Learning Customized Visual Models with Retrieval-Augmented Knowledge

Haotian Liu[†§♠] Kilho Son[‡] Jianwei Yang[‡] Ce Liu[‡] Jianfeng Gao[‡] Yong Jae Lee[†¶] Chunyuan Li[‡¶♠]
[†] **University of Wisconsin–Madison** [‡] **Microsoft**
{lht,yongjaelee}@cs.wisc.edu {kilhoson,jianwyan,ce.liu,jfgao,chunyl}@microsoft.com
https://react-vl.github.io

Abstract

Image-text contrastive learning models such as CLIP have demonstrated strong task transfer ability. The high generality and usability of these visual models is achieved via a web-scale data collection process to ensure broad concept coverage, followed by expensive pre-training to feed all the knowledge into model weights. Alternatively, we propose **REACT**, ***RE****trieval-****A****ugmented* ***C****us****T****omization, a framework to acquire the relevant web knowledge to build customized visual models for target domains. We retrieve the most relevant image-text pairs (∼3% of CLIP pre-training data) from the web-scale database as external knowledge and propose to customize the model by only training new modularized blocks while freezing all the original weights. The effectiveness of* REACT *is demonstrated via extensive experiments on classification, retrieval, detection and segmentation tasks, including zero, few, and full-shot settings. Particularly, on the zero-shot classification task, compared with CLIP, it achieves up to 5.4% improvement on ImageNet and 3.7% on the* ELEVATER *benchmark (20 datasets).*

1. Introduction

It has been a fundamental research problem in computer vision (CV) to build a transferable visual system that can easily adapt to a wide range of downstream tasks. With remarkable advances in deep learning, a de facto solution to achieve this is to train deep neural networks on a large amount of data to pursue the so-called *generic* visual representations. This dates back to the standard supervised training on ImageNet [10], whose superb representation power is further demonstrated in BiT [23]/ViT [12] by scaling up the training to JFT300M [50]. Along the way, recent efforts have been applied to the popular image self-supervised learning [6, 16, 17] to reduce the demand for labeled data. The third approach is image-text contrastive learning trained on billion-scale web-crawled image-text pairs. Such models, like CLIP [43] and ALIGN [20], are able to achieve great performance on different downstream domains, without the need of any human labels.

Excellent empirical performance has been achieved with the above three pre-training methods, by following the well established two-stage *pre-training then adaptation* pipeline: model pre-training from scratch on large data, then model adaptation directly on downstream tasks. Specifically, the pre-trained models are adapted to downstream tasks by considering the available task-specific samples only: either evaluated in a zero-shot task transfer manner, or updated using linear probing (LP) [43], finetuning (FT) [27], or prompt tuning [44,71]. Following this two-stage pipeline, most research has reverted to the faith that building transferable visual systems is equivalent to developing more generic visual models by feeding all knowledge in the model pre-training stage. Therefore, the community has been witnessing a trend in exploring scaling success of pre-training model and data size with less care on the target domain, hoping that the model can adapt to any downstream scenario.

In this paper, we argue that the conventional two-stage pipeline above is over-simplified and less efficient, in achieving the goal of building a transferable visual system in real-world settings. Instead, we propose a *customization* stage in between the pre-training and adaptation, where customization is implemented by systematically leveraging retrieved external knowledge. The inspiration comes from how humans are specialized in society for better generalization: instead of trying to memorize all concepts, humans are trained/prepared in a relevant subject to master a certain skill, while maintaining the basic skills in pre-training.

To this end, we explore a systematic approach to acquire and learn with external knowledge sources from a large image-text corpus for model customization. The process of collecting external image-text knowledge is fully automatic without extra human annotation. The acquired knowledge typically contains richer information about the concept: relevant images that never appear in the downstream training

♠ core contribution; ¶ equal advising; § work initiated during an internship at Microsoft.

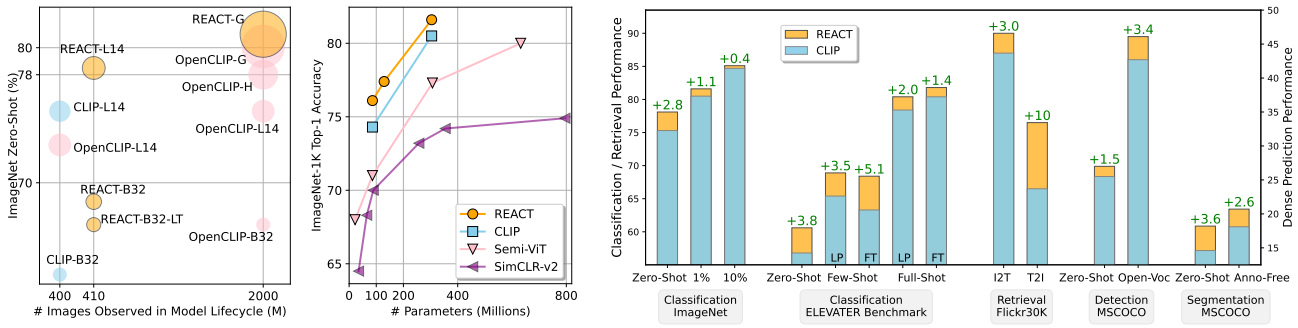


Figure 1. REACT achieves the best zero-shot ImageNet performance among public checkpoints (Left), achieves new SoTA on semi-supervised ImageNet classification in the 1% labeled data setting (Middle), and consistently transfer better than CLIP on across a variety of tasks, including ImageNet classification, zero/few/full-shot classification on 20 datasets in ELEVATER benchmark, image-text retrieval, object detection and segmentation (Right). Please see the detailed numbers and settings in the experimental section. For the left figure, circle size indicates model size.

and evaluation set, and richer text descriptions about concept semantics. Such multi-modal knowledge sources are generally available on the web, and further open-sourced like LAION [45, 46]. They cover a variety of domains, making it possible to develop customized visual models for task-level transfer. Similar retrieval-augmented intuitions have been exploited in computer vision for class-level transfer [32], but not yet for task-level transfer (similar to that of CLIP). Our main findings/contributions can be summarized as follows.

We propose to explore the potential of the web-scale image-text corpus as external knowledge to significantly improve task-level transfer performance on the target domain at an affordable cost. A simple and effective strategy is proposed. To begin with, we build a large-scale multi-modal indexing system to retrieve the relevant image-text pairs using CLIP features and approximate nearest neighbor search. For a CV problem, the task instruction is often sufficiently specified with text such as class names, which allows us to utilize them as queries to retrieve the relevant image-text pair knowledge from the indexing system. *No images from the CV problem are needed.* To efficiently build the customized visual model, we propose a novel modularized learning strategy: only updating the additional trainable weights on the retrieved knowledge, and freezing the original model weights. Hence, the model masters the new skill without forgetting basic skills.

The generality and effectiveness of the proposed customization strategy is demonstrated on four CV problems. We instantiate it with CLIP, and develop the customized visual models for image classification on ImageNet and 20 datasets in ELEVATER [27], image-text retrieval on COCO [30]/Flickr [41], as well as object detection and semantic segmentation on COCO [30]. The knowledge bases are considered as LAION [46] and larger web-crawled multi-modal data. The retrieval-augmented knowledge (∼3% image-text pairs compared with the original training data) significantly improves the model's zero-shot performance without the need of accessing any images on downstream tasks. See Figure 1 for highlighted results. For example, our ViT-L/14 checkpoint achieves 78.5% zero-shot accuracy on ImageNet [10], surpassing all public checkpoints from CLIP [43] and OpenCLIP [18], including those with larger model size and trained on a much larger LAION-2B [45]. The new customized models demonstrate higher few/full-shot performance than the generic model counterparts.

Our retrieval system, codebase, and pre-trained models are publicly available. To make this line of research more accessible, our retrieved subsets for both ELEVATER and ImageNet will also be made available, with an easy-to-use toolkit to download the subsets without storing the whole dataset locally. It poses a feasible direction for leveraging the ever-increasing data from the Internet for customized visual recognition, especially for the low-resource regimes.

2. Related Work

Vision-Language Pretraining. Learning transferable visual representations from natural language supervision is an emerging research area. The pioneering works of CLIP [43] and ALIGN [20] make use of contrastive learning to pretrain models on billion-scale web-crawled image-text pairs. There are an increasing number of studies to improve their generality from various modeling perspectives, including training objectives [11, 13, 14, 38, 65, 68], scaling techniques [9, 40, 65], data efficiency [25, 28], and leveraging multilingual correlations [9, 19]. In academia, several works demonstrate techniques to improve the learned semantic representations on datasets at a smaller scale (*e.g.* CC3M [47], CC12M [4], YFCC15M [43, 51]), by exploring pretraining on a unified image-text-label space [59], token-level contrastive loss [61], and auxiliary within-modality contrastive loss [29, 37, 58, 63]. Complementary to the above works, we build on top of existing pre-trained generic models, and aim to improve the model's performance by customizing them using retrieved *relevant* image-text pairs.

Retrieval-Augmented Models. In natural language processing, several works augment large language models with external data encoded with structured language and relation representations [3, 15, 22, 26, 31, 39, 66]. Motivated by retrieval-augmented models in NLP, several recent works leverage visual and / or textual knowledge to improve classification [32], question answering [7, 35, 56, 60], image generation [2, 8, 49, 72], and multi-modal tasks simultaneously [62]. RAC [32] improves long-tail classification by retrieving from a non-parametric memory consisting of pre-encoded images and text. K-LITE [48] enhances the text prompts with the retrieved external knowledge that is encoded in natural language. Our paper leverages the paired knowledge of image-text and aims to improve task transfer performance for core vision problems such as classification, retrieval, detection and segmentation.

Adaptation of Vision-Language models. CLIP demonstrates impressive zero-shot and linear probing performance on different downstream domains. Several works explore improving the domain adaptation performance on CLIP models. ELEVATER [27] leverages the text encoder outputs to initialize the task-specific linear head to improve the linear probe and finetuning performance of CLIP. Inspired by prompting techniques in NLP, recent works [44, 71] make use of learnable prompts that are trained on a few samples on downstream tasks. Similar to these works, this paper aims to improve CLIP's performance on downstream tasks, while making use of relevant image-text pairs data to improve the model's performance, without access to the downstream images. Furthermore, when downstream samples are available, they are complimentary to our method.

3. Retrieval-Augmented Customization

3.1. Preliminaries

Computer vision models have achieved strong transfer performance, when learning with large-scale image data only [17], image-label data [23] and/or image-caption data [43, 59, 67]. Without loss of generality, we follow [59] and define a unified triplet-wise format $(\boldsymbol{x}, \boldsymbol{t}, y)$ for image-text-label data, where $\boldsymbol{x} \in \mathcal{X}$ is an image, $\boldsymbol{t} \in \mathcal{T}$ is its language description, and $y \in \mathcal{Y}$ is a label indicating the index of the unique language description in the dataset. In a general form, the language description is a text sequence $\boldsymbol{t} = [t_1, \cdots, t_L]$. It ranges from simple category names representing visual concepts when L is small, to more free-form and semantic-rich sentences such as captions when L is relatively large.

A typical transfer learning pipeline follows the procedure of *pre-training then adaptation*: (i) With large-scale pre-training, an image encoder foundation model $f_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ is first trained to represent image $\boldsymbol{x}$ as a visual feature vector $\tilde{\boldsymbol{v}} \in \mathbb{R}^{P\times 1}$: $\tilde{\boldsymbol{v}} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$. For recent language-image models [43], a dual-encoder architecture is often employed, where an additional text encoder $f_{\boldsymbol{\phi}}(\boldsymbol{t})$ parameterized by $\boldsymbol{\phi}$ represents the sentence $\tilde{\boldsymbol{u}} \in \mathbb{R}^{P\times 1}$: $\tilde{\boldsymbol{u}} = f_{\boldsymbol{\phi}}(\boldsymbol{t})$. (ii) Given a downstream task, model adaptation is typically performed using the available task-specific information, or *task instruction* $\mathcal{I}$. For example, the task-level transfer of a language-image model is described as:

- *Zero-shot.* In a customized setting, the simplest task definition can be provided as a set of category names for visual recognition, leading to the task instruction $\mathcal{I}_0 = \{\boldsymbol{t}\}$. No training image $\boldsymbol{x}$ is available, not to mention the corresponding label y.
- *Few/Full-shot.* The users may spend annotation cost to curate N image-label pairs as the training instances, making the task instruction more specific, $\mathcal{I}_{\text{F}} = \{(\boldsymbol{x}_n, \boldsymbol{t}_n, y_n)\}_{n=1}^{N}$, which allows updating the image encoder model $f_{\boldsymbol{\theta}}$ for better adaptation performance.

In this paper, we assume there exists a web-scale image-text corpus as the external knowledge source $\mathcal{S} = \{(\boldsymbol{x}_m, \boldsymbol{t}_m)\}_{m=1}^{M}$, where M is the database size, *e.g.* 400M for LAION [46]. One may use the task instruction $\mathcal{I}$ as a query to seek additional relevant knowledge to build a more transferable visual system. Given the downstream task instruction $\mathcal{I}$ and an external knowledge source $\mathcal{S}$, our goal is to learn customized visual-semantic representations, which are readily transferable to the downstream task of interest, whose training and evaluation images are not observed during the customization process. To this end, we propose REACT. We illustrate the high-level idea in Figure 2, and describe the process as follows.

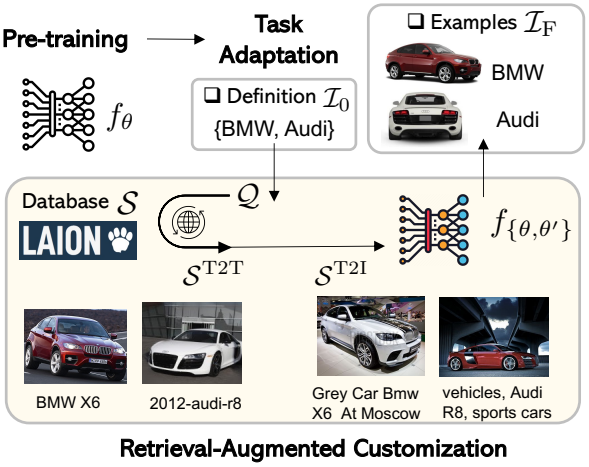


Figure 2. Illustration of the proposed REACT framework.

3.2. Multi-modal External Knowledge

Knowledge Base Construction. We explore web-scale image-text data as the multi-modal knowledge base $\mathcal{S}$ in this paper. Ideally, one may consider the entire web as the knowledge base, and use Google or Bing search to retrieve the relevant knowledge. We consider two large static datasets

with image-text pairs. To control the experiment complexity and ensure reproducibility, we use LAION-400M [46], a publicly available database with 400M pairs, for most of the experiments. To further study the scaling influence of the retrieval base, we conduct comparisons on Web-800M, a privately collected web database with 800M pairs.

To facilitate an efficient knowledge acquisition process, we use pre-trained contrastive models (*e.g.* CLIP) as the feature extractor, and build a cross-modal retrieval system using FAISS [21]. We use its Hierarchical Navigable Small World (HNSW) approximate k-NN lookup [34] to balance performance and efficiency. Please see appendix for more details. After the retrieval system is built on the designated retrieval pool, it can be efficiently used for retrieving relevant image-text pairs for *various* downstream domains.

Retrieval-Augmented Task Instruction. To facilitate the same interface for various customized visual tasks in the wild, it is desirable to have the same uniform task instruction schema. In NLP, all task instructions can follow the same uniform schema, composed of *task definition* and *positive/negative examples* [36, 55]. Here, the task definition defines a given task in natural language, completely specifying how an input is expected to be mapped to an output text. We note a coherence connection between this NLP task schema and the customized zero/few/full-shot CV settings in Section 3.1. Following a similar schema, the minimum requirement to specify a visual task is the task definition $\mathcal{I}_0$, where category names illustrate the target visual concepts in natural language. Though adding human-annotated examples is a natural way to clarify the task and yield the complete schema $\mathcal{I}_{\text{F}}$, extra cost is introduced.

It is of high interest to clarify the task using relevant examples, without human curating cost. Therefore, we propose to augment the task instruction with the retrieved examples from the external multi-modal knowledge base $\mathcal{S}$. For each concept $\boldsymbol{t} \in \mathcal{I}_0$ in a given task, we first represent it in natural language $\boldsymbol{q} = g_{prompt}(\boldsymbol{t})$ using the language prompt as in [43], through inserting the concept into a set of task-specific templates $\mathcal{P}$. The task definition is expanded in its natural language form:

$$\mathcal{Q} = \{\boldsymbol{q} \mid \boldsymbol{q} = g_{prompt}(\boldsymbol{t}), \forall \boldsymbol{t} \in \mathcal{I}_0, prompt \in \mathcal{P}\}. \quad (1)$$

Next, we perform our knowledge retrieval process to acquire the relevant image-text pair $\boldsymbol{s} = g_{retrieve}(\boldsymbol{q})$ from the source $\mathcal{S}$. Two types of retrieval processes are considered to acquire the top-K pairs:

- Text-to-Text (T2T) retrieval allows us to retrieve more relevant examples as they have a better match with our target concept. The T2T-retrieved set for $\mathcal{I}_0$ is:

$$\mathcal{S}^{T2T} = \{(\boldsymbol{x}, \boldsymbol{t}) \in \mathcal{S} : \underset{\boldsymbol{t} \in \mathbb{T}, |\mathbb{T}|=K}{\text{argmax}} f_{\boldsymbol{\phi}}(\boldsymbol{t})^\top f_{\boldsymbol{\phi}}(\boldsymbol{q}), \forall \boldsymbol{q} \in \mathcal{Q}\} \quad (2)$$

- Text-to-image (T2I) retrieval allows us to have more diversity in the text descriptions in our retrieved examples. The T2I-retrieved set for $\mathcal{I}_0$ is:

$$\mathcal{S}^{T2I} = \{(\boldsymbol{x}, \boldsymbol{t}) \in \mathcal{S} : \underset{\boldsymbol{x} \in \mathbb{X}, |\mathbb{X}|=K}{\text{argmax}} f_{\boldsymbol{\theta}}(\boldsymbol{x})^\top f_{\boldsymbol{\phi}}(\boldsymbol{q}), \forall \boldsymbol{q} \in \mathcal{Q}\} \quad (3)$$

Both $\mathcal{S}^{T2T}$ and $\mathcal{S}^{T2I}$ are retrieved examples to augment the task definition $\mathcal{I}_0$, without accessing the images in the training or validation set of the task. Compared to $\mathcal{I}_{\text{F}}$, they are "free" external knowledge to clarify the task and can be used to build a more transferable system.

3.3. Model Customization

After retrieving the relevant multi-modal examples, one may employ the naive customization solution by fine-tuning the full-model initialized from pre-trained weights, as in Figure 3(a). Alternatively, we propose an affordable solution to endow pre-trained models with a new capability to leverage this external knowledge. The pre-trained generic visual models have gained strong transfer abilities and access to a large amount of internal knowledge stored in the model weights. We freeze the weights of these models so that their initial capacity remains unchanged. To bridge these pre-trained models harmoniously to the customized domain, we consider *locked-text gated-image tuning* with the following two techniques, illustrated in Figure 3(d).

Modularized Image Encoder. In order to provide sufficient expressivity to the model and make it able to adapt well on retrieved knowledge, we insert gated self-attention dense blocks in between the original layers of the image encoder, and train the new blocks from scratch. Those blocks are made of a self-attention layer, that attends the early layer inputs, followed by an extra dense feed-forward layer. Please see a visual illustration of this gated block in the rightmost of Figure 3(d). We denote the parameters of all new modules as $\boldsymbol{\theta}'$. This design is inspired by the gated cross-attention-dense blocks in Flamingo [1] and frozen multi-modal model [52]. The difference is that the trainable module is introduced in Flamingo to enable cross-modal conditioning, while we adapt it for model growing in new customized domains.

Frozen Text Encoder. The text encoder in language-image contrast models represents the task semantic space. To maintain it, we propose *locked-text tuning*, which freezes the text model weights so that the generic task encoding knowledge remains locked; see Figure 3(c). This is in contrast with *locked-image tuning* (LiT) [68] in Figure 3(b), where the image encoder is frozen and the text encoder is fine-tuned, which teaches a text model to read out good representations from a pre-trained image model for new tasks.

We extract the normalized feature vectors in a hyper-sphere using $\boldsymbol{u}_i = \frac{f_{\{\boldsymbol{\theta}, \boldsymbol{\theta}'\}}(\boldsymbol{x}_i)}{\|f_{\{\boldsymbol{\theta}, \boldsymbol{\theta}'\}}(\boldsymbol{x}_i)\|}$ and $\boldsymbol{v}_j = \frac{f_{\boldsymbol{\phi}}(\boldsymbol{t}_j)}{\|f_{\boldsymbol{\phi}}(\boldsymbol{t}_j)\|}$. To customize the model wrt task definition $\mathcal{I}_0$, we update $\boldsymbol{\theta}'$ using a bidirectional learning objective between images and

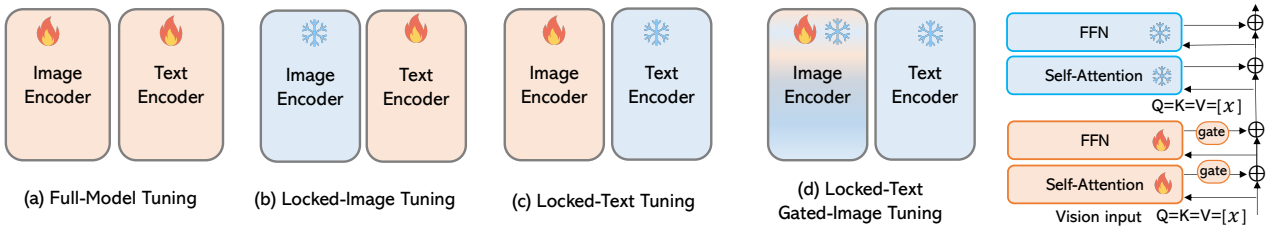


Figure 3. Illustrative comparisons across different model tuning methods. (a) and (b) are existing baseline tuning methods. For model customization in a target domain, we found that (c) and (d) work better. One layer of the proposed modularized image encoder in locked-text gated-image tuning is illustrated in right side.

language on the retrieved knowledge pool $\mathcal{S}^{T2T}$ and/or $\mathcal{S}^{T2I}$:

$$\min_{\{\boldsymbol{\theta}'\}} \ \mathcal{L}_{\mathrm{C}} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}, \text{ with } \mathcal{B} \sim \mathcal{S}^{T2T} \text{ or } \mathcal{S}^{T2I} \tag{4}$$

$$\mathcal{L}_{i2t} = -\sum_{i\in\mathcal{B}} \frac{1}{|\mathcal{P}(i)|} \sum_{k\in\mathcal{P}(i)} \log \frac{\exp(\tau \boldsymbol{u}_i^\top \boldsymbol{v}_k)}{\sum_{j\in\mathcal{B}} \exp(\tau \boldsymbol{u}_i^\top \boldsymbol{v}_j)} \quad \text{and} \tag{5}$$

$$\mathcal{L}_{t2i} = -\sum_{j\in\mathcal{B}} \frac{1}{|\mathcal{Q}(j)|} \sum_{k\in\mathcal{Q}(j)} \log \frac{\exp(\tau \boldsymbol{u}_k^\top \boldsymbol{v}_j)}{\sum_{i\in\mathcal{B}} \exp(\tau \boldsymbol{u}_i^\top \boldsymbol{v}_j)} \tag{6}$$

where τ is a temperature hyper-parameter controlling the strength of penalties on hard negative samples, and $\mathcal{P}(i) = \{k|k \in \mathcal{B}, \boldsymbol{v}_k^\top \boldsymbol{v}_i \geq \gamma\}$, $\mathcal{Q}(j) = \{k|k \in \mathcal{B}, \boldsymbol{v}_k^\top \boldsymbol{v}_j \geq \gamma\}$. We set $\gamma = 0.9$ for classification tasks to force image-text pairs sharing the similar text to be positive. Note (4) is a general form; it reduces to UniCL [59] when $\gamma = 1.0$; it further reduces to the training objective of CLIP [43] when there is a one-to-one mapping between an image and its paired caption in a batch, *i.e.* $\mathcal{P}(i) = \{i\}$ and $\mathcal{Q}(j) = \{j\}$.

In our empirical study we find that locked pre-trained image and text encoders with trainable gated modules in image encoder work best. Once the customized visual models are trained with the retrieved knowledge, we transfer it to the downstream domain for zero/few/full-shot evaluation.

4. Experiments

In this section, we conduct experiments to answer three research questions: (1) What are the unique advantages of retrieval-augmented image-text knowledge for task transfer? (2) How does our design choice of locked-text gated-image tuning compare to existing methods for model customization? (3) Is customization still beneficial in settings where the training data in downstream tasks are observed, *i.e.*, in few-shot or full-shot settings? (4) Does customization scales well to dense prediction tasks like detection/segmentation?

We evaluate our models on four CV problems: image classification, image-text retrieval, object detection, and semantic segmentation. We first consider ImageNet [10] for zero-shot task transfer. We then further evaluate our model on ELEVATER [27], which is an open-set image classification benchmark that contains 20 datasets. We also conduct experiments on image-text retrieval with MSCOCO [30] and Flickr [64] dataset. Finally, we evaluate on object detection and semantic segmentation with MSCOCO [30] dataset.

One of the most intriguing benefits of REACT is that it does not need access to any images from the downstream task. Therefore, we first evaluate on task-level zero-shot transfer, which requires no images in the target to be observed [27, 43, 48]. This setting is different from traditional class-level zero-shot [57], where both the category and images in evaluation should not be observed in training. We argue that ImageNet concepts have been observed in CLIP (Sec. 2.2 of [43]) and other web-scale trained models [28], as WordNet synsets and common words in English Wikipedia are explicitly added in the query list when searching for (image, text) pairs in their training data construction process.

4.1. Image Classification

4.1.1 Zero-Shot on ImageNet-1K

As shown in Table 1, by customizing the generic model CLIP/OpenCLIP on 10M retrieved image-text pairs from LAION-400M, REACT achieves a significant and consistent gain (up to 5.4%) on zero-shot image classification on ImageNet-1K, with different backbones and original pre-training datasets. There are three interesting findings.

F1: REACT *can benefit from model's own pre-training data.* Compared to OpenCLIP [18] (ViT-B/32) trained on LAION-400M, by training on 10M relevant pairs from the *same* LAION-400M dataset, REACT improves over OpenCLIP by 3.5%. Note that the model purely uses the image-text pairs that it has seen during its pre-training, and does *not* see any extra data. This shows that REACT can more adequately adapt to the target domain during the model customization stage, suggesting a favorable property that no new data is required for customization.

F2: REACT *efficiently explores new image-text sources, even for large models.* We customize CLIP [43] ViT-L/14 on 10M retrieved relevant image-text pairs, and the model achieves a 2.8% improvement to 78.1%. This surpasses the checkpoint with a much larger ViT-H/14 backbone and trained on a much larger LAION-2B dataset. This suggests that REACT is a more sample-efficient approach to improve the model performance on the domain-of-interest.

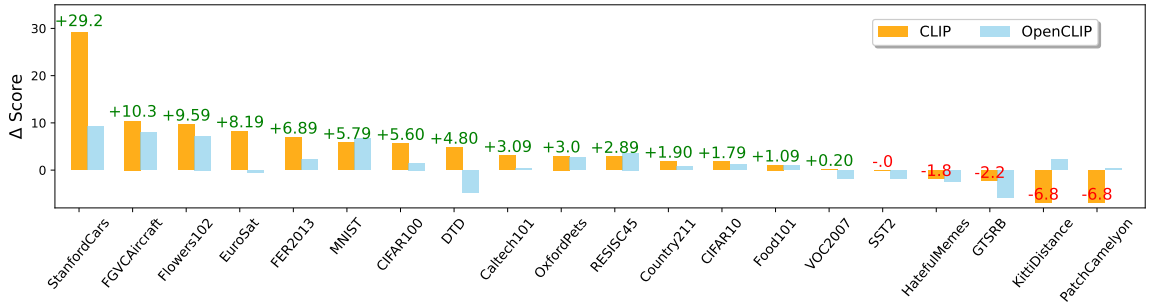


Figure 4. Zero-shot comparison on ELEVATER ICinW 20 datasets. REACT (B32) improves over the base checkpoints on most datasets.

f_θ	Pretrain Data	Retrieved Data Dataset	Retrieved Data Size	Method	ImageNet-1K Zero-Shot
B/32	WIT-400M	–	–	CLIP	63.2
		L-400M	10M	REACT	**68.6** (+5.4)
	LAION-400M	–	–	OpenCLIP	62.9
		L-400M	10M	REACT	**66.4** (+3.5)
L/14	WIT-400M	–	–	CLIP	75.3
		L-400M	6M	REACT	**77.6** (+2.3)
		L-400M	10M	REACT	**78.1** (+2.8)
		W-800M†	6M	REACT	**78.5** (+3.2)
	LAION-400M	–	–	OpenCLIP	72.8
	LAION-2B	–	–	OpenCLIP	75.3
H/14	LAION-2B	–	–	OpenCLIP	78.0
G/14	LAION-2B	–	–	OpenCLIP	80.1
		L-2B	12M	REACT	**81.0** (+0.9)

Table 1. Comparison of zero-shot task transfer with public checkpoints from CLIP [43] and OpenCLIP [18]. LAION [45, 46] is abbreviated as "L" in the table. Web-800M†: a privately collected web database with 800M image-text pairs. By continue pretraining on only ∼10M retrieved data, REACT outperforms *all* public CLIP/OpenCLIP checkpoints.

F3: Scaling up the retrieval pool increases performance. We perform REACT in a privately collected dataset with over 800M pairs, and train a customized model on 6M retrieved pairs. The performance is increased to 78.5%, yielding 0.9% gain compared with 6M pairs retrieved from LAION-400M. This suggests that REACT scales well with the larger retrieval pool. It showcases REACT as a cost-efficient approach to leveraging the ever-increasing web image-text corpus.

4.1.2 Zero-, Few-, and Full-Shot on ELEVATER

As a proxy for performing vision tasks for many customized scenarios in the wild, we consider the *image classification in the wild* (ICinW) benchmark in ELEVATER [27]. It consists of 20 datasets from a diverse selection of domains and covers a wide range of concepts, totaling 1151 classes.

We perform multi-modal knowledge retrieval for 20 datasets together – the retrieved samples are around 10M image-text pairs in total, on which one single customized visual model is trained. After the process, we feed the customized model to different downstream tasks separately. For each downstream dataset, we use the official ELEVATER toolkit to obtain the train/val/test splits, and perform zero-shot, few-shot, and full-shot evaluation.

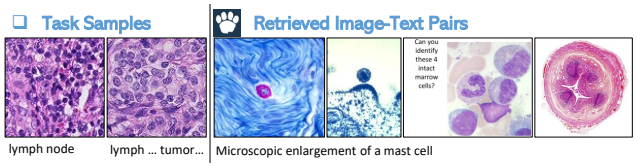
Method	Zero-Shot	Few-shot LP	Few-shot FT	Full-shot LP	Full-shot FT
CLIP	56.8	65.4	63.3	78.4	80.4
REACT	**60.6**	**68.9**	**68.4**	**80.4**	**81.8**
Gains	(+3.8)	(+3.5)	(+5.1)	(+2.0)	(+1.4)

Table 2. The average scores of image classification performance on 20 datasets in ELEVATER. REACT consistently outperforms CLIP in both data-limited and data-rich regimes.

We report the average scores in Table 2. It achieves 3.8% improvement in the zero-shot setting, even when we do not perform a separate customization for different datasets. This demonstrates the robustness of our customization process. Further, we see the consistent improvement in few-shot and full-shot settings, including both linear probe (LP) and fine-tuning (FT). This result is encouraging, as it demonstrates that when we have access to some or all data from the downstream task, the proposed model customization stage remains beneficial. Therefore, *we advocate model customization process in both data-limited and data-rich settings*.

Breakdown Analysis. Next, we ask why does the retrieved image-text knowledge improve the zero-shot task transfer performance on a broad range of datasets? We compare the breakdown performance on all 20 datasets in Figure 4 for the zero-shot settings. Out of 20 datasets, the retrieval-augmented knowledge shows superior/comparable/inferior performance to the baseline on 15/1/4 datasets for CLIP and 14/0/6 datasets for OpenCLIP, respectively. Most of the improved and failure datasets are consistent for both checkpoints. For the top two datasets that gains the most, *i.e.* StanfordCars and FGVC Aircraft, relevant image-text knowledge is retrieved from the web-crawled data LAION-400M to describe the concepts; see Fig. 5a. Interestingly, this observation is complementary to K-LITE [48], which failed on these two datasets, because no knowledge was extracted from Wiktionary for them, as it often requires domain-specific knowledge and even visual knowledge to best define a car brand (*e.g.* BMW X6 SUV or Audi R8) or an aircraft model type (*e.g.* DC-10 or A321).

(a) Success examples. The two datasets with largest improvement in Fig. 4: Stanford-Cars [24] and FGVC-Aircraft [33]. There is a high concept converage for these datasets in LAION, resulting in a relevant and diverse retrieved set.



(b) Failure case. The dataset with the largest degradation in Fig. 4: PatchCamelyon [54]. LAION-400M has a low concept coverage on this domain, and the retrieved samples are in a different distribution from the target set.

Figure 5. Success and failure cases in ELEVATER benchmark. We show class name and the caption of the first retrieved image-text pairs, others are similar and omitted due to limited space.

Limitations. As shown in Fig. 4, REACT struggles on the PatchCamelyon dataset, a cancer cell recognition benchmark. We visualize the retrieved samples and the samples from the original training set in Fig. 5b. The retrieved images are either instruction photos and from another sensing method, which exhibits a different visual distribution from PatchCamelyon. This suggests the importance of ensuring the retrieval quality for the domain-of-interest.

4.2. Image-Text Retrieval

To demonstrate the generality of REACT, we consider Flickr30K [64] and MSCOCO [30] image-text retrieval tasks, in both zero-shot and full-shot settings. We use the standard image-text contrastive objective [43]. For image-text retrieval task, following [20, 43], we use the CLIP-L/14 with 336x336 input resolution in both zero-shot, customization, and fine-tuning stage. We use the captions from MSCOCO as queries to retrieve 6M image-text pairs and perform customization. Note that *none* of the caption queries are used in the model training stage.

As shown in Table 3, REACT improves the generic CLIP counterparts on both zero-shot and full-shot retrieval for Flickr30K and MSCOCO datasets. The gain on zero-shot task transfer is large. On Flickr30K, it achieves 3.4%/10.0% recall improvement for I2T and T2I retrieval, respectively. Afer fine-tuning on full training data, REACT still improves over the baseline slightly. It provides another piece of evidence for REACT in data-rich settings. Furthermore, we conduct the same customization procedure of REACT on a large checkpoint Bletchley [53] with 864M parameters, and observe consistent gains over both datasets. It demonstrates that REACT scales well with model size on retrieval tasks.

	Method	Flickr30K Img → Text R@1	Flickr30K Img → Text R@5	Flickr30K Text → Img R@1	Flickr30K Text → Img R@5	MSCOCO Img → Text R@1	MSCOCO Img → Text R@5	MSCOCO Text → Img R@1	MSCOCO Text → Img R@5
Zero-Shot	ImgBert [42]	70.7	90.2	54.3	79.6	44.0	71.2	32.3	59.0
	ALIGN [20]	88.6	98.7	75.7	**93.8**	58.6	83.0	45.6	69.8
	CLIP [43]	88.0	98.7	68.7	90.6	58.4	81.5	37.8	62.4
	CLIP†	87.0	98.3	66.5	88.0	59.2	80.7	37.8	62.4
	REACT	**90.4**	**99.1**	**76.5**	**93.7**	**63.3**	**85.1**	**47.5**	**72.0**
	Bletchley†	90.8	98.2	78.0	94.0	66.7	85.6	48.9	72.7
	REACT	**92.1**	**98.7**	**79.2**	**94.7**	**67.7**	**85.9**	**50.5**	**74.4**
Fine-tuned	GPO [5]	88.7	98.9	76.1	94.5	68.1	90.2	52.7	80.2
	ALIGN [20]	95.3	**99.8**	84.9	97.4	77.0	93.5	59.9	83.3
	CLIP†	96.4	**99.8**	86.5	**97.9**	78.3	93.8	60.9	83.8
	REACT	**96.6**	**99.9**	**86.8**	**98.0**	**78.7**	**94.0**	**61.1**	**84.1**

Table 3. Image-text retrieval results on Flickr30K [41] and MSCOCO [30] datasets. CLIP†, Bletchley†: our evaluation.

	Pretrain Method	Backbone	Region Proposals	MSCOCO AP_{50} Novel	Base	All
Zero-Shot	CLIP	ResNet-50	GT	58.6	58.2	58.3
	REACT	ResNet-50	GT	**58.9** (+0.3)	**59.4** (+1.2)	**59.3** (+1.0)
	CLIP	ResNet-50	RPN	29.7	24.0	25.5
	REACT	ResNet-50	RPN	**31.6** (+1.9)	**25.4** (+1.4)	**27.0** (+1.5)
OVD	CLIP	ResNet-50	–	14.2	52.8	42.7
	REACT	ResNet-50	–	**20.6** (+6.4)	**55.1** (+2.3)	**46.1** (+3.4)

Table 4. Zero-shot and open-vocabulary object detection results on MSCOCO [30] dataset using RegionCLIP [69] pipeline.

4.3. Dense Prediction Tasks

Although REACT is optimized with the image-level contrastive loss during the customization stage, we find it beneficial for dense prediction tasks as well. We showcase its application to dense prediction tasks on object detection and semantic segmentation.

Object Detection. For object detection, we choose the state-of-the-art RegionCLIP [69] as our framework. We conduct experiments in two settings: zero-shot inference and open-vocabulary object detection (OVD) on MSCOCO dataset. We perform the model customization following the same setting as Sec. 4.2. Following RegionCLIP, we conduct experiments on ResNet50 backbone. The results are shown in Table 4. REACT consistently improves over CLIP checkpoint under all settings.

For zero-shot inference, when ground-truth region proposal is used, REACT improves over CLIP by +1.0 on overall AP50; when the pretrained RPN is used, REACT demonstrates +1.5/+1.4/+1.9 AP50 improvements on novel, base, and all classes, respectively.

For OVD, we can see that with the REACT, the detector yields improved performance on base with +2.3 AP50, and

Method	mIoU zero-shot	mIoU w/ refine	mIoU w/ finetune
MaskCLIP [70]	12.5	14.6	18.1
REACT (Locked-Text)	**14.4** (+1.9)	**18.2** (+3.6)	**20.7** (+2.6)
REACT (Locked-Text Gated-Image)	**14.5** (+2.0)	**16.3** (+1.7)	**19.2** (+1.1)

Table 5. Zero-shot and annotation-free semantic segmentation results on COCO Stuff [30] using MaskCLIP [70] (ViT-B/16).

importantly, it significantly improves novel categories with +6.4 AP50. This suggests that the injected knowledge during the model customization stage improves the learned fine-grained visual feature that is beneficial to both seen and unseen categories for object detection, when the downstream coarse-grained data is available. This is favored, because (1) the weakly-supervised data such as the coarse-grained image-text pairs requires much less human annotation cost than fine-grained bounding box annotation, (2) the paired data in REACT is free, as it is retrieved from the web, where COCO image-text pairs are not used in customized training.

Semantic Segmentation. For semantic segmentation, we choose the state-of-the-art MaskCLIP [70] as the framework. It investigates three evaluation settings for segmentation. First, it makes use of the pretrained CLIP checkpoint to discover the alignment between grid visual features and the text prompt features, so as to perform zero-shot semantic segmentation. Second, to further improve the performance, MaskCLIP [70] proposes two techniques for refining its zero-shot predictions: key smoothing and prompt denoising. Lastly, when the training images are available, without the need to access the training labels, it further proposes MaskCLIP+ [70] to perform full-shot finetuning on the target training set using pseudo-labels. Following MaskCLIP [70], we use ViT-B/16 checkpoints, and use their official code base to train and evaluate models. We report results in Table 5.

On all of the three settings, REACT demonstrates improvements over the MaskCLIP. Notably, when refinement techniques are used, REACT with locked-text tuning demonstrates a significant 3.6% gain in mIoU. Surprisingly, without seeing the downstream COCO images, the zero-shot evaluation of REACT (18.2) even slightly outperforms MaskCLIP+ (18.1), which is finetuned on the downstream training COCO images with self-training.

Summary. These results are encouraging, as it shows that the customized knowledge from REACT transfers well to dense prediction tasks like detection and segmentation.

4.4. Ablation Studies

We ablate REACT on ImageNet with CLIP ViT-B/32 checkpoint, with 10M retrieved image-text pairs from LAION-400M. See more ablations in supplementary.

Tuning strategy. We ablate the design choices in the model customization stage: (1) direct tuning the pre-trained weights *vs*. training gated blocks from scratch; (2) updating visual *vs*. text encoder. We report results in Table 6. First, a frozen text

	Method	Visual	Text	# Train.	IN1K Acc.	COCO R@1 T2I	COCO R@1 I2T
Direct	CLIP [43]	✗	✗	–	63.2	48.8	29.9
	Locked-Image [68]	✗	✓	63.4M	63.7	50.5	34.2
	Locked-Text	✓	✗	88.1M	66.9	51.1	36.1
	Full-model	✓	✓	151.3M	64.3	54.3	37.9
Gated		✗	✓	18.9M	63.0	49.5	33.5
	Locked-Text Gated-Image	✓	✗	42.5M	**68.6**	53.4	38.1
		✓	✓	89.8M	**68.7**	54.3	39.9

Table 6. Ablation: tuning strategy. For our model customization purpose, we advocate locked-text (gated-image) tuning methods in gray rows. ✓: trainable, ✗: locked.

encoder consistently outperforms a frozen visual encoder. We conjecture the phenomenon is due to that the retrieved texts have a much more limited space, comparing to text space in the original pre-training set (*e.g*. LAION-400M), as the concepts are limited to the query classes from the target domain. Therefore, fine-tuning the text encoder may tend to collapse the pre-trained semantic space.

We advocate two tuning methods for model customization. Locked-text gated-image tuning has a strong adaptation power, and is efficient in the model customization stage, with fewer trainable parameters. Locked-text tuning is also an effective way of customizing the models to downstream tasks, without the need of adding extra parameters. By default, we use gated blocks for its superior performance and efficiency.

Retrieval size. We observe that training with a small retrieval size is more likely to overfit. We find that a larger retrieval size generally yields better performance, and saturates at around 6-10M image-text pairs.

Retrieval Size	0	1M	3M	6M	10M
ImageNet-1K Accuracy	63.2	64.8	66.9	**68.6**	**68.6**

5. Conclusion

We presented REACT, a plug-and-play framework for leveraging large-scale image-text corpus as external knowledge to efficiently customize models on downstream tasks. Extensive experiments demonstrate its generality and effectiveness in image classification, image-text retrieval, object detection, and semantic segmentation, on more than 20 different downstream datasets. We highly advocate the model customization stage for building more transferable visual system for different downstream tasks.

Acknowledgement. This work was supported in part by NSF CAREER IIS2150012, NASA 80NSSC21K0295, and Institute of Information & communications Technology Planning & Evaluation(IITP) grants funded by the Korea government(MSIT) (No. 2022- 0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration) and (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 4

[2] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-augmented diffusion models. *arXiv preprint arXiv:2204.11824*, 2022. 3

[3] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*, 2021. 3

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 2

[5] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15789–15798, 2021. 7

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1

[7] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*, 2022. 3

[8] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 3

[9] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 5

[11] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022. 2

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

[13] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *arXiv preprint arXiv:2204.14095*, 2022. 2

[14] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurams, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022. 2

[15] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*, 2020. 3

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 3

[18] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. *Zenodo*, 2021. 2, 5, 6

[19] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125*, 2021. 2

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

[21] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 4

[22] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019. 3

[23] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *ECCV*, 2020. 1, 3

[24] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 7

[25] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. *arXiv preprint arXiv:2209.13430*, 2022. 2

[26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*, 2020. 3

[27] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. ELEVATER: A

benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS Track on Datasets and Benchmarks*, 2022. 1, 2, 3, 5, 6

[28] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2, 5

[29] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *CVPR*, 2023. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 5, 7, 8

[31] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-BERT: Enabling language representation with knowledge graph. In *AAAI*, 2020. 3

[32] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6959–6969, 2022. 2, 3

[33] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, arXiv, 2013. 7

[34] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018. 4

[35] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In *CVPR*, 2021. 3

[36] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021. 4

[37] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021. 2

[38] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *arXiv preprint arXiv:2206.02770*, 2022. 2

[39] Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*, 2019. 3

[40] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv: 2111.10050*, 2021. 2

[41] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2, 7

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

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[44] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Prefix conditioning unifies language and label supervision. *arXiv preprint arXiv:2206.01125*, 2022. 1, 3

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2, 6

[46] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2, 3, 4, 6

[47] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2

[48] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, Kurt Keutzer, Trevor Darrell, and Jianfeng Gao. K-LITE: Learning transferable visual models with external knowledge. In *NeurIPS*, 2022. 3, 5, 6

[49] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022. 3

[50] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017. 1

[51] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016. 2

[52] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 2021. 4

[53] Turing Bletchley. https://www.microsoft.com/en-us/research/blog/turing-bletchley-a-universal-image-language-representation-model-by-microsoft/. 7

[54] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital

pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018. 7

[55] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2022. 4

[56] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-modal answer validation for knowledge-based VQA. *arXiv preprint arXiv:2103.12248*, 2021. 3

[57] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *PAMI*, 2018. 5

[58] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 2

[59] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Lu Yuan, Ce Liu, and Jianfeng Gao. Unified contrastive learning in image-text-label space. *CVPR*, 2022. 2, 3, 5

[60] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of GPT-3 for few-shot knowledge-based VQA. *arXiv preprint arXiv:2109.05014*, 2021. 3

[61] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[62] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022. 3

[63] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *European Conference on Computer Vision*, pages 69–87. Springer, 2022. 2

[64] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5, 7

[65] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[66] Wenhao Yu, Chenguang Zhu, Yuwei Fang, Donghan Yu, Shuohang Wang, Yichong Xu, Michael Zeng, and Meng Jiang. Dict-bert: Enhancing language model pre-training with dictionary. *arXiv preprint arXiv:2110.06490*, 2021. 3

[67] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[68] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2, 4, 8

[69] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. *CVPR*, 2022. 7

[70] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 8

[71] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 3

[72] Yufan Zhou, Chunyuan Li, Changyou Chen, Jianfeng Gao, and Jinhui Xu. Lafite2: Few-shot text-to-image generation. *arXiv preprint arXiv:2210.14124*, 2022. 3