The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Unifying Vision-Language Representation Space with Single-Tower Transformer

Jiho Jang[1*†], Chaerin Kong[1*†], Donghyeon Jeon[2], Seonhoon Kim[3†], Nojun Kwak[1]

[1]Seoul National University
[2]NAVER
[3]Coupang
{geographic,veztylord,nojunk}@snu.ac.kr, donghyeon.jeon@navercorp.com, sekim625@coupang.com

Abstract

Contrastive learning is a form of distance learning that aims to learn invariant features from two related representations. In this work, we explore the hypothesis that an image and caption can be regarded as two different views of the underlying mutual information, and train a model to learn a unified vision-language representation space that encodes both modalities at once in a modality-agnostic manner. We first identify difficulties in learning a one-tower model for vision-language pretraining (VLP), and propose One Representation (OneR) as a simple yet effective framework for our goal. We discover intriguing properties that distinguish OneR from the previous works that have modality-specific representation spaces such as zero-shot localization, text-guided visual reasoning and multi-modal retrieval, and present analyses to provide insights into this new form of multi-modal representation learning. Thorough evaluations demonstrate the potential of a unified modality-agnostic VLP framework.

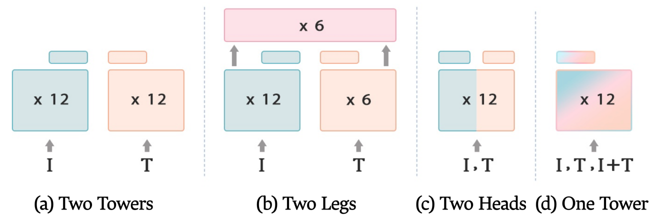


Figure 1: Typical architectures of vision-language models. (a) is the basic form, with one transformer encoder and a projector for each modality. (b) adds fusion encoder blocks on top. (c) uses a single transformer encoder, but has separate projections. (d) unifies the two modalities with a generic one-tower model (OneR).

Introduction

Self-supervised learning (SSL) is the core driving force behind recent boom in large scale training (Devlin et al. 2018; Radford et al. 2018) as it provides means to leverage a huge stack of unlabeled data handily accessible from the web. In the computer vision community, contrastive learning is one of the most popular SSL frameworks that essentially aims to maximize the mutual information between two related representations, *i.e.,* views. When training with images, this is realized by first generating different views from random augmentations and encouraging the model to learn the augmentation-invariant features.

Meanwhile, the seminal work of CLIP (Radford et al. 2021) has declared the opening of the Vision-Language Pretraining (VLP) era, where many works (Li et al. 2022; Mu et al. 2021; Li et al. 2021; Yang et al. 2022; Yu et al. 2022; Yuan et al. 2021; Zhu et al. 2022) have leveraged the contrastive objective for connecting images and their descriptions. However, they fundamentally differ from the aforementioned SSL contrastive framework in that they learn two separate representation spaces each for vision and language, respectively. The features from each modality are compared only after sufficient abstraction operations, typically done with self-attention layers in transformers and separate learnable projections. This renders them short for modality-agnostic representation learning, a promising research direction towards a generic perceptual agent.

A modality-agnostic representation learner should be capable of both 1) mapping visual and linguistic information into a unified representation space at the global sequence level and 2) mixing 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 linguistic description (*e.g. the phrase "a photo of panda"*) contains common information, which can be viewed as two different representations of implicit underlying information, analogous to the augmented views of an image. Hence, we apply contrastive SSL approach, MoCo-v3 (Chen, Xie, and He 2021), in VLP setting to congregate relevant semantics, either from visual signals, linguistic symbols, or their mixture, in a *single unified representation space*. This way, our model learns to associate visual signals with structured symbols from the lowest level, breaking the boundaries between the two.

As shown in Fig. 1, our approach is distinguished from the conventional counterparts that acknowledge the innate differences between the two modalities and encode rele-

* These authors contributed equally.
† Work done at NAVER.

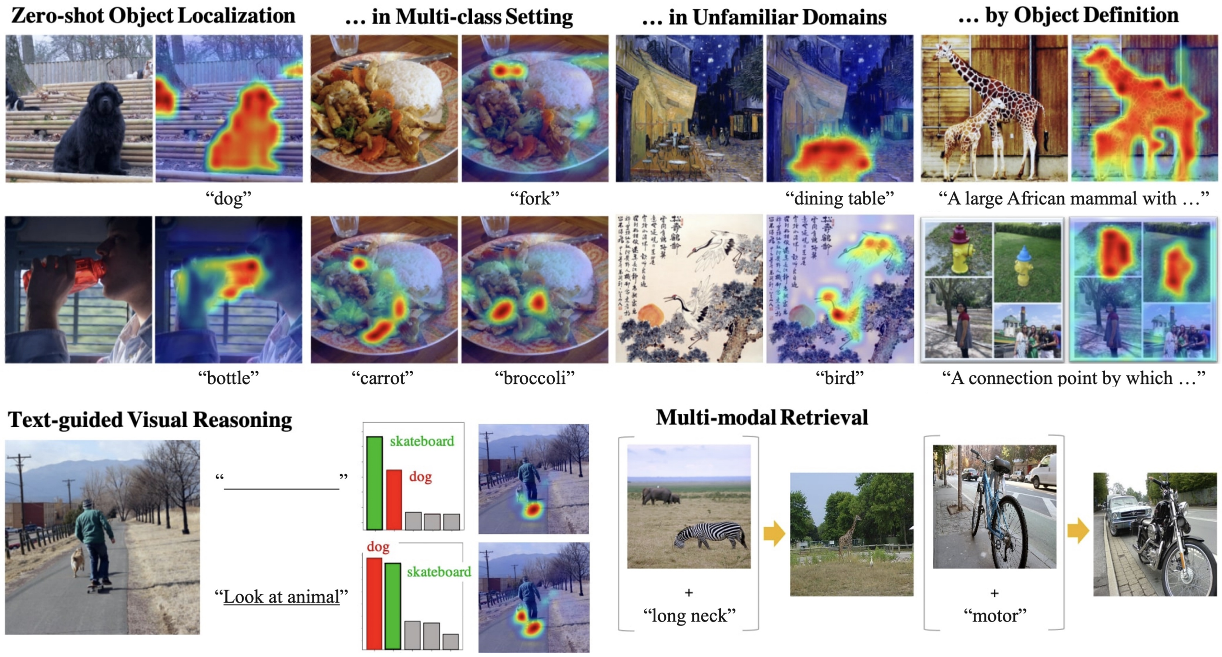


Figure 2: A truly unified vision-language representation space displays intriguing properties. (*top*) Visualization of embedding similarities between image patches and the text prompt. (*bottom left*) Steering image classification with additional text input provided as simple token sequence concatenation. 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.

vant inductive bias into the model architecture. We adopt a generic single-tower model, thus a single representation space, to handle two different modalities at once. We empirically demonstrate that the failure of naive single-tower image-text contrastive learning is due to the modality gap, and propose *cross-modal mixup* as a simple yet effective remedy. Furthermore, we train our model to learn to aggregate information within each sequence in a modality-agnostic manner by forwarding concatenation of image and text for contrastive loss computation. This allows our model to form integrated representations even from mixed inputs of image and text, achieving both of our previous desiderata. We name our framework OneR, short for One Representation that suits both modalities.

Aside from the academic pursuit of general intelligence, unifying multi-modal representation space with a single generic model has been shown to have benefits in scalability and cross-modal/cross-task transferability (Wang et al. 2021b; Mustafa et al. 2022). We further observe that our OneR's capacity to associate low-level visual signals to language symbols makes it an excellent zero-shot object localizer, and we can steer its visual reasoning with 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 (Li et al. 2021)). We note that these properties do not rely on any modality-specific heads, segment tokens, nor special cross-attention modules, but are natural outcomes of embedding

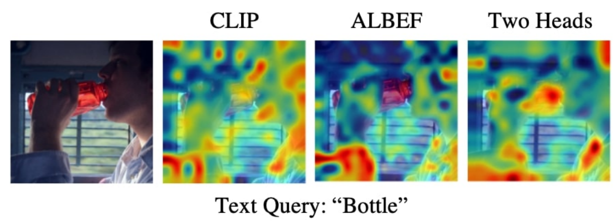
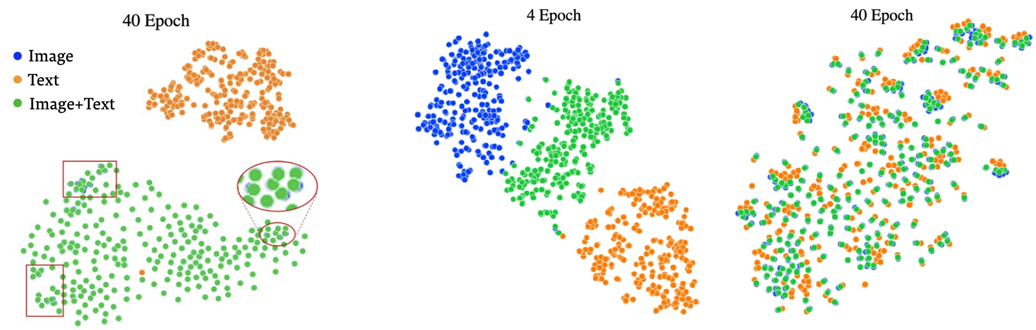


Figure 3: Patch embedding similarity map w.r.t. the text query. This clearly shows that two towers (*e.g.,* CLIP), two legs (*e.g.,* ALBEF) and two heads all learn modality-specific features spaces, forbidding similarity operations between embeddings. Projections are not applicable since they are only suited for the `[CLS]` token.

similarity and input concatenation.

Our key contributions can be summarized as:

- We analyze the collapse of naive single-tower vision-language contrastive learning, and propose cross-modal mixup to mitigate the modality gap.
- 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 conduct extensive qualitative and quantitative evaluations to demonstrate the advantage of our approach, which includes distinguished capabilities in zero-shot ob-



(a) Naive single-tower ITC.

(b) OneR at the beginning and the end of the training.

Figure 4: T-SNE (Van der Maaten and Hinton 2008) 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*).

ject localization, text-guided visual reasoning and multi-modal retrieval (See Fig. 2).

Overcoming Modality Gap with Cross-Modal Mixup

A typical vision-language pretraining framework with contrastive objective employs batch-dependent InfoNCE (Oord, Li, and Vinyals 2018) that pulls positive {image, text} pairs together and pushes others apart. We state this image text contrastive (ITC) loss as

$$\mathcal{L}_{ITC} = ctr(\mathcal{F}(I), \mathcal{F}(T)), \tag{1}$$

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. 1a, 1b) with separate modality-specific encoders (Radford et al. 2021; Li et al. 2021), but we have observed training failure for a generic single-tower model (Fig. 1d, Tab. 1). Visualization of the representation space in Fig. 4a indicates the presence of a severe modality gap, as visual signals and linguistic symbols are significantly dissimilar. Hence, the model fails to blend these two distant modalities in a unified representation space, being unable to encode positive {image, text} pair close together.

Cross-Modal Mixup

Mixup (Zhang et al. 2017) was originally introduced in the vision community as a data-augmentation routine that improves classification performance, model robustness and generalization by extending the training data distribution with linear interpolation. Recently, a concurrent work (Hao et al. 2022) has incorporated mixup into VLP in a similar spirit, applying 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

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

where we use an online model and its momentum (EMA) counterpart for feature extraction in practice[1]. This straightforward approach to mitigate 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. 6.

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**

Table 1: Zero-shot Imagenet (Deng et al. 2009) evaluations. Note that all models are one tower except for the second row. Adding XMC enables one tower contrastive learning, and enforcing modality-blind token attentions further improves the performance. Masked modeling is included in all ablation models.

[1]Note that ctr by definition in Eq. (2) uses two separate feature extractors (online and EMA) symmetrically.

Towards Modality-Agnostic Representations

In the previous section, we have identified modality gap as the primary obstacle for learning a unified vision-language representation space, and proposed XMC loss to reconcile the 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 whether it uses image or text as the context (*i.e., key and value in self-attention*). To enforce such behavior, we devise Contextual Invariance Contrastive (CIC) loss and incorporate it into our framework.

Contextual Modality Invariance

The high level idea is to encourage the model representation from an image context to be close to that from the text context. Specifically, from a pair, we choose either the image or the text to be the *query*. Then, at one side, we use image tokens for *key* and *value*, while on the other side, we use the text tokens. 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. Combining it with XMC in Eq. (2), the formal definition becomes

$$\mathcal{L}_{CIC} = ctr(\frac{\mathcal{F}(I|T)+\mathcal{F}(T|I)}{2}, \frac{\mathcal{F}(I|I)+\mathcal{F}(T|T)}{2}), \quad (3)$$

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 Eq. (1) and Eq. (2) is an abbreviated expression equivalent to $\mathcal{F}(X|X)$.

Contextual Mixup Contrast (CMC)

As apparent from Tab. 1, CIC improves overall performance by encouraging the model to not only 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 replace the left contrastive term in Eq. (3) with simple concatenation of $\{$image, text$\}$ ($\mathcal{F}(I,T)$) and train the model to optimize Contextual Mixup Contrastive (CMC) objective instead.

$$\mathcal{L}_{CMC} = ctr(\mathcal{F}(I,T|I,T), \frac{\mathcal{F}(I|I)+\mathcal{F}(T|T)}{2}) \quad (4)$$

This is a generalized form that further integrates XMC and CIC, 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-aided visual reasoning (Table 3) and multi-modal retrieval (Fig. 2). 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 supplementary.

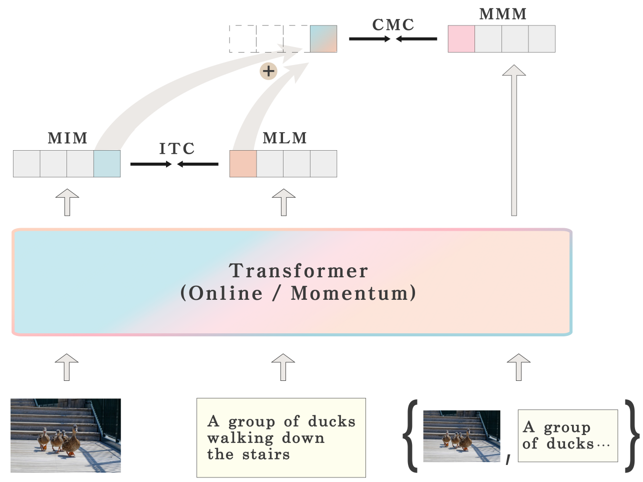


Figure 5: Overview of OneR. Image-text contrastive and contextual mixup contrastive objective provide guidance in parallel with masked modeling for three input types: image, language and multi-modal (image+text).

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.

One Representation (OneR)

Fig. 5 illustrates the overall pipeline of OneR. Model input can be one of *image*, *text* or *image+text*, and CMC objective in Eq. (4) is combined with the traditional image-text contrastive (ITC) loss. Masked modeling is also carried out for all three input types (*i.e., image, text and multi-modal*). Our framework employs no modality-specific architectural component except for the initial token embedding layer, making our model generic and modality-agnostic with minimal inductive bias. Tab. 2 summarizes the overall formulations.

Experiments

Training Setup

Datasets Following prior works (Li et al. 2021; Yang et al. 2022; Gan et al. 2020), we train OneR on the combination of CC3M (Sharma et al. 2018), SBU Captions (Ordonez, Kulkarni, and Berg 2011), Visual Genome (Krishna et al. 2017) and COCO (Lin et al. 2014), which sums up to 4M images and 5.1M image-text pairs. Ablation models are trained on CC3M.

Implementation Details We adopt the model architecture of Masked AutoEncoder (He et al. 2022) with BERT (Devlin et al. 2018) 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 [2]. 1D and 2D

[2]Two-legged models typically initialize their encoders with a

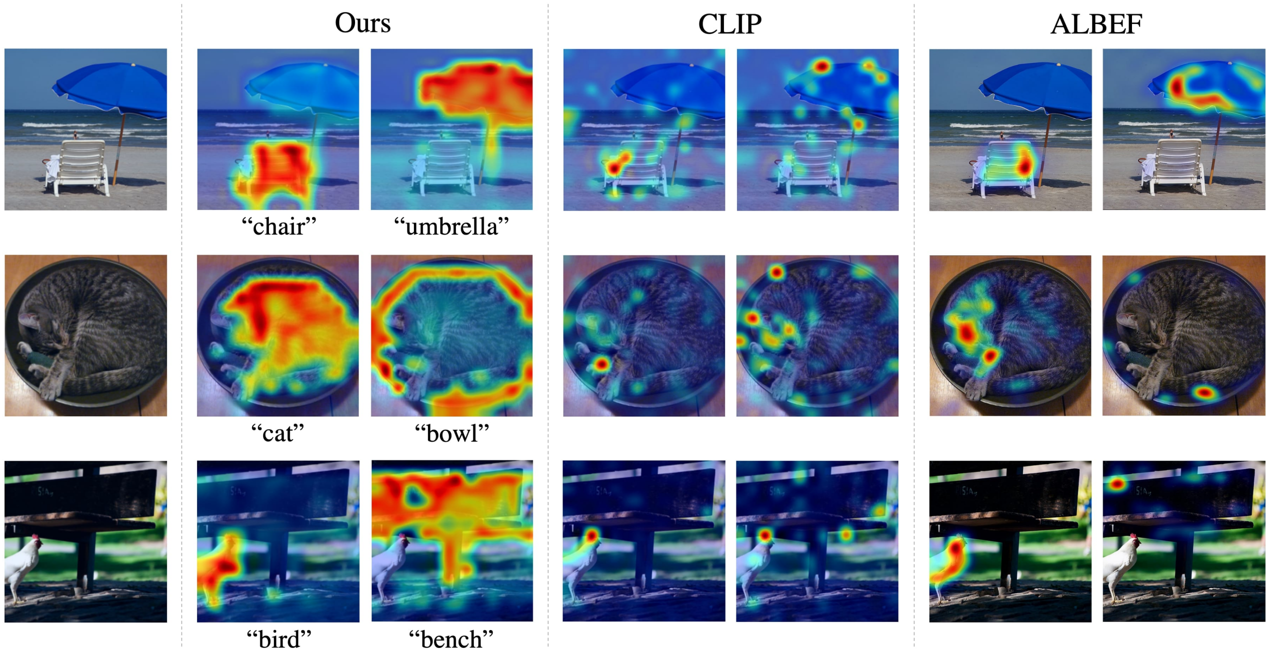


Figure 6: Qualitative evaluation for object-level scene understanding. We simply compute token similarities for OneR, and Grad-CAM is used for CLIP and ALBEF. It is visually apparent that OneR correctly associates low-level visual signals to its corresponding language symbol, resulting in segmentation-map-like patch similarity maps.

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

Table 3: 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.

Cross-modal Transfer	Arch.	INet Acc.	MS COCO TR@1	MS COCO IR@1
SBU	two heads	**7.28**	**8.88**	5.73
	one tower	6.49	8.60	**5.77**
SBU + CC3M caption	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 4: 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.

sinusoidal positional embeddings 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.,* [SEP] or [SEG]) are further removed from typical one tower baselines in order to train a fully modality-agnostic representation learner. We train our model with 32 A100 GPUs for 40 epochs under PyTorch framework. Details on hyperparameters are listed in the supplementary.

Properties of One Representation

Zero-shot Localization Conventional vision-language transformers typically rely on [CLS] cross-attention map or Grad-CAM (Selvaraju et al. 2017) 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. 2), 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. 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 token level similarity map conveys no meaningful information, as illustrated in Fig. 3.

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

pretrained ViT and a pretrained language model such as BERT, which makes the training much simpler.

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@1	R@5	R@1	R@5	R@1	R@5
ImageBert†	One Tower	O	6M	44.0	71.2	32.3	59.0	**66.4**	**89.8**	**50.5**	**78.7**
ViLT	One Tower	O	4M	56.5	82.6	40.4	70.0	61.5	86.3	42.7	72.9
Uni-Perceiver	One Tower	X	44.3M	<u>57.7</u>	<u>85.6</u>	<u>46.3</u>	**75.0**	64.7	<u>87.8</u>	<u>48.3</u>	75.9
OneR	One Tower	X	4M	**62.9**	**86.3**	**47.0**	<u>74.7</u>	<u>66.1</u>	<u>87.8</u>	<u>48.3</u>	<u>76.0</u>
CLIP	Two Towers	X	400M	58.4	81.5	37.8	62.4	-	-	-	-
FLAVA	Two Legs	O	70M	42.7	76.8	38.4	67.5	-	-	-	-
ALBEF	Two Legs	O	4M	68.7	89.5	50.1	76.4	73.1	91.4	56.8	81.5
TCL	Two Legs	O	4M	71.4	90.8	53.5	79.0	75.6	92.8	59.0	83.2

Table 5: Quantitative evaluations on COCO image and text retrieval. Two-legs models generally perform better as they have modality-specific encoders and more parameters. Note that previous vision-language models typically initialize their weights from a pretrained model such as Imagenet ViT or Bert to help training (*Pre.*). OneR, on the other hand, achieves the best zero-shot performance among one-tower models without any initialization prior, and compares on par after fine-tuning. † indicates the use of an additional object detection module.

the best output. Looking at Fig. 6, 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.

Text-guided Visual Reasoning As illustrated in Fig. 2, 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 Table 3, 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. Then we concatenate each to the image sequence and compute similarity once more, similar to sample re-ranking. The intuition is that when *image+text* input is given, image patches that attend strongly 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*.

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. Table 4 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 available than image-text pairs in the web, which we leave for future works.

Quantitative Evaluations

Table 5 shows the quantitative comparison with state-of-the-art methods on widely used image-text retrieval benchmark.

Method	INet	MS-COCO			
	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 (two heads)	17.5	10.4	26.8	10.7	26.4
ITC	1.6	0.8	2.5	0.7	2.2
+ XMC	22.1	25.2	48.1	15.2	33.6
+ XMC + CIC	22.9	25.4	48.1	16.3	35.5
+ CMC (OneR)	**23.7**	**25.5**	**48.2**	**16.9**	**36.9**

Table 6: Method ablation. Our proposed components consistently improve the performance, with the final CMC outperforming the two-tower baseline that uses more parameters and intra-modal contrastive loss. Additional ablations are presented in the supplementary.

Models with modality-specific encoders 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 competent outcome without any initialization prior commonly used in the literature. 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.

In Table 6, we present full ablations for our framework. Naive ITC with one 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, which is explained further in the supplementary, resulting in the best performance that surpasses competent two-tower baselines.

Visual Reasoning Analysis

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

Robustness Fig. 7 shows an example of how OneR recog-

Figure 7: 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*).

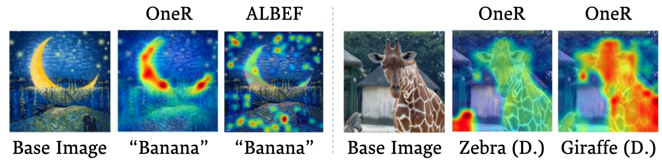


Figure 8: (*left*) Patch embedding similarity (OneR) and Grad-Cam (ALBEF). (*right*) Patch embedding similarity map w.r.t. definitions of zebra and giraffe.

nizes 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. We present additional results in the supplementary materials, including inference in unfamiliar domains.

Multi-level vision-language connection Looking at Fig. 8, OneR recognizes the *moon* as being visually similar to *banana* in terms of embedding similarity, while ALBEF condenses the global semantic in `[CLS]`, resulting in a randomly scattered Grad-CAM. 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.

Related Works

Vision-Language Pretraining CLIP (Radford et al. 2021) first demonstrated the effectiveness of large-scale vision-language contrastive learning. ALIGN (Jia et al. 2021) scaled up the training with noisy alt-text pair data. Another line of works (Li et al. 2020b,a; Chen et al. 2020b; Gan et al. 2020; Kim, Son, and Kim 2021) leveraged an off-the-shelf object detector to extract visual concepts first, which were then used to train the multi-modal transformer. In an attempt to learn cross-modal interactions, ALBEF (Li et al. 2021), TCL (Yang et al. 2022), FLAVA (Singh et al. 2022), and Florence (Yuan et al. 2021) adopted multi-modal fusion layers on top of modality-specific transformer encoders. Another group of works (Li et al. 2022; Yu et al. 2022; Wang et al. 2021b; Mokady, Hertz, and Bermano 2021) explored generative modeling, typically in the form of image captioning, to further improve performances on challenging tasks such as visual question answering.

Unified VL Framework Uni-Perceiver (Zhu et al. 2022) adopted a single-tower transformer architecture to tackle different V-L tasks in a unified manner. Unified-IO (Lu et al. 2022) further used a pretrained VQ-VAE to model a wide range of tasks with a generic sequence-to-sequence framework. These works have demonstrated promising direction towards a unified perception system, but the fact that they employ multi-task pretraining strategy renders them less scalable compared to CLIP or ALIGN. UFO (Wang et al. 2021a) has shown that a single transformer model suffices for typical VLP when combined with two modality-specific projectors. LIMoE (Mustafa et al. 2022), a concurrent work of ours, also explores single-tower (two heads) VLP but with a new set of inductive biases, *i.e.,* mixture of experts. OneR, in contrast, learns a common embedding space without any modality-specific components, which empowers the model with unique capabilities previously demonstrated.

Self-supervised Learning Self-supervised learning first bloomed in the NLP domain as masked language modeling (MLM) and language modeling (LM) enabled pretraining large language models with huge stock of unlabeled text corpus (Devlin et al. 2018; Radford et al. 2018; Lewis et al. 2019; Liu et al. 2019). In the vision community, contrastive learning has led the rise of SSL. MoCo (He et al. 2020) and SimCLR (Chen et al. 2020a) are the pioneers to demonstrate the potential of contrastive representation learning, which we adapt for VLP setting. BYOL (Grill et al. 2020) and SimSiam (Chen and He 2021) explored new settings with no negative samples that mitigate the batch size dependency. Recent works (Caron et al. 2021; Chen, Xie, and He 2021; Jang et al. 2021) actively employ ViT (Dosovitskiy et al. 2020) to improve the performance and discover new properties. This architecture is also widely used in VLP as it can model data from different modalities in an integrated manner.

Conclusion

Modality-agnostic representation learning is a meaningful step towards 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.

Acknowledgements

This work was supported by NRF grant (2021R1A2C3006659) and IITP grants (No.2022-0-00953, No.2021-0-01343), all of which are funded by Korean Government. The authors would like to express our sincere appreciation to Jungwon Choi for her exceptional efforts in creating the figures for this paper.

References

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

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

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

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

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020b. Uniter: Learning universal image-text representations. 104–120.

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

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

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

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

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

Hao, X.; Zhu, Y.; Appalaraju, S.; Zhang, A.; Zhang, W.; Li, B.; and Li, M. 2022. MixGen: A New Multi-Modal Data Augmentation. *arXiv preprint arXiv:2206.08358*.

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

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

Jang, J.; Kim, S.; Yoo, K.; Kim, J.; and Kwak, N. 2021. Self-Distilled Self-Supervised Representation Learning. *arXiv preprint arXiv:2111.12958*.

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

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

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

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

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

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

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

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

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

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

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

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

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

Mustafa, B.; Riquelme, C.; Puigcerver, J.; Jenatton, R.; and Houlsby, N. 2022. Multimodal Contrastive Learning with LIMoE: the Language-Image Mixture of Experts. *arXiv preprint arXiv:2206.02770*.

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

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

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

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training. *OpenAI research covers:1806*.

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

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. 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)*, 2556–2565.

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

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

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

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

Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; and Huang, J. 2022. Vision-Language Pre-Training with Triple Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15671–15680.

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

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

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

Zhu, X.; Zhu, J.; Li, H.; Wu, X.; Li, H.; Wang, X.; and Dai, J. 2022. 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*, 16804–16815.