M.S. THESIS

Exploring Unified Vision-Language Representation Space with One-tower CLIP

단일 타워 CLIP을 이용해 통합된 시각 언어 표현 공간 탐색

BY

JIHO JANG

February 2023

Intelligence and Information
Department of Intelligence and Information
Graduate School of Convergence Science and Technology
SEOUL NATIONAL UNIVERSITY

M.S. THESIS

Exploring Unified Vision-Language Representation Space with One-tower CLIP

단일 타워 CLIP을 이용해 통합된 시각 언어 표현 공간 탐색

BY

JIHO JANG

February 2023

Intelligence and Information
Department of Intelligence and Information
Graduate School of Convergence Science and Technology
SEOUL NATIONAL UNIVERSITY

Exploring Unified Vision-Language Representation Space with One-tower CLIP

단일 타워 CLIP을 이용해 통합된 시각 언어 표현 공간 탐색

지도교수 곽 노 준

이 논문을 공학석사 학위논문으로 제출함

2023년 2월

서울대학교 대학원

융합과학부 지능형융합시스템전공

장 지 호

장지호의 공학석사 학위 논문을 인준함

2023년 2월

위 원 장:	박 재 흥	(인)
부위원장:	곽 노 준	(인)
위 원:	전 동 석	(인)

Abstract

Contrastive learning is widely adopted in self-supervised representation learning (SSL) to learn common attributes from similar sample pairs. In this paper, we boldly hypothesize that an image and its caption can be simply regarded as two different views of an underlying semantic, and aim to build a unified vision-language representation space by inducing a one-tower transformer that can encode both type of data samples in a modality-agnostic manner. We show that applying typical SSL frameworks to vision-language pretraining (VLP) naively fails to train a generic one-tower model due to a severe modality gap, and propose One Representation (OneR) to mitigate the disparity. We explore emerging properties of OneR distinguished from prior works with modality-specific representation spaces such as zero-shot object localization, text-guided visual reasoning, and multi-modal retrieval, and analyze our novel multi-modal representation learning. Comprehensive evaluations demonstrate the potential of a modality-agnostic VLP framework that has unified representation space.

keywords: Vision-language Pretraining, Multi-modal, Self-supervised Learning, Representration Learning, Transformer, Deep Learning

student number: 2021-22933

Contents

Abstract **i**

Contents **ii**

List of Tables **iv**

List of Figures **vi**

1 Introduction **1**

2 Related Works **5**

2.1 Self-Supervised Learning 5

2.2 Vision-Language Pretraining 6

2.3 Unified Vision-Language Framework 7

3 Overcoming Modality Gap **9**

3.1 Cross-Modal Mixup 10

4 Modality-agnostic Representations **12**

4.1 Contextual Modality Invariance 12

4.2 Contextual Mixup Contrast 13

4.3 Theoretical Explanation of CMC 14

4.4 One Representation 16

5 Experiment **18**

5.1 Experimental Setup 18

5.1.1 Datasets 18

5.1.2 Implementation Details 18

5.2 Qualitative Results 20

5.2.1 Zero-shot Localization 20

5.2.2 Text-guided Visual Reasoning 22

5.2.3 Multi-modal Retrieval 24

5.3 Visual Reasoning Analysis 24

5.3.1 Robustness 24

5.3.2 Multi-level vision-language connection 25

5.4 Quantitative Results 26

5.4.1 Image-text Retrieval 26

5.4.2 Cross-modal Knowledge Transfer 27

5.5 Ablation Study 28

5.5.1 Proposed Loss Ablation 28

5.5.2 Masked Modeling Ablation 29

6 Discussion **30**

7 Conclusion **32**

Abstract (In Korean) **40**

List of Tables

3.1 Experiments about proposed losses and summary of each objectives. Note that all models are one tower except for the second row in (a). Adding XMC enables one tower contrastive learning, and enforcing modality-blind token attentions further improves the performance. Masked modeling is included in all experiments. 11

5.1 Evaluation with bootstrapped language guidance. We can feed predicted class labels in simple concatenation to the input image to further improve accuracy. Note that this is not possible with two-tower or two-leg models, as the former does not accept mixture inputs and the latter forms a separate feature space after fusion, forbidding the similarity operation. 23

5.2 Quantitative evaluations on COCO image and text retrieval. Two-legs models generally perform better as they have modality-specific encoders and more parameters. *Pre.* means that vision-language models initialize their weights from a pretrained model. *Single tower* architecture refers to the fact that it uses a same transformer for both modalities but does not have a unified representation like one tower. † indicates the use of an additional object detection module. 26

5.3 Cross-modal knowledge transfer. Under a unified representation space, additional training in one modality benefits performance in the other modality with bigger margins. TR and IR is for text and image retrieval, respectively. 27

5.4 Method ablation. Our proposed components consistently improve performance, with the final CMC outperforming the two-tower baseline that uses more parameters and intra-modal contrastive loss (SLIP) with large margin in retrieval task, especially. 28

5.5 Ablations on masked modeling objectives. 29

List of Figures

1.1 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 transformer blocks on top of (a). (c) uses a single transformer encoder with modality specific projectors. (d) unifies the two modalities with a generic one-tower model (OneR). 3

2.1 Illustration of SimCLR. Random augmentation makes two different views of an image, and then contrastive learning trains the encoder and projector to extract the augmentation invariant information from the views. The output of the encoder is used for downstream tasks. 6

2.2 Typical architectures of vision-language models. (a) is the basic form, with one transformer encoder and a projector for each modality. (b) adds transformer blocks on top of (a). (c) uses a single-tower transformer, but has separate projections. (d) unifies the two modalities with a generic one-tower model (OneR). 7

3.1 T-SNE [40] representation visualization. Single-tower model trained with naive image-text contrastive objective fails to blend two distant modalities (*left*). Note that image features (blue dots) almost perfectly overlap with concatenation features (green dots), possibly due to sequence length bias *(best viewed zoom-in)*. Cross-modal mixup maps embeddings from two disjoint modalities to a common middle ground, and the corresponding image, text and image+text embeddings are well clustered after 40 epochs of training (*right*). 10

4.1 Graphical illustration of the proposed contrastive components. Blue dots represent the momentum features and red dots indicate the online network features. Note that these can be swapped in practice. 14

4.2 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). 17

5.1 A truly unified vision-language representation space displays intriguing properties. (*top*) Visualization of embedding similarities between image patches and the text prompt. (*bottom left*) Steering image classification with additional text input provided as simple token sequence concatenation. Here, we plot the attention map of `[CLS]`. (*bottom right*) This mixture input can also control image retrieval by combining the information from two modalities. 20

5.2 Patch embedding similarity map w.r.t. the text query. This clearly shows that two towers (*e.g.,* CLIP), two legs (*e.g.,* ALBEF) and two heads all learn modality-specific features spaces, forbidding similarity operations between embeddings. Projections are not applicable since they are only suited for the `[CLS]` token. . . . 21

5.3 Qualitative evaluation for object-level scene understanding. We simply compute token similarities for OneR, and Grad-CAM is used for CLIP and ALBEF. It is visually apparent that OneR correctly associates low-level visual signals to its corresponding language symbol, resulting in segmentation-map-like patch similarity maps. 22

5.4 Additional zero-shot localization results. 23

5.5 As OneR learns to associate low-level visual signals to the language, it shows robust visual reasoning even with a relatively small pretraining dataset. Above, OneR robustly recognizes *bicycle* from different visual clues (*e.g., handles, wheels or the body*). 24

5.6 (*left*) Patch embedding similarity (OneR) and Grad-Cam (ALBEF). (*right*) Patch embedding similarity map w.r.t. definitions of zebra and giraffe. 25

5.7 Additional zero-shot localization results by the definition. We compute cosine similarity between image patches and the text sentence (definition). 26

Chapter 1

Introduction

Recent boom of large scale pretraining has been triggered by self-supervised learning (SSL) as it provides means to leverage a huge stack of unlabeled data handily aggregated from the web. While language modeling [8, 36] is a pretext task of prevalence in the domain of natural language, contrastive learning is one of the most popular SSL framework in the computer vision field that essentially aims to maximize mutual information between related data pairs. When training with images, this is realized by first generating several distinct views of an image through random augmentation [6] and encouraging the model to enhance features' similarities between them.

Meanwhile, after the seminal work of CLIP [35] has declared the opening of Vision-Language Pretraining (VLP) era, many works [21, 31, 22, 44, 45, 46, 48] have exploited the contrastive objective for connecting representations between images and their descriptions. However, they are fundamentally different from the aforementioned SSL contrastive frameworks in that they make two separate representation spaces for vision and language, respectively. The representations for contrastive learning is computed only after sufficient abstraction operations,

typically done with modality-specific transformers and learnable projections. This renders them short for modality-agnostic representation learning, a promising research direction towards a generic perception model.

Our ultimate goal for training the model to have a unified representation space in modality-agnostic manner should require capabilities of both 1) mapping visual and textual information into a unified representation space at the global sequence level and 2) processing information within an input sequence in a modality-blind manner with generic token level attentions. First we hypothesize that an image (*e.g. a photo of panda*) and its caption (*e.g. the phrase "a photo of panda"*) contains common semantics, which can be regarded as two different views of implicit underlying information, analogous to the randomly augmented images in vision SSL frameworks. Hence, the contrastive SSL approach can be applied in vision language pretraining to congregate relevant semantics, either from visual signals, linguistic descriptions, or their mixture, into a *single unified representation space*. This way, our model learns to extract the modality-agnostic information by associating visual signals with structured symbols from the lowest level, breaking the boundaries between the two.

As shown in Fig. 2.2, while the conventional VLP works acknowledge the nature of each modality and inject relevant inductive bias into the model architecture, our approach adopt a generic single-tower transformer without modality-specific component to handle two different modalities at once and induce a single representation space. We empirically demonstrate that the modality gap is the main obstacle for training one-tower model in a naive image-text contrastive learning framework and propose *cross-modal mixup* as a simple yet effective remedy. Furthermore, we feed our model concatenation of image and its caption for contrastive loss computation to enable the model to aggregate information

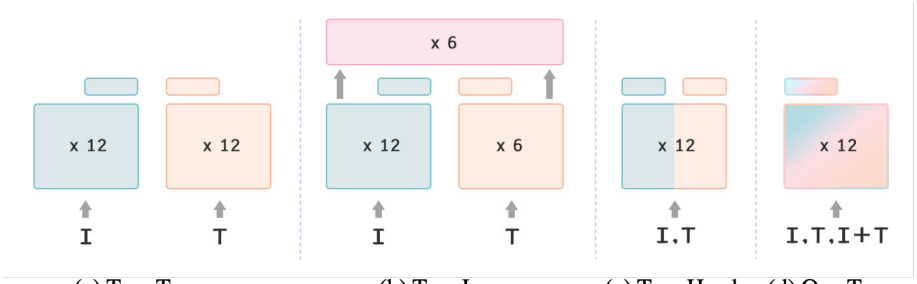


Figure 1.1: Typical architectures of vision-language models. (a) is the basic form, with one transformer encoder and a projector for each modality. (b) adds fusion encoder transformer blocks on top of (a). (c) uses a single transformer encoder with modality specific projectors. (d) unifies the two modalities with a generic one-tower model (OneR).

within each sequence in a modality-agnostic manner. This allows our model to form integrated representation space even from concatenated inputs of image and text, achieving both of our previous desiderata. We name our framework OneR, short for One Representation that suits both modalities.

Apart from the academic pursuit of general intelligence, single tower architecture has been shown to have benefits in scalability and cross-modal/cross-task transferability [42, 32]. Furthermore, we observe that our OneR's capacity to associate low-level visual signals to language symbols in unified representation space makes it an excellent zero-shot object localizer, and visual reasoning can be steered by auxiliary language guidance thanks to its natural ability to process *image+text* mixture inputs. The fact that mixture inputs are mapped to the same *One Representation* space further renders operations like multi-modal retrieval straightforward unlike two-leg baselines (*e.g.,* ALBEF [22]). We note that these properties do not rely on any modality-specific heads, segment tokens, nor spe-

cial cross-attention modules, but are natural outcomes of embedding similarity and input concatenation.

Our key contributions can be summarized as:

- We identify the failure of naive one-tower vision-language contrastive learning caused by the modality gap, and propose cross-modal mixup to mitigate it.
- We present OneR, a simple modality-agnostic representation learning framework that combines cross-modal mixup with contextual modality invariance to form a unified embedding space.
- We show extensive qualitative and quantitative results to demonstrate the advantages of our method, which includes distinguished capabilities in zero-shot object localization, text-guided visual reasoning and multimodal retrieval (See Fig. 5.1).

Chapter 2

Related Works

2.1 Self-Supervised Learning

Self-supervised learning first thrived in Natural Language Processing as masked language modeling (MLM) and autoregressive language modeling (LM) enabled pretraining large scale language models with huge stock of unlabeled text corpus, which is crawled from the Internet [8, 36, 20, 26].

In the vision domain, contrastive learning is representative framework of SSL. MoCo [15] and SimCLR [2] are the pioneers to demonstrate the strong performance of contrastive representation learning, which we adapt in our VLP frameworks. BYOL [12] and SimSiam [3] explored a new training algorithm that exploits positive samples only to mitigate the batch size dependency. Recently, various works [1, 4, 16] actively employ the vision transformer [9] to improve the performance and discover new properties. This architecture is modified for VLP as it can model data from different modalities in an elegant and integrated manner.

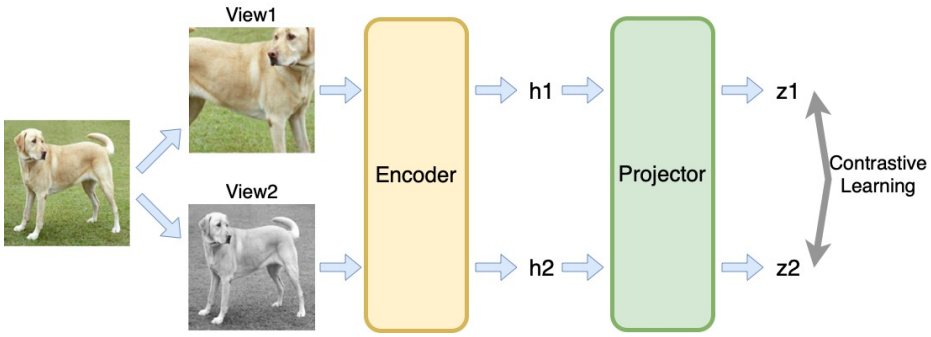


Figure 2.1: Illustration of SimCLR. Random augmentation makes two different views of an image, and then contrastive learning trains the encoder and projector to extract the augmentation invariant information from the views. The output of the encoder is used for downstream tasks.

2.2 Vision-Language Pretraining

Motivated by the success of contrastive based self-supervised learning, CLIP [35] first proposed contrastive Vision-Language Pretraining framework equipped with large scale paired image-text dataset. This novel framework can train a Vision-Language model very efficiently and achieve competitive performance compared to those of fully supervised baselines. ALIGN [17] scaled up the train dataset with noisy images and alt-text pair data. In another line of works [24, 23, 5, 10], off-the-shelf object detectors are leveraged to extract visual features first, and then used to train the multi-modal transformer. In an attempt to learn cross-modal interactions, ALBEF [22], TCL [44], FLAVA [39], and Florence [46] adopted transformer blocks with cross-attention as fusion layers on top of modality-specific transformer encoders. These models show impressive performance on various vision-language tasks, as they are capable of processing both single-modal and multi-modal inputs. Another group of works [21, 45,

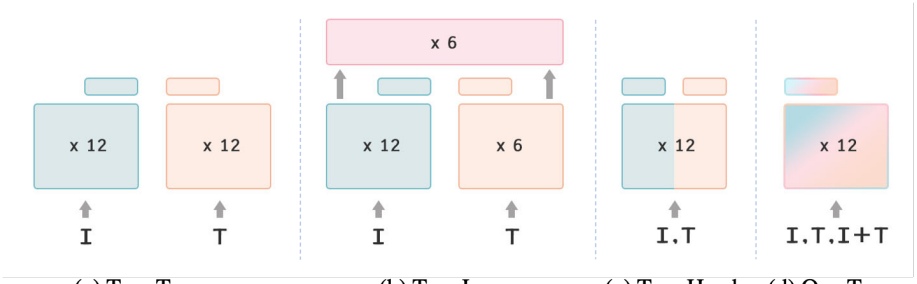


Figure 2.2: Typical architectures of vision-language models. (a) is the basic form, with one transformer encoder and a projector for each modality. (b) adds transformer blocks on top of (a). (c) uses a single-tower transformer, but has separate projections. (d) unifies the two modalities with a generic one-tower model (OneR).

42, 30] explored autoregressive generative model, typically in the form of image captioning, to further advances state-of-the-art performances on challenging tasks such as visual question answering.

2.3 Unified Vision-Language Framework

Some works have also aimed at integrating modality-specific transformers that can deal with diverse problem settings with minimal inductive bias alongside the efforts to push the state-of-the-art further. A single-tower transformer architecture was adopted by Uni-Perceiver [48] to handle a variety of V-L applications. Using a sequence-to-sequence framework, Unified-IO [29] unified input/output formats further using a VQ-VAE. Although these works have pointed towards a unified perception system, they use modality-specific components as well as a multi-task pretraining strategy, which pools data from different tasks to train

the network. The approach has the disadvantage of being less scalable than simpler contrastive frameworks, such as CLIP and ALIGN, which only rely on weakly linked image-text pairs. UFO [41] has shown that a single transformer model suffices for typical vision-language pretraining, but falls short towards a *unified vision-language representation space* as they attach two independent projectors to map the modalities together. As a concurrent work, LIMoE [32] explores single-tower (two heads) VLP with new inductive biases, i.e., mixture of experts, encoded into the architecture. OneR, in contrast, learns a common embedding space without any modality-specific components, which empowers the model with unique capabilities previously demonstrated.

Chapter 3

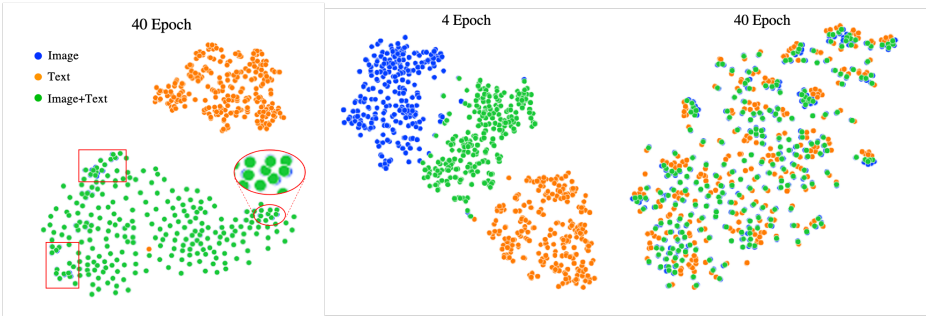
Overcoming Modality Gap

Vision-Language Pretraining frameworks that utilize a contrastive objective typically use batch-dependent InfoNCEs [33] that push negative pairs apart and pull positive pairs together. We state this image text contrastive (ITC) loss as

$$\mathcal{L}_{ITC} = ctr(\mathcal{F}(I), \mathcal{F}(T)), \tag{3.1}$$

where $ctr(A, B) = (NCE(A, B) + NCE(B, A))/2$ employs the generic InfoNCE formulation, $NCE(l, r)$, with the right term (r) being the EMA (exponential moving average) model output in our setting. $\mathcal{F}(X)$ refers to the final transformer hidden state, and I, T stands for image and text respectively.

This formulation works well in two-tower settings (Fig. 2.2a, 2.2b) with separate modality-specific encoders [35, 22], but we have observed training failure for a generic single-tower model (Fig. 2.2d, Tab. 3.1a) without any modality-specific components. In Fig. 3.1a, visualization of the representation space displays a severe modality gap, as visual signals and linguistic symbols are significantly dissimilar in their data structure. Hence, the model is not able to encode positive {image, text} pairs close together, being incapable of mixing these two



(a) Naive single-tower ITC. (b) OneR at the beginning and the end of the training.

Figure 3.1: T-SNE [40] representation visualization. Single-tower model trained with naive image-text contrastive objective fails to blend two distant modalities (*left*). Note that image features (blue dots) almost perfectly overlap with concatenation features (green dots), possibly due to sequence length bias *(best viewed zoom-in)*. Cross-modal mixup maps embeddings from two disjoint modalities to a common middle ground, and the corresponding image, text and image+text embeddings are well clustered after 40 epochs of training (*right*).

distant modalities in a unified representation space.

3.1 Cross-Modal Mixup

Mixup [47] was initially developed in the vision community as a data augmentation strategy to improve classification performance, robustness, and generalization by interpolating linearly from training data distributions. In a concurrent study [13], mixup has been incorporated into VLP in a similar manner, using mixup augmentation within each modality separately. Different from this, we boldly apply mixup across modality, not as a means to augment the training data but as a projection to map image and text embeddings to a common middle

ground. We find it to be an extremely simple yet effective starting point to evade the image-text modality gap, from which the traditional contrastive learning successfully guides the model for instance discrimination. The formal definition of our cross-modal mixup constrastive (XMC) loss can be stated as

$$\mathcal{L}_{XMC} = ctr(\frac{\mathcal{F}(I) + \mathcal{F}(T)}{2}, \frac{\mathcal{F}(I) + \mathcal{F}(T)}{2}), \tag{3.2}$$

where we use an online model and its momentum (EMA) counterpart for feature extraction in practice as we mentioned above[1]. Using this straightforward approach with the ITC loss to mitigate the modality gap works surprisingly well, blending representations from the two distant modalities into a single embedding space successfully and thereby stabilizing training. Full quantitative evaluations are presented in Tab. 5.4.

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**

(a) Zero-shot ImageNet [7] evaluations.

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	$\mathcal{F}(I, T\|I, T)$	$(\mathcal{F}(I) + \mathcal{F}(T))/2$

(b) Summary of the contrastive objectives.

Table 3.1: Experiments about proposed losses and summary of each objectives. Note that all models are one tower except for the second row in (a). Adding XMC enables one tower contrastive learning, and enforcing modality-blind token attentions further improves the performance. Masked modeling is included in all experiments.

[1]Note that ctr by definition in eqn. 3.2 uses two separate feature extractors (online and EMA) symmetrically.

Chapter 4

Modality-agnostic Representations

As we revealed in the previous section, the modality gap is the major obstacle to learning a unified vision-language representation space, and XMC loss is proposed to reconcile distant modalities. Stepping further, under the hypothesis that paired image and text contain similar information, a modality-agnostic representation should depend only on the content of the underlying information, not the modality (format; text or image) it is expressed in. In other words, the final embedding should be similar regardless of whether the context is an image or a text (*i.e., key and value in self-attention*) In order to enforce such behavior, we devise Contextual Invariance Contrastive (CIC) loss and incorporate it into our framework.

4.1 Contextual Modality Invariance

The high-level idea is to encourage the model to bring the representation from an image context to be close to that from the text context. To be specific, from a pair, either the image or the text is selected as the *query*. Then, at one side, we use image tokens for *key* and *value*, while on the other side, we use the text

tokens. The CIC penalizes the distance between the final representations from each side, guiding the model to extract similar information regardless of the modality of the context. The formal definition is

$$\mathcal{L}_{CIC} = ctr(\frac{\mathcal{F}(I|T) + \mathcal{F}(T|I)}{2}, \frac{\mathcal{F}(I|I) + \mathcal{F}(T|T)}{2}), \tag{4.1}$$

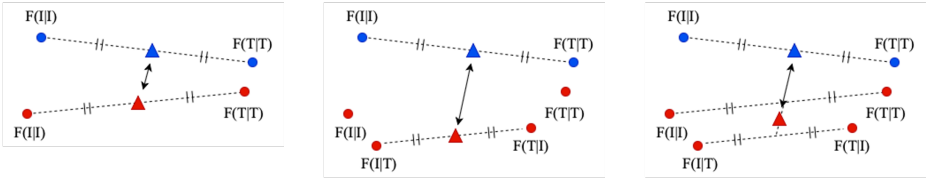
where $\mathcal{F}(X|Y)$ refers to the final embedding of X (query) given Y as the context (key and value). We note that $\mathcal{F}(X)$ in eqn. 3.1 and eqn. 3.2 is an abbreviated expression equivalent to $\mathcal{F}(X|X)$.

4.2 Contextual Mixup Contrast

As is evident from Tab. 3.1a, adding CIC loss improves overall performance by encouraging the model not only to embed paired image and text close together but also utilize information from image and text tokens in a similar fashion from the lowest level. To maximally leverage CIC's generic information aggregation capacity, we adapt our model for mixed-modality input scenario. Formally, we can incorporate the XMC and CIC into a single loss by replacing the left term in Eqn. 4.1 with a simple concatenation of {image, text} ($\mathcal{F}(I, T)$) and train the model to optimize Contextual Mixup Contrastive (CMC) objective instead.

$$\mathcal{L}_{CMC} = ctr(\mathcal{F}(I, T|I, T), \frac{\mathcal{F}(I|I) + \mathcal{F}(T|T)}{2}) \tag{4.2}$$

This is a generalized form which explicitly guides the model to embed mixed-modality inputs to the unified V-L representation space after adequate integration of information from two different modalities. We utilize this property for text-guided visual reasoning (Tab. 5.1) and multi-modal retrieval (Fig. 5.1). The high-level idea is that the self-attention feature of concatenated input can be roughly decomposed to self-attention feature of each plus the cross-attention features, and the theoretical verification is provided in the following section.



(a) XMC (b) CIC (c) CMC

Figure 4.1: Graphical illustration of the proposed contrastive components. Blue dots represent the momentum features and red dots indicate the online network features. Note that these can be swapped in practice.

4.3 Theoretical Explanation of CMC

In this paper, we propose XMC and CIC loss as

$$\mathcal{L}_{XMC} = ctr(\frac{\mathcal{F}(I|I) + \mathcal{F}(T|T)}{2}, \frac{\mathcal{F}(I|I) + \mathcal{F}(T|T)}{2}) \tag{4.3}$$

$$\mathcal{L}_{CIC} = ctr(\frac{\mathcal{F}(I|T) + \mathcal{F}(T|I)}{2}, \frac{\mathcal{F}(I|I) + \mathcal{F}(T|T)}{2}). \tag{4.4}$$

We combine these two components to obtain the concise formulation of CMC as illustrated below.

$$\frac{\mathcal{L}_{XMC} + \mathcal{L}_{CIC}}{2} \simeq \mathcal{A}(\mathcal{L}_{XMC}, \mathcal{L}_{CIC}) \tag{4.5}$$

$$= ctr(\frac{\mathcal{F}(I|I) + \mathcal{F}(I|T) + \mathcal{F}(T|T) + \mathcal{F}(T|I)}{4}, \frac{\mathcal{F}(I|I) + \mathcal{F}(T|T)}{2}) \tag{4.6}$$

where $\mathcal{A}(\cdot, \cdot)$ is a kind of average operation (this will be different from arithmetic mean, harmonic mean or geometric mean) which will be approximately equal to the arithmetic mean.

Then, we define the attention module $f(X|Y)$ as

$$f(X|Y) = S(\frac{Q_X K_Y^T}{\sqrt{d_k}})V_Y, \tag{4.7}$$

where S is the softmax operation along each row.

$$f(X|X,Y) = S(\frac{Q_X cat(K_X, K_Y)^T}{\sqrt{d_k}})cat(V_X, V_Y) \tag{4.8}$$

$$= S(\frac{cat(Q_X K_X^T, Q_X K_Y^T)}{\sqrt{d_k}})cat(V_X, V_Y) \tag{4.9}$$

$$= cat(\lambda_X S(\frac{Q_X K_X^T}{\sqrt{d_k}}), (I - \lambda_X)S(\frac{Q_X K_Y^T}{\sqrt{d_k}}))cat(V_X, V_Y) \tag{4.10}$$

$$= \lambda_X S(\frac{Q_X K_X^T}{\sqrt{d_k}})V_X + (I - \lambda_X)S(\frac{Q_X K_Y^T}{\sqrt{d_k}})V_Y \tag{4.11}$$

$$= \lambda_X f(X|X) + (I - \lambda_X)f(X|Y) \tag{4.12}$$

Here, $cat(\cdot, \cdot)$ is the concatenate operation, I is the identity matrix, and λ_X is a diagonal matrix which can be defined as

$$\lambda_X^{(i)} = \frac{\sum_{j=1}^{l_X} exp(Q_X^{(i)} K_X^{(j)})}{\sum_{j=1}^{l_X} exp(Q_X^{(i)} K_X^{(j)}) + \sum_{j=1}^{l_Y} exp(Q_X^{(i)} K_Y^{(j)})} \tag{4.13}$$

$$\lambda_X = \begin{bmatrix} \lambda_X^{(1)} & & \\ & \ddots & \\ & & \lambda_X^{(l_X)} \end{bmatrix} \tag{4.14}$$

with l_X and l_Y being the sequence length of X and Y, respectively. So far, we have decomposed the softmax of concatenated input into weighted sum of two terms. If we consider the final transformer output $\mathcal{F}(X|Y)$ as the average pooling of the attention module $f(X|Y)$,

$$\mathcal{F}(X|Y) = \frac{1}{l_X} 1^T f(X|Y), \tag{4.15}$$

where 1 is the all-one vector, then we can further decompose the final self-attention output of the concatenated input as four different terms with corre-

sponding weights as follows

$$\mathcal{F}(X,Y|X,Y) = \frac{1}{l_X + l_Y} 1^T f(X,Y|X,Y) \tag{4.16}$$

$$= \alpha \frac{1}{l_X} 1^T f(X|X,Y) + (1-\alpha) \frac{1}{l_Y} 1^T f(Y|X,Y) \tag{4.17}$$

$$= \alpha\lambda_X \frac{1}{l_X} 1^T f(X|X) + \alpha(1-\lambda_X) \frac{1}{l_X} 1^T f(X|Y) \tag{4.18}$$

$$+ (1-\alpha)\lambda_Y \frac{1}{l_Y} 1^T f(Y|Y) + (1-\alpha)(1-\lambda_Y) \frac{1}{l_Y} 1^T f(Y|X) \tag{4.19}$$

$$= \beta_1 \mathcal{F}(X|X) + \beta_2 \mathcal{F}(X|Y) + \beta_3 \mathcal{F}(Y|Y) + \beta_4 \mathcal{F}(Y|X). \tag{4.20}$$

Here, α is the sequence length ratio $\frac{l_X}{l_X+l_Y}$, and $\beta_1+\beta_2+\beta_3+\beta_4 = 1$. Note that we substitute Eqn.(4.12) into Eq.(4.17) to obtain the result. In order to simplify the formulation, we assume $\beta_1 = \beta_2 = \beta_3 = \beta_4$ in practice, which allows substituting Eqn.(4.20) to Eq.(4.6) to obtain the final equivalence.

$$\mathcal{A}(\mathcal{L}_{XMC}, \mathcal{L}_{CIC}) \simeq ctr(\mathcal{F}(I,T|I,T), \frac{\mathcal{F}(I|I) + \mathcal{F}(T|T)}{2}) = \mathcal{L}_{CMC}. \tag{4.21}$$

We note that this is not a rigorous theoretical proof for our CMC formulation. Rather, it is to show how we combine XMC with CIC to formulate CMC, a concise form that explicitly trains the model for mixed-modality input scenario.

4.4 One Representation

Fig. 4.2 illustrates the overall pipeline of OneR. Model input can be one of *image*, *text* or *image+text*, and CMC objective in Eqn. 4.2 is combined with the traditional image-text contrastive (ITC) loss. Masked modeling is also carried out

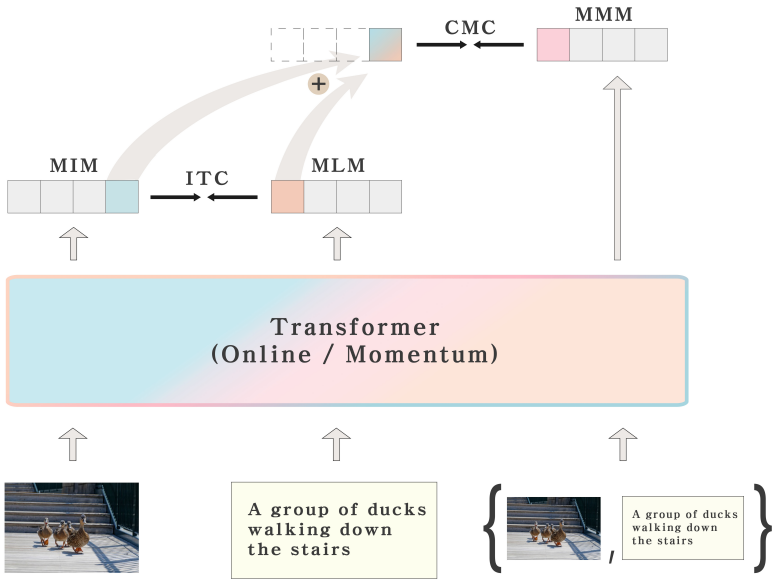


Figure 4.2: Overview of OneR. Image-text contrastive and contextual mixup contrastive objective provide guidance in parallel with masked modeling for three input types: image, language and multi-modal (image+text).

for all three input types (*i.e., image, text and mixture of them*). Our framework employs no modality-specific architectural component except for the initial token embedding layer (i.e. patch projector and word embedding layer), making our model generic and modality-agnostic with minimal inductive bias. Tab. 3.1b summarizes the overall formulations.

Chapter 5

Experiment

In this section, we will describe our experimental setting and show qualitative and quantitative results that demonstrate the advantages of OneR.

5.1 Experimental Setup

5.1.1 Datasets

Following prior works [22, 44, 10], we train OneR on the combination of CC3M [38], SBU Captions [34], Visual Genome [18] and COCO [25], which sums up to 4M images and 5.1M image-text pairs. All ablation models are trained on CC3M.

5.1.2 Implementation Details

We adopt the model architecture of Masked AutoEncoder [14] with BERT [8] word embeddings and language modeling head. Unlike most prior works on VLP, we initialize our entire model *from scratch*, as neither ViT nor BERT suits our goal towards a unified VL representation space. 1D and 2D sinusoidal positional encodings are added to text and image respectively, and a single `[CLS]`

token is prepended to all three input types. Special modality indicator tokens (*e.g.,* separate token `[SEP]` or segment token `[SEG]`) are further removed from typical one tower baselines in order to train a fully modality-agnostic representation learner.

Overall, we follow the MoCo-v3 settings [4], but with learnable ConvStem [43] as the image patch projector. A 3-layer MLP projector and a 2-layer predictor are used as in [2], and momentum ratio was fixed to 0.996 throughout the whole training process. We choose the base learning rate of 1e-4 with linear scaling rule [19, 11] that adapts the learning rate as lr$\times$BatchSize/256. The first 4 epochs are for warm up and cosine scheduling [27] decays the lr for a total of 40 epochs. After 40 epochs of training on 224$\times$224 resolution images, we further train with 384$\times$384 upsampled resolution for additional 5 epochs with positional embeddings interpolated correspondingly. Batch size is 4,096 for 224$\times$224 stage and 1,024 for 384$\times$384. We optimize with AdamW [28] under the weight decay of 0.1.

Unlike recent works [22, 44] that go through additional forward passes for masked modeling during training, OneR computes the contrastive loss and the masked modeling loss simultaneously in a single forward pass. Only the online network learns masked modeling, thus clean inputs are fed to the momentum model. MLM masking ratio is set to 0.15 as done in [22, 8], and MIM ratio is raised from 0.1 to 0.5. In our early experiments, when MIM is combined with contrastive learning, a high masking ratio seems to make the instance discriminate task too difficult, especially in the early stages of learning.

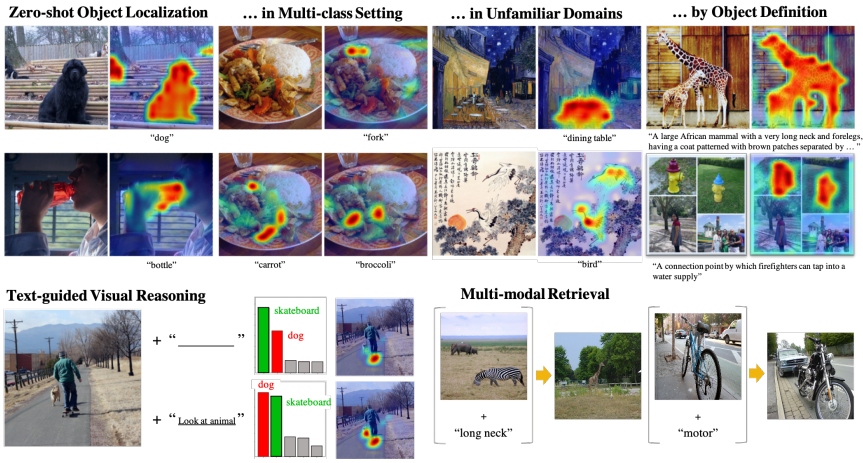


Figure 5.1: A truly unified vision-language representation space displays intriguing properties. (*top*) Visualization of embedding similarities between image patches and the text prompt. (*bottom left*) Steering image classification with additional text input provided as simple token sequence concatenation. Here, we plot the attention map of `[CLS]`. (*bottom right*) This mixture input can also control image retrieval by combining the information from two modalities.

5.2 Qualitative Results

5.2.1 Zero-shot Localization

Conventional vision-language transformers typically rely on cross-attention map or Grad-CAM [37] for visualization. However, the former attributes the global semantics to each local region, rendering it unsuitable for complex scene understanding such as multi-class localization (Fig. 5.1), while the latter requires a separately devised procedure that involves back propagation. One of the most distinguished qualities of OneR is its natural proficiency for object localization. Throughout the paper, we simply compute the *cosine similarities* between image patch embeddings and the average-pooled text embedding for visualization.

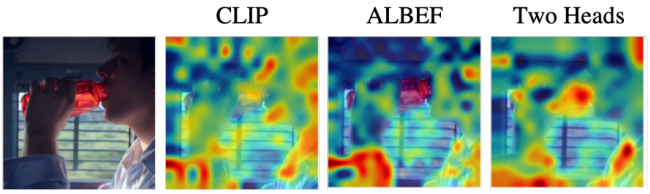


Figure 5.2: Patch embedding similarity map w.r.t. the text query. This clearly shows that two towers (*e.g.,* CLIP), two legs (*e.g.,* ALBEF) and two heads all learn modality-specific features spaces, forbidding similarity operations between embeddings. Projections are not applicable since they are only suited for the `[CLS]` token.

This is possible only because OneR maps both visual and textual information to a unified embedding space where their feature similarity correctly indicates the semantic relevance. Otherwise, the cosine similarity map between separately embedded tokens cannot convey meaningful information, as illustrated in Fig. 5.2.

We present qualitative comparison on zero-shot localization with two competitive baselines, CLIP and ALBEF, where Grad-CAM is used for their visualizations as it yields the best output. Looking at Fig. 5.3, we can see that Grad-CAM of ALBEF better captures the spatial details compared to CLIP, but OneR has the most fine-grained visual reasoning, resulting in almost segmentation-map-like patch similarity maps. This clearly shows that OneR has the capacity to relate low-level visual signals to their corresponding linguistic concepts in a unified vision-language representation space.

Additional examples of patch embedding similarity visualization are presented in Fig. 5.4.

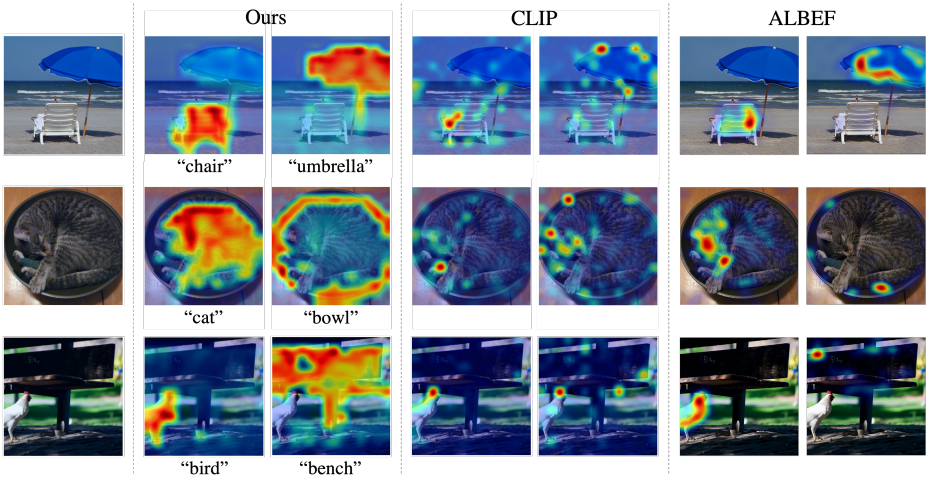


Figure 5.3: Qualitative evaluation for object-level scene understanding. We simply compute token similarities for OneR, and Grad-CAM is used for CLIP and ALBEF. It is visually apparent that OneR correctly associates low-level visual signals to its corresponding language symbol, resulting in segmentation-map-like patch similarity maps.

5.2.2 Text-guided Visual Reasoning

As illustrated in Fig. 5.1, OneR's ability to understand *image+text* mixture input opens up possibilities for diverse forms of multi-modal reasoning. For example, we can simply concatenate additional text to the image input sequence to guide its visual representation, which can be particularly useful in a complex scene understanding setting where an image contains more than one dominant semantic. In such cases, we can *tell* the model where to focus to suit our goals. We provide quantitative results to further demonstrate this property in Tab. 5.1, where we bootstrap with language guidance to improve zero-shot classification accuracy. Specifically, for each image, we retrieve top-10 class labels upon embedding similarity. After that, we concatenate each to the image sequence and

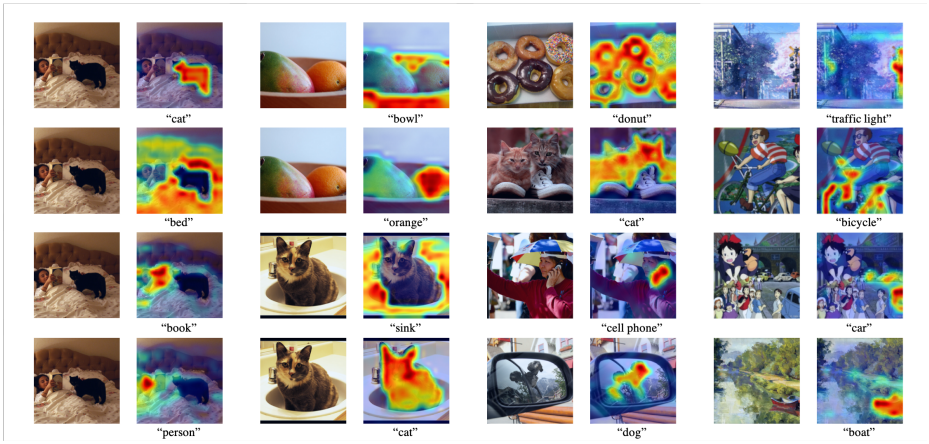


Figure 5.4: Additional zero-shot localization results.

Bootstrapped Language Guidance	ImageNet 0-shot		CIFAR100 0-shot	
	top-1	top-5	top-1	top-5
OneR (4M)	27.33	50.17	31.45	57.52
OneR-Bootstrapped (4M)	**28.00**	**50.69**	**32.23**	**58.24**

Table 5.1: Evaluation with bootstrapped language guidance. We can feed predicted class labels in simple concatenation to the input image to further improve accuracy. Note that this is not possible with two-tower or two-leg models, as the former does not accept mixture inputs and the latter forms a separate feature space after fusion, forbidding the similarity operation.

compute similarity once more, similar to sample re-ranking. The intuition is that when *image+text* input is given, the image patches that strongly attend to the text label are strengthened by the attention mechanism, resulting in clearer representations. We note that we do not provide any external guidance during this procedure, which makes these gains essentially *free*.

5.2.3 Multi-modal Retrieval

Unlike the existing two-tower model, OneR can also perform image-text to image and image-text to text retrieval because all three input types exist in the same representation space as in Fig. 5.1. While two-leg model combine image and text information using fusion layer, the concatenated representation is independent from image and text representation.

5.3 Visual Reasoning Analysis

We further analyze the visual reasoning mechanism of OneR to provide insights into the properties of unified vision-language representation space.

Figure 5.5: As OneR learns to associate low-level visual signals to the language, it shows robust visual reasoning even with a relatively small pretraining dataset. Above, OneR robustly recognizes *bicycle* from different visual clues (*e.g., handles, wheels or the body*).

5.3.1 Robustness

Fig. 5.5 shows an example of how OneR recognizes an object (bicycle, in this case) with different visual clues. OneR recognizes a bicycle even from partial

images of handles or wheels, which we believe is key to its robustness in visual understanding.

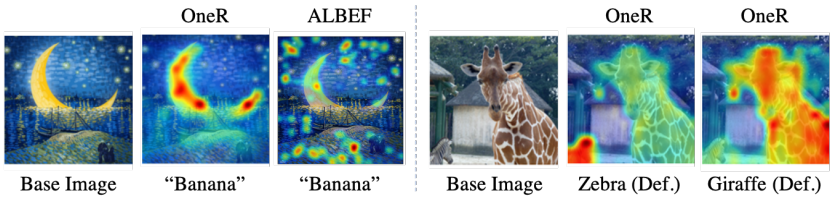


Figure 5.6: (*left*) Patch embedding similarity (OneR) and Grad-Cam (ALBEF). (*right*) Patch embedding similarity map w.r.t. definitions of zebra and giraffe.

5.3.2 Multi-level vision-language connection

Looking at Fig. 5.6, OneR recognizes the *moon* as being visually similar to *banana* in terms of embedding cosine similarity, while ALBEF condenses the global semantic in `[CLS]`, resulting in a randomly scattered Grad-CAM, which means ALBEF does not perceive banana in the image. Although this can be viewed as a failure case of OneR, it reveals how OneR perceives the visual signals. On the right, we can see that *zebra* and *giraffe* are visually similar, and their definitions contain similar phrases such as ‘an African mammal’, resulting in some overlaps in the two similarity maps. However, after abstracting the linguistic semantics, the model correctly identifies each, which shows its ability to process high-level semantics as well. Overall, OneR learns both low-level and high-level vision-language connections, making it a competent modality-agnostic representation learner. Additional examples of zero-shot localization with the definition are presented in Fig. 5.7.

Figure 5.7: Additional zero-shot localization results by the definition. We compute cosine similarity between image patches and the text sentence (definition).

Method	Architecture	Pre.	#Images	Zero-shot MS-COCO (5K)						Fine-tuned MS-COCO (5K)					
				Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
				R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ImageBert†	Single Tower	O	6M	44.0	71.2	80.4	32.3	59.0	70.2	**66.4**	**89.8**	**94.4**	**50.5**	**78.7**	**87.1**
ViLT	Single Tower	O	4M	56.5	82.6	89.6	40.4	70.0	81.1	61.5	86.3	92.7	42.7	72.9	83.1
Uni-Perceiver	Single Tower	X	44.3M	57.7	85.6	92.3	46.3	**75.0**	84.0	64.7	87.8	93.7	48.3	75.9	84.5
OneR	One Tower	X	4M	**62.9**	**86.3**	**92.5**	**47.0**	74.7	**84.1**	66.1	87.8	93.2	48.3	76.0	85.2
CLIP	Two Towers	X	400M	58.4	81.5	88.1	37.8	62.4	72.2	-	-	-	-	-	-
FLAVA	Two Legs	O	70M	42.7	76.8	-	38.4	67.5	-	-	-	-	-	-	-
ALBEF	Two Legs	O	4M	68.7	89.5	94.7	50.1	76.4	84.5	73.1	91.4	96.0	56.8	81.5	89.2
TCL	Two Legs	O	4M	71.4	90.8	95.4	53.5	79.0	87.1	75.6	92.8	96.7	59.0	83.2	89.9

Table 5.2: Quantitative evaluations on COCO image and text retrieval. Two-legs models generally perform better as they have modality-specific encoders and more parameters. *Pre.* means that vision-language models initialize their weights from a pretrained model. *Single tower* architecture refers to the fact that it uses a same transformer for both modalities but does not have a unified representation like one tower. † indicates the use of an additional object detection module.

5.4 Quantitative Results

5.4.1 Image-text Retrieval

Tab. 5.2 shows the quantitative comparison with state-of-the-art methods on the widely used image-text retrieval benchmark. Models with modality-specific en-

coders typically show better performance, as they have more parameters and architectural inductive bias. Among one-tower baselines, OneR shows the best zero-shot performance, sometimes with significant margins. We note that OneR achieves such a competent outcome without any initialization prior commonly used in the literature, such as pretrained BERT or ViT. This shows that vision and language modalities *can* be effectively encoded in a single representation space with minimal inductive bias, once the aforementioned obstacle (*i.e.,* innate modality gap) is overcome.

5.4.2 Cross-modal Knowledge Transfer

We hypothesize that under a unified vision-language representation space, additional training on one modality should benefit performance in the other modality. Tab. 5.3 validates our conjectures, as additional training with language data results in greater gains for the unified one-tower model. This could indicate better scalability of one-tower models, as there is much more single-modality data

Cross-modal Transfer	Architecture	0-shot INet	MS COCO	
		top-1	TR@1	IR@1
SBU	two heads	**7.28**	**8.88**	5.73
	one tower	6.49	8.60	**5.77**
SBU + CC3M (caption only)	two heads	**8.59**	10.41	6.87
	one tower	8.54	**11.31**	**7.20**
Gain	two heads	1.31	1.53	1.14
	one tower	**2.07**	**2.71**	**1.43**

Table 5.3: Cross-modal knowledge transfer. Under a unified representation space, additional training in one modality benefits performance in the other modality with bigger margins. TR and IR is for text and image retrieval, respectively.

available than image-text pairs in the web, which we leave for future works.

5.5 Ablation Study

5.5.1 Proposed Loss Ablation

In Tab. 5.4, we present ablation experiments for our framework. Naive ITC with single tower fails due to the modality gap, and adding modality-specific projectors can be the minimal architectural modification that works, but still lags behind our method. CMC combines XMC and CIC into a concise formulation, resulting in the best performance that surpasses the competent two-tower baselines.

Method	Imagenet	MS-COCO			
	Top-1 Acc.	TR@1	TR@5	IR@1	IR@5
CLIP	17.1	15.0	34.8	10.9	26.7
SLIP	23.0	21.7	45.1	15.6	35.2
ITC	1.6	0.8	2.5	0.7	2.2
ITC (two heads)	17.5	10.4	26.8	10.7	26.4
ITC + XMC	22.1	25.2	48.1	15.2	33.6
ITC + XMC + CIC	22.9	25.4	48.1	16.3	35.5
ITC + CMC (OneR)	**23.7**	**25.5**	**48.2**	**16.9**	**36.9**

Table 5.4: Method ablation. Our proposed components consistently improve performance, with the final CMC outperforming the two-tower baseline that uses more parameters and intra-modal contrastive loss (SLIP) with large margin in retrieval task, especially.

5.5.2 Masked Modeling Ablation

We also present masked modeling ablations to distinguish how masked modeling affects our framework in Tab. 5.5. We observe that training with masked image modeling (MIM) and masked language modeling (MLM) helps the performance but is not the most crucial component.

Method	Imagenet	MS-COCO					
	Top-1 Acc.	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10
OneR	23.7	25.5	48.2	60.2	16.9	36.9	47.9
OneR - MIM	23.4	24.7	46.9	58.7	16.9	36.8	47.7
OneR - MIM - MLM	22.9	23.3	47.2	58.1	13.6	33.3	45.6

Table 5.5: Ablations on masked modeling objectives.

Chapter 6

Discussion

Random Augmentation In our early experiments, while random augmentation is a critical component to make different views in vision SSL frameworks, we have observed that adding random augmentations does not help the overall model performance. We conjecture this could be due to the difficulty of the task: learning a single vision-language representation space. As we have discussed in this paper, our basic hypothesis is that a paired image and text can be seen as two different (but closely related) views of an underlying common semantic. Hence, additionally performing strong augmentations on one side could be unnecessary. Nevertheless, we believe that there is much room for more sophisticated designs to incorporate data augmentation into our framework, which we leave to future works.

Momentum Teacher As we remove strong image augmentations, the presence of a momentum teacher is critical to the performance of OneR. Although it is a relatively common belief that contrastive learning with a momentum teacher network improves the performance [22, 44], we observe that it matters more in a unified single-tower setting. We speculate that momentum teacher works

as a augmentation along weight space in XMC, which uses the exactly same representation as a positive pair if only using an online network. Exploring such behaviors could be another promising research direction.

Chapter 7

Conclusion

Modality-agnostic representation learning is a meaningful step toward a generic perceptual agent that understands the environment in a similar way as humans do. In this work, we explore the difficulties of unifying modalities into a single representation space and introduce OneR as a generic framework that shows unique qualities as a modality-agnostic representation learner.

Bibliography

[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[3] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[4] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[5] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Learning universal image-text representations. 2019.

[6] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 702–703, 2020.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020.

[11] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[12] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[13] X. Hao, Y. Zhu, S. Appalaraju, A. Zhang, W. Zhang, B. Li, and M. Li. Mixgen: A new multi-modal data augmentation. *arXiv preprint arXiv:2206.08358*, 2022.

[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[15] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[16] J. Jang, S. Kim, K. Yoo, J. Kim, and N. Kwak. Self-distilled self-supervised representation learning. *arXiv preprint arXiv:2111.12958*, 2021.

[17] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[18] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[19] A. Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

[20] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[21] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[22] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[23] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[24] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pre-training approach. *arXiv preprint arXiv:1907.11692*, 2019.

[27] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[28] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[29] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

[30] R. Mokady, A. Hertz, and A. H. Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[31] N. Mu, A. Kirillov, D. Wagner, and S. Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

[32] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *arXiv preprint arXiv:2206.02770*, 2022.

[33] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[34] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[36] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[38] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[39] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[40] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[41] J. Wang, X. Hu, Z. Gan, Z. Yang, X. Dai, Z. Liu, Y. Lu, and L. Wang. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021.

[42] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[43] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick. Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, 34:30392–30400, 2021.

[44] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[45] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[46] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[47] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[48] X. Zhu, J. Zhu, H. Li, X. Wu, H. Li, X. Wang, and J. Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815, 2022.

초 록

Contrastive learning은 자기지도학습(Self-supervised learning, SSL)에서 널리 채택되어 비슷한 데이터에서 공통된 특징을 추출하도록 하는 학습방법론이다. 본 논문에서, 우리는 이미지와 이에 대응되는 설명문을 공통된 정보를 바탕으로 다르게 표현된 데이터로 가정하고, 단일 타워의 트랜스포머를 활용하여 이미지와 텍스트를 하나의 표현 공간으로 매핑하려고 한다. 기존의 자기지도학습 방법론들을 단순히 시각 언어 사전학습에 적용하는 것은 표현 양식의 차이로 인한 어려움이 존재하고, 이를 해결하기 위해 One Representation (OneR) 을 제안한다. OneR은 시각과 언어 각각에 특정한 표현공간을 가지는 이전의 연구들과 달리 흥미로운 특성들이 나타나며, 이를 zero-shot 시각화, 자연어 기반의 시각적 이해 및 멀티모달 검색을 통해 보인다. 또한, 포괄적인 평가를 통해 통합된 표현 공간을 가지며, 표현 양식에 구애받지 않은 시각 언어 사전학습방법론의 잠재력을 보여주며 이에 대한 분석을 제공한다.

주요어: 시각 언어 사전학습, 멀티모달, 자기지도학습, 표현학습, 트랜스포머, 딥러닝

학번: 2021-22933