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 also set new state-of-the-art results on Flickr30K and MSCOCO 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, Mountain View, CA, USA.. Correspondence to: Chao Jia <chaojia@google.com>, Yinfei Yang <yinfeiy@google.com>.

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 ∼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. (2019); 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

Pre-training

Contrastive Learning

Visual Tasks

Text Encoder   Image Encoder

Noisy Image-Text Data

ImageNet (Deng et al. 2009)
figure credit to (Krizhevsky et al. 2012)

P_T

Adapt on D_T Test θ_T

92.4

96.3

79.7

Visual Task Adaptation Benchmark (VTAB)
(Zhai et al. 2019)

Fine-grained Image-Text Retrieval   Flickr30k (Plummer et al. 2015), **MSCOCO**(Chen et al. 2015), ...

"Roppongi Hills Spider at night"

"original picture of monet haystack"

"monet haystack png"

"haystack series monet art institute of chicago"
...

"snow"

(A) Text -> Image Retrieval   **(B) Image -> Text Retrieval**   **(C) Image + Text -> Image Retrieval**
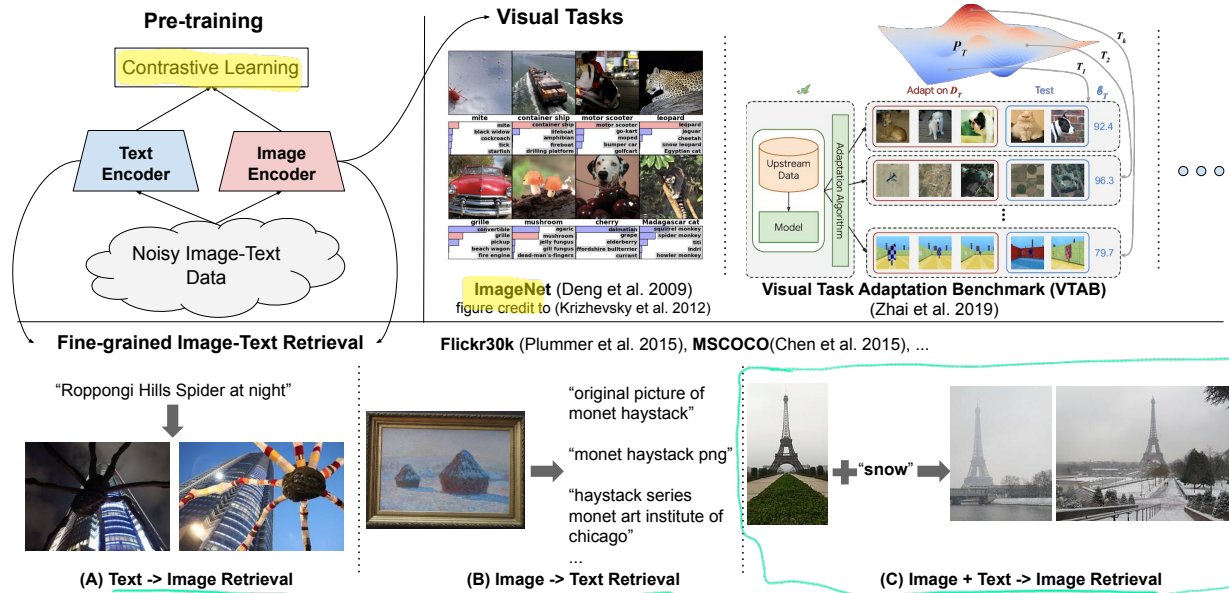
Figure 1. A summary of our method, ALIGN. Visual and language representations are jointly learned from noisy image alt-text data. The representations can be used for vision-only or vision-language task transfer. Without any fine-tuning, ALIGN powers cross-modal search including image-to-text search, 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 **Ima**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 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. 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 (Hénaff et al., 2019; Chen et al., 2020b; 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 for 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 detector or dense feature maps) (Socher et al., 2014; Karpathy et al., 2014; Kiros et al.; Li et al., 2019; Chen et al., 2020a) have been proposed. Recently more advanced models emerge with cross-modal attention layers (Lu et al.,

2019; Chen et al., 2020c) 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.

"motorcycle front wheel" *"thumbnail for version as of 21 57 29 june 2010"* "file frankfurt airport skyline 2017 05 jpg"

"file london barge race 2 jpg" "moustache seamless wallpaper design" "st oswalds way and shops"

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).

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 mggqpuwe-qdpd&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

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

and the other for text-to-image classification

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

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 also transfer the image encoder to downstream visual classification tasks. For this purpose, we use the ImageNet ILSVRC-2012 benchmark (Deng et al., 2009) 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 Results

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), 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

---

[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 Image-BERT (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 ($\times 100$) on Criss-crossed 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 $\pm$ std	SIS avg $\pm$ std	SITS avg $\pm$ 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**

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. Visual Classification Results

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 4 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-Pseuedo-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 5 reports the mean accuracy

Table 4. 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

(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 5. 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 6 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 6. 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

---

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

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 metrics 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 B7 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 and B5 globally-pooled features to match the output dimension of B7 (640). A similar linear layer is added to all text encoders.

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. As expected, scaling up image encoder capacity is more important for vision tasks (e.g., even with BERT-Mini text tower, B7 performs better than B5 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 train EfficientNet-L2 with BERT-Large as reported in Section 5. Table 7 compares ALIGN-L2 with ALIGN-B7 (both with BERT-Large) on MSCOCO image-text retrieval and ImageNet classification results, showing additional quality gain from B7 to L2.

We then study key architecture hyperparameters including

---

[3]For each image in the validation set of ILSVRC-2012, we retrieve their nearest neighbors from the training set using pre-trained image embeddings. Recall@K metric is calculated based on if the groundtruth label of the query image appears in the top-K retrieved images.
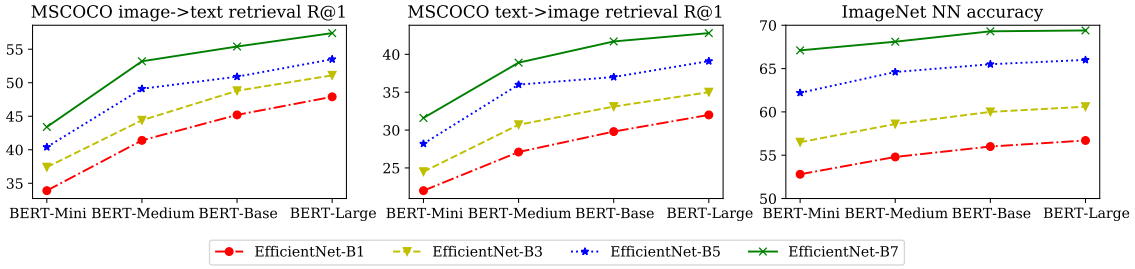
Figure 3. Zero-shot image-text retrieval and ImageNet KNN accuracy@1 with different image and text encoder sizes.

Table 7. Comparison between ALIGN-B7 and ALIGN-L2.

Image Encoder (+BERT-Large)	MSCOCO **zero-shot**		MSCOCO **fine-tuned**		ImageNet acc@1 **(fine-tuned)**
	image→text R@1	text→image R@1	image→text R@1	text→image R@1	
EfficientNet-B7	58.3	43.1	75.0	57.4	86.92
EfficientNet-L2	58.6	45.6	77.0	59.9	88.64

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.

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		ImageNet 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		ImageNet 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

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
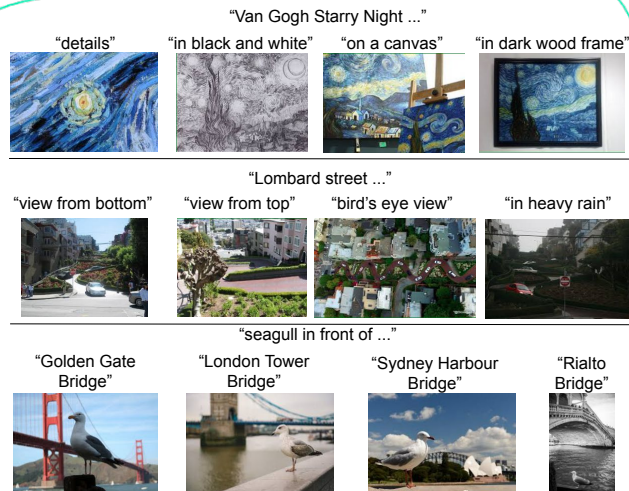
generalize to novel complex concepts.

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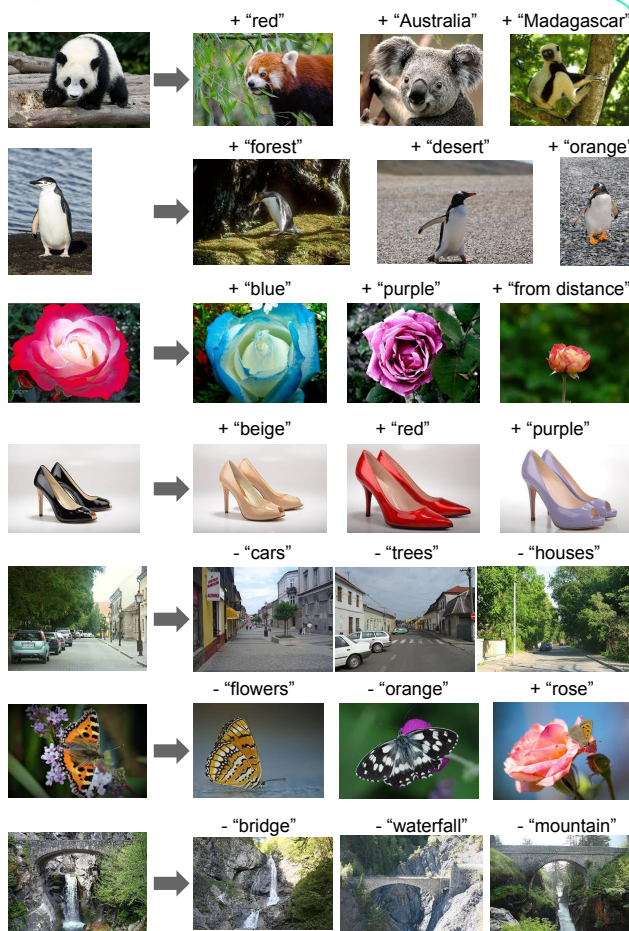

Figure 5. Image retrieval with image±text queries. We add (or subtract) text query embedding to (or from) the image query embedding, and then use the resulting embedding to retrieve relevant images using cosine similarity.

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 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 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. 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.

9. 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 de-

---

[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.

mographic 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

References

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.

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.

Hénaff, O. J., Srinivas, A., Fauw, J. D., Razavi, A., Doersch, C., Eslami, S. M. A., and van den Oord, A. Data-efficient image recognition with contrastive predictive coding. In *arXiv preprint arXiv:1905.09272*, 2019.

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, 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, 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*, 2019.

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.

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.

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.

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.

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.

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

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.

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.

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.