Relaxing Contrastiveness in Multimodal Representation Learning

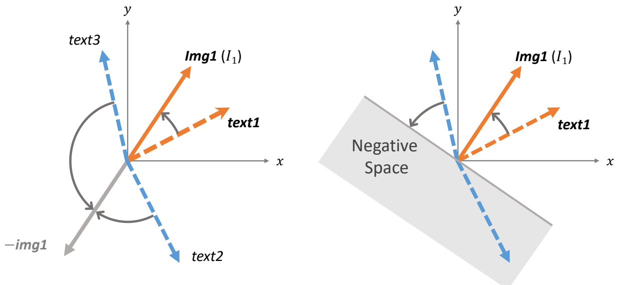
Zudi Lin[1]* Erhan Bas[2]* Kunwar Yashraj Singh[3] Gurumurthy Swaminathan[3] Rahul Bhotika[4]*
[1]Amazon Alexa Science [2]Scale AI [3]AWS AI Labs [4]Optum Labs
{linzud,sinkunwa,gurumurs}@amazon.com

Abstract

Multimodal representation learning for images with paired raw texts can improve the usability and generality of the learned semantic concepts while significantly reducing annotation costs. In this paper, we explore the design space of loss functions in visual-linguistic pretraining frameworks and propose a novel ***Re****laxed* ***Co****ntrastive (****ReCo****) objective, which act as a drop-in replacement of the widely used InfoNCE loss. The key insight of ReCo is to allow a relaxed negative space by not penalizing unpaired multimodal samples (*i.e.*, negative pairs) that are already orthogonal or negatively correlated. Unlike the widely-used InfoNCE, which keeps repelling negative pairs as long as they are not anti-correlated, ReCo by design embraces more diversity and flexibility of the learned embeddings. We conduct extensive experiments using ReCo with state-of-the-art models by pretraining on the MIMIC-CXR dataset that consists of chest radiographs and free-text radiology reports, and evaluating on the CheXpert dataset for multimodal retrieval and disease classification. Our ReCo achieves an absolute improvement of* 2.9% *over the InfoNCE baseline on the CheXpert Retrieval dataset in average retrieval precision and reports better or comparable performance in the linear evaluation and finetuning for classification. We further show that ReCo outperforms InfoNCE on the Flickr30K dataset by* 1.7% *in retrieval Recall@1, demonstrating the generalizability of our approach to natural images.*

1. Introduction

As the most common imaging modality for clinical purposes, chest radiography (chest X-ray) is widely used in the screening and diagnosis of lung and heart abnormalities. However, collecting structured expert annotation from radiologists is expensive and time-consuming [19]. With datasets like MIMIC-CXR [21], which consists of both chest radiographs and corresponding free-text radiology reports, learning generalizable multimodal representations without structured labels becomes a favorable and promising direction [35, 2]. In comparison with recent self-supervised visual representation learning approaches [3, 15, 24, 1, 12, 4, 34] that learns representation from solely unlabeled images, the strength of the multimodal framework is the ability to leverage a broader source of supervision through the semantically denser information in the text.



(a) InfoNCE (b) ReCo (Ours)

Figure 1. Illustration of different contrastive learning losses. We show the optimization goal of **(a)** InfoNCE and **(b)** our proposed *Relaxed Contrastive* (ReCo) loss in a 2D embedding space. Given an image embedding (I_1), InfoNCE forces unpaired text embeddings to be anti-correlated with I_1, while ReCo relaxes the *contrastiveness* by not penalizing text embeddings that are already orthogonal or negatively correlated with I_1.

Visual-linguistic pre-training comes with different forms, including image captioning [8, 30] and learning pretext tasks like reconstructing masked tokens and image regions [23, 31, 5]. Recent works follow the *constrastive* learning methodology, which maps images and texts into a shared embedding space where paired image and text embeddings are attracted while the unpaired are repelled under some similarity measure [35, 28, 20]. The contrastive learning framework is relatively straightforward compared to visual-linguistic models with cross-attention modules [23, 31, 5] and can be scaled up to millions [28] or billions [20] of (image, text) pairs. However, despite the impressive performance achieved by big data, big model, and big compute, most multimodal contrastive learning frameworks by default optimize the InfoNCE [25] objective and

*Work was done when affiliated with the AWS AI Labs.

rarely interrogate the impact of loss functions on the learned multimodal embedding space.

Specifically, given a batch of N (image, text) pairs, an image (called a *query*) only has one paired text treated as the "positive" example, while all other $N-1$ texts are treated as "negatives". The InfoNCE loss maximizes agreement between the query and its positive pair and minimizes agreement with the negatives by applying a multi-class cross-entropy loss[1]. Since most self-supervised learning frameworks (both visual and multimodal) use the *cosine* similarity as the measure of agreement [3, 15, 35, 28], InfoNCE geometrically forces the positive to be aligned with the query while the negatives to be anti-correlated with it (Fig. 1a). However, there is an intrinsically asymmetric distribution: one query only has one positive pair but has $N-1$ negative pairs in the batch. From the perspective of an image, there are $N-1$ unpaired sentences describing distinct semantic information but are all repelled to the reverse direction of the query, which can restrict the diversity and flexibility of the learned representations.

To tackle the challenge, we propose a ***Re****laxed* ***Co****ntrastive* (**ReCo**) loss that follows the contrastive learning scheme but alleviates the *contrastiveness* for negative pairs in the embedding space. The new loss aligns positive pairs by directly maximizing cosine similarity. More importantly, instead of forcing unpaired samples to be anti-correlated with the query as in InfoNCE, ReCo does *not* penalize negatives that are already orthogonal to or negatively correlated with the query (Fig. 1b). ReCo embraces more diversity and flexibility of the embeddings, which is important for medical datasets with semantically complex textual descriptions. Conceptually, the *asymmetric* design for positive and negative space is consistent with the imbalanced distribution of paired and unpaired samples in a batch. We use a scalar to trade-off positive and negative loss terms and remove the temperature parameter for scaling the similarity in InfoNCE. Empirically, we observe that ReCo results in a more right-tailed similarity distribution, improving the distinguishability between positive and negative pairs.

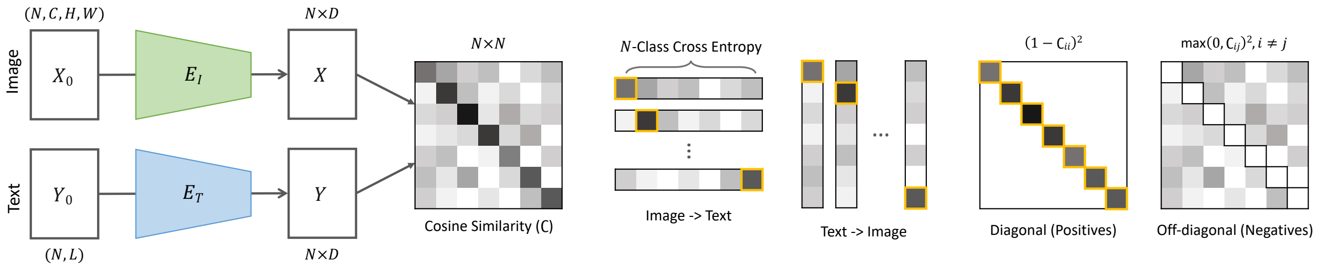
We conduct extensive experiments to demonstrate the effectiveness of ReCo on the recognition tasks of chest radiographs. When trained on the MIMIC-CXR [21] database and evaluated on the CheXpert Retrieval [35] dataset, ReCo significantly improves the average retrieval precision by 2.9% over the InfoNCE baseline with the same architecture and training protocols. The vision encoder optimized with ReCo also achieves better or comparable performance in the linear evaluation and finetuning for disease classification [19]. Furthermore, by finetuning on the strong CLIP [28] model pretrained with 400 million (image, text) pairs, our ReCo loss can still improve the Recall@1 metric by 1.7% on Flickr30K [27] with the same architecture finetuned with InfoNCE loss, demonstrating its generalizability to other domains beyond chest radiographs.

2. Related Work

Unsupervised representation learning. Recent advances in unsupervised visual representation learning focus on learning image representations invariant to transformations [3, 15, 24, 1, 12, 4, 34]. The optimization is done by attracting two augmentations of the same sample and repulsing distinct samples using the *InfoNCE* loss [25] after projecting them via a Siamese encoder [6, 3]. Improvements include using momentum encoder [15] and clustering [24, 1]. Empirical and theoretical results suggest that the InfoNCE loss asymptotically optimizes the alignment and uniformity metrics [33]. Recent results also show that the contrasting negative pairs can be removed [12, 4] or the loss can be changed to a redundancy reduction metric [34]. Practitioners can then optimize a linear classifier on top of the pre-trained backbone for classification or use it for other transfer learning tasks. In the language domain, pre-training includes learning pretext tasks like masked token and next sentence prediction [9], as well as generative pre-training [29]. The contrastive learning idea is also applied for sentence embeddings [11]. In this work, we focus on the joint learning of visual and language representations with primary application on medical data. Unlike unimodal pre-training, multimodal frameworks do not use a Siamese architecture as the inputs come from two different modalities. In addition, compared to unimodal architectures where the projection layers are removed in downstream tasks, the projection heads in multimodal architectures are usually preserved for applications like text-to-image retrieval [35, 28, 20].

Vision-language pretraining. Joint representation learning in the multimodal domain appears in different forms. The first line of works optimize a vision encoder and a language decoder for the image captioning task and transfer the learned visual representations to downstream applications [8, 30]. The second line of literature jointly learns multimodal pretext tasks like reconstructing masked image regions and language tokens, as well as directly predicting the alignment between image and text [23, 31, 5]. However, the cross-modal attention modules that emerge in those methods make them less efficient in practical retrieval systems. The third stream, which is closer to the *contrastive* methodology in visual representation learning, uses a dual-encoder architecture to directly map image and text data into a shared embedding space, where the agreement between paired samples is maximized while the agreement between the unpaired samples is minimized [35, 28, 20]. The independent encoders improve the flexibility in downstream

[1] InfoNCE is symmetrically applied for aligning image to text. Without loss of generality, we only discuss the text-to-image part here.



(a) Multimodal Learning Architecture (b) InfoNCE (c) **ReCo** (Ours)

Figure 2. Multimodal learning framework and losses. **(a)** Image and text encoders (E_I and E_T) map multimodal inputs into a shared embedding space and calculate the cosine similarity matrix. **(b)** InfoNCE [25] splits the rows and columns and apply a N-class cross-entropy loss for each vector. **(c)** Our ReCo splits the diagonal and off-diagonal elements and applies L_2 losses to make the paired close and unpaired orthogonal or negatively correlated. Yellow frames (□) denote the similarity between positive pairs.

recognition tasks. However, recent works focus on the data scale and model architectures but pay less attention to the energy function that shapes the embedding space. Our contribution is a new loss, *ReCo*, that relaxes the repelling force between negative pairs in pre-training, which improves the state-of-the-art multimodal contrastive learning frameworks like ConVIRT [35] and CLIP [28] without changing architectures and training data.

3. Relaxation of Contrastiveness

Problem setup. The goal is to learn meaningful representations using paired image and text. As shown in Fig. 2a, the image and text encoders E_I and E_T project a batch of inputs from different modalities into a shared embedding space. Both encoders consist of a backbone model and a projection head. In recent works, the text backbone is usually a transformer [32], and the image backbone can be either a CNN or a transformer [35, 28, 20]. The projection head can be an MLP with non-linearity [3, 35], or just a linear layer [28]. We show results on both types of projection heads in the experiments.

Suppose the batch size is N and the embedding dimension is D, E_I generates a $D \times N$ image embedding matrix $\mathcal{U} = [u_1, u_2, \ldots, u_N]$ where u_i is a D-dimensional vector, while the text encoder E_T generates a text embedding matrix $\mathcal{V} = [v_1, v_2, \ldots, v_N]$. For clarity, we call (u_i, v_i) a *positive* pair and $(u_i, v_j), i \neq j$ a *negative* pair. Then, *cosine similarity* is commonly used to measure the agreement between image and text in the embedding space. Specifically, the $N \times N$ cosine similarity matrix $\mathcal{C}$ is defined as:

$$\mathcal{C}_{ij} = \frac{\langle u_i, v_j \rangle}{\|u_i\| \|v_j\|} = \frac{u_i^\intercal v_j}{\|u_i\| \|v_j\|} \tag{1}$$

The range of $\mathcal{C}_{ij}$ is $[-1, 1]$, where 1 means two vectors are aligned while -1 means reverse-aligned, regardless of the magnitudes. Since u_i and v_i are usually ℓ_2 normalized, the similarity matrix can be calculated with $\mathcal{C} = \mathcal{U}^\intercal \mathcal{V}$.

InfoNCE loss. Recent multimodal pre-training approaches [28, 35, 20] mainly use the InfoNCE [25] loss (NCE stands for Noise-Contrastive Estimation [13]), which is also widely used in unimodal contrastive learning frameworks [3, 15]. With the cosine similarity matrix $\mathcal{C}$, InfoNCE for image-to-text alignment is:

$$\mathcal{L}_{\mathcal{NCE}} \triangleq -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathcal{C}_{ii}/\tau)}{\sum_{j=1}^{N} \exp(\mathcal{C}_{ij}/\tau)} \tag{2}$$

where $\tau > 0$ is a temperature parameter to scale the cosine similarity values, which can be a hyper-parameter [35] or a learnable part of the model [28]. Intuitively this loss can be regarded as a N-way classifier loss which maximizes cosine similarity between u_i and its true pair v_i and minimizes cosine similarity for $v_{j,j \neq i}$ (Fig. 2b). Symmetrically, the loss is also applied to $\mathcal{C}^\intercal = \mathcal{V}^\intercal \mathcal{U}$ for matching a text embedding to the corresponding image embedding. Both loss terms are added as the final energy function to be optimized.

We show an intuitive geometric interpretation of InfoNCE in Fig. 1a. For a query u_i, making the positive pair (u_i, v_i) aligned while making negative pairs $(u_i, v_j), i \neq j$ anti-correlated minimizes the loss for that query (Eqn. 2), which pushes $v_j, i \neq j$ to the reverse direction of u_i. However, there is an intrinsic *asymmetry* in the visual-linguistic contrastive learning framework: for one query in a batch of N pairs, there is only one positive pair but $N - 1$ negative pairs. Even for a relatively small batch size of 64, the ratio of positive pairs is less than 2%. However, InfoNCE keeps repelling the negatives to the reverse direction of the query. Although in practice InfoNCE will not push the cosine similarity to be the global maximum and minimum and make pairs strictly aligned and anti-aligned (using temperature $\tau < 1$ actually make the distribution more concentrated [17]), we believe the *contrastiveness* can be relaxed to improve the flexibility of the learned embeddings.

ReCo loss. To alleviate the contrastiveness in InfoNCE, we propose a novel ***relaxed contrastive*** (ReCo) loss (denoted as $\mathcal{L}_{\mathcal{RC}}$) that relaxes the negative space of any given query (Fig. 1b). Specifically, ReCo considers the diagonal and off-

Algorithm 1 PyTorch-style pseudocode of ReCo

```
# f, g: image and text encoder networks
# N, D: batch size and embedding dimension
#
# diagonal: diagonal elements of a matrix
# off_diagonal: off-diagonal elements of a matrix
# lambda: weight on the negative pairs

for (x0, y0) in loader: # load a batch with N pairs
    # compute embeddings for two modalities
    x = f(x0) # NxD image embeddings
    y = g(y0) # NxD text embeddings

    # l2 normalize along the feature dimension
    x_norm = x / x.norm(1, keepdim=True) # NxD
    y_norm = y / y.norm(1, keepdim=True) # NxD

    # cosine similarity matrix
    c = x_norm @ y_norm.T # NxN

    # relexed contrastive loss
    l_pos = diagonal(c).add_(-1).pow_(2).sum()
    l_neg = max(off_diagonal(c), 0).pow_(2).sum()
    loss = l_pos + lambda * l_neg

    # optimization step
    loss.backward()
    optimizer.step()
```

diagonal parts of the cosine similarity matrix $\mathcal{C}$ separately (Fig. 2c), which corresponds to the similarities of positive and negative pairs in a batch:

$$\mathcal{L}_{\mathcal{RC}} \triangleq \underbrace{\sum_i (1-\mathcal{C}_{ii})^2}_{\text{positive term}} + \lambda \underbrace{\sum_i \sum_{j\neq i} \mathbf{max}(0,\mathcal{C}_{ij})^2}_{\text{negative term}} \qquad (3)$$

where λ is a positive constant balancing the importance of positive and negative terms. The key of our proposed ReCo is the $\mathbf{max}$ operator, which means the energy function does *not* penalize negative pairs that are already orthogonal or negatively correlated. We can also interpret ReCo as a loss that adaptively puts attention to all positive pairs and only *challenging* negative pairs. The temperature parameter (τ) for scaling the similarity score in InfoNCE is no longer required in ReCo. Using popular deep-learning frameworks like PyTorch [26], ReCo can be implemented with few lines of code (Alg. 1) and easily incorporated into the standard multimodal contrastive training framework.

Connection and comparison to other losses. In unimodal representation learning, Wang and Isola [33] shows that InfoNCE (with a increasing batch size) asymptotically optimizes $\mathcal{L} = \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}$, where $\mathcal{L}_{\text{align}}$ is the distance between positive pairs and $\mathcal{L}_{\text{uniform}}$ is the uniformity of the feature distribution. The *alignment* term ($\mathcal{L}_{\text{align}}$) is similar to the positive term in ReCo, as maximizing the cosine similarity is equivalent to minimizing the mean squared error of ℓ_2 normalized vectors, up to a scale of 2. The *uniformity* term is defined as:

$$\mathcal{L}_{\text{uniform}} \triangleq \log \mathbb{E}\left[e^{-t\|u-v\|_2^2}\right],\ t>0 \qquad (4)$$

which pushes all embeddings away from each other to make them roughly uniformly distributed on the (D-1)-dimensional unit hypersphere with sufficient samples. This interpretation is consistent with empirical observations that larger numbers of negative pairs lead to better visual representation learning results [15, 3]. However, both ConVIRT [35] and our studies show that increasing negative pairs does *not* improve learned visual-linguistic representations, demonstrating the unique challenges in the multimodal domain. The negative term in ReCo (Eqn. 3) still pushes unpaired embeddings away but establishes a threshold so that when a negative pair is already far enough (cosine similarity ≤ 0), it alleviates the contrastiveness and no longer consider it in the loss calculation.

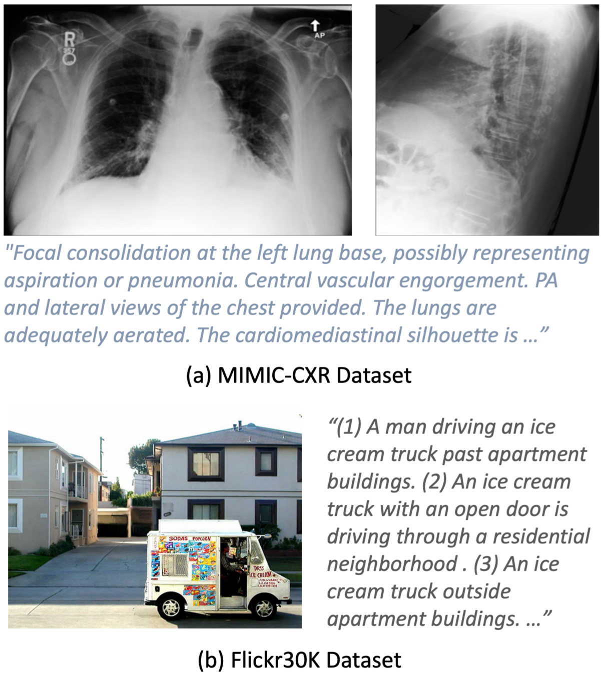
We also discuss differences of ReCo with *Barlow Twins* (BT) [34]. Although the formulations appear similar, the key difference is that BT operates on the decorrelation of feature dimensions, while ReCo (and InfoNCE) operate on the alignment of embeddings. Specifically, given two $D\times N$ embedding matrix $\mathcal{U}$ and $\mathcal{V}$, InfoNCE and ReCo are applied to the embedding similarity matrix $\mathcal{C}_{N\times N} = \mathcal{U}^\intercal\mathcal{V}$, while BT is instead applied to the *cross-correlation* matrix:

$$\mathcal{C}'_{D\times D} = \mathcal{U}\mathcal{V}^\intercal \qquad (5)$$

where $\mathcal{U}$ and $\mathcal{V}$ are ℓ_2 normalized along the *batch* dimension instead of the feature dimension as Eqn. 1. ReCo and BT have this essential dissimilarity because, in visual representation learning, researchers usually discard the embedding space and only keep the backbone for downstream tasks. While in a visual-linguistic framework, we keep embedding space for cross-modal applications like text-image retrieval. Besides, when considering adapting BT to the cosine similarity matrix for multimodal alignment, we see that the minimization of $\max(0, \mathcal{C}_{ij})^2$ in our ReCo is to make negative pairs orthogonal or negatively correlated ($\mathcal{C}_{ij} \leq 0$), instead of strictly orthogonal as in BT ($\mathcal{C}_{ij} = 0$). In experiments, we will show that using $\mathbf{max}$ is not a random choice but necessary for robustly improving the multimodal representation learning performance on different datasets.

The formulation of ReCo also appears similar to the max-margin contrastive loss [14] considering the conversion between ℓ_2 distance and cosine similarity. Our contribution is to adapt such an objective in the multimodal setting without siamese encoders [6] and significantly improve the widely-used InfoNCE. In comparison with VSE++ [10] that explicitly compare the positive with the hardest negative in a batch, ReCo reduces the gradient contribution from easy negatives, which implicitly emphasize hard negatives.

In summary, since the multimodal embeddings reside on a unit hypersphere, ReCo relaxes the negative space of a given query to a hyper-hemisphere to embrace more flexibility and diversity. We also tested a generalized cosine similarity using wedge product for additional flexibility, which



"Focal consolidation at the left lung base, possibly representing aspiration or pneumonia. Central vascular engorgement. PA and lateral views of the chest provided. The lungs are adequately aerated. The cardiomediastinal silhouette is …"

(a) MIMIC-CXR Dataset

"(1) A man driving an ice cream truck past apartment buildings. (2) An ice cream truck with an open door is driving through a residential neighborhood . (3) An ice cream truck outside apartment buildings. …"

(b) Flickr30K Dataset

Figure 3. Examples from tested multimodal datasets. **(a)** Our main focus is the representation learning on the MIMIC-CXR [21] database organized by *studies*, where each study consists of one or more chest X-ray images with a radiology report in free-text form. **(b)** To show the generalizability of our approach, we also experiment with Flickr30K [27] where each image is associated with five semantically similar captions.

is discussed in the supplementary material.

4. Experiments

After pre-training on the MIMIC-CXR dataset of chest radiographs and radiology reports (Fig. 3a), we show the effectiveness of ReCo in retrieval (Sec. 4.1) and perform ablation studies to understand the impact of hyper-parameters (Sec. 4.2). We also test the tranfer learning performance for disease classification through linear evaluation and fine-tuning (Sec. 4.3). We further apply ReCo to Flickr30K (Fig. 3b) to demonstrate its generalization. (Sec. 4.4).

4.1. Text and Image Retrieval

Dataset and evaluation metrics. We use the MIMIC-CXR [21] database for training, which is a collection of chest radiograph images paired with their textual reports. This dataset contains a total of about 217k image-text pairs (organized by studies), with each pair containing an average of 1.7 images and 6.0 sentences. We show one example in Fig. 3a. We randomly sample one image and one sentence from each study to construct a positive pair during training.

Instead of directly splitting the MIMIC-CXR database for evaluation, we follow ConVIRT [35] to use the *CheXpert 8×200 Retrieval* dataset [35] for performance comparison by reporting the validation performance. For simplicity, we denote it as the CheXpert retrieval dataset in the following text. Each image or sentence in this dataset is associated with 1 of 8 category labels provided by a board-certified radiologist. CheXpert retrieval dataset has 40 query sentences (5 for each category) and 80 query images (10 for each category), and 1,600 candidate images (200 for each category). Therefore we can evaluate the performance on both text-image and image-image retrieval.

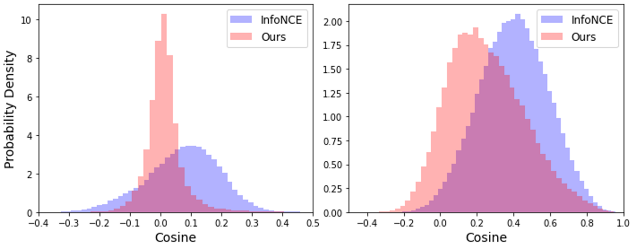
For the retrieval performance, we use Prec@k with $k \in \{5, 10, 50\}$. For a given query (can be a sentence or an image), we rank the similarity score of the query with all candidate images. Then for the k candidates with the highest scores, the precision is $\frac{n}{k}$ where n is the number of ground-truth matches. We average it over all queries to get the Prec@k score. We also use the average of image-image and text-image scores to measure the overall performance.

Implementation details. We follow the ConVIRT [35] implementation. The vision encoder is a ResNet-50 [16] model pretrained on ImageNet [7]. We repeat the gray-scale radiograph images along the channel dimension to make them compatible with the vision encoder. The language encoder is a pretrained ClinicalBERT [18] with 12 transformer layers. The token embedding layer and the first 6 transformer layers are frozen during training. Each encoder is paired with a projection head, a two-layer MLP with ReLU non-linearity for the hidden layer. The embedding dimension is 512. The two encoders are jointly trained with the Adam [22] optimizer using an initial learning rate of 10^{-4}, weight decay of 10^{-6}, as well as cosine learning rate scheduler. We added a small denominator $\epsilon = 10^{-7}$ when calculating the cosine similarity to avoid dividing by 0. We use a batch size of 64. We set negative weight $\lambda = 0.6$ (Eqn. 3) for ReCo. Training is done on a single NVIDIA V100 GPU (16GB) within two days for 300K iterations. Hyper-parameter choices are justified in ablation studies.

Results. We show a quantitative comparison in Table 1. Random initialization gives an average precision of 12.5% for the 8 categories. ConVIRT [35] is optimized with InfoNCE (Eqn. 2) and currently has the state-of-the-art performance on this dataset. Since the code of ConVIRT is not publicly released, we replicate their settings with some help from the ConVIRT authors and denote it as *InfoNCE* in our experiments. The results demonstrate that the contrastive methodology that directly aligns multimodal embeddings significantly outperforms other learning schemes (please refer to ConVIRT [35] for details). We then show that by changing the InfoNCE energy function to our ReCo without any modification to model architecture and training protocols, both image-image and text-image retrieval scores are improved consistently by a large margin. Specifically, upon the InfoNCE baseline with an average precision of 48.6%,

Table 1. Performance comparison on retrieval. We show the unimodal and multimodal retrieval results of InfoNCE and ReCo models, averaged over four runs with different random seeds. With the same architecture and training protocol, our ReCo improves InfoNCE by **2.9**% in average retrieval precision on the CheXpert 8×200 Retrieval dataset [35] (models are trained on MIMIC-CXR [21]).

	Image-Image Retrieval (↑)			Text-Image Retrieval (↑)			Average (↑)
Method	Prec@5	Prec@10	Prec@50	Prec@5	Prec@10	Prec@50	
Random	12.5	12.5	12.5	12.5	12.5	12.5	12.5
ImageNet	14.8	14.4	15.0	–	–	–	–
Results reported in Zhang et al. [35]							
Caption-Transformer	29.8	28.0	23.0	–	–	–	–
Caption-LSTM	34.8	32.9	28.1	–	–	–	–
Contrastive-Binary	38.8	36.6	29.7	15.5	14.5	13.7	24.8
ConVIRT	45.0	42.9	35.7	60.0	57.5	48.8	48.3
Our experiments (with standard error)							
InfoNCE	43.3$_{\pm 0.5}$	40.2$_{\pm 0.7}$	35.0$_{\pm 0.2}$	63.7$_{\pm 1.4}$	59.2$_{\pm 1.2}$	50.1$_{\pm 0.9}$	48.6$_{\pm 0.6}$
ReCo (**Ours**)	**45.6**$_{\pm 0.7}$	**44.1**$_{\pm 0.9}$	**35.7**$_{\pm 0.6}$	**67.4**$_{\pm 1.9}$	**62.8**$_{\pm 1.0}$	**53.1**$_{\pm 0.5}$	**51.5**$_{\pm 0.6}$
	+2.3	+3.9	+0.7	+3.7	+3.6	+3.0	+2.9



(a) Text-Image Similarity

(b) Image-Image Similarity

Figure 4. Similarity distributions of learned embeddings on the Chexpert Retrieval dataset. We show both the **(a)** text-image and **(b)** image-image cosince similarity histograms for models trained with InfoNCE and our ReCo loss.

ReCo improves the performance to 51.5%, an absolute improvement of **2.9**%.

To understand the differences in the learned embedding space, we show the cosine similarity distributions for both losses in Fig. 4. Qualitatively, for the text-image similarity, which is directly optimized by the losses, ReCo pushes more pairs to the orthogonal and negative subspace and has a more right-tailed distribution than InfoNCE (Fig. 4a). This distribution is consistent with our motivation that the loss should account for the *imbalance* of positive and negative pairs. We also notice that although the losses are not directly applied for the image-image alignment task, ReCo also yields a more right-tailed distribution for image-image similarity than InfoNCE (Fig. 4b).

4.2. Ablation Studies

Batch size. Different from visual contrastive learning frameworks where larger batch sizes generally leads to better performance [15, 3] (*e.g.*, $B = 65536$ in He *et al.* [15]), ConVIRT tested $B \in \{16, 32, 128\}$ on MIMIC-CXR and show that increasing batch size decreases the multimodal retrieval performance [35]. We conduct a similar study for $B \in \{32, 64, 96\}$ using the InfoNCE loss and show that $B = 64$ achieves the best average retrieval precision of 48.6% (Fig. 5a). We thus set $B = 64$ as default for the following experiments for both InfoNCE and our ReCo as it denotes a strong InfoNCE baseline.

Embedding dimension. We study the influence of embedding dimensions on the InfoNCE model. Please note that the vision encoder is a ResNet-50 [16] that has 2048 dimensions after global average pooling, while the language encoder is a ClinicalBERT [18] whose dimension is 768 for each output token. We show that for a wide range of $D \in [512, 1536]$, the performance is quite stable with a small gap of 0.3% in average retrieval precision between the best and worst model. We also noticed that when D is relatively small (256) or large (1792 and 2048), there is an obvious performance drop. We argue these might be underfitting and overfitting problems, respectively. Following ConVIRT [35], we fix the embedding dimension to 512 for the following experiments.

Off-diagonal weight λ. Our proposed ReCo loss (Eqn. 3) removes the temperature parameter in the InfoNCE loss and uses a scalar weight λ to trade-off the contribution of diagonal (positive) and off-diagonal (negative) terms. In this study, we solely benchmark the influence of λ and keep other hyper-parameters identical to the baseline InfoNCE configurations described above ($B = 64$ and $D = 512$). We tested a wide range of $\lambda \in [0.1, 0.8]$ with a 0.1 interval. Besides achieving an average retrieval precision of 51.5% when $\lambda = 0.6$, our ReCo loss significantly outperforms the InfoNCE baseline with all off-diagonal weights we tested

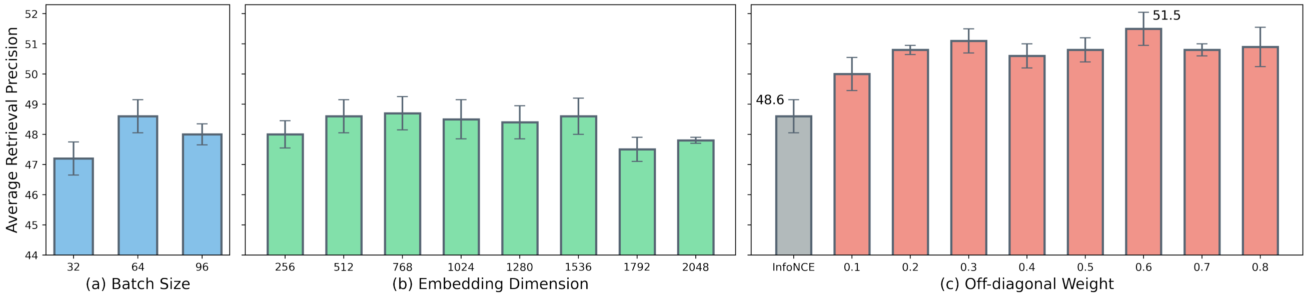


Figure 5. Ablation studies on model hyper-parameters. We show that the baseline InfoNCE model **(a)** performs best using a training batch size of 64 with an embedding dimension of 512, and **(b)** has stable performance for embedding dimensions from 512 to 1536. Therefore we fix those two hyper-parameters ($B = 64$ and $D = 512$) for other experiments. **(c)** Models trained with our proposed ReCo loss consistently outperform the InfoNCE baseline for a wide range of off-diagonal weights $\lambda \in [0.1, 0.8]$.

(Fig. 5c). The minimal average precision of ReCo is 50.0% when $\lambda = 0.1$, which is still 1.4% higher than the InfoNCE baseline at 48.6%. The observations demonstrate the robustness of the proposed ReCo loss.

4.3. Linear Evaluation and Finetuning

Setup. Following previous work in self-supervised visual [3, 15] and multimodal [35] representation learning, we use both linear evaluation protocol and full finetuning to evaluate the learned visual encoders. Specifically, we keep the backbone of the visual encoder (remove the projection MLP) and add a single fully-connected layer for classification. In linear evaluation, only the fully-connected layer is learnable while the rest are frozen (batch normalization layers are in inference mode to use previous running statistics). The whole model is learnable in finetuning. Following ConVIRT [35], we conduct experiments on CheXpert [19], which is a *multi-label* classification task as one image can belong to more than one class (*i.e.*, more than one disease is observed from the radiology image). We use the binary cross-entropy loss averaged over classes for training. From the models trained with different random seeds in Table 1, we choose the ones whose retrieval precision is closest to the mean retrieval score for this experiment. Random image augmentations are applied for both scenarios. We set the learning rate to 0.01 for linear evaluation and 10^{-4} for finetuning without tweaking.

Table 2. Linear evaluation and finetuning results. We show the AUC scores for *multi-label* classification on the CheXpert [19] dataset with pretrained image encoders from multimodal learning. In linear evaluation, only a fully-connected layer is optimized.

	Linear Evaluation			Finetuning		
Method	1%	10%	all	1%	10%	all
Results reported in Zhang et al. [35]						
Random Init.	58.2	63.7	66.2	70.4	81.1	85.8
ImageNet Init.	75.7	79.7	81.0	80.1	84.8	87.6
ConVIRT	85.9	86.8	87.3	87.0	88.1	88.1
Our experiments						
ImageNet Init.	69.90	74.13	77.43	74.90	82.98	87.37
InfoNCE	86.65	88.25	88.38	**87.38**	88.30	88.55
ReCo (**Ours**)	**86.90**	**88.28**	**88.57**	87.07	**88.33**	**88.58**

Results. We have several observations in Table 2. First, contrastive learning on radiology data (both InfoNCE and ReCo) significantly outperforms the model pretrained on ImageNet [7]. This is expected and indicates that when there is a large domain gap, *in-domain* pre-training is required for satisfactory downstream performance. Second, ReCo outperforms the InfoNCE baseline in the linear evaluation and has comparable results in finetuning, indicating its effectiveness in learning not only better *joint* embeddings but also meaningful unimodal representations. Third, for the encoder trained with ReCo, the linear protocol achieves an average AUC of 88.57%, which almost reaches the performance of full-finetuning of 88.58%. Considering that optimizing a linear classifier is several magnitudes faster than full finetuning, we suggest practitioners use this configuration in real-world applications with a budget.

4.4. Flickr30K Experiments

Setup. To demonstrate generalization of the proposed ReCo loss, we also conduct experiments on the Flickr30K [27] dataset, which contains about 32,000 natural-scene images, each paired with 5 captions describing the image. We take 1,000 images (with 5,000 captions) as the validation set and use the rest for training. The evaluation metric is Recall@k, $k \in \{1, 5, 10\}$, which corresponds to whether at least one ground truth is included in the top k retrievals from the validation set. We report both image and text retrieval results in this experiment.

For the model, we use a pretrained CLIP [28] architecture called "RN50", which is a hybrid architecture that uses a customized ResNet-50 as image encoder and a 12-

Table 3. Multimodal retrieval performance on the Flickr30K. CLIP-ZS means the zero-shot results of the pretrained CLIP [28] model. Our proposed ReCo loss consistently outperforms InfoNCE even when finetuning on the CLIP pretrained model.

	Image Retrieval			Text Retrieval		
Method	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ZS [28]	55.0	80.5	86.9	76.0	93.4	97.0
InfoNCE	70.3	91.3	95.3	87.4	97.7	98.9
ReCo (**Ours**)	**71.9**	**91.9**	**95.7**	**89.2**	**98.4**	**99.1**

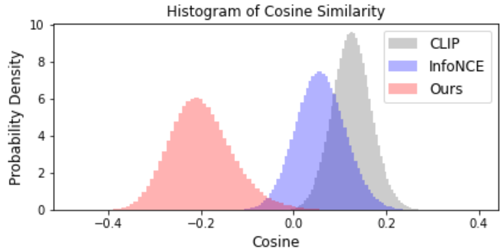


Figure 6. Similarity distributions of learned embeddings on Flickr30K. We show the image-text cosine similarity histograms for pretrained CLIP [28] as well as the models finetuned with InfoNCE and our ReCo loss on the Flickr30K validation set.

layer transformer with 8 heads for each block as the text encoder. This multimodal architecture is similar to what we have tested for the MIMIC-CXR dataset (Sec. 4.1). However, one major difference is that the projection heads for MIMIC experiments are two-layer MLPs with ReLU nonlinearity, while the projection heads are linear layers for the CLIP architecture. The model was pretrained on a huge dataset with 400 million (image, text) pairs. We finetune it on the Flickr30K training set with a mini-batch size of 64, an initial learning rate of 10^{-6} and weight decay of 10^{-2} for a total of 80K iterations.

Results. Table 3 shows the image-text and text-image retrieval scores on the Flickr30K validation set. Starting from a strong baseline CLIP-ZS (ZS stands for zero-shot), finetuning with InfoNCE loss improves the top-1 recall (R@1) from 55.0 to 70.3 for image retrieval and 76.0 to 87.4 for text retrieval. With exactly the same training protocol, changing InfoNCE to ReCo further improves the R@1 scores to 71.9 and 89.2, achieving absolute improvements of **1.6**% and **1.8**% for image and text retrieval, respectively. The results demonstrate that ReCo can robustly improve the multimodal representation learning performance with different types of datasets (Fig. 3) and model architectures.

To understand how ReCo changes the structure of the embedding space, we visualize the cosine similarity distributions between images and captions with different models in Fig. 6. Different from the observation in radiology images (Fig. 4a), ReCo significantly increases the smoothness of the similarity distribution and pushes it more to the negative range instead of stopping when achieving orthogonality, showing the unique characteristics of different datasets. The ReCo similarity distribution is still relatively right-tailed compared to InfoNCE and CLIP-ZS but less evident than the CheXpert retrieval dataset. Next, we discuss another experiment exhibiting dataset differences.

Orthogonality constraint. Inspired by Barlow Twins [34], we also tested a loss that enforces *orthogonality* for negative pairs by removing the max operation in the proposed ReCo objective (Eqn. 3):

$$\mathcal{L}_{OC} \triangleq \sum_i (1-\mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j\neq i} {\mathcal{C}_{ij}}^2 \qquad (6)$$

where $\lambda > 0$ is the weight for negative terms. We follow the training and evaluation protocol on the MIMIC-CXR [21] dataset and show that $\mathcal{L}_{OC}$ achieves an average retrieval precision of 51.3%, slightly lower than ReCo but still outperforming the InfoNCE baseline on the CheXpert Retrieval dataset. However, when applying $\mathcal{L}_{OC}$ to Flickr30K [27], the model overfits quickly and results in recall scores lower than even the zero-shot results of pretrained CLIP [28]. The multimodal similarity distributions suggest that ReCo pushes more pairs to be orthogonal (Fig. 4a) on the CheXpert Retrieval dataset. Therefore, removing the max function has a small influence on the results. However, most multimodal pairs in Flickr30K are far from orthogonality after applying ReCo (Fig. 6). Therefore the orthogonality constraint disrupts the embedding space and results in performance degradation. Those observations further demonstrate the criticality of the max operator in the generalization of ReCo to other datasets.

5. Conclusion and Future Work

In this work, we improve the InfoNCE loss in multimodal learning by introducing a novel loss function, *ReCo*, which alleviates the contrastiveness for negative pairs and achieves better retrieval and transfer learning performance with different architectures on different datasets. Compared with visual representation learning frameworks [3, 15, 24, 1, 12, 4, 34], visual-linguistic models can undoubtedly enjoy performance gain with the semantically denser natural language supervision [35, 28, 20]. However, the limitation is that multimodal systems usually require the pairing of samples in different modalities, while unimodal architectures can work with raw data in a single modality. One natural extension of our work is to apply ReCo to other modalities beyond image and text, including but not limited to video, audio, and genomics. In addition, recent works indicate that the contrasting negative pairs are not necessary for learning meaningful representations in vision [12, 4]. We expect that exploring similar designs for the multimodal setting will also lead to interesting discoveries.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[2] Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–539. Springer, 2020.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

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

[5] 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.

[6] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li 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] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021.

[9] 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.

[10] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

[12] 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. In *NeurIPS*, 2020.

[13] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[14] Raia Hadsell et al. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[18] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[20] 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.

[21] Alistair E. W. Johnson, T. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, M. Lungren, Chih ying Deng, R. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 2019.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] 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.

[24] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. *arXiv preprint arXiv:1912.01991*, 2019.

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

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[27] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74–93, 2015.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[29] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018.

[30] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 153–170. Springer, 2020.

[31] 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.

[32] 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.

[33] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[34] J. Zbontar, L. Jing, Ishan Misra, Y. LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

[35] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text, 2020.