Multi-modal Alignment using Representation Codebook

Jiali Duan[1*] Liqun Chen[1*] Son Tran[1] Jinyu Yang[2] Yi Xu[1] Belinda Zeng[1] Trishul Chilimbi[1]
[1] Amazon [2] University of Texas at Arlington
{duajiali,liquchen,sontran,yxaamzn,zengb,trishulc}@amazon.com {viyjy}@mavs.uta.edu

Abstract

Aligning signals from different modalities is an important step in vision-language representation learning as it affects the performance of later stages such as cross-modality fusion. Since image and text typically reside in different regions of the feature space, directly aligning them at instance level is challenging especially when features are still evolving during training. In this paper, we propose to align at a higher and more stable level using cluster representation. Specifically, we treat image and text as two "views" of the same entity, and encode them into a joint vision-language coding space spanned by a dictionary of cluster centers (codebook). We contrast positive and negative samples via their cluster assignments while simultaneously optimizing the cluster centers. To further smooth out the learning process, we adopt a teacher-student distillation paradigm, where the momentum teacher of one view guides the student learning of the other. We evaluated our approach on common vision language benchmarks and obtain new SoTA on zero-shot cross modality retrieval while being competitive on various other transfer tasks.

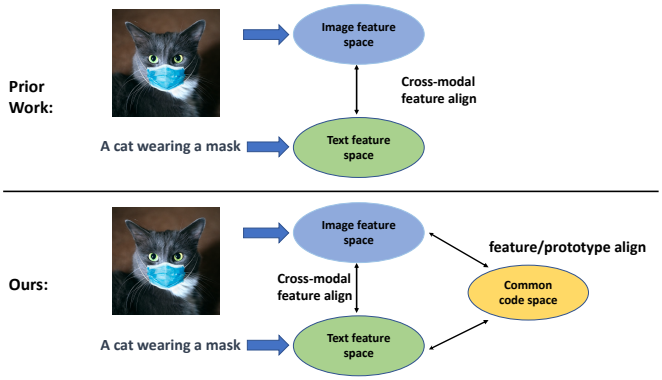

Figure 1. We propose to use a learnable codebook to better align the image and text modalities. The codebook serves as a "bridge" between the image and text features. Each codeword can be interpreted as a prototype, which enables contrasting image and text at the cluster level. We then solve an optimal transport [1] problem to optimize the distance between each modality to the prototypes, which in turn optimizes the alignment between the two modalities. Prototype vectors are learned along with the feature encoders in our V&L framework.

1. Introduction

Vision language (V&L) representation learning is the problem of learning a unified feature embedding using both image and text signals. Pretrained V&L models have a great diversity of applications in various downstream tasks across different settings, e.g. via transfer learning [8, 29, 49]. The main tasks in V&L pretraining include aligning the feature spaces of different modalities (multi-modal alignment [8, 26, 29, 32]) and capturing the interaction across modalities (cross-modal fusion, [13, 45]). Late fusion approaches such as CLIP [38] and ALIGN [22] focused on the first task, while early fusion approaches such as OSCAR [29], VinVL [49] and VilLT [23] focused on the second one. In this work, we adopt a hybrid approach similar to ALBEF [26], where features from image and text modalities were first aligned and then fused using a transformer encoder. The main focus of our work is on the feature alignment stage, which is challenging due to the fact that image and text inputs have very different characteristics. Existing approaches such as CLIP [38] and ALIGN [22] have to rely on large training resources and on massive amount of data to obtain good alignments (400M and 1.8B image-text pairs respectively).

In this work, we propose a more efficient alignment strategy by using a codebook that quantizes the common text-image feature space into codewords. These codewords or cluster centers provide a more stable means for contrastive reasoning compared to individual text or visual features. We took the inspiration from SwAV [4], which was developed for self-supervised visual representation learning. In [4], two augmented versions (views) of the same input image were passed through a deep network for feature extraction. Visual embedding was learned by optimizing an objective function that enforces the consistency between the feature from one view and the assigned cluster from the other view. SwAV achieved impressive performance in various transfer tasks (see [4]). Here, we carried out contrastive reasoning across modalities (image-text) instead of cross image views. De-

tails are in Section 3.1, but in a nutshell, we use a learnable codebook for both image and text modalities and train our model to predict the codeword assignment using either text or visual information. Effectively, visual and text features are lined up via aligning with the common codewords during training. See Figure 1 for an illustration.

The codebook can be considered as a quantized sample of the underlying output feature distribution. It is end-to-end learnable together with the model parameters. To avoid abrupt changes during training, we further employ momentum distillation, which has been widely used in previous self-supervised learning works such as BYOL [17], DINO [5], MoCo [19]. In brief, similar to ALBEF [26], for each of the image, text and fusion encoders, there is a corresponding encoder that is updated through moving average without gradient back propagation. These momentum encoders serve as teachers to guide the self-supervised learning process. Different from ALBEF [26], we use the teachers to guide codebook learning as well as for the cross-modal and intra-modal alignment.

The above two components are wired up to support the stable update of the codebook which, in turn, provides an efficient regularization mean for cross modality alignment. Experiment results (Section 4) show that our approach is competitive with state of the art across various benchmarks even when comparing with approach that use massive amount of data such as CLIP [38] and ALIGN [22]. In summary, our main contributions are as follows,

- We propose a codebook-based approach for efficient vision-language alignment learning. It is an extension from self-supervised vision representation learning (SSL) to the multimodal setting.

- We introduce a new distillation algorithm that helps unimodal and crossmodal contrastive optimization as well as helps stablize codebook learning.

The rest of the paper is organized as follows. We introduce related work to ours in Section 2. In Section 3, we describe our framework, called **Co**debook Learning with **Dis**tillation (CODIS), and its two components, multimodal codebook learning and teacher-student distillation. Experimental results are presented in Section 4. Section 5 concludes the paper.

2. Related Work

Vision-Language Pre-training (V&L) V&L pretraining is an active research area with many recent works. We review here the works that are most relevant to ours. Architecture wise, previous approaches can be broadly clasified into two categories early fusion and late fusion. In early-fusion approaches [8, 23, 29, 42], image and text are transformed into sequences (tokenization) and passed to a single encoder (typically Transformer-based) for embedding generation. Thus multimodal signals are fused in the early stage. Whereas in late-fusion works [22, 38], separate encoders are used for image and text. Extracted features are typically fused during the later fine tuning stage. Our work is a hybrid between these two approaches, similar to ALBEF [26]. The main difference between ALBEF and ours is the codebook and various related contrastive losses.

In vision language learning, codebook has been used in a number of recent works, mostly for image tokenization. BEiT [2] constructed a dictionary of visual words, then used it to form mask image modeling task in the same fashion as mask language modeling. SOHO [21] integrated visual dictionary to the main model and jointly trained both of them. Both works quantized the visual input space. In our work, codebook is used to quantize the joint output space , where multimodal views are aligned via optimal transport [1]. Other concurrent works to ours include [26, 28]. They both align cross-modal instances using InfoNCE [34]. In contrast, we enforce both unimodal and cross-modal alignment, both at the instance level and at the cluster level.

Self-supervised Contrastive Learning The goal of contrastive learning [18] is to attract positive sample pairs and repulse the negative sample pairs. Recently, it has been widely used in computer vision for unsupervised and self-supervised representation learning [5, 7, 19]. Contrastive reasoning is typically formed based on two augmented views of the same input image. One of the main challenge is feature collapsing, and in practice, a large number of negative samples are required, through either large batch size [7] or memory banks [19, 46], to alleviate this problem. Several recent works have shown that one can learn unsupervised features without discriminating instances. Deep clustering [3] and SwAV [4] incorporate online clustering into Siamese networks. In BYOL [17], features are trained by matching them to representations obtained by a momentum encoder. DINO [5] instantiates the momentum encoder with a vision-transformer and adopts a teacher-student distillation paradigm [14, 20, 48]. Our alignment techniques and momentum update were inspired by these works and can be considered as extensions to the multimodal setting.

3. Method

Our goal is to learn explicit alignment between image and text features to facilitate multimodal interactions. We illustrate CODIS in Figure 2 and propose a pseudo-code implementation in Algorithm 1. It shares some similarities with self-supervised contrastive learning [4, 19]. We treat image and text modalities as two views and adopt a teacher-student distillation paradigm [5, 17] to enforce unimodal and cross-modal alignment. To overcome the gap between multimodal

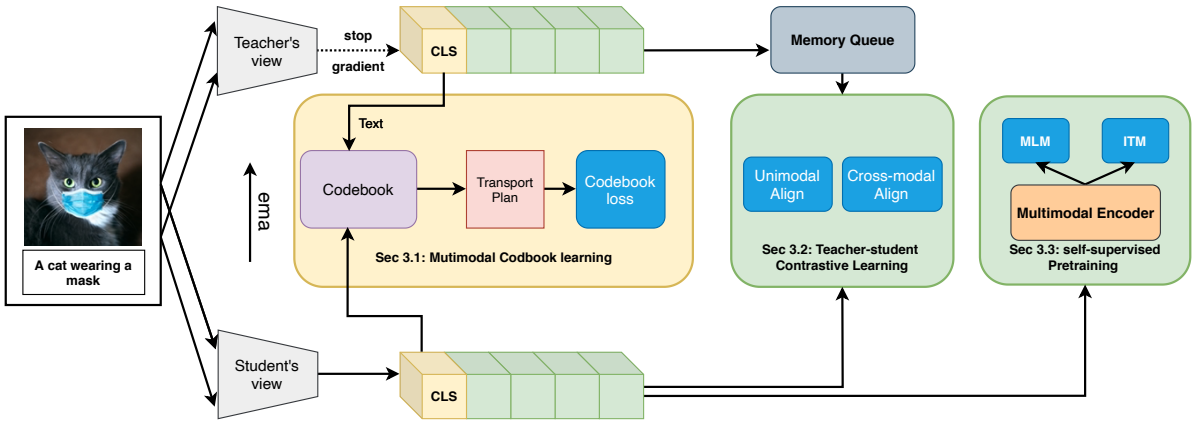

Figure 2. Overview of our framework. For simplicity, we only display a pair of teacher-student encoders (e.g., teacher for the image and student for the text) and similarly for the memory queue. The teacher is updated with an exponential moving average of the student (from the same modality). The codebook helps bridge the gap between the different modalities. The entire framework is end-to-end optimized.

.

distributions, we also learn a codebook, which serves as a bridge to help align features between different modalities. We organize the content of this section as follows.

In Section 3.1, we present multimodal codebook learning, how it's optimized and how to leverage it to resolve distribution mismatch between multimodal inputs. In Section 3.2, we introduce how to achieve unimodal and cross-modal alignment under the teacher-student distillation learning formulation. Finally, we explain how our proposed two components integrate into the V&L framework in Section 3.3.

3.1. Multimodal Codebook Learning

We propose to learn a codebook to facilitate aligning multimodal semantics. It's a collection of learnable prototypes or codewords. We use them interchangeably in this paper. With codebook, we encode image and text into a joint vision-language embedding space and learn the alignment by contrasting their prototype assignments. The codebook can also be interpreted as underlying feature distribution for the paired data [6]. In this way, by aligning features from each modality with the codebook, we implicitly align multimodal features indirectly. In other words, the codebook serves as a "bridge" between the modalities (See Figure 1).

We denote the learnable codebook as $\mathbf{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\} \in \mathcal{R}^{d_c \times K}$, where d_c is the dimension for each code and K equals to the number of codewords (i.e., 4K). We set $d_c = 256$, same as the dimension of projected image/text features. Each $\boldsymbol{c} \in \mathbf{C}$ is a prototype.

Given N image or text feature vectors $\mathbf{Z}^m = [\boldsymbol{z}_1^m, \ldots, \boldsymbol{z}_N^m]$ (superscript m denotes features extracted from the momentum teacher encoder), we compute an optimal cost mapping from the feature vectors to the prototypes. We denote such mapping as a transport plan $\mathbf{T}$, obtained using Optimal Transport [1, 6]. Without loss of generality, we

Algorithm 1 CODIS pseudocode

```
# gs, gt: student/teacher networks for image
# fs, ft: student/teacher networks for text
# C: codebook d-by-K
# Qv, Qt: image/text queue, d-by-M
# tmp, learnable temperature
for (img, txt) in loader: # a minibatch with N samples
    # teacher/student's image view
    img_t, img_s = gt(img), gs(img) # N-by-d

    # teacher/student's text view
    txt_t, txt_s = ft(txt), fs(txt) # N-by-d

    # calculate codebook loss
    I2P, T2P = img_t@C, txt_t@C, # N-by-K
    Tg, Tf = IPOT(1-I2P), IPOT(1-T2P) # refer to Algo 2
    L_ot = Trace(I2P.t()@Tg).sum() + Trace(T2P.t()@Tf).sum()
    L_code = H(img_s@C, Tg) + H(txt_s@C, Tf) + L_ot

    # calculate alignment loss
    L_cross = H(img_s@Qt, img_t@Qt) + H(txt_s@Qv, txt_t@Qv)
    L_unimo = H(img_s@Qv, img_t@Qv) + H(txt_s@Qt, txt_t@Qt)
    L_align = L_cross + L_unimo

    # enqueue/dequeue
    update_queue(Qv, img_t, Qt, txt_t)

    # pretraining loss
    L_pretrain = L_itm + L_mlm

    loss = L_code + L_align + L_pretrain
    loss.backward() # back-propagate

    # student, teacher updates
    update(gs, fs) # SGD
    ema(gs, gt, fs, ft) # momemtum update


def H(s, t):
    t = t.detach() # stop gradient
    s = softmax(s / tmp, dim=1)
    return - (t * log(s)).sum(dim=1).mean()
```

denote $\boldsymbol{z}$ as the projected features for either image or text and optimize the following objective,

$$\mathcal{L}_{\text{ot}} = \min_{\mathbf{T} \in \Pi(\mathbf{u}, \mathbf{v})} \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbf{T}_{ij} \cdot d(\boldsymbol{z}_i^m, \boldsymbol{c}_j) = \min_{\mathbf{T} \in \Pi(\mathbf{u}, \mathbf{v})} \langle \mathbf{T}, \mathbf{D} \rangle , \tag{1}$$

Algorithm 2 IPOT Algorithm.

Input: distance/similarity matrix $\mathbf{Z}$, $\mathbf{C}$, ϵ, probability vectors $\boldsymbol{\mu}, \boldsymbol{\nu}$
$\boldsymbol{\sigma} = \frac{1}{n}\mathbf{1_n}$, $\mathbf{T}^{(1)} = \mathbf{1}\mathbf{1}^\top$
$D_{ij} = d(\boldsymbol{z}_i, \boldsymbol{c}_j)$,$\mathbf{A}_{ij} = \mathrm{e}^{-\frac{\mathbf{D}_{ij}}{\epsilon}}$
for $t = 1, 2, 3 \ldots$ **do**
  $\mathbf{Q} = \mathbf{A} \odot \mathbf{T}^{(t)}$ // $\odot$ is Hadamard product
  for $k = 1, 2, 3, \ldots K$ **do**
    $\boldsymbol{\delta} = \frac{\boldsymbol{\mu}}{n\mathbf{Q}\boldsymbol{\sigma}}, \boldsymbol{\sigma} = \frac{\boldsymbol{\nu}}{n\mathbf{Q}^\top\boldsymbol{\delta}}$
  end for
  $\mathbf{T}^{(t+1)} = \mathrm{diag}(\boldsymbol{\delta})\mathbf{Q}\mathrm{diag}(\boldsymbol{\sigma})$
end for
Return $\mathbf{T}$

where $\Pi(\mathbf{u}, \mathbf{v}) = \{\mathbf{T} \in \mathbb{R}_+^{N\times K} | \mathbf{T}\mathbf{1}_K = \frac{1}{N}\mathbf{1}_N, \mathbf{T}^\top\mathbf{1}_N = \frac{1}{K}\mathbf{1}_K\}$, $\mathbf{1}_N$ denotes an N-dimensional all-one vector. $\mathbf{D}$ is the cost matrix given by $\mathbf{D}_{ij} = d(\boldsymbol{z}_i^m, \boldsymbol{c}_j)$ ($d(\cdot,\cdot) = 1 - \cos(\cdot,\cdot)$) and $\langle \mathbf{T}, \mathbf{D}\rangle = \mathrm{Tr}(\mathbf{T}^\top\mathbf{D})$ represents the Frobenius dot-product. We use Tg and Tf for the optimal transport plan for image and text in Algorithm 1, and $1 - I2P$ corresponds to the cost matrix $\mathbf{D}$ for image modality. It's similar for text.

To solve for the optimal transport plan, we adopt an iterative algorithm shown in Algorithm 2. It takes normalized feature matrix $\mathbf{Z}$, codebook $\mathbf{C}$ as input and output an optimal tranpsort plan $\mathbf{T}$. Internally, the algorithm tries to minimize the optimal transport (OT) distance, optimized to pick similar $\boldsymbol{c}_j, j \in [1, \ldots, K]$ for each $\boldsymbol{z}_i$ based on score $\mathbf{T}[i,:]$ (i^{th} row of $\mathbf{T}$). In other words, $\mathbf{T}$ can be viewed as a distance metric between prototypes and features. When solved, OT yields a sparse solution $\mathbf{T}^*$ containing at most $(2r-1)$ ($r = \max(N, K)$ non-zero elements, leading to a robust and meaningful alignment [11].

In the codebook loss that we are going to formulate, $\mathbf{T}$ will be used as ground-truth signals to guide the feature-to-prototype alignment. We use cross entropy loss and adopt a teacher-student distillation approach to construct the loss for optimizing the codebook as well as the feature encoders,

$$\begin{aligned} \mathcal{L}_{\text{t2p}}(\mathbf{Z}_t, \mathbf{C}, \mathbf{T}_{i2p}) &= H(\mathbf{P}_{t2p}, \mathbf{T}_{i2p}), \\ \mathcal{L}_{\text{i2p}}(\mathbf{Z}_v, \mathbf{C}, \mathbf{T}_{t2p}) &= H(\mathbf{P}_{i2p}, \mathbf{T}_{t2p}), \\ \mathbf{P}_{t2p} = \textbf{SoftMax}(\mathbf{Z}_t\mathbf{C}/\gamma), &\mathbf{P}_{i2p} = \textbf{SoftMax}(\mathbf{Z}_v\mathbf{C}/\gamma) \end{aligned} \tag{2}$$

where $\mathbf{P}$ is the predicted metric calculated with the features from the student encoders while $\mathbf{T}$ is calculated with features from the teacher encoders using Algorithm 2. The reason is that the teacher encoders are updated via exponential moving average, which helps avoid abrupt changes in codebook learning.

We additionally add a regularization term $\mathcal{L}_{\text{ot}}$. The overall loss for multimodal codebook learning is as follows,

$$\begin{aligned} \mathcal{L}_{\text{code}} = &\mathcal{L}_{\text{ot}}(\mathbf{Z}_v^m, \mathbf{C}) + \mathcal{L}_{\text{ot}}(\mathbf{Z}_t^m, \mathbf{C}) \\ &+ \mathcal{L}_{\text{t2p}}(\mathbf{Z}_t, \mathbf{C}, \mathbf{T}_{t2p}) + \mathcal{L}_{\text{i2p}}(\mathbf{Z}_v, \mathbf{C}, \mathbf{T}_{i2p}) \end{aligned} \tag{3}$$

As shown in Figure 3, codebook acts as a bridge between the image and text modality, as both text to prototype loss

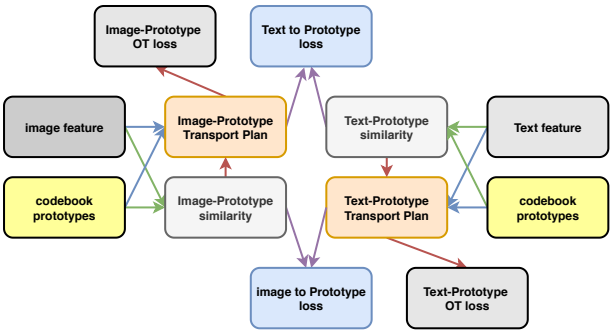

Figure 3. This is the diagram illustrating how to calculate four codebook losses. "→": softmax operator. "→": IPOT algorithm. "→": OT loss. "→": cross entropy.

($\mathcal{L}_{\text{t2p}}$) or image to prototype loss ($\mathcal{L}_{\text{i2p}}$) chain features from both modalities. For example, Text to Prototype loss chains Image-Prototype Transport Plan and Text-Prototype Similarity and vice versa. More importantly, learning codebook allows contrasting features across modalities at the prototype level, i.e, feature distribution matching. When calculating the transport plan, we use the teacher features as they provide a more stable supervision signal to guide the learning of the student. The calculated losses will be backpropagated to update both the codebook and student encoders.

3.2. Teacher-student Distillation Learning

This loss is designed to align the features from two unimodal encoders, which is inspired by the recent success of SSL learning [5, 19]. Our motivation is that image and text can be treated as two "views" of the same entity, and we adopt a teacher-student distillation paradigm to align them. Since the raw feature directly from unimodal encoders are in different feature spaces, we learn a joint embedding space of dimension 256, $\boldsymbol{z}_v \in \mathcal{R}^{256}$, $\boldsymbol{z}_t \in \mathcal{R}^{256}$ for image and text student features. Following [19, 26], we store features from the teacher encoders $\boldsymbol{z}_v^m \in \mathcal{R}^{256}$, $\boldsymbol{z}_t^m \in \mathcal{R}^{256}$ in memory queues $\mathbf{Q}_v$, $\mathbf{Q}_t$ for image and text respectively.

For a pair of image and text, we can calculate the cross-modal similarity and intra-modal similarity as follows:

$$\begin{aligned} \boldsymbol{p}_{t2i}(T) &= \exp\frac{\boldsymbol{z}_t\boldsymbol{z}_v^{m\top}}{\gamma} / \sum_{\boldsymbol{z}_v^{m'}\in\mathbf{Q}_v} \exp\frac{\boldsymbol{z}_t\boldsymbol{z}_v^{m'\top}}{\gamma} \\ \boldsymbol{p}_{i2t}(I) &= \exp\frac{\boldsymbol{z}_v\boldsymbol{z}_t^{m\top}}{\gamma} / \sum_{\boldsymbol{z}_t^{m'}\in\mathbf{Q}_t} \exp\frac{\boldsymbol{z}_v\boldsymbol{z}_t^{m'\top}}{\gamma} \\ \boldsymbol{p}_{i2i}(I) &= \exp\frac{\boldsymbol{z}_v\boldsymbol{z}_v^{m\top}}{\gamma} / \sum_{\boldsymbol{z}_v^{m'}\in\mathbf{Q}_v} \exp\frac{\boldsymbol{z}_v\boldsymbol{z}_v^{m'\top}}{\gamma} \\ \boldsymbol{p}_{t2t}(T) &= \exp\frac{\boldsymbol{z}_t\boldsymbol{z}_t^{m\top}}{\gamma} / \sum_{\boldsymbol{z}_t^{m'}\in\mathbf{Q}_t} \exp\frac{\boldsymbol{z}_t\boldsymbol{z}_t^{m'\top}}{\gamma} \end{aligned} \tag{4}$$

where pseudo image negatives for estimating $\boldsymbol{p}_{t2i}(T)$ is sampled from the image queue $\mathbf{Q}_v$ and similarly for $\boldsymbol{p}_{i2t}(I)$. In addition to [26], we also considered unimodal (intra) alignment. Intuitively, enhancing unimodal feature representation lays a better foundation for cross-modal alignment.

To further smooth out the learning process, we use the features from the momentum teacher to provide the soft distillation target, $\boldsymbol{y}_{i2t}, \boldsymbol{y}_{t2i}, \boldsymbol{y}_{t2t}, \boldsymbol{y}_{i2i}$ (refer to Algorithm 1 for details). The loss for intra/cross-modal alignment is defined as,

$$\mathcal{L}_{ica} = \mathbb{E}_{I,T\sim\boldsymbol{p}_{\text{data}}} [H(\boldsymbol{p}_{t2t}, \boldsymbol{y}_{t2t}) + H(\boldsymbol{p}_{i2i}, \boldsymbol{y}_{i2i}) + H(\boldsymbol{p}_{t2i}, \boldsymbol{y}_{t2i}) + H(\boldsymbol{p}_{i2t}, \boldsymbol{y}_{i2t})] \quad (5)$$

where H is cross entropy. This objective can also be viewed as knowledge distillation, between teacher encoders and student encoders from the same modality (i.e., $H(\boldsymbol{p}_{t2t}, \boldsymbol{y}_{t2t})$ and $H(\boldsymbol{p}_{i2i}, \boldsymbol{y}_{i2i})$, as well as between teacher encoders and student encoders from different modality (i.e., $H(\boldsymbol{p}_{t2i}, \boldsymbol{y}_{t2i})$ and $H(\boldsymbol{p}_{i2t}, \boldsymbol{y}_{i2t})$). Parameters for the teacher encoder is an exponential moving average of the student, detached from gradient update. We adopt momentum update similar to [19] to update the teacher encoders:

$$f_t = \alpha f_t + (1-\alpha) f_s, g_t = \alpha g_t + (1-\alpha) g_s \quad (6)$$

α is the momentum parameter. In practice, we set $\alpha = 0.995$, in order to smoothly update teacher encoders.

3.3. Self-supervised Pre-training

In this section, we will first introduce two commonly used objectives for multimodal training frameworks: (i) masked language modeling loss (MLM) and (ii) image-text matching (ITM) on the multimodal encoder. Then we discuss how codebook and teacher-student distillation components are integrated. We denote the image and text features extracted by student network as $\{v_{cls}, v_1, ..., v_m\}$ and $\{t_{cls}, t_1, ..., t_n\}$, respectively. Specifically, v_{cls} is the image [CLS] token, $\{v_1, ..., v_m\}$ are image patch embeddings. Similarly, t_{cls} indicate the text [CLS] token, $\{t_1, ..., t_n\}$ are word embeddings.

3.3.1 Image-Text Matching (ITM) Loss

To fuse vision and language representations, we adopt ITM that is widely used in modern V&L frameworks. Given an arbitrary pair of image and text, ITM predicts whether they are aligned (positive pairs) or not (negative pairs). This procedure can be formulated as a binary classification problem.

Specifically, [CLS] token from the fusion encoder is used as the joint representation of the image-text pair. ITM head is a fully connected layer to predict the matching probability $\boldsymbol{p}_{\text{itm}}$. We assume that each image-text pair (I_i, T_i) sampled from the pre-training datasets is a positive example and construct negative examples through the following strategy: For each image I_i within the batch, we sample one negative text T_j from the same batch based on the contrastive similarity distribution. So that text that is more similar to this image will have a higher chance to get sampled. Similarly, one hard negative image will be sampled for each text T_i. We denote y_{itm} as the ground-truth labels indicating whether the image-text pair is positive or negative.

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{I,T\sim\boldsymbol{p}_{\text{data}}} H(\boldsymbol{p}_{\text{itm}}, \boldsymbol{y}_{\text{itm}}) \quad (7)$$

where H is the cross entropy operator.

3.3.2 Masked Language Modeling (MLM) Loss

We follow the design of MLM loss from BERT [12], which aims to predict the ground-truth labels of masked text tokens y_{mlm}. Specifically, we randomly mask out 15% of input text tokens, those masked tokens are replaced with special token [MASK]. Different from BERT, our MLM loss is conditioned on both surrounding text tokens and image representations. Assume the predicted token probability is $\boldsymbol{p}_{\text{mlm}}$, we construct the loss objective as follows,

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{I,\hat{T}\sim\boldsymbol{p}_{\text{data}}} H(p_{\text{mlm}}, \boldsymbol{y}_{\text{mlm}}) \quad (8)$$

where $\hat{T}$ is the text token sequence after masking.

3.4. Summary

We simultaneously optimize the codebook and the student encoders within the framework in an end-to-end manner, employing the losses discussed in previous sections as follows,

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itm}} + \mathcal{L}_{\text{ica}} + \mathcal{L}_{\text{code}} \quad (9)$$

among which MLM and ITM loss have been widely used in many V&L methods particularly those "early-fusion" frameworks. The ica loss is the main objective function for "late-fusion" V&L frameworks. CODIS combines the merits of both "early-fusion" and "late-fusion" approaches, by explicitly learning alignment along with fusion.

Intra-cross alignment ($\mathcal{L}_{\text{ica}}$) loss described in Section 3.2 can be viewed as an instance-to-instance alignment loss, similar to the one in [26]. The difference is we consider both intra and cross modal alignment. We assume that a stronger unimodal representation can lay a solid foundation for cross-modal representation. Empirical evidence is provided in Section 4.5. The codebook loss ($\mathcal{L}_{\text{code}}$) designed in Section 3.1 measures the the distance between the transport plan and similarity matrix. It contrasts features at the prototype level and can be interpreted as distance metric matching [3, 6]. Combining these two help avoid prototype collapsing problem, as online prototype clustering requires careful tuning [4]. Finally, The supervision signals for both intra-cross alignment loss and codebook loss require features from the momentum teacher and we adopt a teacher-student

distillation approach. This can be seen as a generalization of unimodal SSL into the multimodal setting, under the V&L framework.

4. Experiments

To evaluate our approach, we conduct extensive studies on commonly used benchmarks and present experimental comparisons against state-of-the-art V&L methods as shown in this section.

4.1. Pre-training Datasets

We follow previous experimental protocols [8, 26] for fair comparisons. We use COCO [30], Visual Genome (VG) [24], Conceptual Captions (CC) [41], and SBU Captions [35] as the pre-training dataset in our study, where a total of 4.0M unique images and 5.1M image-text pairs are covered.

4.2. Downstream Tasks

Image-Text Retrieval consists of two tasks: (1) image as query and text as targets (TR); (2) text as query and image as targets (IR). The pre-trained model is evaluated on Flickr30K [36] and COCO [30] by following both fine-tuning and zero-shot settings. For the fine-tuning setting, the pre-trained model is fine-tuned on the training data and evaluated on the validation/test data. For the zero-shot setting, the pre-trained model is directly evaluated on the test data without any further training. In particular, for zero-shot retrieval on Flickr30K, we follow the procedure proposed in [26] (zero-shot evaluating on Flickr with the model fine-tuned using MSCOCO).

Visual Question Answering (VQA) [16] aims to predict the answer given an image and a question (in text format), which requires an understanding of vision, language and commonsense knowledge to answer. We consider this task as a generation problem by following the same setting in [26]. Specifically, an answer decoder is fine-tuned to generate the answer from the 3,192 candidates.

Visual Entailment (SNLI-VE) [47] predicts whether an given image semantically entails a given text, which is a three-classes classification problem. Specifically, the class or relationship between any given image-text pair can be entailment, neutral, or contradictory. Compared with VQA, this task requires fine-grained reasoning.

Visual Reasoning (NLVR^2) [43] determines whether a natural language caption is true about a pair of photographs. We evaluate our model on NLVR^2 dataset which contains 107,292 examples of human-written English sentences paired with web photographs. Since this task takes a text and two images as input, we extend our model by following [26].

4.2.1 Implementation Details

All of our experiments were performed on 8 NVIDIA A100 GPUs. We adopt ViT-B/16 [13] as our vision encoder. The text encoder uses $\text{BERT}_{\text{base}}$ with 123.7M parameters. We set queue size to be $65,536$, codebook size as 4000 and moving average $\alpha = 0.995$. For the pre-training stage, the model is trained for 30 epochs with a batch size of 512. We use mini-batch AdamW optimizer [31] with a weight decay of 0.02. The learning rate is initialized as $1e-5$ and first warmed-up to $1e-4$ after 1,000 iterations. Then it's decreased with a cosine decay strategy to $1e-5$. For data augmentation, we randomly crop each image and resize its size to 256×256, and apply RandAugment [10]. During fine-tuning, the image resolution is increased to 384×384 and the positional encoding is interpolated according to the number of image patches.

4.3. Evaluation on Image-Text Retrieval

For the image-text retrieval tasks, we conduct two different scenarios for evaluation: "zero-shot" retrieval task and "after-finetuning" retrieval task, following the setting in [8, 26, 29]. We compare with both early-fusion methods such as [8, 23, 29] and late-fusion methods such as [22, 39]. ALBEF [26] is an hybrid approach that also performs feature alignment along with fusion. Results in Table 1 and 2 show consistent improvements of our approach against prior state-of-the-arts.

"Zero-shot": As shown from Table 1, CODIS outperforms existing baselines with a clear margin across the two datasets, for both image and text retrieval tasks, especially at R@1. Compared to the best-performing early-fusion approach [8], we obtain a margin of $11.0\%/11.9\%$ TR/IR in terms of R@1 on Flickr30K. When compared to highest late-fusion approach [22], there's a rise of $3.1\%/2.4\%$ TR/IR in R@1 on Flickr30K and $12.5\%/6.5\%$ increase of TR/IR in R@1 on MSCOCO, despite the fact that ALIGN [22] uses 1.8B data in training (approx. 360× more image-text pairs than our model). Our approach also outperforms ALBEF 4M [26] with a clear margin of 1.2%/1.3% in terms of R@1 for TR/IR on Flickr30K and 2.5%/2.0% R@1 for TR/IR on MSCOCO, revealing that our model can further benefit from codebook representation learning.

"After-finetuning": This task showcases the ability of V&L pretraining via transfer learning. For small datasets such as Flickr30K, performance gap tends to reduce as the model converges. However, our approach still achieves the best result in most of the metrics and the largest margins occur for R@1, especially on MSCOCO. Compared against the closest performing method ALBEF [26], CODIS obtains an improvement of $2.2\%/1.9\%$ TR/IR in R@1 on MSCOCO, which is a similar gap as in the zero-shot setting between the two approaches, providing evidence to the effectiveness of CODIS for transfer learning.

Method	Flickr30K (1K)						MSCOCO (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 [37]	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
Unicoder-VL [25]	64.3	85.8	92.3	48.4	76.0	85.2	-	-	-	-	-	-
UNITER [8]	80.7	95.7	98.0	66.2	88.4	92.9	-	-	-	-	-	-
ViLT [23]	73.2	93.6	96.5	55.0	82.5	89.8	56.5	82.6	89.6	40.4	70.0	81.1
CLIP [38]	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
ALIGN [22]	88.6	98.7	**99.7**	75.7	93.8	**96.8**	58.6	83.0	89.7	45.6	69.8	78.6
ALBEF 4M [26]	90.5	98.8	**99.7**	76.8	93.7	96.7	68.6	89.5	94.7	50.1	76.4	84.5
Ours	**91.7**	**99.0**	99.6	**78.1**	**94.1**	96.6	**71.1**	**90.6**	**95.1**	**52.1**	**78.0**	**85.9**

Table 1. Performance comparison of zero-shot image-text retrieval on Flickr30K and COCO datasets.

Method	Flickr30K (1K)						MSCOCO (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 [37]	87.0	97.6	99.2	73.1	92.6	96.0	66.4	89.8	94.4	50.5	78.7	87.1
UNITER [8]	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
VILLA [15]	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
OSCAR [29]	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
ViLT [23]	83.5	96.7	98.6	64.4	88.7	93.8	61.5	86.3	92.7	42.7	72.9	83.1
UNIMO [28]	89.7	98.4	99.1	74.6	93.4	96.0	-	-	-	-	-	-
SOHO [21]	86.5	98.1	99.3	72.5	92.7	96.1	66.4	88.2	93.8	50.6	78.0	86.7
ALBEF 4M [26]	94.3	**99.4**	99.8	82.8	96.7	**98.4**	73.1	91.4	96.0	56.8	81.5	89.2
Ours	**95.1**	**99.4**	**99.9**	**83.3**	96.1	97.8	**75.3**	**92.6**	**96.6**	**58.7**	**82.8**	**89.7**

Table 2. Performance comparison of fine-tuned image-text retrieval on Flickr30K and COCO datasets.

Method	VQA		NLVR^2		SNLI-VE	
	test-dev	test-std	dev	test-P	val	test
VisualBERT [27]	70.80	71.00	67.40	67.00	-	-
VL-BERT [32]	71.16	-	-	-	-	-
LXMERT [44]	72.42	72.54	74.90	74.50	-	-
12-in-1 [33]	73.15	-	-	78.87	-	76.95
UNITER [8]	72.70	72.91	77.18	77.85	78.59	78.28
VL-BART/T5 [9]	-	71.3	-	73.6	-	-
ViLT [23]	70.94	-	75.24	76.21	-	-
OSCAR [29]	73.16	73.44	78.07	78.36	-	-
VILLA [15]	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF 4M [26]	74.54	74.70	80.24	80.50	80.14	80.30
Ours	**74.86**	**74.97**	**80.50**	**80.84**	**80.47**	**80.40**

Table 3. Comparison with variety of state-of-the-art methods on downstream vision-language tasks: VQA, NVLR^2, SNLI-VE.

4.4. Evaluation on VQA, NLVR and VE

Following previous approaches [8, 26], we further report performances of CODIS on various other vision-language tasks such as VQA, NLVR and VE. It's worth noting that some results are not directly comparable as [8] additionally uses out-of-domain data, [29] leverages additional object tags and [15] with adversarial data augmentation. Nevertheless, we observe consistent improvement of our method on all tasks across different datasets in Table 3.

4.5. Ablation Study

In this section, we do ablation studies on the performance of our approach with different variants of CODIS. To get a clear understanding about the effects of each component, we perform comparisons under the zero-shot setting without any finetuning. Note that the setting here for Flickr30K is different than the one in Section 4.3, as the latter reports numbers based on the finetuned model on MSCOCO (5K). Refer to [8] for more details.

Results are summarized in Table 4. By removing the effect of codebook, we provide two baselines that perform alignment at the instance level, namely cross-modal alignment only and intra + cross alignment. The former is an equivalent of ALBEF [26], as both consider only alignment across modalities. The performances consistently increase for all R@1 TR/IR metrics (+0.9%/+1.52% in R@1 for TR/IR on Flickr and +1.26%/+0.42% on in R@1 for TR/IR on MSCOCO) by involving intra-modal alignment, i.e., enhancing unimodal representations.

We observe a consistent improvement over the two baselines when codebook is considered. In this genre, we provide three variants of CODIS designs. The 1st and 3rd row compare the effects of intra-modal alignment whereas the 2nd and 3rd row studies the effects of using student and teacher features for computing the codebook loss. This experiment

Objective functions	Flickr30K (1K)						MSCOCO (5K)					
	Text Retrieval			Image Retrieval			Text Retrieval			Text 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
MLM+ITM+ITC (cross align only)	84.90	97.20	99.00	68.18	88.58	93.02	68.6	89.5	94.70	50.10	76.40	84.50
MLM+ITM+ITC (intra + cross align)	85.80	96.80	98.10	69.70	89.60	93.48	69.86	89.48	94.42	50.52	77.02	85.17
MLM+ITM+ITC (cross align) + codebook (teacher feature)	86.00	97.00	98.20	70.18	90.66	94.44	70.74	89.54	94.88	51.39	77.86	85.60
MLM+ITM+ITC (intra + cross align) + codebook (student feature)	86.30	96.90	98.30	70.34	90.0	93.84	71.12	89.62	94.78	51.40	77.42	85.53
MLM+ITM+ITC (intra + cross align) + codebook (teacher feature)	**86.70**	**97.30**	98.70	**71.40**	**90.82**	**94.62**	**71.10**	**90.60**	**95.10**	**52.10**	**78.00**	**85.90**

Table 4. Performance comparison of zero-shot image-text retrieval on Flickr30K and COCO datasets for ablation study.

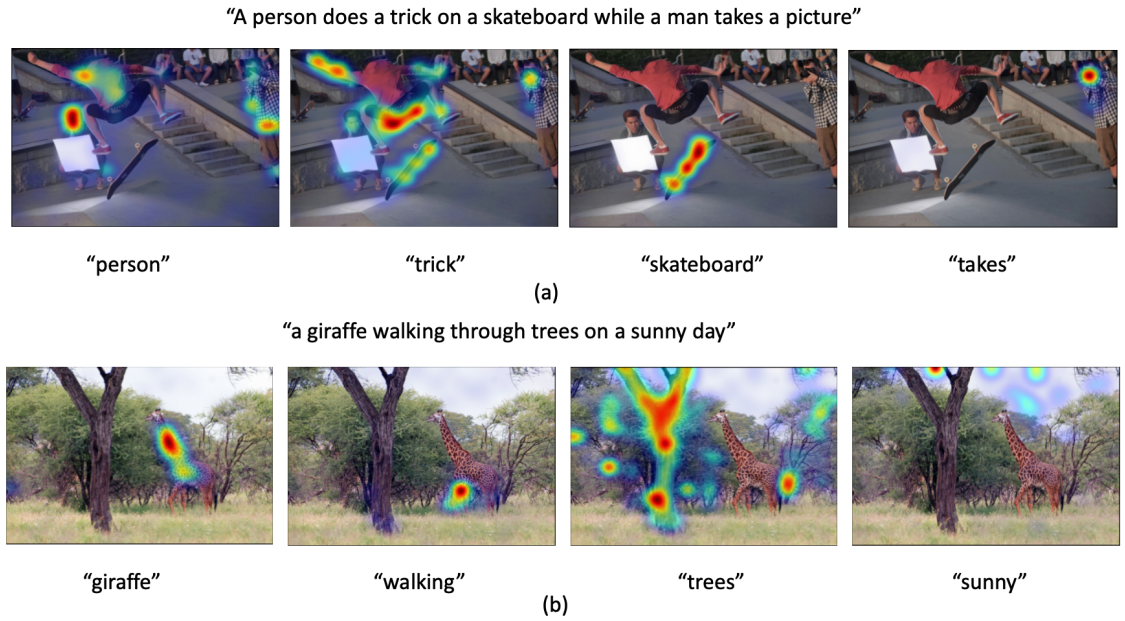

Figure 4. Grad-CAM visualization on the cross-attention maps corresponding to individual words

also serves to support the validity by combining teacher-student distillation with codebook representation learning. Combining the two contributions, CODIS improves the first baseline by a clear margin of $1.8\%/3.22\%$ absolute R@1 for TR/IR on Flickr and $2.5\%/2.0\%$ in R@1 for TR/IR on MSCOCO.

4.6. Cross-attention visualization

We visualize the cross-attention maps using Grad-CAM [40] following [26] to provide qualitative assessment of CODIS. Figure 4 shows that CODIS is able to associate language with "regions of interest" by attending to meaningful objects and locations, visually reflecting the quality of our model in multimodal alignment. For example, in the first row of the figure, the model attends to all men when word "person" is given, while for words such as "tricks" and "takes", the model performs surprisingly well, by "focusing" exclusively on the related persons. In the second example, we choose a scene where multiple correspondences exist (e.g., trees and sunny day). The model seems to allocate more attention to trees closest to the camera and can differentiate trees from grass. It's also interesting to observe that the model switches its "attention" from the upper-body of the giraffe to its feet when the word changes from "giraffe" to "walking", demonstrating the model's capability in understanding the semantic relations between image and text.

5. Conclusion and Future Work

Vision and language pretraining is attracting growing attention of the computer vision community and has exhibited great potential across a diversity of vision-language downstream tasks. One of the keys to the success of V&L is to improve multimodal alignment. In this paper, we propose multimodal alignment using representation codebook, which acts as a medium between the modalities. We also make a connection between self-supervised learning and V&L pretraining, by generalizing teacher-student distillation learning to the multimodal setting under the V&L framework. Our work is a step toward more principled multimodal alignment. We hope to inspire more works in this direction.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008. 1, 2, 3

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 2, 5

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 1, 2, 5

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 2, 4

[6] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020. 3, 5

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1, 2, 6, 7

[9] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. *arXiv preprint arXiv:2102.02779*, 2021. 7

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc 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. 6

[11] Fernando De Goes et al. An optimal transport approach to robust reconstruction and simplification of 2d shapes. In *Computer Graphics Forum*, 2011. 4

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 6

[14] Jiali Duan, Yen-Liang Lin, Son Tran, Larry S Davis, and C-C Jay Kuo. Slade: A self-training framework for distance metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9644–9653, 2021. 2

[15] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020. 7

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 6

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 2

[18] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross 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. 2, 4, 5

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[21] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021. 2, 7

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 1, 2, 6, 7

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021. 1, 2, 6, 7

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David 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. 6

[25] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. 7

[26] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align

before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint arXiv:2107.07651*, 2021. 1, 2, 4, 5, 6, 7, 8

[27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 7

[28] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020. 2, 7

[29] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 2, 6, 7

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 1, 7

[33] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020. 7

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[35] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011. 6

[36] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 6

[37] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020. 7

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 7

[39] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 6

[40] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv 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. 8

[41] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu 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. 6

[42] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[43] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 6

[44] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 7

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[46] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[47] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 6

[48] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 2

[49] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 1