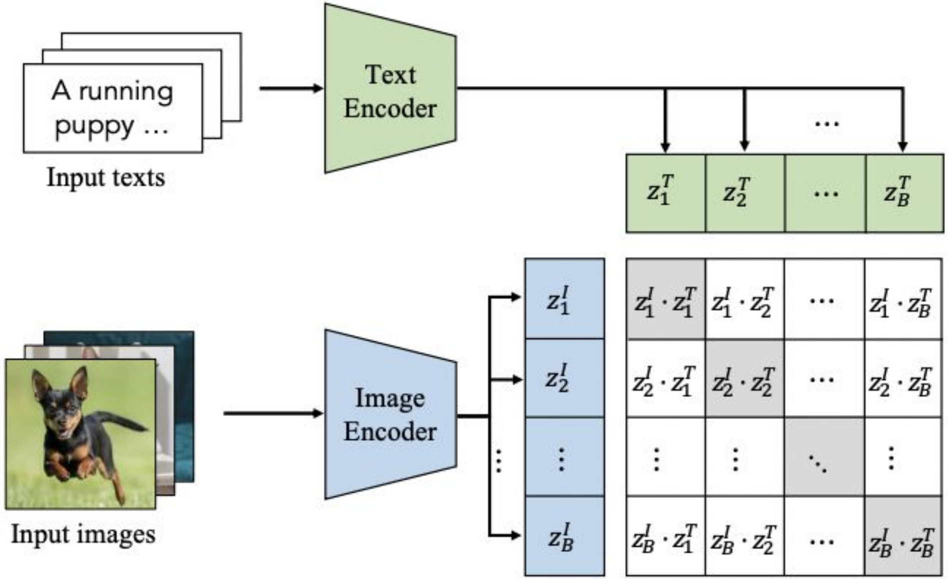


Fig. 6: Illustration of the image-text contrastive learning in CLIP [10]. Figure is reproduced from [10].

2. CLIP 발전시키는 방향 (Image-Text CL)

1. Large-scale datasets: “데이터셋을 더 키우자"
2. Small-scale datasets: “데이터셋을 (알짜배기만) 효율적으로 사용하자"
3. Various semantic level: “not only 거시적(coarse) but also 미시적 (dense)”
4. Augmentation
5. Unifying vision & language encoder

2. CLIP 발전시키는 방향 (Image-Text CL)

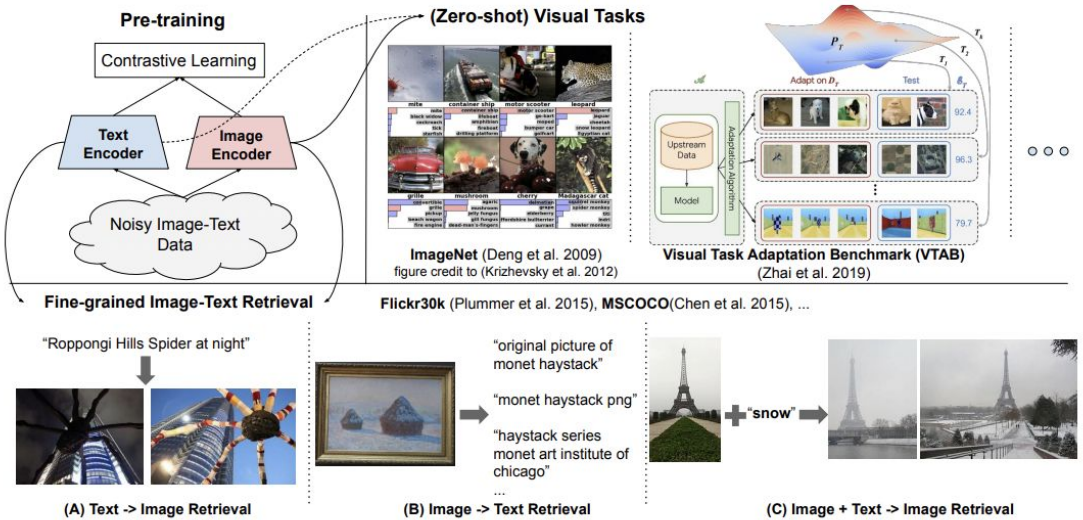
1. Large-scale datasets
   - ALIGN (ICML 2021),
2. Small-scale datasets
   - DeCLIP (ICLR 2022), OTTER (ICLR 2022), ZeroVL (ECCV 2022)
3. Various semantic levels
   - FILIP (ICLR 2022), PyramidCLIP (NeurIPS 2022)
4. Augmentation
   - LaCLIP (NeurIPS 2023), ALIP (ICCV 2023), RA-CLIP (CVPR 2023)
5. Unifying vision & language encoder
   - CLIPPO (CVPR 2023), OneR (AAAI 2023)

Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision

2. CLIP 발전시키는 방향 (Image-Text CL)

(1) Large-scale datasets

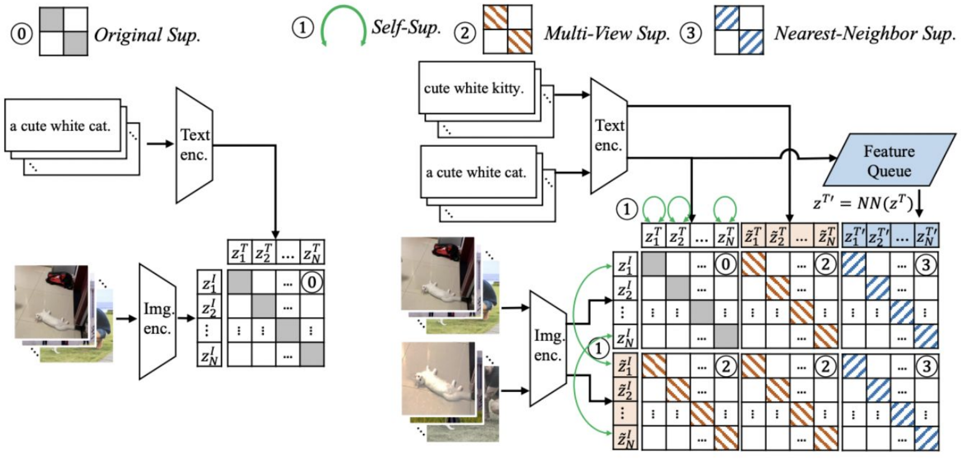
- ALIGN (ICML 2021)
  - Noisy pair도 활용
  - 다양한 task



Pre-training
Contrastive Learning
Text Encoder
Image Encoder
Noisy Image-Text Data
(Zero-shot) Visual Tasks
ImageNet (Deng et al. 2009)
figure credit to (Krizhevsky et al. 2012)
Upstream Data
Model
Adaptation Algorithm
Adapt on D_T
Test
92.4
96.3
79.7
Visual Task Adaptation Benchmark (VTAB)
(Zhai et al. 2019)
Fine-grained Image-Text Retrieval
Flickr30k (Plummer et al. 2015), MSCOCO(Chen et al. 2015), ...
"Roppongi Hills Spider at night"
"original picture of monet haystack"
"monet haystack png"
"haystack series monet art institute of chicago"
...
"snow"
(A) Text -> Image Retrieval
(B) Image -> Text Retrieval
(C) Image + Text -> Image Retrieval

Figure 1. A summary of our method, ALIGN. Visual and language representations are jointly learned from noisy image alt-text data. The representations can be used for vision-only or vision-language task transfer. Without any fine-tuning, ALIGN powers zero-shot visual classification and cross-modal search including image-to-text search, text-to-image search and even search with joint image+text queries.

2. CLIP 발전시키는 방향 (Ima

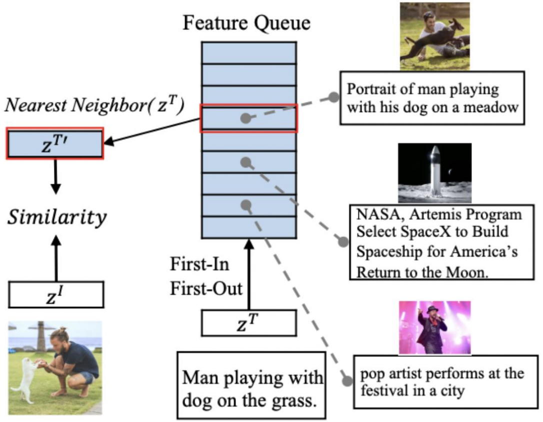


(a) CLIP & ALIGN

(b) DeCLIP

(2) Small-scale dataset

- DeCLIP (ICLR 2022)
  - CLIP 한계점: Data-hungry
  - 해결책: Data-efficient CLIP!
  - How? Image-text pair의 supervision을 활용하
  - Details:
    - (1) Self-supervision **within** each modality
    - (2) Multi-view supervision **across** modalities
    - (3) NN supervision from other similar pairs


Feature Queue
Nearest Neighbor(z^T)
$z^{T'}$
Similarity
z^I
First-In
First-Out
z^T
Man playing with dog on the grass.
Portrait of man playing with his dog on a meadow
NASA, Artemis Program Select SpaceX to Build Spaceship for America's Return to the Moon.
pop artist performs at the festival in a city

Data Efficient Language-supervised Zero-shot Recognition with Optimal Transport Distillation

2. CLIP 발전시키는 방향 (Image-Text CL)

(2) Small-scale dataset

- OTTER (ICLR 2022)
    - CLIP 한계점: Data-hungry (이유: Noisy pairs)
    - 해결책: 3M장만으로 만들
- ZeroVL (ECCV 2022)

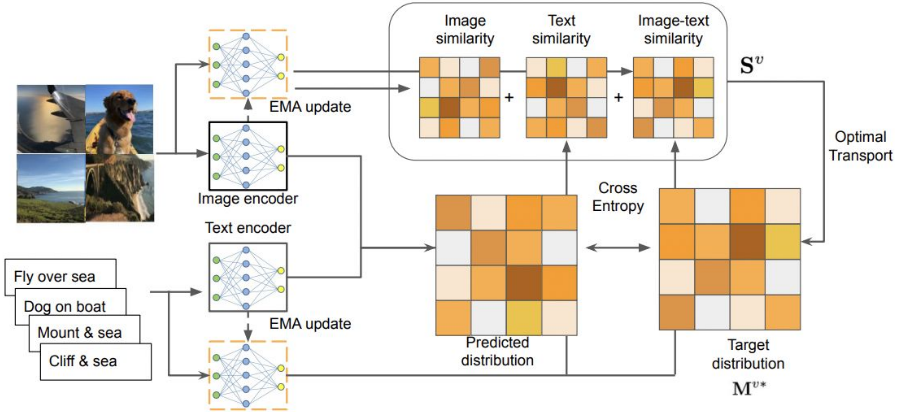


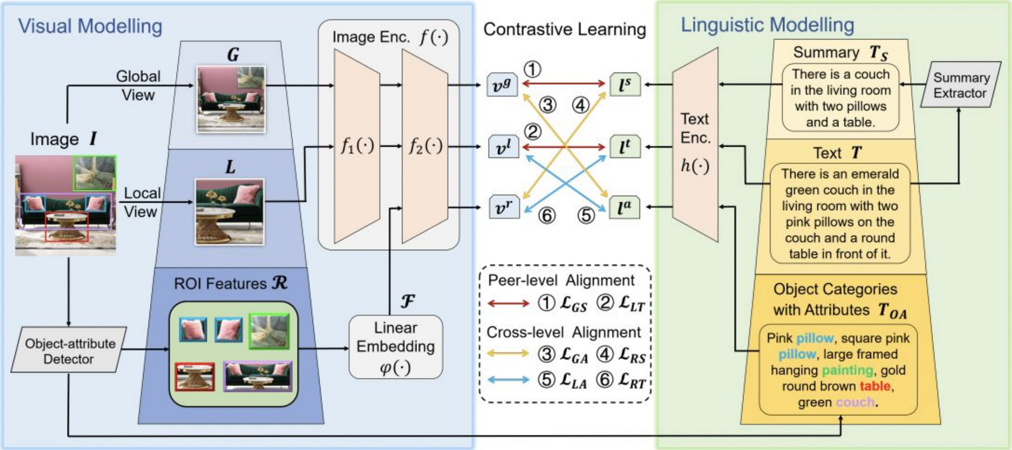
Figure 2: Architecture of OTTER. We use image and text embeddings to compute similarity matrices $\mathbf{S}^v$ (and $\mathbf{S}^t$), which is then used to solve for matching probabilities $\mathbf{M}^{v*}$ (and $\mathbf{M}^{t*}$) as targets.

PyramidCLIP: Hierarchical Feature Alignment for Vision-language Model Pretraining

2. CLIP 발전시키는 방향 (Image-Text CL)

(3) Various Semantic Level

- PyramidCLIP (NeurIPS 2022)
    - (1) Cross-level
    - (2) Peer-level



Visual Modelling
Global View
Image I
Local View
G
L
ROI Features $\mathcal{R}$
Object-attribute Detector
Image Enc. $f(\cdot)$
$f_1(\cdot)$
$f_2(\cdot)$
$\mathcal{F}$
Linear Embedding $\varphi(\cdot)$
Contrastive Learning
v^g
v^l
v^r
l^s
l^t
l^a
① ② ③ ④ ⑤ ⑥
Peer-level Alignment
① $\mathcal{L}_{GS}$ ② $\mathcal{L}_{LT}$
Cross-level Alignment
③ $\mathcal{L}_{GA}$ ④ $\mathcal{L}_{RS}$
⑤ $\mathcal{L}_{LA}$ ⑥ $\mathcal{L}_{RT}$
Linguistic Modelling
Text Enc. $h(\cdot)$
Summary T_S
There is a couch in the living room with two pillows and a table.
Summary Extractor
Text T
There is an emerald green couch in the living room with two pink pillows on the couch and a round table in front of it.
Object Categories with Attributes T_{OA}
Pink pillow, square pink pillow, large framed hanging painting, gold round brown table, green couch.

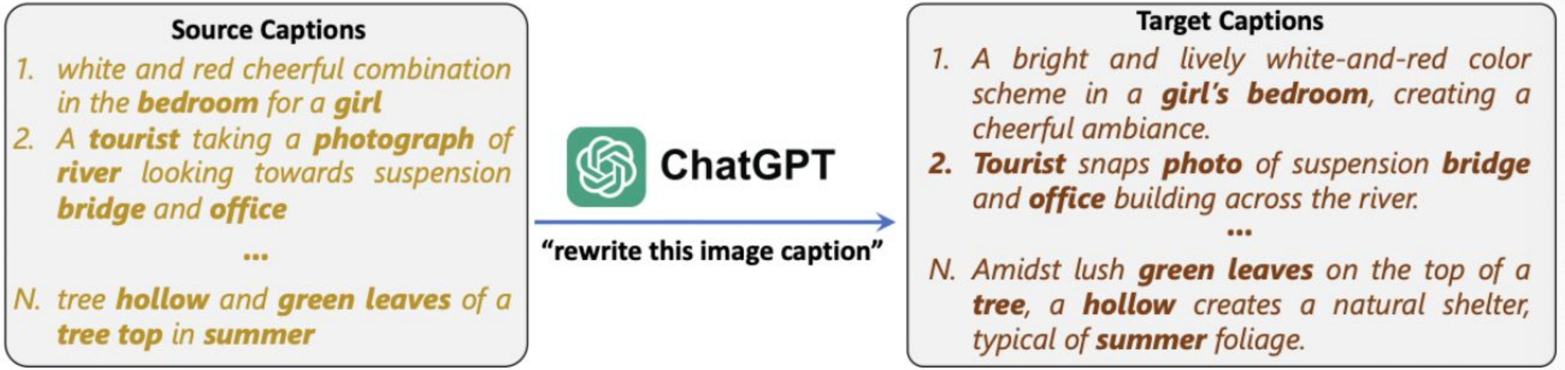
Figure 2: Overall architecture of the proposed PyramidCLIP which is a dual-stream network. The input elements of visual modelling and linguistic modelling both have three-level semantics. The elements of the two modalities interact through peer-level semantics alignment and cross-level relation alignment.

Improving CLIP Training with Language Rewrites

2. CLIP 발전시키는 방향 (Image-Text CL)

(4) Augmentation

- LaCLIP (NeurIPS 2023)
    - La = Language augmented
    - Solution: Synthetic caption via LLM => Exhibit diversity



Source Captions

1. *white and red cheerful combination in the* ***bedroom*** *for a* ***girl***
2. *A* ***tourist*** *taking a* ***photograph*** *of* ***river*** *looking towards suspension* ***bridge*** *and* ***office***

...

N. *tree* ***hollow*** *and* ***green leaves*** *of a* ***tree top*** *in* ***summer***

Target Captions

1. *A bright and lively white-and-red color scheme in a* ***girl's bedroom****, creating a cheerful ambiance.*
2. ***Tourist*** *snaps* ***photo*** *of suspension* ***bridge*** *and* ***office*** *building across the river.*

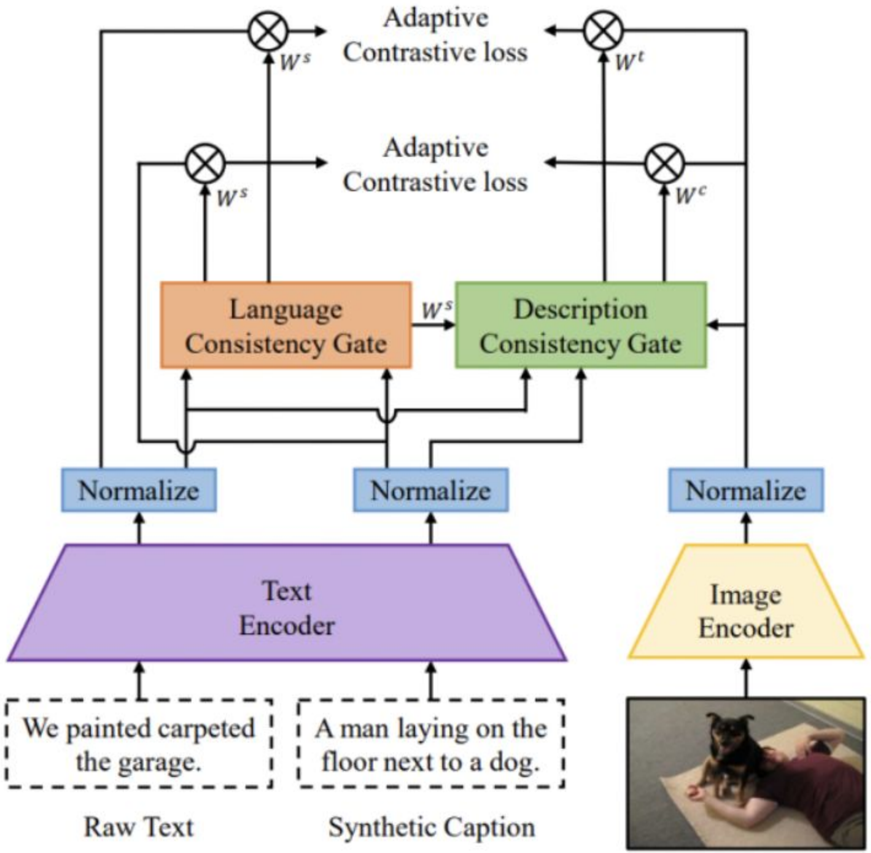
...

N. *Amidst lush* ***green leaves*** *on the top of a* ***tree****, a* ***hollow*** *creates a natural shelter, typical of* ***summer*** *foliage.*

2. CLIP 발전시키는 방향 (Image-Text CL)

(4) Augmentation

- ALIP (ICCV 2023)
  - ALIP = Adaptive LIP
  - CLIP 한계점: Noisy data
    (= Unmatched image & text)
  - Solution: Use both….
    - (1) Raw text
    - (2) Synthetic caption


Adaptive Contrastive loss
W^s
W^t
Adaptive Contrastive loss
W^s
W^c
Language Consistency Gate
W^s
Description Consistency Gate
Normalize
Normalize
Normalize
Text Encoder
Image Encoder
We painted carpeted the garage.
A man laying on the floor next to a dog.
Raw Text
Synthetic Caption

2. CLIP 발전시키는 방향 (Image-Text CL)

(4) Augmentation

- RA-CLIP (CVPR 2023)
    - RA = Retrieval Augmented
    - CLIP 한계점: Data-hungry
    - Solution: Augment embeddings by “online” retrieval
        - Step 1) Use “hold-out” reference set
        - Step 2) Given input, retrieve relevant pairs from reference set

2. CLIP 발전시키는 방향 (Image-Text CL)

(4) Augmentation

- RA-CLIP (CVPR 2023)

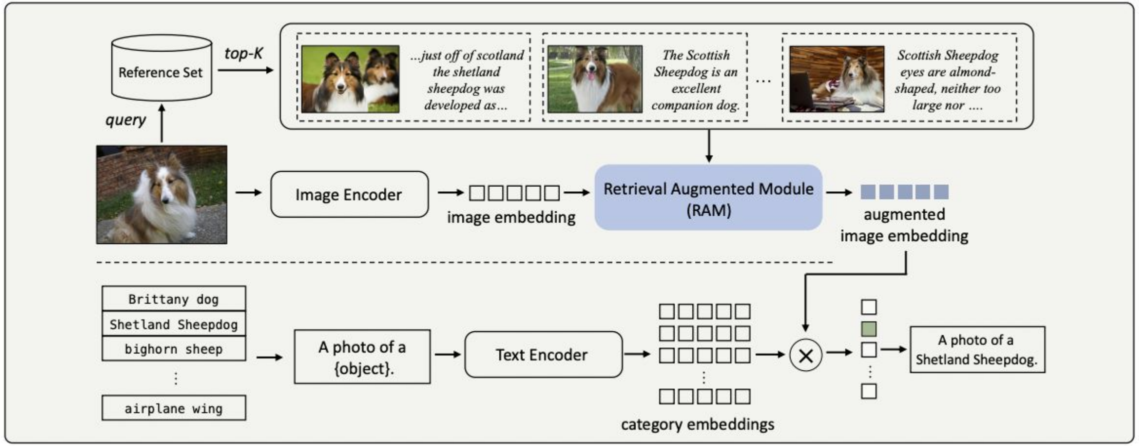
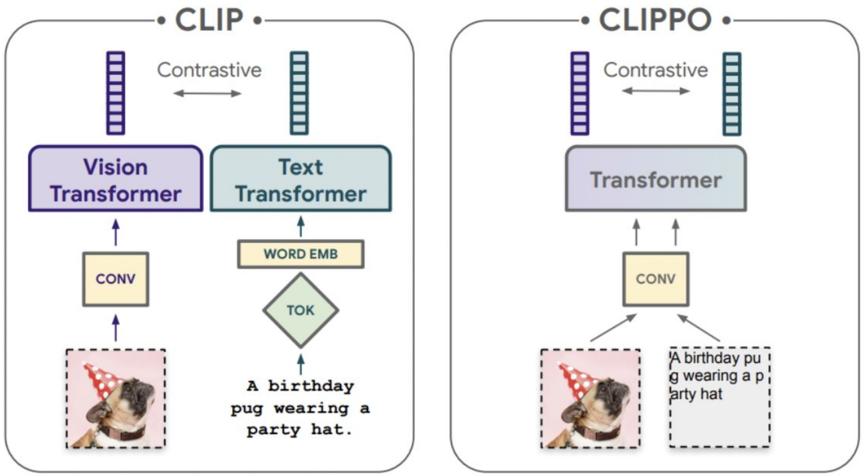

Reference Set
top-K
query
...just off of scotland the shetland sheepdog was developed as...
The Scottish Sheepdog is an excellent companion dog.
...
Scottish Sheepdog eyes are almond-shaped, neither too large nor ....
Image Encoder
image embedding
Retrieval Augmented Module (RAM)
augmented image embedding
Brittany dog
Shetland Sheepdog
bighorn sheep
airplane wing
A photo of a {object}.
Text Encoder
category embeddings
A photo of a Shetland Sheepdog.

Figure 2. Overview of the proposed RA-CLIP. Given an input image, RA-CLIP retrieves K most similar image-text pairs from the reference set, which provide informative descriptions for the input image, thus we feed them into the Retrieval Augmented Module (RAM) to enrich the representation of input image.

CLIPPO: Image-and-Language Understanding from Pixels Only

2. CLIP 발전시키는 방향 (Image



(5) Unifying vision & language encoder

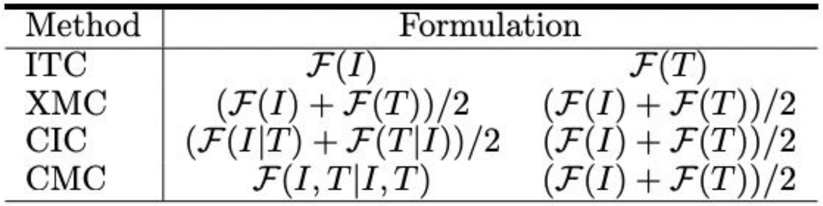
- CLIPPO (CVPR 2023)
  - CLIP-Pixels Only

    (text & image 모두 “픽셀" 취급)
  - CLIP의 한계점: Independent text & image towers
  - 목적: data modality간의 efficient communication을 위해!
  - 해결책: By rendering the alt-text as image!

2. CLIP 발전시키는 방향 (Image-Text CL)

(5) Unifying vision & language encoder

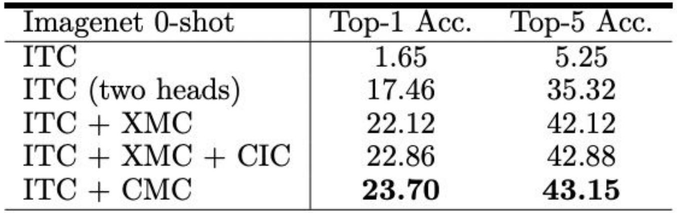
- OneR (AAAI 2023)
  - One-tower Model
  - 기본 가정: Image & Text는 같은 본질이 다른 식으로 발현된 것일 뿐
  - 네 가지 Task
    - Image-text Contrastive (ITC)
    - Cross-modal Mixup Contrastive (XMC)
    - Contextual Invariance Contrastive (CIC)
    - Contextual Mixup Contrast (CMC)



Method	Formulation	
ITC	$\mathcal{F}(I)$	$\mathcal{F}(T)$
XMC	$(\mathcal{F}(I)+\mathcal{F}(T))/2$	$(\mathcal{F}(I)+\mathcal{F}(T))/2$
CIC	$(\mathcal{F}(I\|T)+\mathcal{F}(T\|I))/2$	$(\mathcal{F}(I)+\mathcal{F}(T))/2$
CMC	$\mathcal{F}(I,T\|I,T)$	$(\mathcal{F}(I)+\mathcal{F}(T))/2$

Table 2: Summary of the contrastive objectives.

Unifying Vision-Language Represe



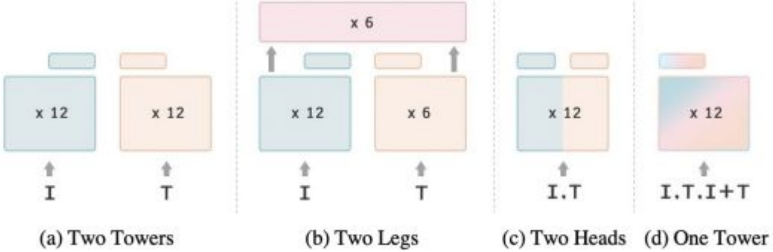
Imagenet 0-shot	Top-1 Acc.	Top-5 Acc.
ITC	1.65	5.25
ITC (two heads)	17.46	35.32
ITC + XMC	22.12	42.12
ITC + XMC + CIC	22.86	42.88
ITC + CMC	**23.70**	**43.15**

r

2. CLIP 발전시키는 방향 (Image-Tex

(5) Unifying vision & language encoder

- OneR (AAAI 2023)



(a) Two Towers (b) Two Legs (c) Two Heads (d) One Tower

Figure 2: Typical architectures of vision-language models. (a) is the basic form, with one transformer encoder and a projector for each modality. (b) adds fusion encoder blocks on top. (c) uses a single transformer encoder, but has separate projections. (d) unifies the two modalities with a generic one-tower model (OneR).

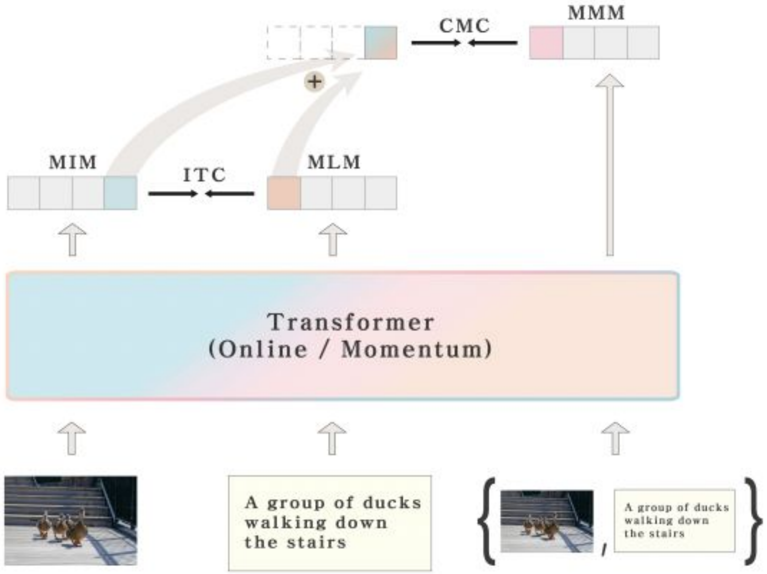
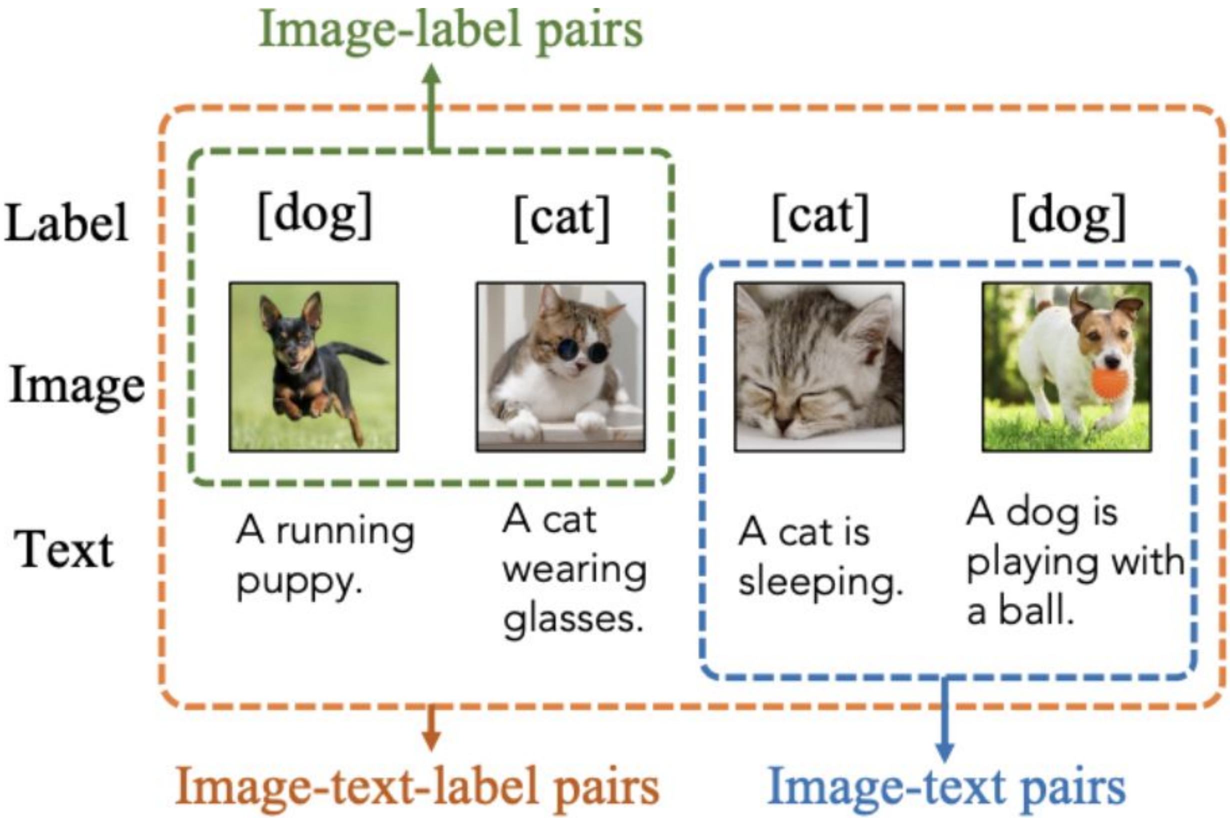

CMC
MMM
MIM
ITC
MLM
Transformer
(Online / Momentum)
A group of ducks walking down the stairs
A group of ducks walking down the stairs

Figure 5: Overview of OneR. Image-text contrastive and contextual mixup contrastive objective provide guidance in parallel with masked modeling for three input types: image, language and multi-modal (image+text).

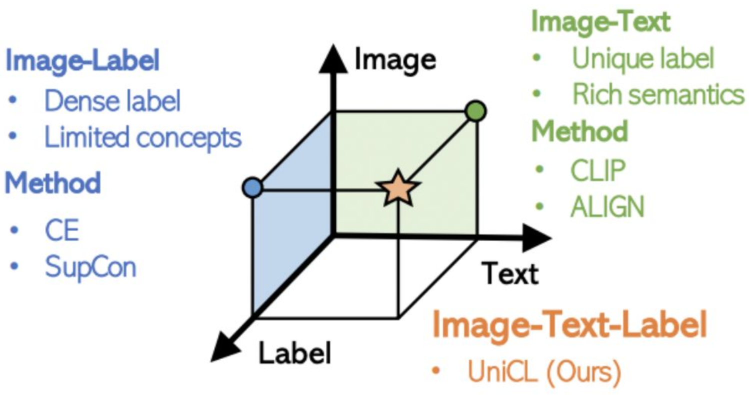
3. CLIP 발전시키는 방향 (Image-Text-Label CL)

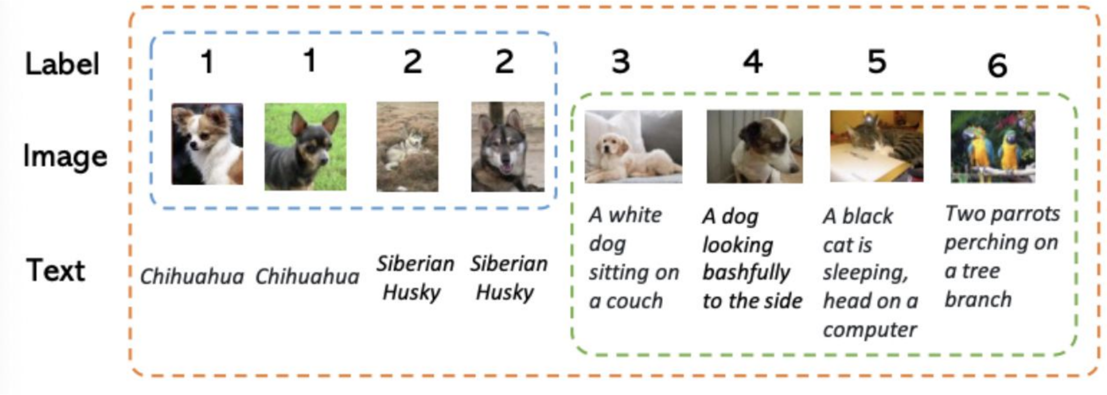


3. CLIP 발전시키는 방향 (Image-Text-Label CL)

(1) UniCL (CVPR 2022)

- UniCL = Unified CL




Label
Image
Text
1
1
2
2
3
4
5
6
Chihuahua
Chihuahua
Siberian Husky
Siberian Husky
A white dog sitting on a couch
A dog looking bashfully to the side
A black cat is sleeping, head on a computer
Two parrots perching on a tree branch

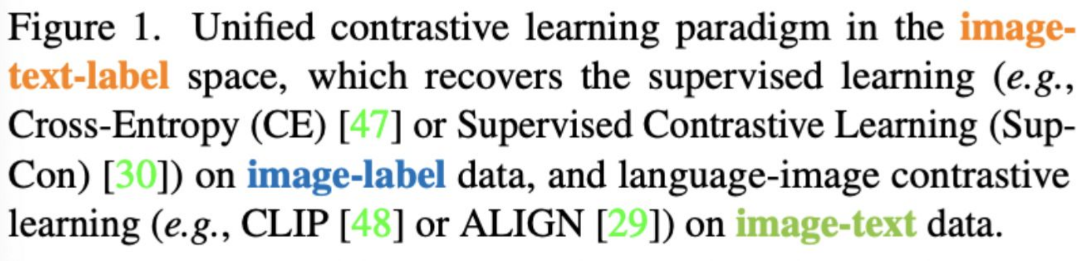


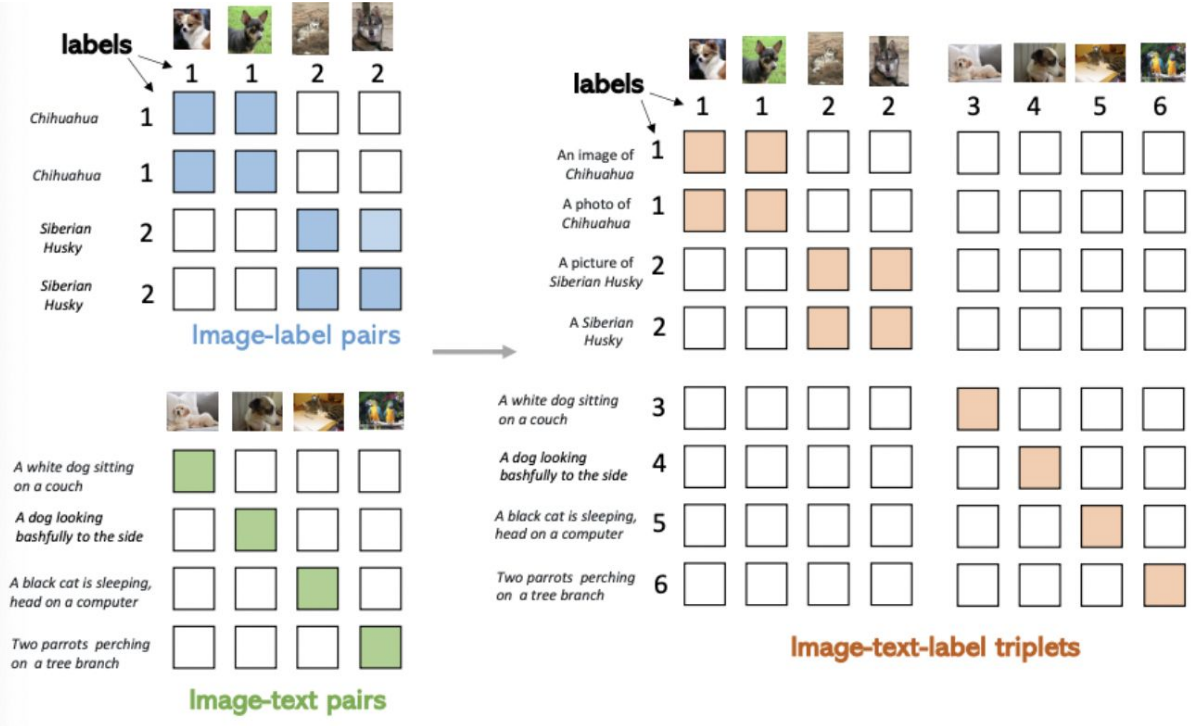
Figure 1. Unified contrastive learning paradigm in the **image-text-label** space, which recovers the supervised learning (*e.g.*, Cross-Entropy (CE) [47] or Supervised Contrastive Learning (SupCon) [30]) on **image-label** data, and language-image contrastive learning (*e.g.*, CLIP [48] or ALIGN [29]) on **image-text** data.

Unified Contrastive Learning in Image-Text-Label Space

3. CLIP 발전시키는 방향 (Image-Text-Label CL)

(1) UniCL (CVPR 2022)

- UniCL = Unified CL


labels
1 1 2 2
Chihuahua 1
Chihuahua 1
Siberian Husky 2
Siberian Husky 2
Image-label pairs
A white dog sitting on a couch
A dog looking bashfully to the side
A black cat is sleeping, head on a computer
Two parrots perching on a tree branch
Image-text pairs
labels
1 1 2 2 3 4 5 6
An image of Chihuahua 1
A photo of Chihuahua 1
A picture of Siberian Husky 2
A Siberian Husky 2
A white dog sitting on a couch 3
A dog looking bashfully to the side 4
A black cat is sleeping, head on a computer 5
Two parrots perching on a tree branch 6
Image-text-label triplets