RA-CLIP: Retrieval Augmented Contrastive Language-Image Pre-training

Chen-Wei Xie*, Siyang Sun*, Xiong Xiong*, Yun Zheng, Deli Zhao, Jingren Zhou
Alibaba Group
{eniac.xcw,siyang.ssy,moxiong.xx,zhengyun.zy}@alibaba-inc.com
zhaodeli@gmail.com, jingren.zhou@alibaba-inc.com

Abstract

Contrastive Language-Image Pre-training (CLIP) is attracting increasing attention for its impressive zero-shot recognition performance on different down-stream tasks. However, training CLIP is data-hungry and requires lots of image-text pairs to memorize various semantic concepts. In this paper, we propose a novel and efficient framework: Retrieval Augmented Contrastive Language-Image Pre-training (RA-CLIP) to augment embeddings by online retrieval. Specifically, we sample part of image-text data as a hold-out reference set. Given an input image, relevant image-text pairs are retrieved from the reference set to enrich the representation of input image. This process can be considered as an open-book exam: with the reference set as a cheat sheet, the proposed method doesn't need to memorize all visual concepts in the training data. It explores how to recognize visual concepts by exploiting correspondence between images and texts in the cheat sheet. The proposed RA-CLIP implements this idea and comprehensive experiments are conducted to show how RA-CLIP works. Performances on 10 image classification datasets and 2 object detection datasets show that RA-CLIP outperforms vanilla CLIP baseline by a large margin on zero-shot image classification task (+12.7%), linear probe image classification task (+6.9%) and zero-shot ROI classification task (+2.8%).

1. Introduction

Traditional visual representation learning systems are trained to predict a fixed set of predetermined image categories [12, 16, 22, 34]. This limits their transferability since additional labeled training data are required to recognize new visual concepts. Recently, vision-language pre-training approaches such as CLIP [29] emerge as a promising alternative which introduces text description as supervision. CLIP aligns image modality and text modality by learning a modality-shared representation. During pre-training, CLIP learns to pull matched image-text pairs together and push non-matched pairs apart. After pre-training, CLIP can be transferred to zero-shot image classification task: categories can be referred by textual descriptions, and the image classification task can be converted to image-to-text retrieval task. Experimental results show that CLIP performs well on zero-shot image classification task, e.g., for ImageNet zero-shot classification task, CLIP can match the accuracy of ImageNet pre-trained ResNet50, even that CLIP doesn't use any of the 1.28 million training examples of ImageNet for training.

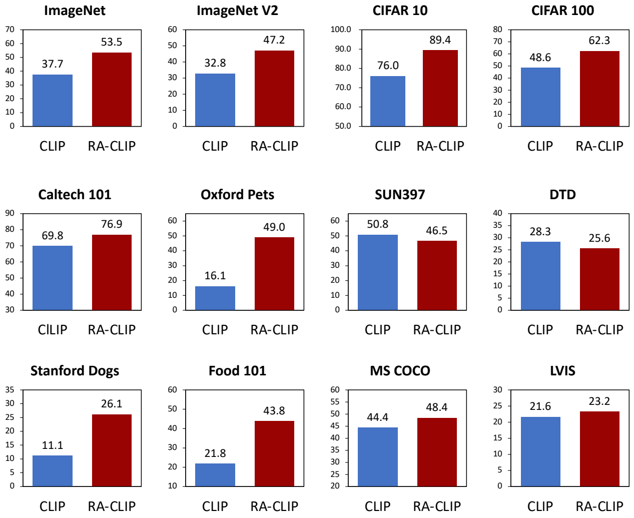

Figure 1. Transferring the CLIP and RA-CLIP to 12 down-stream visual recognition datasets for zero-shot evaluation. Our RA-CLIP achieves better results in 10 out of 12 datasets, and brings +12.7% averaged improvements on the 10 image classification datasets and 2.8% averaged improvements on the 2 object detection datasets..

Despite the impressive zero-shot performance, CLIP requires lots of image-text pairs to train encoders and memorize various semantic concepts, which limits its applications since it is not affordable for most laboratories and companies. Recent works [24, 25] try to alleviate this limitation by taking full advantage of existing data and

[*] indicates equal contribution.

train encoders to memorize concepts as many as possible, e.g., DeCLIP [24] explores widespread supervision from given image-text pairs, and SLIP [25] introduces self-supervised learning which helps encoders learn better visual representation.

In this paper, we propose a novel and efficient way to make use of image-text pairs. We sample part of image-text data as a hold-out *reference set*. Given an input image, our model first retrieves similar images from the reference set with an unsupervised pre-trained image retrieval model, then we use the relationship between retrieved images and texts to augment the representation of input image. A heuristic explanation of the idea is that, it can be considered as an open-book exam: our model doesn't have to memorize all visual concepts in the training data, but learns to recognize visual concepts with the help of a cheat sheet (i.e., the reference set). We propose a framework called Retrieval Augmented Contrastive Language-Image Pre-training (RA-CLIP) to implement this idea. Although using the same amount of image-text data with the vanilla CLIP, RA-CLIP achieves better zero-shot classification performance and linear probe classification performance.

Our contributions are three-fold:

- For contrastive language-image pre-training (CLIP), we present a novel and efficient utilization of image-text pairs. Concretely, we construct a hold-out reference set composed by image-text pairs. Given an input image, we find relevant image-text pairs and use them to help us build better representation for the input image.

- We propose Retrieval Augmented Contrastive Language-Image Pre-training (RA-CLIP), a framework to implement the idea described above. We conduct comprehensive experiments to validate the effectiveness of each block. Visualization results are also provided to explain how RA-CLIP works.

- We compare the proposed RA-CLIP with previous methods on a dozen of commonly used visual recognition benchmarks. Experimental results show that our proposed method significantly outperforms vanilla CLIP baseline and other recently proposed methods.

2. Related Work

2.1. Contrastive Language Image Pre-training

CLIP [29] introduces a new paradigm for visual representation learning. Given image-text pairs as training data, CLIP learns a visual encoder and a textual encoder to align images and text sentences. Different from previous visual recognition system which can only recognize categories specified by the training set, CLIP can be flexibly transferred to new categories. Given new categories, CLIP extends them into text sentences and feed them into the text encoder to obtain categories embeddings. Experimental results show that CLIP can match the accuracy of ImageNet trained ResNet50, even that CLIP didn't use any of the 1.28 million training examples of ImageNet during pre-training. Despite the impressive transfer performance, CLIP is quite data-hungry, it needs tens of millions or more image-text pairs for pre-training [1, 18, 29, 41, 42]. To decrease the amount of training data, recent works [24, 25] try to introduce as much as supervision from existing datasets to help CLIP train better encoders.

In this paper, we propose a novel and efficient methods to make better use of image-text pairs. Different from previous methods that only use image-text pairs to train encoders, we also use them to construct a reference set which can provide supplementary information for the encoder. Given an input image, we find relevant images along with their text descriptions from the reference set, these image-text pairs contain information that can help us describe the input image, thus we use them to augment original CLIP's image representation.

2.2. Knowledge-Enhanced Models

Structured external knowledge has also been included to improve the zero-shot performance, K-Lite [33] enriches entities in natural language with WordNet and Wiktionary, the motivation is straightforward, some concepts like "zebra" may have less training example in the training set, by introducing a knowledge base, "zebra" can be extended to a more general description: "Zebras are African equines with distinctive black-and-white striped coats". CLIP-event [23] also introduces structured external knowledge to construct hard negative text descriptions for images during training. ASIF [26] builds a relative representation for image and text respectively and aligns image modality and text modality with image-text pairs.

Recently large language models have shown impressive zero-shot and few-shot performance on various downstream NLP tasks, massive parameters have to be introduced to encode the knowledge that will be used. Previous works Atlas [17] and RETRO [2] introduce an external non-parametric knowledge base to store the information, which significantly reduces the parameters of their model, e.g., Atlas outperforms a 540B model with 50× fewer parameters.

The methods mentioned above mainly focus on providing more information for individual modalities. Different from them, we utilize the correspondence between image-text pairs in the reference set, which helps us align visual concepts and textual concepts.

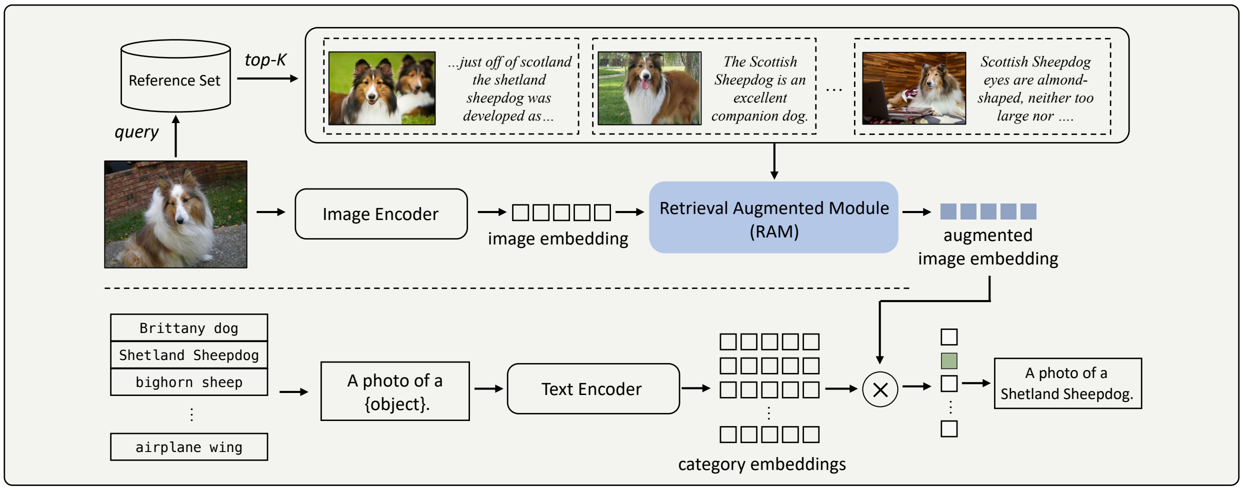
Reference Set
top-K
query
...just off of scotland the shetland sheepdog was developed as...
The Scottish Sheepdog is an excellent companion dog.
...
Scottish Sheepdog eyes are almond-shaped, neither too large nor ....
Image Encoder
image embedding
Retrieval Augmented Module (RAM)
augmented image embedding
Brittany dog
Shetland Sheepdog
bighorn sheep
⋮
airplane wing
A photo of a {object}.
Text Encoder
⋮
category embeddings
×
⋮
A photo of a Shetland Sheepdog.

Figure 2. Overview of the proposed RA-CLIP. Given an input image, RA-CLIP retrieves K most similar image-text pairs from the reference set, which provide informative descriptions for the input image, thus we feed them into the Retrieval Augmented Module (RAM) to enrich the representation of input image.

2.3. Zero-shot Visual Recognition

Zero-shot visual recognition can be categorized into two generations: the traditional class-level zero-shot setting and recently popular task-level zero-shot setting [33]. The traditional class-level zero-shot aims at recognizing objects that belong to categories that are not included in the training set. The task-level zero-shot like CLIP [29] is more practical in real-world applications, the model is trained by hundreds of millions of image-text pairs and is directly evaluated on image classification task, by prompting category names into sentence descriptions [1, 18, 29, 41, 42].

Memorizing visual and textual concepts for down-stream zero-shot tasks usually requires lots of image-text pairs. In this paper, we propose that, besides using image-text pairs to train image encoder and text encoder, we can also introduce a reference set to help the model align visual concepts and textual concepts.

3. Method

In this section, we present the details of our proposed Retrieval Augmented Language Image Pre-training (RA-CLIP). An overview of RA-CLIP is summarized in Figure 2. Given N image-text pairs $\{I_i, T_i\}_{i=1}^{N}$ as training data, where I_i is the i-th image in the dataset and T_i is its corresponding text description. Different from previous Contrastive Language Image Pre-training (CLIP) method that only uses $\{I_i, T_i\}_{i=1}^{N}$ to train encoders, we propose a novel utilization of the image-text pairs. Specifically, we split $\{I_i, T_i\}_{i=1}^{N}$ into two disjoint sets: training set $\mathcal{T}$ and reference set $\mathcal{R}$. Given I_i as input image, we first feed it into the image encoder to obtain the initial image embedding $\mathbf{v}_i$, following previous method CLIP. After that, we retrieve K most similar images $\left\{\mathbf{r}_k^I\right\}_{k=1}^{K}$ from $\mathcal{R}$ as well as their corresponding text sentences $\left\{\mathbf{r}_k^T\right\}_{k=1}^{K}$. $\left\{\mathbf{r}_k^I\right\}_{k=1}^{K}$ and $\left\{\mathbf{r}_k^T\right\}_{k=1}^{K}$ provide informative description for I_i, so we use them to augment $\mathbf{v}_i$ by a proposed Retrieval Augmented Module (RAM). The augmented image embedding is output as $\mathbf{v}_i'$. The text branch of RA-CLIP is like the one of CLIP.

3.1. Dual-Encoder Architecture

The dual-encoder architecture consists of an image encoder and a text encoder. The image encoder used in our experiments is a Vision Transformer [12]. Given an image I_i, we take the embedding of the [CLS] token and normalize it by its L2-norm to obtain the representation of I_i, which is a d-dimension feature vector $\mathbf{v}_i$. The text encoder is also a Transformer. Given a sequence of input text T_i, the text encoder transforms and normalizes T_i into another d-dimension feature vector $\mathbf{t}_i$.

3.2. Overview of RA-CLIP

Reference Set Construction. Multi-modal image-text pairs naturally establish the connection between images and texts. Different images and text descriptions of the same concept can provide informative description of this concept. Inspired by this observation, we construct the reference set $\mathcal{R}$ by random sampling 1.6M (about 1/10 amount of our total training data) image-text pairs from original train set. Ablation experiments will show that commonly used image-text datasets such as YFCC [35], CC12M [5] and LAION [32] can be used as reference data. We also exploit how the amount of image-text pairs effect the final performance. Please refer to Section 4.2.2 for more details.

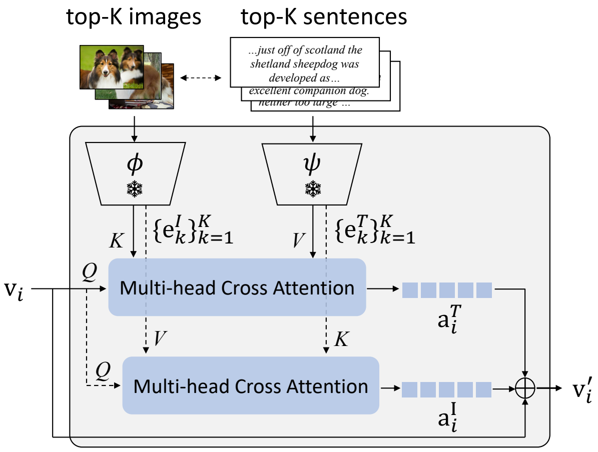

Figure 3. The proposed Retrieval Augmented Module (RAM).

Reference Image-Text Retrieval. Given an input image I_i, RA-CLIP retrieves K most similar images $\{\mathbf{r}_k^I\}_{k=1}^K$ from the reference set as well as their corresponding text descriptions $\{\mathbf{r}_k^T\}_{k=1}^K$. Concretely, we first extract the embeddings of the images in reference set, with an unsupervised pre-trained image encoder (e.g., DINO) for a fairer comparison. This process can be pre-computed offline. Given an input image I_i as query, we also extract the embedding of I_i with the same model, then use the embedding to retrieve K most similar images from the reference set. The retrieval process can be conducted efficiently by existing libraries like Faiss [19]. Finally, we get K images $\{\mathbf{r}_k^I\}_{k=1}^K$ and their corresponding text sentences $\{\mathbf{r}_k^T\}_{k=1}^K$, which are sent to the Retrieval Augmented Module (RAM) to obtain augmented image representation.

Retrieval Augmented Module (RAM). We propose a novel Retrieval Augmented Module (RAM), which first extracts the embeddings of each retrieved $\mathbf{r}_k^I$ and $\mathbf{r}_k^T$:

$$\begin{aligned} \mathbf{e}_k^I &= \phi(\mathbf{r}_k^I), \\ \mathbf{e}_k^T &= \psi(\mathbf{r}_k^T). \end{aligned} \tag{1}$$

Note that ϕ and ψ are pre-trained uni-modal encoders for image and text respectively. They are frozen during the training stage, so Equation 1 can be pre-computed before training.

After that, RAM uses $\{\mathbf{e}_k^I\}_{k=1}^K$ and $\{\mathbf{e}_k^T\}_{k=1}^K$ to augment $\mathbf{v}_i$. RAM first introduces a Multi-head Attention [37] block which takes $\mathbf{v}_i$ as *query*, $\{\mathbf{e}_k^I\}_{k=1}^K$ as *key* and $\{\mathbf{e}_k^T\}_{k=1}^K$ as *value* to obtain text-augmented embedding $\mathbf{a}_i^T$:

$$\mathbf{a}_i^T = MultiheadAttn(\mathbf{v}_i, \{\mathbf{e}_k^I\}_{k=1}^K, \{\mathbf{e}_k^T\}_{k=1}^K). \tag{2}$$

In Equation 2, $\mathbf{v}_i$ learns to weight and aggregate $\{\mathbf{e}_k^T\}_{k=1}^K$ according to the relationship between $\mathbf{v}_i$ and $\{\mathbf{e}_k^I\}_{k=1}^K$. The aggregated text embedding $\mathbf{a}_i^T$ will be used to augment $\mathbf{v}_i$.

Likewise, RAM also computes image-augmented embedding $\mathbf{a}_i^I$ by aggregating $\{\mathbf{e}_k^I\}_{k=1}^K$ in a similar way:

$$\mathbf{a}_i^I = MultiheadAttn(\mathbf{v}_i, \{\mathbf{e}_k^T\}_{k=1}^K, \{\mathbf{e}_k^I\}_{k=1}^K). \tag{3}$$

Finally, RAM outputs the augmented image embedding $\mathbf{v}_i'$:

$$\mathbf{v}_i' = \mathbf{v}_i + \mathbf{a}_i^T + \mathbf{a}_i^I, \tag{4}$$

as illustrated in Figure 3.

3.3. Loss Function.

At the training stage, we follow previous contrastive learning methods [29] and use InfoNCE loss [36] to train our framework. Concretely, given a batch of N image-text pairs $\{I_i, T_i\}_{i=1}^N$ as training data, RA-CLIP produces the augmented image embeddings and the text embeddings: $\{\mathbf{v}_i', \mathbf{t}_i\}_{i=1}^N$, then we compute text-to-image contrastive loss $\mathcal{L}_{t2v}$ and image-to-text contrastive loss $\mathcal{L}_{v2t}$:

$$\mathcal{L}_{v2t} = -log(\frac{exp(\sigma(\mathbf{t}_i, \mathbf{v}_i')/\tau)}{\sum_{j=1}^N exp(\sigma(\mathbf{t}_i, \mathbf{v}_j')/\tau)}), \tag{5}$$

$$\mathcal{L}_{t2v} = -log(\frac{exp(\sigma(\mathbf{v}_i', \mathbf{t}_i)/\tau)}{\sum_{j=1}^N exp(\sigma(\mathbf{v}_i', \mathbf{t}_j)/\tau)}), \tag{6}$$

where N is the batch size, and τ is a temperature factor, which is initialized as 0.07 and trained end-to-end. σ computes the cosine similarity between two vectors. The total loss of our framework is:

$$\mathcal{L} = \mathcal{L}_{t2v} + \mathcal{L}_{v2t}. \tag{7}$$

4. Experiment

Following previous works [24, 25, 39], we pre-train our framework on YFCC dataset [35], then evaluate the performance of zero-shot classification and linear probe classification on commonly used visual recognition benchmarks. Besides that, we also follow RegionCLIP [43] and evaluate the zero-shot ROI classification performance on two widely used object detection datasets, i.e., MS COCO [7] and LVIS [15].

We first describe the implementation details in Section 4.1, including datasets, evaluation metrics, model architecture and optimization details. Then we conduct ablation experiments in Section 4.2 to validate the effectiveness of our proposed method and analyse the effectiveness of each component. Finally, we compare our proposed method with recently proposed CLIP variants like MS-CLIP [39], DeCLIP [24] and SLIP [25] in Section 4.3.

Table 1. Ablation experiments on zero-shot ImageNet classification.

ID	Method	Init. of Image Enc.	Init. of Text Enc.	Pretrain Dataset	Reference Dataset	ϕ	ψ	ImageNet Top-1
1	CLIP	ViT rand.	BERT	YFCC	✗	✗	✗	37.7
2	CLIP	DINO-S	SentenceT	YFCC	✗	✗	✗	21.0
3	CLIP	ViT IN1K	BERT	YFCC	✗	✗	✗	46.1
4	CLIP	ViT rand.	BERT	YFCC+CC	✗	✗	✗	42.1
5	RA-CLIP	ViT rand.	BERT	YFCC	YFCC	SentenceT	DINO-S	53.5
6	RA-CLIP	ViT rand.	BERT	YFCC	CC	SentenceT	DINO-S	54.5
7	RA-CLIP	ViT rand.	BERT	YFCC	LAION	SentenceT	DINO-S	54.2
8	RA-CLIP	ViT rand.	BERT	YFCC	CC	Text Encoder	DINO-S	54.4

4.1. Implementation Details

Dataset All the experiments are conducted on public available datasets. For a fair comparison, we follow previous works [24, 25, 39] and use YFCC [35] for pre-training. We follow CLIP [29] and use the 15 million English subset of YFCC for pre-training. For the reference set used in our framework, we test three different image-text datasets, including the YFCC15M described above, CC12M [5] and LAION [32]. By default, we random sample 1,600,000 image-text pairs (about 1/10 amount of our pre-train dataset) from corresponding dataset to construct the reference set. 10 widely used visual recognition datasets are used to evaluate the zero-shot classification performance and linear probe classification performance, including ImageNet [10], ImageNet V2 [30], CIFAR 10 [21], CIFAR 100 [21], Caltech 101 [13], Oxford Pets [27], SUN397 [38], Food 101 [3], DTD [8] and Stanford Dogs [20]. Following RegionCLIP [43], two object detection datasets are used to evaluate the zero-shot ROI classification performance, i.e., COCO [7] and LVIS [15]. More details about the datasets can be found in our supplement.

Metrics The metrics used to evaluate image classification on all datasets are top-1 accuracy except for Oxford Pets, and Caltech101, which are measured by mean accuracy over classes as in CLIP [29]. The evaluation metric for COCO and LVIS is box Average Precision (AP) as used in RegionCLIP [43].

Architecture By default, the image encoder is a random initialized ViT-B/32 [12], which has 12 layers of Transformer blocks. Each Transformer block has 12 attention heads and the hidden size is set to 768. The text encoder is BERT-base [11], which shares similar scale with the image encoder, i.e., 12 layers of Transformer blocks, 12 attention heads in each Transformer block and 768 hidden size. We also try smaller text encoder from scratch while comparing with previous state-of-the-arts in Section 4.3. The output of the text encoder and image encoder are projected to 384-dim by linear projection. We adopt DINO-S/8 [4] as the image retriever and the uni-modal image encoder ϕ described in Section 3.2, which is pre-trained on ImageNet 1K [10] with *self-supervised* method. Meanwhile, we adopt Sentence Transformer (SentenceT) [31] as the uni-modal text encoder ψ, which is a 6-layer Transformer. For our proposed Retrieval Augmented Module (RAM), we use 6 layers of cross-attention blocks. We truncate the input text tokens so that they have a maximum length of 77. The input image is resized to 224×224.

Optimization We implement our framework with PyTorch [28]. All pre-training experiments are conducted on 8 NVIDIA Tesla A100 GPUs. Batch size is set to 4096. The framework is trained for 32 epochs with LAMB optimizer [40] and an initial learning rate of 2.5e-3. The learning rate follows a cosine decay schedule with 5 epochs of linear warm-up. Weight decay is set to 0.2. For data augmentation, we random crop a 224×224 patch from the input image, then conduct random horizontal flip, random color distortions, random gaussian blur and RandAugment [9], following previous works [6, 24, 25]. Training details of down-stream tasks (e.g., linear probe classification) can be found in our supplement.

4.2. Ablation Study

In this section, we first validate the effectiveness of our proposed method by comparing RA-CLIP and CLIP trained on the same dataset. After that, we build reference sets by sampling image-text pairs from different datasets (YFCC, CC or LAION) and sampling different amounts of image-text pairs from the same dataset to see if the reference set is sensitive to these two factors. Since we introduce pre-trained DINO-S/8 (ϕ in Section 3.2) and SentenceT (ψ in Section 3.2) to extract reference embeddings, we also discuss how these models effect the final performance. At last, ablation experiments with different K-value and different design choices of RAM are provided.

4.2.1 Effectiveness of the Retrieval Augmented CLIP

To validate the effectiveness of our proposed method, we train and test CLIP and RA-CLIP with the same dataset.

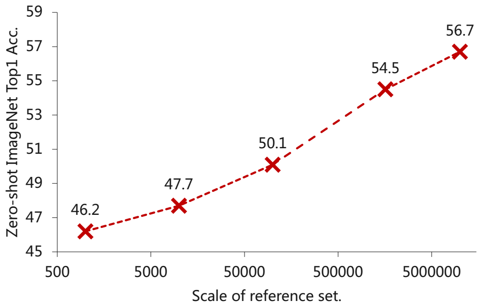

Figure 4. Zero-shot classification performance of RA-CLIP with different amounts of reference data. Five different scales of reference sets are used, i.e., 1K, 10K, 100K, 1.6M, and 10M.

For CLIP, we use YFCC15M for pre-training, and evaluate the zero-shot performance on ImageNet. For RA-CLIP, we random sample 1.6 million image-text pairs (about 1/10 amount of entire dataset) from YFCC15M to construct the reference set, and use the other 13 million image-text pairs for training. Experimental results can be found in Table 1, RA-CLIP (experiment ID 5) outperforms CLIP (experiment ID 1) by a large margin (53.5% vs 37.7%), which validates the effectiveness of RA-CLIP.

We provide a heuristic explanation for this comparison. The pipeline of training and testing CLIP may be considered as a closed-book exam. In the training stage, CLIP trains encoders to learn visual and textual concepts in the training data. In the testing stage, given a test query image, CLIP may return a false prediction if it didn't *memorize* related concepts of the test image. Different from CLIP, the training and testing of RA-CLIP can be considered as an open-book exam. In the training stage, RA-CLIP doesn't have to *memorize* all concepts in the training data, but learns how to recognize an image with a help of a cheat-sheet (i.e., the reference set). In the testing stage, given a test query image, RA-CLIP can *look up* the cheat-sheet (i.e., the reference set) and finds related *knowledge* about the query image, then uses the knowledge to enrich the representation of input image.

4.2.2 Different image-text data as reference set

Since the reference set is the core component of RA-CLIP, we then analyse if RA-CLIP provides significant improvements while using different datasets or different amounts of image-text pairs as reference sets.

Different dataset as reference set. We adopt another two widely used datasets to construct the reference set, i.e., CC12M and LAION. We sample 1.6 million image-text pairs from CC12M to construct the reference set, and use YFCC15M for RA-CLIP pre-training. We also train CLIP on the combination of the reference set and YFCC15M.

Table 2. Different design choices of Retrieval Augmented Module (RAM). All models are trained on YFCC15M with CC1.6M as reference set and evaluated on zero-shot ImageNet classification.

Method	Augment Image	Augment Text	Fusion Type	K	ImageNet Top-1
RAM	✓	✗	$\mathbf{a}_i^T$	64	52.1
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I$	64	51.8
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	64	54.5
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	16	54.3
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	128	53.9
RAM	✓	✓	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	64	53.1

By comparing experiment 6 and experiment 4 in Table 1, we can see that RA-CLIP still outperforms CLIP with significant improvement(54.5% v.s. 42.1%). We then conduct similar experiments with 1.6 million image-text pairs from LAION as reference set, which achieves 54.2% top-1 accuracy on ImageNet. Experiments above show that the commonly used datasets like YFCC15M, CC12M, LAION can be used as reference set and provide similar performance.

Different amounts of reference data. We then test the performance of reference sets with different amounts of image-text pairs. 5 reference sets are constructed by sample 1K, 10K, 100K, 1.6M, 10M image-text pairs from CC12M. YFCC15M is used as training data. The performance of RA-CLIP with these 5 different reference sets is illustrated in Figure 4. We can see that introducing more image-text pairs as reference set usually brings better performance. Since the horizontal axis of Figure 4 is plotted in log-space, we can conclude that performance is getting saturated while introducing more image-text pairs as reference set.

4.2.3 Ablation of pre-trained SentenceT and DINO

Two frozen pre-trained uni-modal encoders are introduced in RA-CLIP to process image-text data of the reference set: SentenceT and DINO-S/8. It is necessary for us to show if the improvement described above is brought by these two encoders. We re-implement CLIP with frozen SentenceT as text encoder and frozen DINO-S/8 as image encoder, linear projection layer is appended after each encoder to obtain final embeddings. We train the CLIP model on YFCC15M, the zero-shot classification performance on ImageNet is 21.0%, referred as experiment 2 in Table 1, which is significantly lower than CLIP baseline (experiment 1). Experimental results show that although the pre-trained SentenceT and DINO-S provide promising results on uni-modal tasks, they provide poor performance on multi-modal alignment task.

We also prove that, SentenceT is not necessary for our final performance. We re-implemented RA-CLIP by

Table 3. Zero-shot image classification performance and linear probe classification performance on 10 downstream datasets (%). RA-CLIP* denotes the model that trains a text encoder with $width$ = 512 and $head$ = 8 from scratch.

Evaluation Type	Method	ImageNet	ImageNetV2	Pets	CIFAR10	CIFAR100	SUN397	Food101	Caltech101	DTD	Dogs	Avg.
Zero-shot	K-Lite [33]	45.3	–	–	–	–	–	–	–	–	–	–
	CLIP [29]	37.7	32.8	16.1	76.0	48.6	50.8	21.8	69.8	28.3	11.1	39.3
	MS-CLIP [39]	36.7	30.2	–	–	–	–	–	–	–	5.6	–
	SLIP [25]	38.3	33.3	28.3	72.2	45.3	45.1	44.7	65.9	21.8	11.8	40.7
	DeCLIP [24]	43.2	36.1	30.2	72.1	39.7	51.6	46.9	70.1	24.2	11.7	42.6
	RA-CLIP*	51.2	45.4	50.5	89.4	61.8	45.7	43.9	76.1	24.6	22.0	51.1
	RA-CLIP	53.5	47.2	49.0	89.4	62.3	46.5	43.8	76.9	25.6	26.1	**52.0**
Linear probe	CLIP [29]	63.5	51.3	69.8	91.7	74.1	64.7	69.1	84.9	66.5	50.5	68.6
	MS-CLIP [39]	68.1	49.8	62.1	87.2	66.7	71.7	76.0	81.6	69.4	46.1	67.9
	SLIP [25]	68.1	52.1	75.4	90.5	75.3	73.5	77.1	87.2	71.1	52.6	72.3
	DeCLIP [24]	69.2	53.1	76.5	88.6	71.6	75.9	79.3	88.0	69.1	49.9	72.1
	RA-CLIP*	73.3	62.3	88.2	94.9	78.4	60.7	65.5	86.8	65.5	75.5	75.1
	RA-CLIP	72.9	61.9	88.2	95.2	78.9	61.1	66.5	87.2	66.6	76.0	**75.5**

replacing SentenceT with the text encoder of RA-CLIP, the zero-shot performance on ImageNet is 54.4% (experiment 8), which is comparable with the one used SentenceT (experiment 6). We also note that, although replacing SentenceT with the text encoder of RA-CLIP can provide similar performance, the text encoder of RA-CLIP is trained end-to-end, thus the embeddings of the reference text cannot be pre-computed off-line before training, which introduces more computation cost. So we use SentenceT by default.

DINO-S/8 in Section 3.2 uses ImageNet 1K for *self-supervised learning*, extracting reference images on-the-fly during training with the image encoder is much more time-consuming due to the time of reading, pre-processing and extracting images. So we test if CLIP can achieve comparable or better performance than RA-CLIP if it uses ImageNet data for pre-training. Concretely, we use the ViT-B/32 pre-trained on ImageNet 21K and fine-tuned on ImageNet-1K to initialize the image encoder of CLIP. Then train the CLIP on YFCC15M, the classification performance is referred as experiment 3 in Table 1. Initializing the image encoder by ImageNet pre-trained model brings significant improvements (46.1% v.s. 37.7%). However, it is not *zero-shot* evaluation anymore. We can see that RA-CLIP (53.5%) still outperforms this strong CLIP baseline.

4.2.4 Different hyper-parameters and design choices

Different K-value in retrieval process. We train RA-CLIP by retrieving different amounts (K=16, 64, 128) of image-text pairs from the reference set for each query image. Experimental results are shown in row 3, 4 and 5 of Table 2. RA-CLIP provides stable performance with different K values, and the model with K=64 performs slightly better than the others.

Different design choices of RAM architecture. We further test different design choices of RAM architecture in Table 2. Firstly, we take $\mathbf{a}_i^T$ as the final augmented representation. In this case, RAM gathers related information from the reference texts and produces final embedding, the result is 52.1%. After that, we fuse $\mathbf{a}_i^T$ and $\mathbf{a}_i^I$, which provides similar performance (51.8%). Finally, we take $\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$ as the final image embedding to fuse information from original image encoder and the reference data, which performs best (54.5%) among the three different variants as expected.

Augment both image and text representation. Since experiments above have validated the effectiveness of RA-CLIP, a straightforward question is: will this idea improve CLIP's text representation? To validate that, we use SentenceT to retrieve relevant image-text pairs for input text sentence, then apply similar augmentation to the text embedding. However, it leads to a slightly worse performance 53.1% as shown in Table 2. We conjecture that text sentence is less informative and the retrieved image-text pairs are more diverse and may not bring correct information.

4.3. Main Results

In this section, we compare RA-CLIP with CLIP and recently proposed CLIP-variant on two down-stream visual recognition tasks: zero-shot image classification and linear probe image classification. Evaluations are conducted on 10 widely used visual recognition benchmarks. Note that some previous works (e.g., DeCLIP [24]) train a text encoder with $width$ = 512 and $head$ = 8 from scratch, thus we follow them and re-implemented the RA-CLIP for a fair comparison, the results are denoted as RA-CLIP*. Besides that, we also follow RegionCLIP to evaluate the zero-shot ROI classification performance on MS COCO and LVIS.

Zero-shot image classification Previous methods usually take YFCC15M or its augmented version as train set.

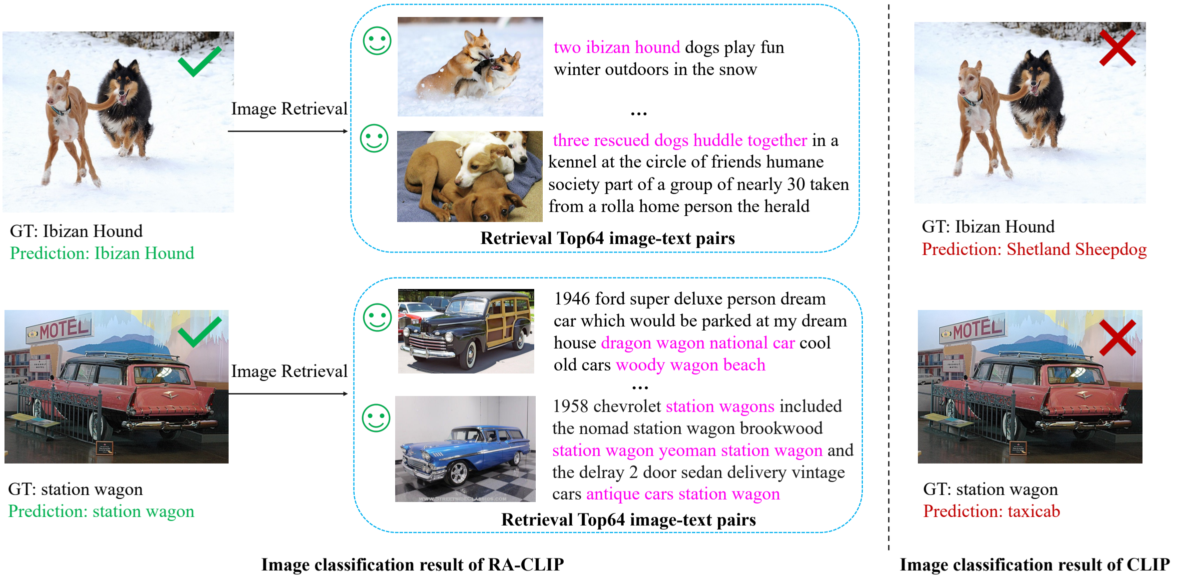
GT: Ibizan Hound
Prediction: Ibizan Hound
Image Retrieval
two ibizan hound dogs play fun winter outdoors in the snow
...
three rescued dogs huddle together in a kennel at the circle of friends humane society part of a group of nearly 30 taken from a rolla home person the herald
Retrieval Top64 image-text pairs
GT: Ibizan Hound
Prediction: Shetland Sheepdog
GT: station wagon
Prediction: station wagon
Image Retrieval
1946 ford super deluxe person dream car which would be parked at my dream house dragon wagon national car cool old cars woody wagon beach
...
1958 chevrolet station wagons included the nomad station wagon brookwood station wagon yeoman station wagon and the delray 2 door sedan delivery vintage cars antique cars station wagon
Retrieval Top64 image-text pairs
GT: station wagon
Prediction: taxicab
Image classification result of RA-CLIP
Image classification result of CLIP

Figure 5. Case study for RA-CLIP in image classification. The retrieved images and texts enrich the representation of input image.

Table 4. Zero-shot ROI classification on COCO and LVIS (%).

Method	LVIS				COCO			
	AP	APs	APm	APl	AP	APs	APm	APl
Regin CLIP	21.6	8.7	31.0	45.7	44.4	21.9	51.0	61.8
Region RA-CLIP	23.2	10.9	34.2	44.9	48.4	29.3	57.9	61.9

We thus random sample 1.6M image-text from YFCC15M as reference set, and use the other 13M image-text pairs as train set. We report the performance of RA-CLIP on all 10 datasets as well as the results of previous methods CLIP [29], K-lite [33], SLIP [25], DeCLIP [24] and MS-CLIP [39] shown in Table 3. RA-CLIP shows superior performance than previous state-of-the-art methods.

We also compare with OpenAI's CLIP model trained on 400 million image-text pairs. To reduce the training cost, we use OpenAI's CLIP-B/32 to initialize the image encoder, text encoders, ϕ and ψ of RA-CLIP, then freeze them and train the other modules on YFCC15M. The model achieves 68.2% top-1 accuracy on ImageNet, significantly better than OpenAI's CLIP-B/32 (63.3%).

In Figure 5, we also provide some visualization to help us understand how RA-CLIP works. The image-text pairs retrieved from the reference set can introduce additional important information that can help RA-CLIP make correct predictions.

Linear probe image classification Following previous methods, we also conduct linear probe image classification experiments to test the transferability of the learned image representation. We fix the pre-trained image encoder of our trained CLIP and RA-CLIP, then train a linear classifier to conduct image classification on down-stream datasets. Experimental results are shown in Table 3. RA-CLIP achieves better performance than previous methods CLIP, MS-CLIP, SLIP and DeCLIP, bringing 2.8% (75.1% v.s. 72.3%) averaged improvements.

Zero-shot ROI classification Following previous work RegionCLIP [43], we transfer the proposed RA-CLIP to object detection task on COCO [7] and LVIS [15] datasets to test the performance of ROI classification. Following RegionCLIP [43], we use ground-truth bounding boxes as region proposals, then apply RA-CLIP and CLIP to classify the proposals, following a RCNN [14] pipeline.

We report results on the validation set of LVIS and COCO in Table 4. We can see that RA-CLIP still outperforms CLIP on these two benchmarks, bringing 1.6% improvement on LVIS and 4.0% improvement on COCO. Note that the proposed RA-CLIP performs much better on small objects and medium objects.

5. Conclusion

In this paper, we propose a novel and efficient utilization of image-text pairs. Different from previous Contrastive Language Image Pre-training methods that only use image-text pairs to train encoders, we also use image-text pairs to construct a reference set. Given an input image, we find relevant images as well as corresponding texts in the reference set and use them to enrich the representation of the input image. Experiments on zero-shot classification and linear probe classification validate the effectiveness of our methods.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *CoRR*, 2022. 2, 3

[2] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. *CoRR*, 2022. 2

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Eur. Conf. Comput. Vis.*, pages 446–461, 2014. 5

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision Transformers. In *Int. Conf. Comput. Vis.*, pages 9650–9660, 2021. 5

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3558–3568, 2021. 3, 5

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Int. Conf. Mach. Learn.*, pages 1597–1607, 2020. 5

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, 2015. 4, 5, 8

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3606–3613, 2014. 5

[9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 702–703, 2020. 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009. 5

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. 5

[12] 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. In *Int. Conf. Learn. Represent.*, 2020. 1, 3, 5

[13] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 178–178, 2004. 5

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 00, pages 580–587, 2014. 8

[15] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5356–5364, 2019. 4, 5, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 1

[17] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *CoRR*, 2022. 2

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Int. Conf. Mach. Learn.*, pages 4904–4916. PMLR, 2021. 2, 3

[19] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2017. 4

[20] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011. 5

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Adv. Neural Inform. Process. Syst.*, 2012. 1

[23] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. CLIP-Event: Connecting text and images with event structures. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16399–16408, 2022. 2

[24] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *Int. Conf. Learn. Represent.*, 2022. 1, 2, 4, 5, 7, 8

[25] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: self-supervision meets language-image pre-training. In *Eur. Conf. Comput. Vis.*, volume 13686, pages 529–544, 2022. 1, 2, 4, 5, 7, 8

[26] Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. ASIF: Coupled data turns unimodal models to multimodal without training. *arXiv preprint arXiv:2210.01738*, 2022. 2

[27] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3498–3505, 2012. 5

[28] 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 *Adv. Neural Inform. Process. Syst.*, pages 8024–8035, 2019. 5

[29] 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. In *Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 5, 7, 8

[30] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Int. Conf. Mach. Learn.*, volume 97, pages 5389–5400, 2019. 5

[31] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019. 5

[32] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *CoRR*, 2021. 3, 5

[33] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, Kurt Keutzer, Trevor Darrell, and Jianfeng Gao. K-LITE: Learning transferable visual models with external knowledge. *Adv. Neural Inform. Process. Syst.*, 2022. 2, 3, 7, 8

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. Learn. Represent.*, 2014. 1

[35] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016. 3, 4, 5

[36] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, 2018. 4

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 2017. 4

[38] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. SUN database: Exploring a large collection of scene categories. *Int. J. Comput. Vis.*, 119(1):3–22, 2016. 5

[39] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *Eur. Conf. Comput. Vis.*, pages 69–87. Springer, 2022. 4, 5, 7, 8

[40] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training BERT in 76 minutes. In *Int. Conf. Learn. Represent.*, 2020. 5

[41] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 2, 3

[42] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *CoRR*, 2021. 2, 3

[43] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-based language-image pretraining. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16793–16803, 2022. 4, 5, 8