Introduction of Visual-Language Pretraining

Paper Reading

Jianglin Lu

https://jianglin954.github.io/
jianglinlu@outlook.com

Outline

1. Preliminaries

2. Image-Language Pretraining

3. Video-Language Pretraining (Advanced)

Outline

1. Preliminaries

2. Image-Language Pretraining

3. Video-Language Pretraining (Advanced)

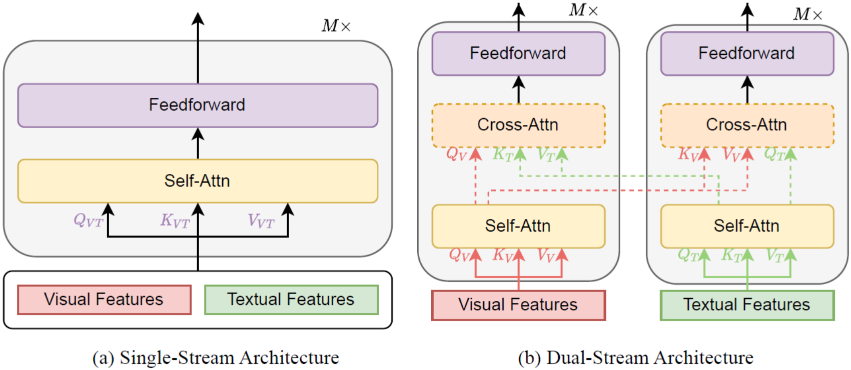
Preliminaries

Substantial works have shown pre-training models are beneficial for downstream uni-modal tasks and avoid training a new model from scratch. So *can such pre-trained models be applied to multi-modal tasks*?

We focus on mainstream *vision-language pre-training (VLP)*, including image-text and video-text pre-training, which mainly learns the semantic correspondence between different modalities by pre-training on large-scale data.

Chen et al. VLP: A Survey on Vision-Language Pre-training. arXiv, 2022

Preliminaries



(a) Single-Stream Architecture

(b) Dual-Stream Architecture

Fig. 1 Illustration of two types of model architectures for VLP.

Chen et al. VLP: A Survey on Vision-Language Pre-training. arXiv, 2022

Preliminaries

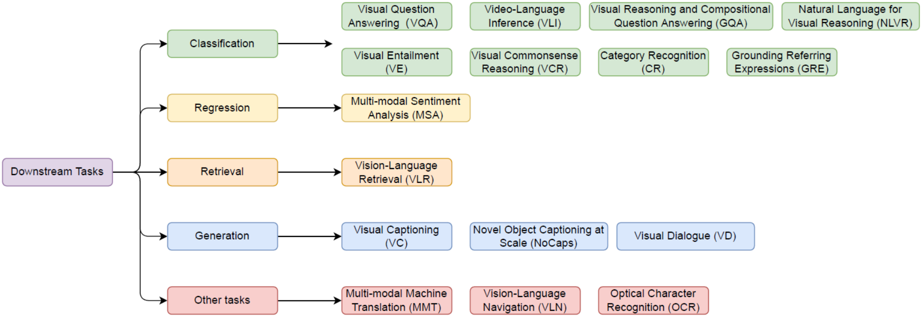

Downstream Tasks
Classification
Visual Question Answering (VQA)
Video-Language Inference (VLI)
Visual Reasoning and Compositional Question Answering (GQA)
Natural Language for Visual Reasoning (NLVR)
Visual Entailment (VE)
Visual Commonsense Reasoning (VCR)
Category Recognition (CR)
Grounding Referring Expressions (GRE)
Regression
Multi-modal Sentiment Analysis (MSA)
Retrieval
Vision-Language Retrieval (VLR)
Generation
Visual Captioning (VC)
Novel Object Captioning at Scale (NoCaps)
Visual Dialogue (VD)
Other tasks
Multi-modal Machine Translation (MMT)
Vision-Language Navigation (VLN)
Optical Character Recognition (OCR)

Fig. 2 Illustration of downstream tasks in VLP.

Chen et al. VLP: A Survey on Vision-Language Pre-training. arXiv, 2022

Outline

1 Preliminaries

2 Image-Language Pretraining

3 Video-Language Pretraining (Advanced)

Image-Language Pretraining

The State-of-the-art Methods

- Contrastive Language-Image Pre-training (**CLIP**), ICML 2021
- A Large-scale ImaGe and Noisy-text embedding (**ALIGN**), ICML 2021
- Align before Fuse (**ALBEF**), NIPS, 2021
- Fine-grained Interactive LIP (**FILIP**), arXiv, 2021
- Data Efficient CLIP (**DeCLIP**), ICLR 2022
- Triple Contrastive Learning (**TCL**), CVPR 2022
- SIngle-stream Multi-Level Alignment (**SIMLA**), ECCV 2022
- Hierarchical Feature Alignment (**PyramidCLIP**), arXiv, 2022

Contrastive Language-Image Pre-training (**CLIP**)

Motivation:

- This restricted form of supervision limits the generality and usability of computer vision systems 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.*
- 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.

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML, 2021

Contrastive Language-Image Pre-training (**CLIP**)

Approaches:

△ We explored training a system to solve the potentially easier proxy task of *predicting only which text as a whole is paired with which image and not the exact words of that text*.

△ CLIP *learns a multi-modal embedding space* by jointly training an image encoder and text encoder to maximize the cosine similarity of the image and text embeddings of the N real pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings.

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML, 2021

Contrastive Language-Image Pre-training (**CLIP**)

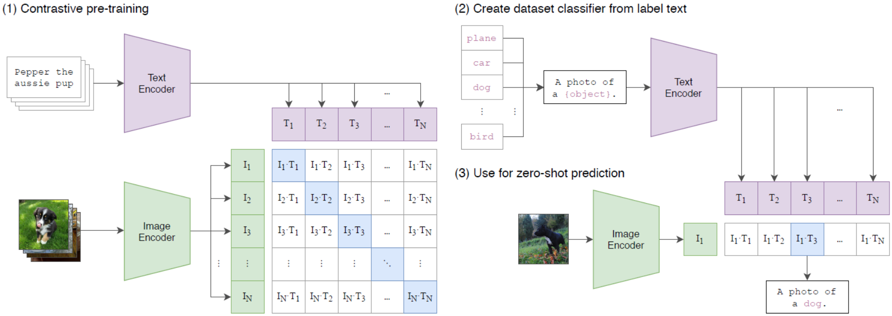
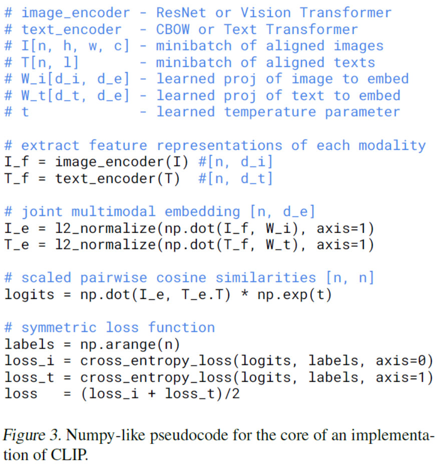

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
T1 T2 T3 ... TN
Image Encoder
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML, 2021

Contrastive Language-Image Pre-training (**CLIP**)



```
# 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]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML, 2021

A Large-scale ImaGe and Noisy-text embedding (**ALIGN**)

Motivation:

- For vision-language, popular datasets like Conceptual Captions, MSCOCO or CLIP all *involve a non-trivial data collection (and cleaning) process*. This costly curation process limits the size of datasets and hence hinders the scaling of trained models.

Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML, 2021

A Large-scale ImaGe and Noisy-text embedding (**ALIGN**)

Approaches:

△ We leverage *a noisy dataset of over one billion image alt-text pairs*, obtained without expensive filtering or post-processing steps in the Conceptual Captions dataset.

△ We use an objective that aligns the visual and language representations in a shared latent embedding space *using a simple dual-encoder architecture*.

△ ALIGN follows the natural distribution of image-text pairs from the raw alt-text data, while CLIP collects the dataset by first constructing an allowlist of high-frequency visual concepts from English Wikipedia. *Strong visual and vision-language representations can be learned with a dataset that does not require expert knowledge to curate*.

Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML, 2021

A Large-scale ImaGe and Noisy-text embedding (**ALIGN**)

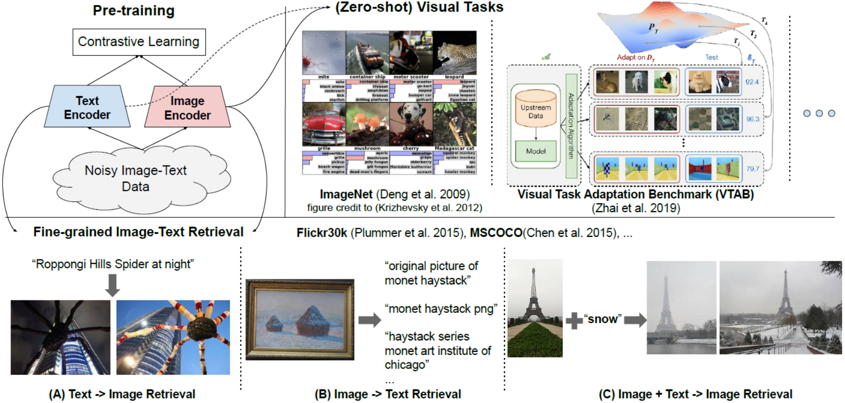

Pre-training
Contrastive Learning
Text Encoder
Image Encoder
Noisy Image-Text Data
(Zero-shot) Visual Tasks
mite
container ship
motor scooter
leopard
grille
mushroom
cherry
Madagascar cat
ImageNet (Deng et al. 2009)
figure credit to (Krizhevsky et al. 2012)
Upstream Data
Model
Adaptation Algorithm
Adapt 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"
(A) Text -> Image Retrieval
"original picture of monet haystack"
"monet haystack png"
"haystack series monet art institute of chicago"
...
(B) Image -> Text Retrieval
"snow"
(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.

Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML, 2021

A Large-scale ImaGe and Noisy-text embedding (**ALIGN**)

We minimize the sum of two losses: one for image-to-text classification:

$$L_{i2t} = -\frac{1}{N}\sum_{i}^{N} \log \frac{\exp\left(x_i^\top y_i/\sigma\right)}{\sum_{j=1}^{N} \exp\left(x_i^\top y_j/\sigma\right)}$$

and the other for text-to-image classification:

$$L_{t2i} = -\frac{1}{N}\sum_{i}^{N} \log \frac{\exp\left(y_i^\top x_i/\sigma\right)}{\sum_{j=1}^{N} \exp\left(y_i^\top x_j/\sigma\right)}$$

Here, x_i and y_j are the normalized embedding of image in the i-th pair and that of text in the j-th pair, respectively. N is the batch size, and σ is the temperature to scale the logits.

Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML, 2021

Align before Fuse (**ALBEF**)

Motivation:

- The image features and the word token embeddings reside in their own spaces, which makes it *challenging for the multimodal encoder to learn to model their interactions*.
- The *object detector is both annotation-expensive and compute-expensive*, because it requires bounding box annotations during pre-training, and high resolution (e.g. 600 × 1000) images during inference.
- The widely used image-text datasets are collected from the web and are inherently noisy, and existing pre-training objectives such as MLM may *overfit to the noisy text and degrade the model's generalization performance*.

Li et al. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. NIPS, 2021

Align before Fuse (**ALBEF**)

Approaches:

△ We introduce a contrastive loss to *align the image and text representations before fusing them through cross-modal attention*, which enables more grounded vision and language representation learning.

△ Our method *does not require bounding box annotations nor high-resolution images*. To improve learning from noisy web data, we propose *momentum distillation, a self-training method which learns from pseudo-targets produced by a momentum model*.

Li et al. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. NIPS, 2021

Align before Fuse (**ALBEF**)

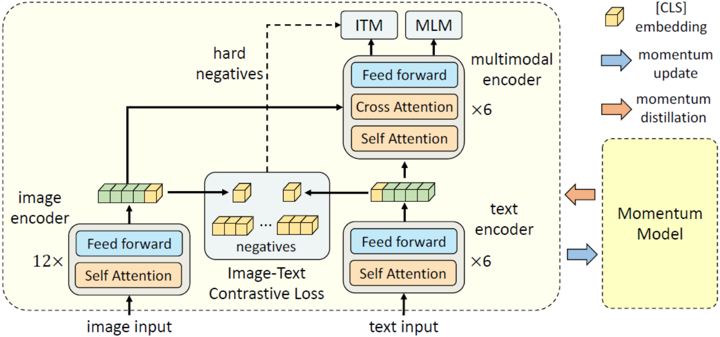


Figure 1: **Illustration of ALBEF.** It consists of an image encoder, a text encoder, and a multimodal encoder. We propose an image-text contrastive loss to align the unimodal representations of an image-text pair before fusion. An image-text matching loss (using in-batch hard negatives mined through contrastive similarity) and a masked-language-modeling loss are applied to learn multimodal interactions between image and text. In order to improve learning with noisy data, we generate pseudo-targets using the momentum model (a moving-average version of the base model) as additional supervision during training.

Li et al. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. NIPS, 2021

Align before Fuse (**ALBEF**)

△ Image-Text Contrastive Learning:

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

and the image-text contrastive loss is defined as the cross-entropy H between $\boldsymbol{p}$ and $\boldsymbol{y}$:

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

△ Masked Language Modeling (MLM):

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

△ Image-Text Matching (ITM):

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

Li et al. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. NIPS, 2021

FILIP: Fine-grained Interactive LIP (**FILIP**)

Motivation:

- Methods that use a pre-trained object detector to extract region-of-interest (ROI) features from images *complicates the pre-training due to pre-computing and storing a large number of ROI features*. In addition, the zero-shot ability of these approaches is usually *limited by the predefined number of classes* and their performance is also *restricted by the quality of the detector*.
- Cross-attention requires to be performed in an encoder-decoder structure, while *the complexity of the self-attention grows quadratically with the length of the prolonged concatenated sequences of both modalities*. During inference, the data from both modalities are intertwined to compute the cross-attention or self-attention, and can not be pre-computed offline as dual-stream models like CLIP and ALIGN.

Yao et al. Fine-grained Interactive Language-Image Pre-training. arXiv, 2021

FILIP: Fine-grained Interactive LIP (**FILIP**)

Approaches:

△ We model the fine-grained semantic alignment through *a novel cross-modal late interaction mechanism in the contrastive loss, instead of using cross or self-attention*. Specifically, our fine-grained contrastive learning uses a *token-wise maximum similarity* between visual and textual tokens to guide the contrastive objective.

△ We discard the padded tokens and *use average instead summation of token-wise maximum similarities* when computing the image-text alignment, which enhances the cross-modal representation learning and stabilizes training.

Yao et al. Fine-grained Interactive Language-Image Pre-training. arXiv, 2021

FILIP: Fine-grained Interactive LIP (**FILIP**)

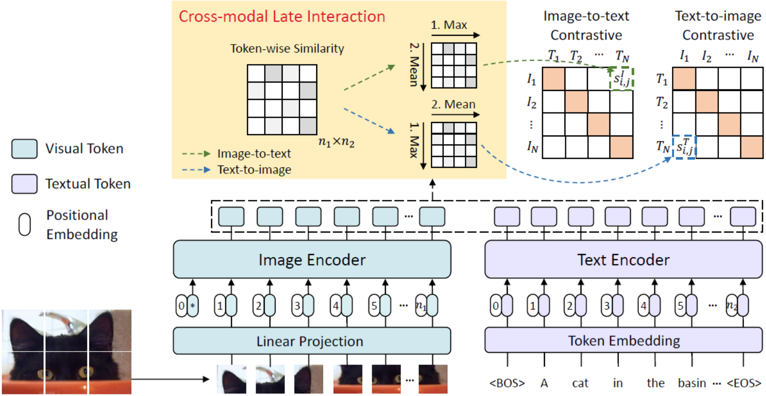


Figure 1: Overall architecture of FILIP, a dual-stream model with Transformer-based image and text encoders. On top of the image and text encoders, the representations of textual tokens and visual tokens are linearly projected to the multi-modal joint space. A novel fine-grained contrastive learning equipped with cross-modal late interaction is proposed, which uses a token-wise maximum similarity between visual and textual tokens.

Yao et al. Fine-grained Interactive Language-Image Pre-training. arXiv, 2021

Data Efficient CLIP (**DeCLIP**)

Motivation:

- CLIP is quite *data-hungry* and requires 400M image-text pairs for pre-training, thereby restricting its adoption.
- The prior arts *only use the single image-text contrastive supervision* while overlooking the widespread supervision within the pairs, thus is inefficient.

Li et al. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. ICLR, 2022

Data Efficient CLIP (**DeCLIP**)

Approaches:

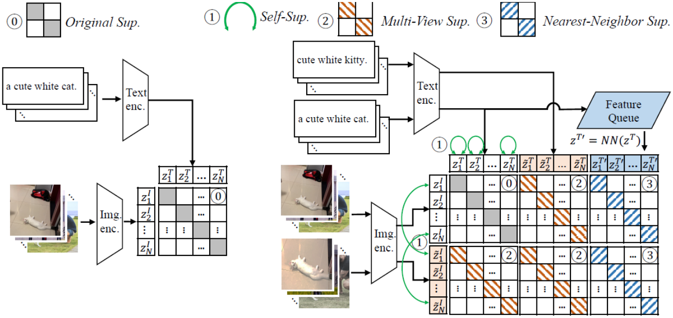
△ Self-supervision (SS) within each modality: For image SS, we adopt the simple yet effective SimSiam. The objective is to maximize the similarity between two augmented image features. For text SS, we adopt the most widely used Masked Language Modeling (MLM) as the pre-text task.

△ Multi-view supervision across modalities: we apply stochastic data augmentations for both images and texts, resulting in two correlated views of each example. Then, the image-text contrastive loss is calculated for all the 2×2 pairs.

△ Nearest-neighbor supervision from other similar pairs: we maintain a first-in-first-out feature queue that is representative of the whole data distribution. We use the nearest-neighbor search in embedding space to get the semantically similar text descriptions.

Li et al. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. ICLR, 2022

Data Efficient CLIP (**DeCLIP**)



0 Original Sup.
1 Self-Sup.
2 Multi-View Sup.
3 Nearest-Neighbor Sup.
a cute white cat.
Text enc.
Img. enc.
z_1^T z_2^T ... z_N^T
z_1^I z_2^I ⋮ z_N^I
cute white kitty.
a cute white cat.
Text enc.
Feature Queue
$z^{T'} = NN(z^T)$
Img. enc.
$\tilde{z}_1^T$ $\tilde{z}_2^T$ ... $\tilde{z}_N^T$
$z_1^{T'}$ $z_2^{T'}$ ... $z_N^{T'}$
$\tilde{z}_1^I$ $\tilde{z}_2^I$ ⋮ $\tilde{z}_N^I$

(a) CLIP & ALIGN

(b) DeCLIP

Figure 4: (a) CLIP and ALIGN jointly train an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. (b) Our DeCLIP overview. (1) means Self-Supervision(SS). For image SS, we maximize the similarity between two augmented views of the same instance. For text SS, we leverage Masked Language Modeling(MLM) within a text sentence. (2) represents cross-modal Multi-View Supervision(MVS). We first have two augmented views of both image and text, then contrast the 2 × 2 image-text pairs. (3) indicates Nearest-Neighbor Supervision(NNS). We sample text NN in the embedding space to serve as additional supervision. The combination of the three supervision leads to efficient multi-modal learning.

Li et al. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. ICLR, 2022

Triple Contrastive Learning (**TCL**)

Motivation:

- Cross-modal alignment (CMA) methods fail to ensure that *similar inputs from the same modality stay close by*. Simply performing cross-modal alignment (CMA) cannot fully guarantee the expressiveness of the learned features that is essential for joint multi-modal representation learning.
- However, global MI maximization *fails to consider localized and structural information in the input*. One potential side-effect is that it encourages the encoder to mainly extract information from certain unrelated/noisy image patches or text tokens that dominate MI.

Yang et al. Vision-Language Pre-Training with Triple Contrastive Lear. CVPR, 2022

Triple Contrastive Learning (**TCL**)

Approaches:

△ Cross-Modal Alignment (CMA): pulls the embeddings of matched image-text pairs together while pushing those of non-matched pairs apart by maximizing global mutual information between matched image and text.

△ Intra-Modal Contrastive (IMC): maximizes agreement between differently augmented views of the same data example through maximizing their global global mutual information.

△ Local MI Maximization (LMI): encourages high mutual information between the global representation and every local region (e.g., image patches and text tokens) of the input, which is designed to remedy the side-effects that are introduced by the global mutual information maximization.

Yang et al. Vision-Language Pre-Training with Triple Contrastive Lear. CVPR, 2022

Triple Contrastive Learning (**TCL**)

Approaches:

△ Image-Text Matching (ITM): predicts whether they are matched (positive examples) or not (negative examples). Assume each image-text pair (I, T) sampled from the pre-training datasets is a positive example (with label 1) and construct negative examples (with label 0) through batch-sampling.

△ Masked Language Modeling (MLM): aims to predict the ground truth labels of masked text tokens T^{msk}. Different from BERT, our MLM is conditioned on both surrounding text tokens of T^{msk} and image representations.

Yang et al. Vision-Language Pre-Training with Triple Contrastive Lear. CVPR, 2022

Triple Contrastive Learning (**TCL**)

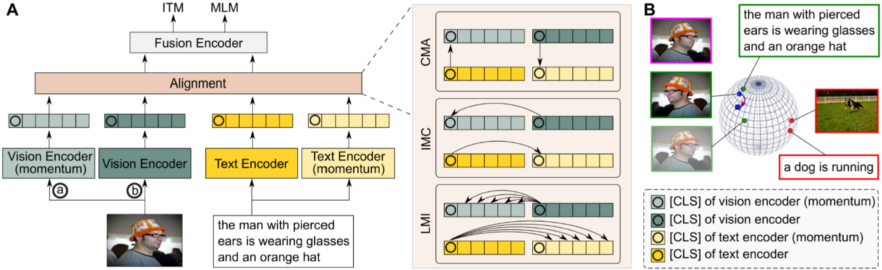

A
ITM
MLM
Fusion Encoder
Alignment
Vision Encoder (momentum)
Vision Encoder
Text Encoder
Text Encoder (momentum)
ⓐ
ⓑ
the man with pierced ears is wearing glasses and an orange hat
CMA
IMC
LMI
B
the man with pierced ears is wearing glasses and an orange hat
a dog is running
[CLS] of vision encoder (momentum)
[CLS] of vision encoder
[CLS] of text encoder (momentum)
[CLS] of text encoder

Figure 1. (A): An overview of our framework which consists of a vision encoder, a text encoder, and a fusion encoder. Each encoder has a paired momentum encoder updated by the momentum-based moving average. For the image input, we apply two separate data augmentation operators (a and b) which are sampled from the same family of augmentations. The alignment module contains three contrastive objectives (i.e., CMA, IMC, and LMI) for both cross-modal and intra-modal representation learning (make it easier for the fusion encoder to learn joint multi-modal embeddings). (B): The motivation of leveraging both cross-modal and intra-modal supervision. The original image (pink) is augmented to two different views (green). For CMA only, the middle image only has a positive text example (green) and treats other texts (red) as negatives. Its embedding (blue cirble) would be close to its positive text example. By incorporating IMC, it has two positive examples (one text and one image) and two sets of negative examples (one from text and one from image) and tends to learn more reasonable embeddings (blue square).

Yang et al. Vision-Language Pre-Training with Triple Contrastive Lear. CVPR, 2022

SIngle-stream Multi-Level Alignment (**SIMLA**)

Motivation:

- The unsupervised contrastive learning paradigm is *extremely data hungry*. Web-collected image captions are often incomplete and do not contain all relevant semantic concepts that the image captures.
- The dual-stream approach focuses primarily on a global alignment between modalities. This architecture design forces the pre-training objective to *only operate on the global image and text representation, making a tighter alignment difficult*.
- The contrastive learning objective *does not explicitly align visual and language concepts*.

Khan et al. Single-Stream Multi-Level Alignment for Vision-Language Pretraining. ECCV, 2022

SIngle-stream Multi-Level Alignment (**SIMLA**)

Approaches:

△ We construct a modular transformer-only architecture for vision-language learning that is *single-stream*, consisting of a single stack of transformer encoder layers. In this way we can *align modalities at multiple levels*, in contrast to typical dual stream architectures.

△ We introduce *a symmetric cross-modality reconstruction task* that trains the model to learn fine-grained alignment between image patches and language tokens.

△ We introduce another training signal by constructing a concept prediction task that *extracts pseudo labels for each image without supervision* and trains the visual encoder to *detect concepts that are missing from the caption but present in the image*.

Khan et al. Single-Stream Multi-Level Alignment for Vision-Language Pretraining. ECCV, 2022

SIngle-stream Multi-Level Alignment (**SIMLA**)

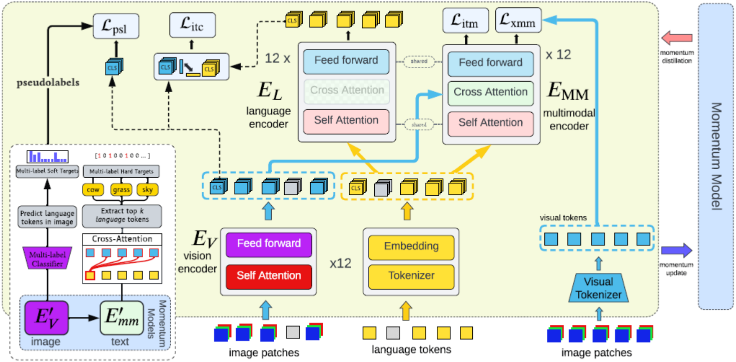


Fig. 1. SIMLA architecture. A language encoder E_l is stacked atop a vision encoder E_v. We add cross attention to E_l, allowing us to reuse it as a multimodal encoder E_{mm} by consuming image embeddings from E_v. Four tasks align images and language at multiple levels, exploiting a momentum model for additional supervision. A D-VAE tokenizes image patches for the cross-modality reconstruction task.

Khan et al. Single-Stream Multi-Level Alignment for Vision-Language Pretraining. ECCV, 2022

Hierarchical Feature Alignment (**PyramidCLIP**)

Motivation:

- Semantic Mismatch:
  - Caption Redundancy: the affiliated text description is redundant and contains irrelevant information.
  - Image Redundancy: the Regionof- Interest (ROI) corresponding to the text is only a sub-region of the image.
  - Cast Deficiency: text misses the descriptions of main objects in the image, while visual modelling needs to reason about the relationship among salient instances.
- Mutual Compatibility:
  - Captions might be compatible to some extent among pairs.

Gao et al. PyramidCLIP: Hierarchical Feature Alignment for Vision-language Model Pretraining. arXiv, 2022

Hierarchical Feature Alignment (**PyramidCLIP**)

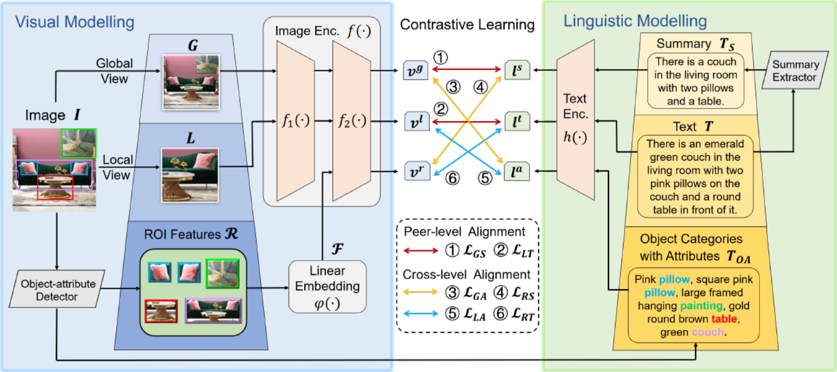
Approaches:

△ We introduce PyramidCLIP, which *constructs an input pyramid with different semantic levels for each modality*, and aligns visual elements and linguistic elements in the form of hierarchy via *peer-level semantics alignment and cross-level relation alignment*.

△ We soften the loss of negative samples (unpaired samples) so as to weaken the strict constraint during the pre-training stage, thus *mitigating the risk of forcing the model to distinguish compatible negative pairs*.

Gao et al. PyramidCLIP: Hierarchical Feature Alignment for Vision-language Model Pretraining. arXiv, 2022

Hierarchical Feature Alignment (**PyramidCLIP**)



There is a couch in the living room with two pillows and a table.

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.

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.

Gao et al. PyramidCLIP: Hierarchical Feature Alignment for Vision-language Model Pretraining. arXiv, 2022

Outline

1. Preliminaries

2. Image-Language Pretraining

3. Video-Language Pretraining (Advanced)

Video-Language Pretraining (Advanced)

The State-of-the-art Methods

- A Unified Video and Language pre-training (**UniVL**), arXiv 2020
- Hierarchical EncodeR for Omnirepresentation (**HERO**), arXiv, 2020
- Video-Language Model Pre-training (**VLM**), arXiv 2021
- Object-aware Transformer (**OA-Trans**), CVPR, 2022
- Survey: Transformer based video-language pre-training, AI Open 2022
- All-in-one Transformer (**All-in-one**), CVPR 2023
- HiTeA Video-Language Pre-training (**HiTeA**), ICCV 2023

Thanks!

Jianglin Lu

https://jianglin954.github.io/

jianglinlu@outlook.com