Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision

Chao Jia [1] **Yinfei Yang** [1] **Ye Xia** [1] **Yi-Ting Chen** [1] **Zarana Parekh** [1] **Hieu Pham** [1] **Quoc V. Le** [1]
Yunhsuan Sung [1] **Zhen Li** [1] **Tom Duerig** [1]

Abstract

Pre-trained representations are becoming crucial for many NLP and perception tasks. While representation learning in NLP has transitioned to training on raw text without human annotations, visual and vision-language representations still rely heavily on curated training datasets that are expensive or require expert knowledge. For vision applications, representations are mostly learned using datasets with explicit class labels such as ImageNet or OpenImages. For vision-language, popular datasets like Conceptual Captions, MSCOCO, or CLIP all involve a non-trivial data collection (and cleaning) process. This costly curation process limits the size of datasets and hence hinders the scaling of trained models. In this paper, we leverage a noisy dataset of over one billion image alt-text pairs, obtained without expensive filtering or post-processing steps in the Conceptual Captions dataset. A simple dual-encoder architecture learns to align visual and language representations of the image and text pairs using a contrastive loss. We show that the scale of our corpus can make up for its noise and leads to state-of-the-art representations even with such a simple learning scheme. Our visual representation achieves strong performance when transferred to classification tasks such as ImageNet and VTAB. The aligned visual and language representations enables zero-shot image classification and also set new state-of-the-art results on Flickr30K and MSCOCO image-text retrieval benchmarks, even when compared with more sophisticated cross-attention models. The representations also enable cross-modality search with complex text and text + image queries.

[1]Google Research. Correspondence to: Chao Jia <chaojia@google.com>, Yinfei Yang <yinfeiy@google.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1. Introduction

In the existing literature, visual and vision-language representation learning are mostly studied separately with different training data sources. In the vision domain, pre-training on large-scale supervised data such as ImageNet (Deng et al., 2009), OpenImages (Kuznetsova et al., 2020), and JFT-300M (Sun et al., 2017; Kolesnikov et al., 2020) has proven to be critical for improving performance on downstream tasks via transfer learning. Curation of such pre-training datasets requires heavy work on data gathering, sampling, and human annotation, and hence is difficult to scale.

Pre-training has also become the de-facto approach in vision-language modeling (Lu et al., 2019; Chen et al., 2020c; Li et al., 2020). However, vision-language pre-training datasets such as Conceptual Captions (Sharma et al., 2018), Visual Genome Dense Captions (Krishna et al., 2016), and ImageBERT (Qi et al., 2020) require even heavier work on human annotation, semantic parsing, cleaning and balancing. As a result, the scales of these datasets are only in the realm of $\sim$10M examples. This is at least an order of magnitude smaller than their counterparts in the vision domain, and much smaller than large corpora of text from the internet for NLP pre-training (e.g., Devlin et al. (2019); Radford et al. (2019); Yang et al. (2019); Liu et al. (2019b); Raffel et al. (2020)).

In this work, we leverage a dataset of over one billion noisy image alt-text pairs to scale visual and vision-language representation learning. We follow the procedures described in the Conceptual Captions dataset (Sharma et al., 2018) to have a large noisy dataset. But instead of applying the complex filtering and post-processing steps as proposed by (Sharma et al., 2018) to clean the dataset, we only apply simple frequency-based filtering. The resulting dataset is noisy, but is two orders of magnitude larger than the Conceptual Captions dataset. We show that visual and vision-language representations pre-trained on our exascale dataset achieve very strong performance on a wide range of tasks.

To train our model, we use an objective that aligns the visual and language representations in a shared latent embedding space using a simple dual-encoder architecture. Similar

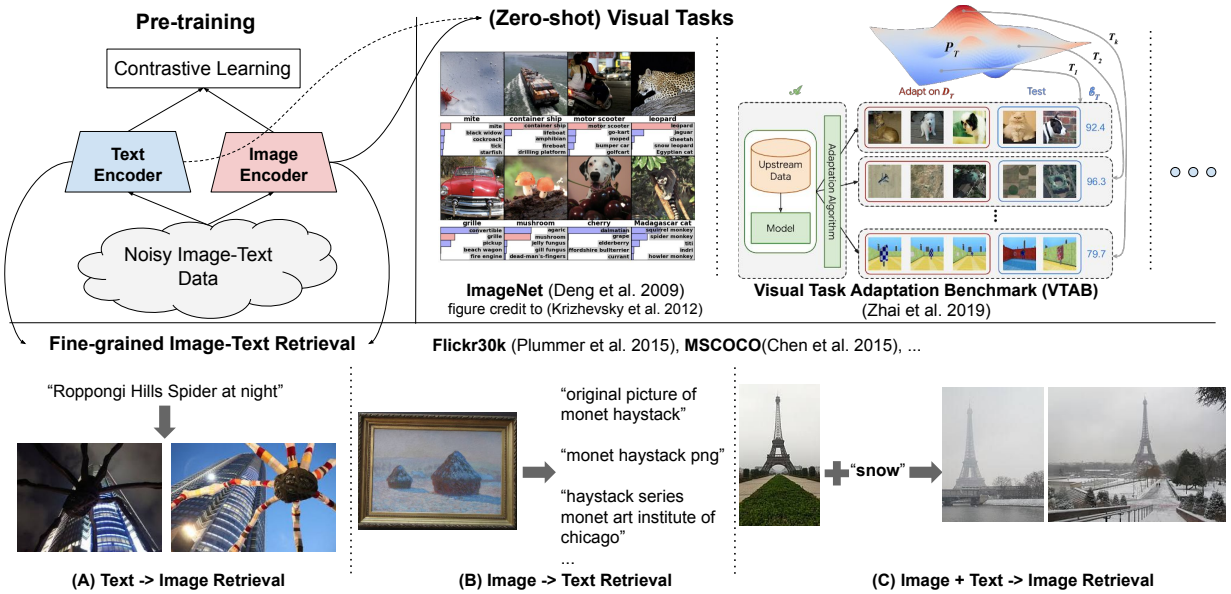

Figure 1. A summary of our method, ALIGN. Visual and language representations are jointly learned from noisy image alt-text data. The representations can be used for vision-only or vision-language task transfer. Without any fine-tuning, ALIGN powers zero-shot visual classification and cross-modal search including image-to-text search, text-to-image search and even search with joint image+text queries.

objectives has been applied to learning visual-semantic embeddings (VSE) (Frome et al., 2013; Faghri et al., 2018). We name our model **ALIGN**: **A** **L**arge-scale **I**ma**G**e and **N**oisy-text embedding. Image and text encoders are learned via a contrastive loss (formulated as normalized softmax) that pushes the embeddings of matched image-text pair together while pushing those of non-matched image-text pair apart. This is one of the most effective loss functions for both self-supervised (Chen et al., 2020b) and supervised (Zhai & Wu, 2019; Musgrave et al., 2020) representation learning. Considering paired texts as fine-grained labels of images, our image-to-text contrastive loss is analogous to the conventional label-based classification objective; and the key difference is that the text encoder generates the "label" weights. The top-left of Figure 1 summarizes the method we use in ALIGN.

The aligned image and text representations are naturally suited for cross-modality matching/retrieval tasks and achieve state-of-the-art (SOTA) results in corresponding benchmarks. For instance, ALIGN outperforms the previous SOTA method by over 7% in most zero-shot and fine-tuned R@1 metrics in Flickr30K and MSCOCO. Moreover, such cross-modality matching naturally enables zero-shot image classification when feeding the classnames into the text encoder, achieving 76.4% top-1 accuracy in ImageNet without using any of its training samples. The image representation itself also achieves superior performance in various downstream visual tasks. For example, ALIGN achieves 88.64% top-1 accuracy in ImageNet. Figure 1-bottom shows the cross-modal retrieval examples that come from a real retrieval system built by ALIGN.

2. Related Work

High-quality visual representations for classification or retrieval are usually pre-trained on large-scale labeled datasets (Mahajan et al., 2018; Kolesnikov et al., 2020; Dosovitskiy et al., 2021; Juan et al., 2020). Recently, self-supervised (Chen et al., 2020b; Tian et al., 2020; He et al., 2020; Misra & Maaten, 2020; Li et al., 2021; Grill et al., 2020; Caron et al., 2020) and semi-supervised learning (Yalniz et al., 2019; Xie et al., 2020; Pham et al., 2020) have been studied as alternative paradigms. However, models trained by these methods so far show limited transferability to downstream tasks (Zoph et al., 2020).

Leveraging images and natural language captions is another direction of learning visual representations. Joulin et al. (2015); Li et al. (2017); Desai & Johnson (2020); Sariyildiz et al. (2020); Zhang et al. (2020) show that a good visual representation can be learned by predicting the captions from images, which inspires our work. These works are however limited to small datasets such as Flickr (Joulin et al., 2015; Li et al., 2017) and COCO Captions (Desai & Johnson, 2020; Sariyildiz et al., 2020), and the resulting models don't produce a vision-language representation that is needed for tasks like cross-modal retrieval.

In the vision-language representation learning domain, visual-semantic embeddings (VSE) (Frome et al., 2013; Faghri et al., 2018) and improved versions (e.g., leveraging object detectors, dense feature maps, or multi-attention layers) (Socher et al., 2014; Karpathy et al., 2014; Kiros et al.; Nam et al., 2017; Li et al., 2019; Messina et al., 2020; Chen et al., 2020a) have been proposed. Recently more

advanced models emerge with cross-modal attention layers (Liu et al., 2019a; Lu et al., 2019; Chen et al., 2020c; Huang et al., 2020b) and show superior performance in image-text matching tasks. However, they are orders of magnitudes slower and hence impractical for image-text retrieval systems in the real world. In contrast, our model inherits the simplest VSE form, but still outperforms all previous cross-attention models in image-text matching benchmarks.

Closely related to our work is CLIP (Radford et al., 2021), which proposes visual representation learning via natural language supervision in a similar contrastive learning setting. Besides using different vision and language encoder architectures, the key difference is on training data: ALIGN follows the natural distribution of image-text pairs from the raw alt-text data, while CLIP collects the dataset by first constructing an allowlist of high-frequency visual concepts from English Wikipedia. We demonstrate that strong visual and vision-language representations can be learned with a dataset that doesn't require expert knowledge to curate.

3. A Large-Scale Noisy Image-Text Dataset

The focus of our work is to scale up visual and vision-language representation learning. For this purpose, we resort to a much larger dataset than existing ones. Specifically, we follow the methodology of constructing Conceptual Captions dataset (Sharma et al., 2018) to get a version of raw English alt-text data (image and alt-text pairs). The Conceptual Captions dataset was cleaned by heavy filtering and post-processing. Here, for the purpose of scaling, we trade quality for scale by relaxing most of the cleaning steps in the original work. Instead, we only apply minimal frequency-based filtering as detailed below. The result is a much larger (1.8B image-text pairs) but noisier dataset. Figure 2 shows some sample image-text pairs from the dataset.

Figure 2. Example image-text pairs randomly sampled from the training dataset of ALIGN. One clearly noisy text annotation is marked in *italics*.

Image-based filtering. Following Sharma et al. (2018), we remove pornographic images and keep only images whose shorter dimension is larger than 200 pixels and aspect ratio is smaller than 3. Images with more than 1000 associated alt-texts are discarded. To ensure that we don't train on test images, we also remove duplicates or near-duplicates of test images in all downstream evaluation datasets (e.g., ILSVRC-2012, Flickr30K, and MSCOCO). See Appendix A for more details.

Text-based filtering. We exclude alt-texts that are shared by more than 10 images. These alt-texts are often irrelevant to the content of the images (e.g., "1920x1080", "alt_img", and "cristina"). We also discard alt-texts that contain any rare token (outside of 100 million most frequent unigrams and bigrams from the raw dataset), and those that are either too short (<3 unigrams) or too long (>20 unigrams). This removes noisy texts like "image_tid 25&id mggqpuweqdpd&cache 0&lan_code 0", or texts that are too generic to be useful.

4. Pre-training and Task Transfer

4.1. Pre-training on Noisy Image-Text Pairs

We pre-train ALIGN using a dual-encoder architecture. The model consists of a pair of image and text encoders with a cosine-similarity combination function at the top. We use EfficientNet with global pooling (without training the 1x1 conv layer in the classification head) as the image encoder and BERT with [CLS] token embedding as the text embedding encoder (we generate 100k wordpiece vocabulary from our training dataset). A fully-connected layer with linear activation is added on top of BERT encoder to match the dimension from the image tower. Both image and text encoders are trained from scratch.

The image and text encoders are optimized via normalized softmax loss (Zhai & Wu, 2019). In training, we treat matched image-text pairs as positive and all other random image-text pairs that can be formed in a training batch as negative.

We minimize the sum of two losses: one for image-to-text classification

$$L_{i2t} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)} \quad (1)$$

and the other for text-to-image classification

$$L_{t2i} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)} \quad (2)$$

Here, x_i and y_j are the normalized embedding of image in the i-th pair and that of text in the j-th pair, respectively. N is the batch size, and σ is the temperature to scale the logits. For in-batch negatives to be more effective, we concatenate embeddings from all computing cores to form a much larger batch. The temperature variable is crucial as both image

and text embeddings are L2-normalized. Instead of manually sweeping for the optimal temperature value, we find that it can be effectively learned together with all the other parameters.

4.2. Transferring to Image-Text Matching & Retrieval

We evaluate ALIGN models on image-to-text and text-to-image retrieval tasks, with and without finetuning. Two benchmark datasets are considered: Flickr30K (Plummer et al., 2015) and MSCOCO (Chen et al., 2015). We also evaluate ALIGN on Crisscrossed Captions (CxC) (Parekh et al., 2021), which is an extension of MSCOCO with additional human semantic similarity judgments for caption-caption, image-image, and image-caption pairs. With extended annotations, CxC enables four intra- and inter-modal retrieval tasks including image-to-text, text-to-image, text-to-text, and image-to-image retrieval, and three semantic similarity tasks including semantic textual similarity (STS), semantic image similarity (SIS), and semantic image-text similarity (SITS). As the training set is identical to the original MSCOCO, we can directly evaluate the MSCOCO fine-tuned ALIGN model on CxC annotations.

4.3. Transferring to Visual Classification

We first apply zero-shot transfer of ALIGN to visual classification tasks on ImageNet ILSVRC-2012 benchmark (Deng et al., 2009) and its variants including ImageNet-R(endition) (Hendrycks et al., 2020) (non-natural images such as art, cartoons, sketches), ImageNet-A(dversarial) (Hendrycks et al., 2021) (more challenging images for ML models), and ImageNet-V2 (Recht et al., 2019). All of these variants follow the same set (or a subset) of ImageNet classes, while the images in ImageNet-R and ImageNet-A are sampled from drastically different distributions from ImageNet.

We also transfer the image encoder to downstream visual classification tasks. For this purpose, we use the ImageNet as well as a handful of smaller fine-grained classification datasets such as Oxford Flowers-102 (Nilsback & Zisserman, 2008), Oxford-IIIT Pets (Parkhi et al., 2012), Stanford Cars (Krause et al., 2013), and Food101 (Bossard et al., 2014). For ImageNet, results from two settings are reported: training the top classification layer only (with frozen ALIGN image encoder) and fully fine-tuned. Only the latter setting is reported for fine-grained classification benchmarks. Following Kolesnikov et al. (2020), we also evaluate the robustness of our model on Visual Task Adaptation Benchmark (VTAB) (Zhai et al., 2019) which consists of 19 diverse (covering subgroups of natural, specialized and structured image classification tasks) visual classification tasks with 1000 training samples each.

5. Experiments and Results

We train our ALIGN models from scratch, using the open-sourced implementation of EfficientNet as the image encoder and BERT as the text encoder. Unless in the ablation study, we use the results of ALIGN where the image encoder is EfficientNet-L2 and the text encoder is BERT-Large. The image encoder is trained at resolution of 289×289 pixels no matter what EfficientNet variant is used. We first resize input images to 346×346 resolution and then perform random crop (with additional random horizontal flip) in training and central crop in evaluation. For BERT we use wordpiece sequence of maximum 64 tokens since the input texts are no longer than 20 unigrams. The softmax temperature variable is initialized as 1.0 (this temperature variable is shared between image-to-text loss and text-to-image loss) and we use 0.1 as label smoothing parameter in the softmax losses. We use LAMB optimizer (You et al., 2020)[1] with weight decay ratio 1e-5. The learning rate is warmed up linearly to 1e-3 from zero in 10k steps, and then linearly decay to zero in 1.2M steps ($\sim$12 epochs). We train the model on 1024 Cloud TPUv3 cores with 16 positive pairs on each core. Therefore the total effective batch size is 16384.

5.1. Image-Text Matching & Retrieval

We evaluate ALIGN on Flickr30K and MSCOCO cross-modal retrieval benchmarks, in both zero-shot and fully fine-tuned settings. We follow (Karpathy & Fei-Fei, 2015) and most existing works to obtain the train/test splits. Specifically, for Flickr30K, we evaluate on the standard 1K test set, and finetune on the 30k training set. For MSCOCO, we evaluate on the 5K test set, and finetune on 82K training plus 30K additional validation images that are not in the 5K validation or 5K test sets.

During fine-tuning, the same loss function is used. But there can be false negatives when the batch size is comparable to the total number of training samples. So we reduce the global batch size from 16384 to 2048. We also reduce the initial learning rate to 1e-5 and train for 3K and 6K steps (with linear decay) respectively on Flickr30K and MSCOCO. All the other hyper-parameters are kept the same as pre-training.

Table 1 shows that, compared to previous works, ALIGN achieves SOTA results in all metrics of Flickr30K and MSCOCO benchmarks. In the zero-shot setting, ALIGN gets more than 7% improvement in image retrieval task compared to the previous SOTA, CLIP (Radford et al., 2021). With fine-tuning, ALIGN outperforms all existing methods by a large margin, including those that employ more complex cross-modal attention layers such as ImageBERT (Qi et al., 2020), UNITER (Chen et al., 2020c),

[1] We tried SGD with momentum and ADAM which are known to work well for CNNs and BERT respectively. LAMB appears to be a better choice for training both image and text encoders.

Table 1. Image-text retrieval results on Flickr30K and MSCOCO datasets (zero-shot and fine-tuned). ALIGN is compared with ImageBERT (Qi et al., 2020), UNITER (Chen et al., 2020c), CLIP (Radford et al., 2021), GPO (Chen et al., 2020a), ERNIE-ViL (Yu et al., 2020), VILLA (Gan et al., 2020), and Oscar (Li et al., 2020).

		Flickr30K (1K test set)						MSCOCO (5K test set)					
		image → text			text → image			image → text			text → image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot	ImageBERT	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
	UNITER	83.6	95.7	97.7	68.7	89.2	93.9	-	-	-	-	-	-
	CLIP	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
	ALIGN	**88.6**	**98.7**	**99.7**	**75.7**	**93.8**	**96.8**	**58.6**	**83.0**	**89.7**	**45.6**	**69.8**	**78.6**
Fine-tuned	GPO	88.7	98.9	99.8	76.1	94.5	97.1	68.1	90.2	-	52.7	80.2	-
	UNITER	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
	ERNIE-ViL	88.1	98.0	99.2	76.7	93.6	96.4	-	-	-	-	-	-
	VILLA	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
	Oscar	-	-	-	-	-	-	73.5	92.2	96.0	57.5	82.8	**89.8**
	ALIGN	**95.3**	**99.8**	**100.0**	**84.9**	**97.4**	**98.6**	**77.0**	**93.5**	**96.9**	**59.9**	**83.3**	**89.8**

Table 2. Multimodal retrieval performance on Crisscrossed Captions (CxC) dataset. ALIGN is compared with VSE++ (Faghri et al., 2018), VSRN (Li et al., 2019), DE_{I2T} (Parekh et al., 2021), and $DE_{T2T+I2T}$ (Parekh et al., 2021).

	image → text			text → image			text → text			image → image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
VSE++	43.1	74.3	84.2	32.5	62.7	75.4	38.7	62.3	72.2	36.4	70.4	81.3
VSRN	52.4	81.9	90.0	40.1	71.1	81.5	41.0	64.8	74.5	44.2	76.7	86.2
DE_{I2T}	53.9	82.7	91.2	39.8	70.2	80.9	26.0	47.1	57.5	38.3	74.1	85.0
$DE_{T2T+I2T}$	55.9	84.2	91.8	41.7	72.3	83.0	42.4	64.9	74.0	38.5	73.6	84.9
ALIGN	**78.1**	**94.3**	**97.4**	**61.8**	**84.9**	**91.1**	**45.4**	**66.8**	**75.2**	**49.4**	**81.4**	**89.1**

Table 3. Spearman's R Bootstrap Correlation (×100) on Crisscrossed Captions (CxC) dataset. ALIGN is compared with VSE++ (Faghri et al., 2018), VSRN (Li et al., 2019), DE_{I2T} (Parekh et al., 2021), and $DE_{T2T+I2T}$ (Parekh et al., 2021).

Model	STS avg ± std	SIS avg ± std	SITS avg ± std	Mean Avg
VSE++	**74.4±0.4**	73.3±0.9	55.2±1.5	67.6
VSRN	73.0±0.4	70.1±1.0	60.4±1.3	67.8
DE_{I2T}	50.9±0.6	**81.3±0.7**	61.6±1.4	64.6
$DE_{T2T+I2T}$	74.2±0.4	74.5±0.9	61.9±1.3	70.2
ALIGN	72.9±0.4	77.2±0.8	**67.6±1.2**	**72.6**

ERNIE-ViL (Yu et al., 2020), VILLA (Gan et al., 2020) and Oscar (Li et al., 2020).

Table 2 reports the performance of ALIGN on Crisscrossed Captions (CxC) retrieval tasks. Again, ALIGN achieves SOTA results in all metrics, especially by a large margin on image-to-text (+22.2% R@1) and text-to-image (20.1% R@1) tasks. Table 3 shows that ALIGN also outperforms the previous SOTA on SITS task with an improvement of 5.7%. One interesting observation is that, despite being much better on inter-modal tasks, ALIGN is not as impressive on intra-modal tasks. For instance, the improvements on text-to-text and image-to-image retrieval tasks (in particular the former) are less significant compared to those on image-to-text and text-to-image tasks. The performance on STS and SIS tasks is also slightly worse than VSE++ and DE_{I2T}. We suspect it is because the training objective of ALIGN focuses on cross-modal (image-text) matching instead of intra-modal matching. Parekh et al. (2021) suggest multitask learning could produce more balanced representations. We leave it to the future work.

5.2. Zero-shot Visual Classification

If we directly feed the texts of classnames into the text encoder, ALIGN is able to classify images into candidate classes via image-text retrieval. Table 4 compares ALIGN with CLIP on Imagenet and its variants. Similar to CLIP, ALIGN shows great robustness on classification tasks with different image distributions. In order to make a fair comparison, we use the same prompt ensembling method as CLIP. Each classname is expanded with a set of prompt templates defined by CLIP such as "A photo of a {classname}". The class embedding is computed by averaging the embeddings of all templates followed by an L2-normalization. We find that such ensembling gives 2.9% improvement on ImageNet top-1 accuracy.

Table 4. Top-1 Accuracy of zero-shot transfer of ALIGN to image classification on ImageNet and its variants.

Model	ImageNet	ImageNet-R	ImageNet-A	ImageNet-V2
CLIP	76.2	88.9	**77.2**	**70.1**
ALIGN	**76.4**	**92.2**	75.8	**70.1**

Table 5. ImageNet classification results. ALIGN is compared with WSL (Mahajan et al., 2018), CLIP (Radford et al., 2021), BiT (Kolesnikov et al., 2020), ViT (Dosovitskiy et al., 2021), NoisyStudent (Xie et al., 2020), and Meta-Pseudo-Labels (Pham et al., 2020).

Model (backbone)	Acc@1 w/ frozen features	Acc@1	Acc@5
WSL (ResNeXt-101 32x48d)	83.6	85.4	97.6
CLIP (ViT-L/14)	85.4	-	-
BiT (ResNet152 x 4)	-	87.54	98.46
NoisyStudent (EfficientNet-L2)	-	88.4	98.7
ViT (ViT-H/14)	-	88.55	-
Meta-Pseudo-Labels (EfficientNet-L2)	-	**90.2**	**98.8**
ALIGN (EfficientNet-L2)	**85.5**	88.64	98.67

5.3. Visual Classification w/ Image Encoder Only

On the ImageNet benchmark, we first freeze the learned visual features and only train the classification head. Afterwards we fine-tune all layers. We use basic data augmentations including random cropping (same as in Szegedy et al. (2015)) and horizontal flip. In evaluation we apply a single central crop with ratio of 0.875. Following Touvron et al. (2019), we use 0.8 scale ratio between training and evaluation to mitigate the resolution discrepancy introduced by random crop. Specifically, train/eval resolution is 289/360 with frozen visual features, and is 475/600 when fine-tuning all variables.

In both stages of training, we use a global batch size of 1024, SGD optimizer with momentum 0.9, and learning rate decayed every 30 epochs with ratio 0.2 (100 epochs in total). Weight decay is set to zero. With frozen visual features, we use the initial learning rate of 0.1. When fine-tuning all layers with use the initial learning rate of 0.01, and use 10x smaller learning rate on the backbone network compared to the classification head.

Table 5 compares ALIGN with previous methods on the ImageNet benchmark. With frozen features, ALIGN slightly outperforms CLIP and achieves SOTA result of 85.5% top-1 accuracy. After fine-tuning ALIGN achieves higher accuracy than BiT and ViT models, and is only worse than Meta Pseudo Labels which requires deeper interaction between ImageNet training and large-scale unlabeled data. Compared to NoisyStudent and Meta-Pseudeo-Labels which also use EfficientNet-L2, ALIGN saves 44% FLOPS by using smaller test resolution (600 instead of 800).

In VTAB eval, we follow a hyper-parameter sweep as shown in the Appendix I in (Zhai et al., 2019) with 50 trials for each task. Each task is trained on 800 images and the hyperparameters are selected using the validation set of 200 images. After the sweep, the selected hyperparameters are used to train on the combined training and validation splits of 1000 images for each task. Table 6 reports the mean accuracy (including the breakdown results on each subgroup) with standard deviation from three fine-tuning runs and shows that ALIGN outperforms BiT-L (Kolesnikov et al., 2020) with similar hyper-parameter selection method applied.

Table 6. VTAB (19 tasks) comparison between ALIGN and BiT-L.

Model	All tasks	Natural	Specialized	Structured
Bit-L	78.72	-	-	-
ALIGN	**79.99±0.15**	83.38	87.56	73.25

To evaluate on smaller fine-grained classification benchmarks, we adopt a simple fine-tuning strategy for all tasks. We use the same data augmentation and optimizer as in ImageNet fine-tuning. Similarly, we first train the classification head and then fine-tune all layers, except with batch norm statistics frozen. The train/eval resolution is fixed at 289/360. We use batch size 256 and weight decay 1e-5. The initial learning rate is set to 1e-2 and 1e-3 respectively, with cosine learning rate decay in 20k steps. Table 7 compares ALIGN with BiT-L (Kolesnikov et al., 2020) and SAM (Foret et al., 2021) which both apply same fine-tuning hyper-parameters for all tasks.[2] For small tasks like these, details in fine-tuning matter. So we list the baseline results in (Foret et al., 2021) without using SAM optimization for a fairer comparison. Our result (average of three runs) is comparable to the SOTA results without tweaking on optimization algorithms.

Table 7. Transfer learning results on Fine-grained Classification Tasks. BiT-L (Kolesnikov et al., 2020) was trained with ResNet152 x 4 whereas SAM-baseline, SAM-final (Foret et al., 2021) and ALIGN were trained with EfficientNet-L2.

Model	Oxford Flowers	Oxford Pets	Stanford Cars	Food101
BiT-L	99.63	96.62	-	-
SAM-baseline	99.60	96.92	95.07	96.03
SAM-final	**99.65**	**97.10**	95.96	**96.18**
ALIGN	**99.65**	96.19	**96.13**	95.88

6. Ablation Study

In the ablation study, we compare model performance mostly on MSCOCO zero-shot retrieval and ImageNet K-Nearest-neighbor (KNN) tasks.[3] We find these two met-

[2] ViT (Dosovitskiy et al., 2021) uses different hyper-parameters for different tasks and hence is not included in comparison.

[3] For each image in the validation set of ImageNet, we retrieve its nearest neighbors from the training set w/ pre-trained image encoder. Recall@K metric is calculated based on if the groundtruth label of the query image appears in the top-K retrieved images.

rics are representative and correlate well with other metrics reported in the section above. If not mentioned, hyper-parameters other than the ablated factor are kept the same as in the baseline model.

6.1. Model Architectures

We first study the performance of ALIGN models using different image and text backbones. We train EfficientNet from B1 to L2 for the image encoder and BERT-Mini to BERT-Large for the text encoder. We add an additional fully-connected layer with linear activation on top of B1, B3, B5 and L2 globally-pooled features to match the output dimension of B7 (640). A similar linear layer is added to all text encoders. We reduce the training steps to 1M in ablation to save some runtime.

Figures 3 shows MSCOCO zero-shot retrieval and ImageNet KNN results with different combinations of image and text backbones. Model quality improves nicely with larger backbones except that the ImageNet KNN metric starts to saturate from BERT-Base to BERT-Large with EfficientNet-B7 and EfficientNet-L2. As expected, scaling up image encoder capacity is more important for vision tasks (e.g., even with BERT-Mini text tower, L2 performs better than B7 with BERT-Large). In image-text retrieval tasks the image and text encoder capacities are equally important. Based on the nice scaling property shown in Figure 3, we only fine-tune the model with EfficientNet-L2 + BERT-Large as reported in Section 5.

We then study key architecture hyperparameters including embedding dimensions, number of random negatives in the batch, and the softmax temperature. Table 8 compares a number of model variants to a baseline model (first row) trained with the following settings: EfficientNet-B5 image encoder, BERT-Base text encoder, embedding dimension 640, all negatives in the batch, and a learnable softmax temperature.

Rows 2-4 of Table 8 show that model performance improves with higher embedding dimensions. Hence, we let the dimension scale with larger EfficientNet backbone (L2 uses 1376). Rows 5 and 6 show that using fewer in-batch negatives (50% and 25%) in the softmax loss will degrade the performance. Rows 7-9 study the effect of the temperature parameter in the softmax loss. Compared to the baseline model that learns the temperature parameter (converged to about 1/64), some hand-selected, fixed temperatures could be slightly better. However, we choose to use the learnable temperature as it performs competitively and makes learning easier. We also notice that the temperature usually quickly decrease to only around 1.2x of the converged values in the first 100k steps, and then slowly converges until the end of training.

Table 8. Ablation study of key architecture parameters. Baseline model (first row) is trained with embedding dimension 640, using all negatives in the batch, and a learnable softmax temperature.

Model	MSCOCO		ImangeNet KNN
	I2T R@1	T2I R@1	R@1
B5 + BERT-base	51.7	**37.5**	64.6
w/ embedding dim=320	50.3	34.1	64.0
w/ embedding dim=160	47.0	34.4	63.7
w/ embedding dim=80	42.0	29.3	61.9
w/ 50% in-batch negs	50.2	37.0	63.8
w/ 25% in-batch negs	48.7	35.8	63.3
w/ softmax temp=1/128	**52.2**	36.5	**64.8**
w/ softmax temp=1/64	**52.2**	37.3	**64.8**
w/ softmax temp=1/32	39.6	26.9	61.2

6.2. Pre-training Datasets

It's also important to understand how the model performs when trained on different datasets with varying size. For this purpose, we train two models: EfficientNet-B7 + BERT-base and EfficientNet-B3 + BERT-mini on three different datasets: full ALIGN training data, 10% randomly sampled ALIGN training data, and Conceptual Captions (CC-3M, around 3M images). CC-3M is much smaller so we train the model with 1/10 of the default number of steps. All models are trained from scratch. As shown in Table 9, a large scale training set is essential to allow scaling up of our models and to achieve better performance. For instance, models trained on ALIGN data clearly outperform those trained on CC-3M data. On CC-3M, B7+BERT-base starts to overfit and performs even worse than B3+BERT-mini. Conversely, a larger model is required to fully utilize the larger dataset – the smaller B3+BERT-mini almost saturate at 10% of ALIGN data, while with the larger B7+BERT-base, there is a clear improvement with full ALIGN data.

Table 9. Ablation study of different training datasets.

Model + Data	MSCOCO		ImangeNet KNN
	I2T R@1	T2I R@1	R@1
B7 + BERT-base			
+ ALIGN full data	55.4	41.7	69.3
+ ALIGN 10% data	52.0	39.2	68.8
+ CC-3M data	18.9	15.5	48.7
B3 + BERT-mini			
+ ALIGN full data	37.4	24.5	56.5
+ ALIGN 10% data	36.7	24.4	55.8
+ CC-3M data	22.1	17.3	48.9

To understand better how data size scaling wins over the increased noise, we further randomly sample 3M, 6M, and 12M ALIGN training data and compare them with the cleaned CC-3M data on B7+BERT-base model. Table 10 shows that while the ALIGN data performs much worse than CC data with the same size (3M), the model quality trained on 6M and 12M ALIGN data rapidly catches up. Despite being noisy, ALIGN data outperforms Conceptual Captions with only 4x size.

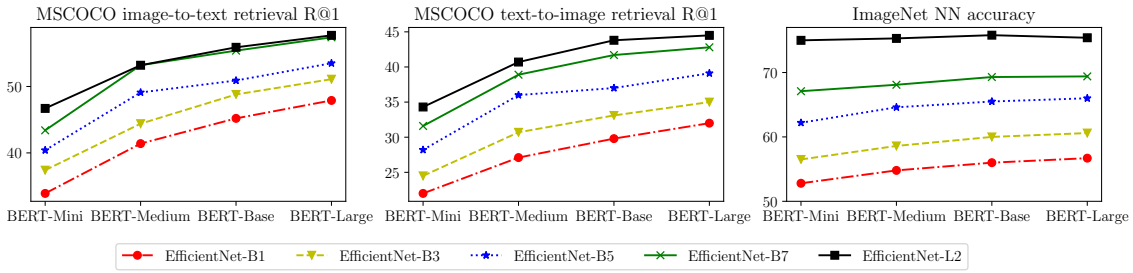

Figure 3. Zero-shot image-text retrieval and ImageNet KNN accuracy@1 with different image and text encoder sizes.

Table 10. Tradeoff between training data size and quality.

Model + Data	MSCOCO I2T R@1	MSCOCO T2I R@1	ImangeNet KNN R@1
B7 + BERT-base			
+ ALIGN 12M data	23.8	17.5	51.4
+ ALIGN 6M data	15.8	11.9	47.9
+ ALIGN 3M data	8.1	6.3	41.3
+ CC-3M data	18.9	15.5	48.7

7. Analysis of Learned Embeddings

We build a simple image retrieval system to study the behaviors of embeddings trained by ALIGN. For demonstration purposes, we use an index consisting of 160M CC-BY licensed images that are separate from our training set. Figure 4 shows the top 1 text-to-image retrieval results for a handful of text queries not existing in the training data. ALIGN can retrieve precise images given detailed descriptions of a scene, or fine-grained or instance-level concepts like landmarks and artworks. These examples demonstrate that our ALIGN model can align images and texts with similar semantics, and that ALIGN can generalize to novel complex concepts.

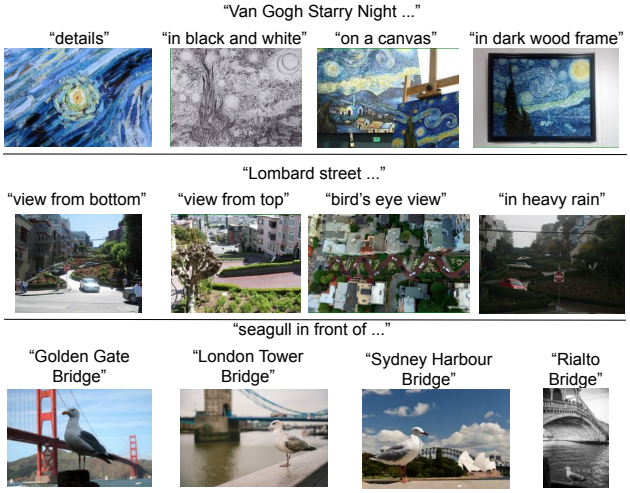

Figure 4. Image retrieval with fine-grained text queries using ALIGN's embeddings.

Previously word2vec (Mikolov et al., 2013a;b) shows that linear relationships between word vectors emerge as a result of training them to predict adjacent words in sentences and paragraphs. We show that linear relationships between

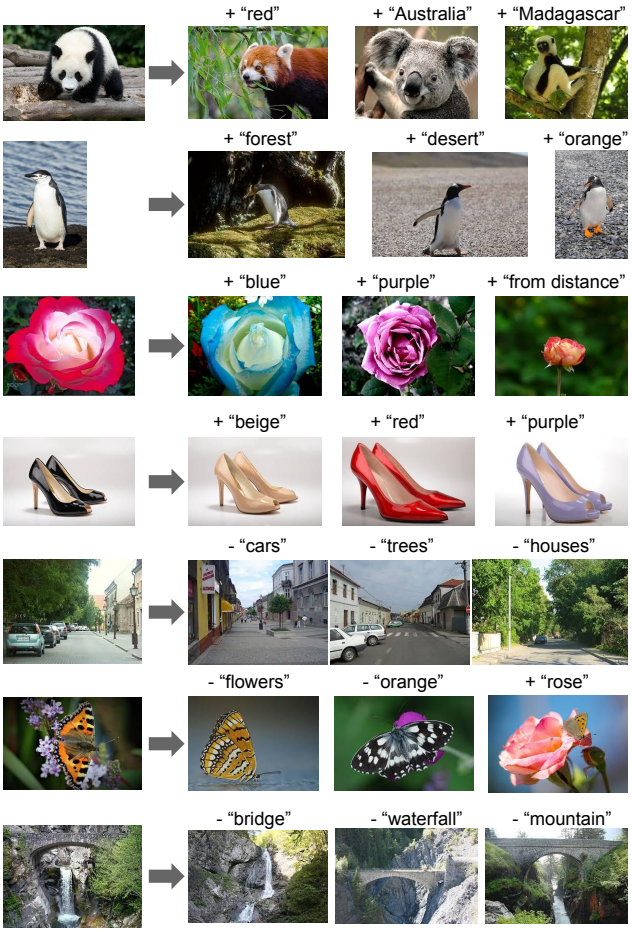

Figure 5. Image retrieval with image±text queries. We add (or subtract) text query embedding to (or from) the image query embedding, and then use the resulting embedding to retrieve relevant images using cosine similarity.

image and text embeddings also emerge in ALIGN. We perform image retrieval using a combined image+text query. Specifically, given a query image and a text string, we add their ALIGN embeddings together and use it to retrieve relevant images.[4] Figure 5 shows results for a variety of

[4]We normalize the text and image embeddings before adding them. We also tried various scale factor and found that a scale of 2 for the text embedding and 1 for the image embedding give best results as shown in the figure, although 1:1 also works well.

image+text queries. These examples not only demonstrate great compositionality of ALIGN embeddings across vision and language domains, but also show the feasibility of a new paradigm of "search with multi-modal query" that would otherwise be hard using only text query or image query. For instance, one could now look for the "Australia" or "Madagascar" equivalence of pandas, or turn a pair of black shoes into identically-looking shoes with the color of "beige". Finally, as shown in the last three rows of Figure 5, removing objects/attributes from a scene is possible by performing subtraction in the embedding space.

8. Multilingual ALIGN Model

One advantage of ALIGN is that the model is trained on noisy web image text data with very simple filters, and none of the filters are language specific. Given that, we further lift the language constraint of the conceptual caption data processing pipeline to extend the dataset to multilingual (covering 100+ languages) and match its size to the English dataset (1.8B image-text pairs). A multilingual model $\text{ALIGN}_{\text{mling}}$ is trained using this data. We created a new mutlilingual wordpiece vocabulary with size 250k to cover all languages. Model training follows the exact English configuration.

We test the multilingual model on Multi30k, a multilingual image text retrieval dataset extends Flickr30K (Plummer et al., 2015) to German (de) (Elliott et al., 2016), French (fr) (Elliott et al., 2017) and Czech (cs) (Barrault et al., 2018). The dataset consists of 31,783 images with 5 captions per image in English and German and 1 caption per image in French and Czech. The train/dev/test splits are defined in Young et al. (2014). We evaluate the zero-shot model performance of ALIGN and compare it with M^3P (Huang et al., 2020a) and UC2 (Zhou et al., 2021). The evaluation metric is mean Recall (mR), which computes the average score of Recall@1, Recall@5 and Recall@10 on image-to-text retrieval and text-to-image retrieval tasks.

Table 11 shows that the zero-shot performance of $\text{ALIGN}_{\text{mling}}$ outperforms M^3P on all languages by a large margin, with the largest +57.8 absolution mR improvement on fr. The zero-shot performance of $\text{ALIGN}_{\text{mling}}$ is even comparable to the fine-tuned (w/ training splits) M^3P and UC2 except on cs. On en, $\text{ALIGN}_{\text{mling}}$ performs slightly worse on its counterpart ALIGN_{EN} (trained on EN-only data.)

Table 11. Multimodal retrieval performance on Multi30K dataset. The metric is the mean Recall (mR).

Model	en	de	fr	cs
zero-shot				
M^3P	57.9	36.8	27.1	20.4
ALIGN_{EN}	**92.2**	-	-	-
$\text{ALIGN}_{\text{mling}}$	90.2	84.1	**84.9**	63.2
w/ fine-tuning				
M^3P	87.7	82.7	73.9	72.2
UC2	88.2	**84.5**	83.9	**81.2**

9. Conclusion

We present a simple method of leveraging large-scale noisy image-text data to scale up visual and vision-language representation learning. Our method avoids heavy work on data curation and annotation, and only requires minimal frequency-based cleaning. On this dataset, we train a simple dual-encoder model using a contrastive loss. The resulting model, named ALIGN, is capable of cross-modal retrieval and significantly outperforms SOTA VSE and cross-attention vision-language models. In visual-only downstream tasks, ALIGN is also comparable to or outperforms SOTA models trained with large-scale labeled data.

10. Social Impacts and Future Work

While this work shows promising results from a methodology perspective with a simple data collection method, additional analysis of the data and the resulting model is necessary before the use of the model in practice. For instance, considerations should be made towards the potential for the use of harmful text data in alt-texts to reinforce such harms. On the fairness front, data balancing efforts may be required to prevent reinforcing stereotypes from the web data. Additional testing and training around sensitive religious or cultural items should be taken to understand and mitigate the impact from possibly mislabeled data.

Further analysis should also be taken to ensure that the demographic distribution of humans and related cultural items like clothing, food, and art do not cause model performance to be skewed. Analysis and balancing would be required if such models will be used in production.

Finally, unintended misuse of such models for surveillance or other nefarious purposes should be prohibited.

Acknowledgements

This work was done with invaluable help from colleagues from Google. We would like to thank Jan Dlabal and Zhe Li for continuous support in training infrastructure, Simon Kornblith for building the zero-shot & robustness model evaluation on ImageNet variants, Xiaohua Zhai for help on conducting VTAB evaluation, Mingxing Tan and Max Moroz for suggestions on EfficientNet training, Aleksei Timofeev for the early idea of multimodal query retrieval, Aaron Michelony and Kaushal Patel for their early work on data generation, and Sergey Ioffe, Jason Baldridge and Krishna Srinivasan for the insightful feedback and discussion.

References

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 304–323, 2018.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020.

Chen, J., Hu, H., Wu, H., Jiang, Y., and Wang, C. Learning the best pooling strategy for visual semantic embedding. In *arXiv preprint arXiv:2011.04305*, 2020a.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of International Conference on Machine Learning*, 2020b.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. In *arXiv preprint arXiv:1504.00325*, 2015.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *Proceedings of European Conference on Computer Vision*, 2020c.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2009.

Desai, K. and Johnson, J. Virtex: Learning visual representations from textual annotations. In *arXiv preprint arXiv:2006.06666*, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations*, 2021.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, 2016.

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, September 2017.

Faghri, F., Fleet, D. J., Kiros, J. R., and Fidler, S. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, 2018.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., and Mikolov, T. Devise: A deep visual-semantic embedding model. In *Proceedings of Neural Information Processing Systems*, 2013.

Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., and Liu, J. Large-scale adversarial training for vision-and-language representation learning. In *Proceedings of Neural Information Processing Systems*, 2020.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *CVPR*, 2021.

Hill, F., Reichart, R., and Korhonen, A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 2015.

Huang, H., Su, L., Qi, D., Duan, N., Cui, E., Bharti, T., Zhang, L., Wang, L., Gao, J., Liu, B., Fu, J., Zhang, D., Liu, X., and Zhou, M. M3p: Learning universal representations via multitask multilingual multimodal pre-training. *arXiv*, abs/2006.02635, 2020a.

Huang, Z., Zeng, Z., Liu, B., Fu, D., and Fu, J. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020b.

Joulin, A., van der Maaten, L., Jabri, A., and Vasilache, N. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, 2015.

Juan, D.-C., Lu, C.-T., Li, Z., Peng, F., Timofeev, A., Chen, Y.-T., Gao, Y., Duerig, T., Tomkins, A., and Ravi, S. Graph-rise: Graph-regularized image semantic embedding. In *Proceedings of ACM International Conference on Web Search and Data Mining*, 2020.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2015.

Karpathy, A., Joulin, A., and Li, F. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems*, 2014.

Kiros, J., Chan, W., and Hinton, G. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big transfer (bit): General visual representation learning. In *Proceedings of European Conference on Computer Vision*, 2020.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of ICCV Workshop on 3D Representation and Recognition*, 2013.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2016.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., and Ferrari, V. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 2020.

Li, A., Jabri, A., Joulin, A., and van der Maaten, L. Learning visual n-grams from web data. In *Proceedings of IEEE International Conference on Computer Vision*, 2017.

Li, J., Zhou, P., Xiong, C., and Hoi, S. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021.

Li, K., Zhang, Y., Li, K., Li, Y., and Fu, Y. Visual semantic reasoning for image-text matching. In *Proceedings of International Conference on Computer Vision*, 2019.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of European Conference on Computer Vision*, 2020.

Liu, F., Liu, Y., Ren, X., He, X., and Sun, X. Aligning visual regions and textual concepts for semantic-grounded image representations. In *Advances in Neural Information Processing Systems*, 2019a.

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

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of Neural Information Processing Systems*, 2019.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. Exploring the limits of weakly supervised pretraining. In *Proceedings of European Conference on Computer Vision*, 2018.

Messina, N., Amato, G., Esuli, A., Falchi, F., Gennaro, C., and Marchand-Maillet, S. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2020.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013b.

Misra, I. and Maaten, L. v. d. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

Musgrave, K., Belongie, S., and Lim, S.-N. A metric learning reality check. In *Proceedings of European Conference on Computer Vision*, 2020.

Nam, H., Ha, J.-W., and Kim, J. Dual attention networks for multimodal reasoning and matching. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2017.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Parekh, Z., Baldridge, J., Cer, D., Waters, A., and Yang, Y. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. In *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Pham, H., Dai, Z., Xie, Q., Luong, M.-T., and Le, Q. V. Meta pseudo labels. In *arXiv preprint arXiv:2003.10580*, 2020.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the International Conference on Computer Vision*, 2015.

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., and Sacheti, A. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint-arXiv:2001.07966*, 2020.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarawl, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019.

Sariyildiz, M. B., Perez, J., and Larlus, D. Learning visual representations with caption annotations. *arXiv preprint arXiv:2008.01392*, 2020.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2018.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2014.

Sun, C., Shrivastava, A., Sigh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the International Conference on Computer Vision*, 2017.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2015.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *European Conference on Computer Vision*, 2020.

Touvron, H., Vedaldi, A., Douze, M., and Jégou, H. Fixing the train-test resolution discrepancy. In *Advances in Neural Information Processing Systems*, 2019.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., and Wu, Y. Learning fine-grained image similarity with deep ranking. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2014.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2020.

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., and Mahajan, D. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, 2019.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. Large batch optimization for deep learning: Training bert in 76 minutes. In *Proceedings of International Conference on Learning Representations*, 2020.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014.

Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., and Wang, H. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.

Zhai, A. and Wu, H.-Y. Classification is a strong baseline for deep metric learning. In *Proceedings of the British Machine Vision Conference*, 2019.

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., and Houlsby, N. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Zhou, M., Zhou, L., Wang, S., Cheng, Y., Li, L., Yu, Z., and Liu, J. UC2: Universal cross-lingual cross-modal vision-and-language pre-training. *arXiv preprint arXiv:2104.00332*, 2021.

Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E. D., and Le, Q. V. Rethinking pre-training and self-training. In *Advances in Neural Information Processing Systems*, 2020.

A. Remove Near-Duplicate Test Images from Training Data

To detect near-duplicate images, we first train a separate high-quality image embedding model following (Wang et al., 2014) with a large-scale labeled dataset as in (Juan et al., 2020), and then generate 4K clusters via k-means based on all training images of the embedding model. For each query image (from the ALIGN dataset) and index image (from test sets of downstream tasks), we find their top-10 nearest clusters based on the embedding distance. Each image is then assigned to $\binom{10}{3}$ buckets (all possible combinations of 3 clusters out of 10). For any query-index image pair that falls into the same bucket, we mark it as near-duplicated if their embedding cosine similarity is larger than 0.975. This threshold is trained on a large-scale dataset built with human rated data and synthesized data with random augmentation.

B. Evaluation on SimLex-999

The image-text co-training could also help the natural language understanding as shown in Kiros et al. (2018). For instance, with language only, it is very hard to learn antonyms. In order to test this capability of ALIGN model, we also evaluate the word representation from ALIGN model[5] on SimLex-999 (Hill et al., 2015), which is a task to compare word similarity for 999 word pairs. We follow Kiros et al. (2018) to report the results on 9 sub-tasks each contains a subset of word pairs: *all, adjectives, nouns, verbs, concreteness quartiles (1-4)*, and *hard*.

Table 12. SimLex-999 results (Spearman's ρ).

	GloVe	Picturebook	**ALIGN**
all	**40.8**	37.3	39.8
adjs	**62.2**	11.7	49.8
nouns	42.8	**48.2**	45.9
verbs	**19.6**	17.3	16.6
conc-q1	**43.3**	14.4	23.9
conc-q2	41.6	27.5	**41.7**
conc-q3	42.3	46.2	**47.6**
conc-q4	40.2	**60.7**	57.8
hard	27.2	28.8	**31.7**

The results are listed in the Table 12 compared to Picturebook (Kiros et al., 2018) and GloVe (Pennington et al., 2014) embeddings. Overall the learned ALIGN perform better than Picturebook but slightly worse than GloVe embeddings. What is interesting is that the ALIGN word embeddings has a similar trend of Picturebook embeddings, with better performance on *nouns* and *most concrete* categories but worse on *adjs* and *less concrete* categories compared to GloVe embeddings. ALIGN word embedding achieves the highest performance on the *hard* category, which similarity is difficult to distinguish from relatedness. This observation confirmed the hypothesis from Kiros et al. (2018) that image-based word embeddings are less likely to confuse similarity with relatedness than text learned distributional-based methods.

[5] As ALIGN uses the wordpiece tokens, one word can be split into multiple pieces. We feed the wordpieces of a word into ALIGN model and use the [CLS] token representation before the project layers as the word embeddings.