# LiT🔥: Zero-Shot Transfer with Locked-image text Tuning

Xiaohua Zhai*† Xiao Wang* Basil Mustafa* Andreas Steiner* Daniel Keysers Alexander Kolesnikov Lucas Beyer*†
Google Research, Brain Team, Zürich

## Abstract

*This paper presents* contrastive-tuning*, a simple method employing contrastive training to align image and text models while still taking advantage of their pre-training. In our empirical study we find that locked pre-trained image models with unlocked text models work best. We call this instance of contrastive-tuning "Locked-image Tuning" (LiT), which just teaches a text model to read out good representations from a pre-trained image model for new tasks. A LiT model gains the capability of zero-shot transfer to new vision tasks, such as image classification or retrieval. The proposed LiT is widely applicable; it works reliably with multiple pre-training methods (supervised and unsupervised) and across diverse architectures (ResNet, Vision Transformers and MLP-Mixer) using three different image-text datasets. With the transformer-based pre-trained ViT-g/14 model, the LiT model achieves 84.5% zero-shot transfer accuracy on the ImageNet test set, and 81.1% on the challenging out-of-distribution ObjectNet test set.*

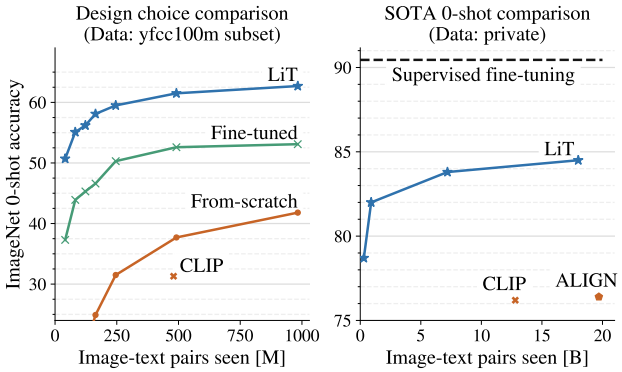


Figure 1. Comparison to the previous SOTA methods. **Left**: results on public YFCC100m subset, with from-scratch, fine-tuned from a pre-trained image model, and LiT with a pre-trained image model. The proposed LiT improves over 30% ImageNet zero-shot transfer accuracy on YFCC100m subset. **Right**: results on privately gathered data, LiT halves the gap between previous from-scratch methods CLIP [45], ALIGN [30] and supervised fine-tuning [12, 68].

## 1. Introduction

Transfer learning [44] has been a successful paradigm in computer vision [32, 33, 42]. Zero-shot learning [35, 36, 65] is an alternative approach aiming to develop models that can handle a new task without task-specific data or adaptation protocols. Recently it was demonstrated that web-sourced paired image-text data can be used to pre-train strong models for zero-shot transfer [30, 45]. Zero-shot *transfer* differs from classical zero-shot learning in that the transfer setup may see relevant supervised information during pre-training; it is zero-shot insofar as no supervised examples are used during the transfer protocol. GPT-3 [3] explored a similar zero-shot transfer setup using model prompting via natural language.

In [30, 45] authors propose a contrastive learning framework where an image model (or image tower) is trained simultaneously with a text model (or text tower). Both towers are trained to minimize a contrastive loss, which encourages representations of paired images and texts to be similar, and representations of non-paired images and texts to be dissimilar. At test time, the resulting model can be used for zero-shot image classification by comparing the image embedding with embeddings of textual class descriptions.

In this paper, we adopt a contrastive learning framework and propose a more data- and compute-efficient strategy named *contrastive-tuning*. The key idea is to tune the text tower using image-text data, while using a pre-trained, strong image model as the image tower. During training, both towers' weights can be locked or unlocked, leading to different design choices that are illustrated in Figure 2. Specifically, we find that locking the image tower works best, as shown in Figure 1. We call this specific instance of contrastive-tuning "Locked-image Tuning" (LiT), which just teaches a text model to read out suitable representations from a pre-trained image model. LiT achieves better results compared with the from-scratch CLIP [45] or ALIGN [30] models. With the pre-trained model ViT-g/14 [68], LiT achieves 84.5% zero-shot transfer accuracy on ImageNet, halving the gap between previous best zero-shot transfer re-

*equal technical contribution, †equal advising

sults [30,45] and supervised fine-tuning results [12,68]. The best LiT model also sets new state-of-the-art on several out-of-distribution (OOD) ImageNet test variants, compared to previous supervised and unsupervised methods. For example, it achieves 81.1% accuracy on the challenging ObjectNet test set [1], outperforming the previous state-of-the-art method [45] by 7.8%.

We believe the reason that LiT works well lies in its decoupling of data sources and techniques for learning image descriptors and vision-language alignment. Image-text data can be great for learning correspondences between natural language and the visual world, but, at the same time, it may not be precise and clean enough to result in state-of-the-art image descriptors. In this paper we carefully investigate this hypothesis and support it with empirical evidence.

The proposed LiT works with both supervised and self-supervised pre-trained models. We verify LiT across three image-text datasets, with Vision Transformer [20], ResNet [32], and MLP-Mixer [60] architectures. We also show that with a self-supervised pre-trained model, i.e. DINO [4] or MoCo-v3 [10], LiT achieves better performance compared to from-scratch contrastive-learning.

Another contribution of this paper is the proposed recipe for high-performance zero-shot models that can be trained using only modest computational resources and public datasets. By re-using already pre-trained models (e.g. publicly released in the literature), the computational resources used to train the image models can be amortized. Furthermore, we explore publicly available datasets such as YFCC100m [59] and CC12M [5]. Combined with the computational efficiency, we hope to facilitate contributions from a wider audience to research in zero-shot transfer.

## 2. Related work

This work is closely related to a vast amount of literature on *transfer learning* in vision [44, 58]. The main idea of transfer learning is to leverage already pre-trained models to solve a new task better and faster, as opposed to less efficient training from-scratch. This paradigm is usually implemented as a two-step procedure: (1) pre-train (once) an initial model on a large dataset of images that are (weakly)-labeled or using self-supervised losses and (2) fine-tune the pre-trained model for a task of interest using supervised data. In the context of modern deep learning, many earlier works [19, 32, 33, 47] used supervised pre-training to learn transferrable feature representations, with the Vision Transformer revisiting and improving this approach [20, 68]. It was shown that scaling up model and dataset sizes simultaneously leads to dramatic improvements in transfer effectiveness [20, 32, 68] and robustness [17]. Crucially, large pre-trained models exhibit outstanding capabilities in learning in the low-data (few-shot) regime [8, 20, 32].

Still, collecting task-specific data and fine-tuning large pre-trained models remains time-consuming and potentially costly in many realistic scenarios. *Zero-shot transfer* is an alternative paradigm that sidesteps the fine-tuning stage entirely and performs classification solely based on a description of the target classes. Early works demonstrated how to train zero-shot classifiers based on attributes [35] or numerical descriptors [36]. Another approach, which we adopt in this work, is to learn an alignment between image and text embedding spaces [6, 15, 21, 22, 31, 70]. This approach has demonstrated that with modern architectures, contrastive learning, and large data sources it is possible to obtain performance that is competitive with the classical two-step approach that involves fine-tuning on the downstream data [30, 45]. Other efforts in this direction explore image-text alignment or masked language (or image region) modeling [11, 37]. The models have been applied to diverse downstream tasks, including visual question answering [23], visual commonsense reasoning [67] and image captioning [40, 41, 55].

*Contrastive learning* techniques are another closely-related research direction. The high-level idea of a contrastive loss is to simplify the learning task by requiring the model to select the correct answers out of a finite set of carefully designed options. Intuitively, this simplification of the task may encourage the model to focus on high-level information in an image instead of generic information, resulting in high quality learned representations. Early works that investigate very specific instances of this idea include [18, 43]. More recently, contrastive learning was formulated and studied in more general settings [7, 24, 61], leading to very promising results. Finally, [30, 45] use contrastive learning for learning from image-text data and derive state-of-the-art zero-shot image classifiers.

## 3. Methods

### 3.1. Contrastive pre-training

Collections of images (potentially noisily) paired with free-form text descriptions have emerged as a powerful resource for training visual models. The key advantage therein is that it is not limited by a finite set of predefined categories and instead describes images using open-ended natural language. As a result, models learned from this data can serve as zero-shot learners for a wide range of tasks, e.g. classification and image/text retrieval.

Contrastive pre-training is one particularly effective approach for training models from image-text data, which was recently proven to work well in practice [30, 45]. We take a closer look at this approach and propose a simple, yet highly effective recipe to significantly enhance contrastive pre-training from image-text data.

The key idea behind the contrastive pre-training approach is to learn two embedding models: an image model

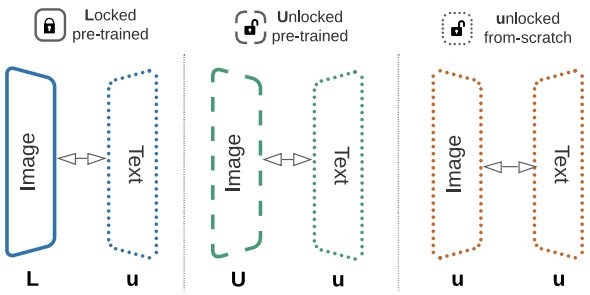


Figure 2. Design choices for contrastive-tuning on image-text data. Two letters are introduced to represent the image tower and text tower setups. `L` stands for locked variables and initialized from a pre-trained model, `U` stands for unlocked and initialized from a pre-trained model, `u` stands for unlocked and randomly initialized. `Lu` is named as "Locked-image Tuning" (LiT).

and a text model, both of which produce representations of the same dimensionality. These models are trained using a contrastive loss. This loss encourages corresponding image-text pairs to have similar embeddings and, conversely, encourages non-corresponding pairs to have distinct embeddings. See [45, 70] for the detailed discussion of the contrastive loss function.

An important detail of this loss function is whether the loss is computed on each accelerator device independently and then accumulated or computed jointly across all devices. We ablate this design choice (Appendix F) and confirm that the latter [30, 45] consistently results in better performance. We therefore use the global loss in all our experiments and ablations.

After image and text towers are trained, they can be readily used for zero-shot classification: class names or descriptions are embedded with the text model. Then, for a given image the label is selected that has the embedding closest to the embedding of the image. This approach also works for image-text retrieval.

### 3.2. Contrastive-tuning

Contrastive pre-training can be viewed as learning two tasks at the same time: (1) learning an image embedding and (2) learning a text embedding to align with the image embedding space. While contrastive pre-training on image-text data works well for solving both of these tasks simultaneously, it may be not the optimal approach.

When not using contrastive pre-training on image-text data, a standard approach to learning image embeddings is to use a large and relatively clean dataset of (semi)-manually labeled images. Large scale and and high quality of such data result in state-of-the-art image embeddings. Some dataset choices for learning powerful image embeddings are ImageNet-21k [14], JFT-300M [56].

However, this common approach has a clear weakness: it is limited to a *predefined set of categories* and, thus, the resulting models can only reason about these categories. In contrast, image-text data does not have this limitation, as it learns from the *free-form text* that potentially spans a broad range of real-life concepts. On the other hand, image-text data that is available may be of lower quality (for learning image embeddings) than carefully curated datasets.

We propose *contrastive-tuning* to combine advantages of both sources of data. One specific way of doing this is to initialise the contrastive pre-training with an image model that was *already pre-trained* using cleaner (semi-)manually labeled data. This way the image-text alignment is learned independently of image embedding, enabling benefit from both data sources.

Beyond using supervised pre-trained image models, the proposed contrastive-tuning is also flexible enough to integrate any models that can produce meaningful representations. We verify this in our experiments using self-supervised pre-trained image models.

Similar lines of reasoning can also be applied to the text tower, as there are many powerful pretrained models that use text-specific data sources and learning techniques.

### 3.3. Design choices and Locked-image Tuning

Introducing pre-trained image or text models into the contrastive learning setting involves several design choices. First, each tower (image and text) can independently be initialized randomly or from a pre-trained model. For a pre-trained model there are at least two variants: we can lock (freeze) it or allow fine-tuning. Note that there are many choices between these two extremes (e.g. partial freezing of selected layers, or custom learning rates), but they are not investigated in this paper.

Pre-trained image-text models may have different representation sizes, while the contrastive loss expects representations of the same size. To compensate, we add an optional linear projection (head) to each tower, which maps the representations to a common dimensionality. Preliminary investigations with tried MLP-based heads did not yield significant improvements over such a simple linear head.

We introduce a two-character notation to discuss the potential design choices outlined above (see Figure 2). Each character encodes the setting chosen for the image model and the text model (in this order). We define three potential settings: `L` (locked weights, a initialized from pre-trained model), `U` (unlocked/trainable weights, initialized from a pre-trained model) and `u` (unlocked/trainable weights, randomly initialized). For example, the notation `Lu` means locked pre-trained image model, and unlocked (trainable) randomly initialized text model. Previous works training models from scratch [30, 45] are `uu`. In our experiments we find the `Lu` setting to work particularly well, so we explicitly name it as *Locked-image Tuning* (LiT🔥).

## 4. Image-text datasets

**CC12M.** The Conceptual Captions dataset [51] extracts, filters & transforms image & alt-text pairs from web pages. We use the latest 12 million image-text pair version, i.e. CC12M [5]. Due to expired URLs, only 10 million image-text pairs were used for our experiments.

**YFCC100m.** The Yahoo Flickr Creative Commons dataset [59] contains 100 million media objects. Of these, 99.2 million are photos that come with rich metadata including camera info, timestamp, title, description, tags, geolocation, and more. [45] defines and uses a subset of 15 million images that have been filtered for English text of high quality, which we call YFCC100m-CLIP. A detailed investigation of this dataset and how best to use it, including whether to filter it, is presented in Appendix E.

**Our dataset.** We collect 4 billion image and alt-text pairs following the same process as ALIGN [30], with the same image-based filtering but simpler text-based filtering. Appendix L shows that reducing text filtering does not harm performance. To avoid misleading evaluation results, we remove from our dataset near-duplicate images of all splits from all datasets we evaluate on. We do not consider the creation of our dataset a main contribution of this paper; we just simplify the data collection process in ALIGN [30] to demonstrate the efficacy of our methods at scale.

## 5. Experiments

In this section, we first compare LiT🔥 to state-of-the-art image-text models. We consider two scenarios: (1) only using public datasets for model training and (2) using privately gathered data. We then present learnings from experimental evaluations of contrastive tuning design choices with various training settings & datasets. We generally perform evaluation on 0-shot ImageNet classification ("0-shot") and MSCOCO image ("T→I") and text ("I→T") retrieval.

### 5.1. Comparison to the previous state-of-the-art

In this section, we present LiT results on our dataset. The image tower is initialized with a ViT-g/14 model pre-trained on JFT-3B [68], which has been de-duplicated against the downstream tasks. We use 32k batch size, and tune for 18 billion image-text pairs seen (roughly 550k steps). See Appendix C for details.

We compare the LiT method with the previous state-of-the-art methods, including CLIP [45] and ALIGN [30]. In Table 1, we report zero-shot classification results on the ImageNet dataset, five out-of-distribution test variants and seven VTAB-natural tasks [69]. Our model significantly outperforms the previous state-of-the-art methods at ImageNet zero-shot classification. The 8.3% and 8.1% improvement over CLIP and ALIGN, respectively, halves the gap between zero-shot transfer results and supervised fine-tuned results [12, 68].

| Dataset | Method | INet | INet-v2 | INet-R | INet-A | ObjNet | ReaL | VTAB-N |
|---|---|---|---|---|---|---|---|---|
| Private | CLIP [45] | 76.2 | 70.1 | 88.9 | 77.2 | 72.3 | - | - |
| | ALIGN [30] | 76.4 | 70.1 | 92.2 | 75.8 | - | - | - |
| | *LiT* | **84.5** | **78.7** | **93.9** | **79.4** | **81.1** | 88.0 | 72.6 |
| Public | CLIP [45] | 31.3 | - | - | - | - | - | - |
| | OpenCLIP [28] | 34.8 | 30.0 | - | - | - | - | - |
| | *LiT* | **75.7** | **66.6** | 60.4 | 37.8 | 54.5 | 82.1 | 63.1 |
| * | ResNet50 [25] | 75.8 | 63.8 | 36.1 | 0.5 | 26.5 | 82.5 | 72.6 |

Table 1. Zero-shot transfer accuracies (%) on ImageNet, five OOD test variants, and seven VTAB-natural tasks. Results are reported on both public datasets and privately gathered data. For reference, we include the ResNet50 model pre-trained on ImageNet, supervised fine-tuned on downstream datasets. We use * to denote multiple datasets during supervised fine-tuning.

**Robustness.** We evaluate robustness on ImageNet-v2 [48], -R [26, 63], -A [27], -ReaL [2], and ObjectNet [1], following CLIP and ALIGN. On all of the OOD variants, our model consistently outperforms the previous models. Notably, the LiT model sets a new state-of-the-art 81.1% accuracy on the ObjectNet test set. The pre-trained ViT-g/14 model [68], achieves 70.5% accuracy on the ObjectNet test set when fine-tuned on ImageNet. This model gets nearly 10% improvement when instead locked-image tuned (LiT) on our image-text dataset.

**Diverse downstream tasks.** We evaluate the LiT models on VTAB, consisting of 19 diverse tasks. We report averaged results on seven VTAB-natural tasks in Table 1. The LiT models achieve promising zero-shot results, comparing to the supervised fine-tuned ResNet50 baseline. In Appendix I.2, we present zero-shot transfer details on VTAB, as well as more results and analysis on the specialized tasks and structured tasks.

**Data & compute efficiency.** Figure 1 shows more results when tuning with fewer seen image-text pairs. With LiT the model achieves 78.7% top-1 accuracy on 0-shot ImageNet transfer, with only 300M image-text pairs seen. In comparison, it took the from-scratch method (i.e. CLIP) 12.8B image-text pairs seen, i.e. 40 times more data pairs, to reach 76.2% top-1 accuracy. With a pre-trained image model, the proposed setup converges significantly faster than the standard from-scratch setups reported in the literature. LiT provides a way to reuse the already pre-trained models in the literature, amortizing the computational resources used to re-generate the image models.

**Results on public datasets.** Given high data efficiency

| Method | ImgNet | ImgNet-v2 | Cifar100 | Pets |
|---|---|---|---|---|
| Lu | 70.1 | 61.7 | 70.9 | 88.1 |
| Uu | 57.2 | 50.2 | 62.1 | 74.8 |
| uu | 50.6 | 43.3 | 47.9 | 70.3 |

Table 2. Evaluation of design choices on our large dataset.

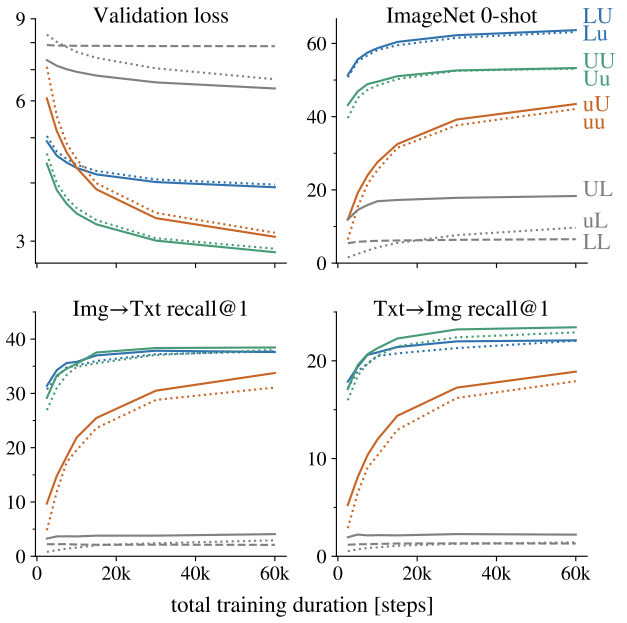


Figure 3. An in-depth study of the possible locking and initialization settings of LiT on the YFCC100m-CLIP dataset. A pre-trained image tower works best, while pre-training of the text tower only helps a little. These are **not** training curves; each point is the final value reached by a training run of that duration.

of LiT, we investigate how well it performs when using only smaller, publicly available models and datasets. Specifically, we tune an ImageNet-21k pre-trained ViT-L/16 model [54] on the union of the *YFCC100m-CLIP* and *CC12M* datasets. More details of the training setup are provided in Appendix D. As a result we achieve unprecedented **75.7%** zero-shot transfer on ImageNet, an absolute improvement of 30.9% over the previously reported state-of-the-art result [28] that uses only public data sources. We also obtain strong results on a wide range of robustness datasets and the VTAB-natural tasks, see Table 1.

## 5.2. Evaluation of design choices

**Small-scale thorough investigation.** We first perform an in-depth study on various combinations of the image and text towers being initialized with pre-trained weights and locked (`L`) or unlocked (`U`) or being randomly initialized and unlocked (`u`). We train each setting many times on the YFCC100m-CLIP dataset, varying the total number of steps from 2 500 to 60 000 in order to understand the setting's trajectory, and sweeping over learning-rates and weight-decays to avoid being misled. Details can be found in Appendix D. Figure 3 shows the best result for each setting for each duration, i.e. each point on the curves is a separate full run for that duration. It is evident that locking the image tower almost always works best and using a pre-trained image tower significantly helps across the board, whereas using a pre-trained text tower only marginally improves performance, and locking the text tower does not work well.

**This still holds in the near-infinite data regime.** One may hypothesize that locking the pre-trained image tower only helps because the YFCC100m-CLIP dataset is *relatively* small (15 million images, compared to 400M [45] or 1.8B [30]), and that a randomly initialized image tower will eventually outperform a locked one on much larger image-text datasets. The trajectory of the `Uu` and `UU` settings in Figure 3 may seem to support this expectation.

Maybe surprisingly, experimental results show that this is not the case, and locking the image tower provides benefits even when contrastively tuning on a very large dataset of image-text pairs. Table 2 shows results of contrastive tuning on our dataset of 4 billion images in three settings: `Lu`, `Uu`, and `uu`. Implementation details can be found in Appendix C. The from-scratch method `uu` unsurprisingly achieves better performance than with smaller datasets such as CC12M and YFCC100m-CLIP.

Initializing the image tower from a pre-trained model provides even better performance and is a relatively straightforward extension of CLIP/ALIGN. Perhaps surprisingly, the frozen setup `Lu`, achieves even better results. While potentially counter-intuitive, another perspective is that LiT simply learns a text tower that extracts knowledge from a strong image embedder. This flexible & performant setup can turn existing vision backbones into a zero-shot learners, by attaching a text-embedding tower.

**Why is locked (`L`) better than unlocked (`U`)?** It is somewhat surprising and counter-intuitive that locking the image tower works better than allowing it to adapt during the contrastive-tuning; Figure 4 gives hints as to why.

The first row shows that locking the image tower leads to substantially worse (contrastive) loss on the dataset used for LiT, while the loss of the locked image variant is substantially better on out-of-distribution datasets such as COCO captions (middle row).

We also measure the *representation quality* of the image model (bottom row) via the performance achieved by a few-shot linear regression on its pre-logits, as is commonly done in the self-supervised representation learning literature. Taken together, these figures reveal that the image representation of a pre-trained image model generalizes very well, but contrastively fine-tuning it worsens the gen-

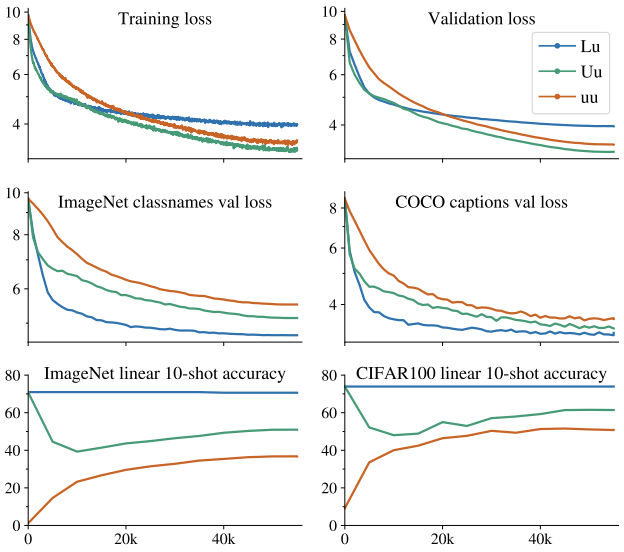


Figure 4. Comparing the loss on the dataset used for LiT (top row) to the loss on out-of-distribution (zero-shot) datasets (middle row) and the "representation quality" as measured by linear few-shot evaluation on the pre-logits (bottom row). This reveals how the different settings behave, see text for details.

erality of the visual representation, leading it to be better on the contrastive dataset, but worse everywhere else. This indicates that locking the image tower during tuning, i.e. LiT, *leads to a text model that is well aligned to an already strong and general image representation*, as opposed to an image-text model that is well aligned but specialized to the dataset used for alignment.

Intermediate variants, such as first locking and later unlocking the image tower or separating learning-rates are explored in Appendix H; we did not find a strictly better setup than LiT and leave this as an open research question.

### 5.3. LiT works better for more generally pre-trained models

One may believe that LiT only works because the image tower is initialized with a backbone that was supervisedly pre-trained for classification, and hence remains a supervised classifier, as opposed to becoming an image-text model. We design a controlled experiment to verify whether that is the case. We find that on the contrary, more generally pre-trained models are better suited for LiT.

We select a set of image models that all use the same ViT-B/16 architecture but were pre-trained in various ways: supervised (AugReg [54]) on ImageNet (IN), on the large but narrow Places [38] dataset, on the much broader ImageNet-21k (IN21k), or fully unsupervised (DINO and MoCo-v3). All but the Places model achieve similar ImageNet top-1 accuracies of around 77% as reported in their respective publications, and can thus be considered *similarly good* models.

| **Model:** ViT-B/16 | **Pre-training** Dataset | Labels? | Full IN | 10-shot | **LiT** 0-shot | I→T | T→I |
|---|---|---|---|---|---|---|---|
| MoCo-v3 [10] | IN | n | 76.7 | 60.6 | 55.4 | 33.5 | 17.6 |
| DINO [4] | IN | n | 78.2 | 61.2 | 55.5 | 33.4 | 18.2 |
| AugReg [54] | IN21k | y | 77.4 | 63.9 | 55.9 | 30.3 | 17.2 |
| AugReg [54] | IN | y | 77.7 | 77.1 | 64.3 | 25.4 | 13.8 |
| AugReg [54] | Places | y | - | 22.5 | 28.5 | 25.1 | 12.9 |

Table 3. The role of pre-training method for the image model: as long as it is general, it does not matter. The background coloring denotes whether a value is similar or far away from the others in that column.

Table 3 shows model performance without LiT (ImageNet 10-shot, and accuracy when fully fine-tuned on ImageNet) alongside achieved performance with LiT on YFCC100m-CLIP (zero-shot ImageNet classification and MS Coco retrieval).

From these results, we conclude that models which are pre-trained in a generic way (e.g. on large amounts of data, or in an unsupervised way) and have similar representation quality, become similarly good image-text models after locked-image tuning (LiT). However, this also shows that a narrowly pre-trained model (AugReg-IN and AugReg-Places) will perform misleadingly well on its narrow task (0-shot IN for AugReg-IN), but significantly fall behind on more general image-text tasks (MSCOCO captions). These findings highlight the importance of a generally pre-trained model and varied set of evaluation tasks.

**Is this specific to ViT image models?** No. Here we fixed the architecture to avoid confounders, but Appendix A explores other architectures.

### 5.4. Which text model to use?

While related work has so far focused on the image model, the text model plays an important yet underexplored role in contrastive image-text learning. We consider four possible transformer-based text models [62]—the transformer from ViT-B [20] which also resembles that used in CLIP [45], T5-base [46], mT5-base [66], and the classic BERT-base [16]—and whether to initialise them randomly, or from a pre-trained checkpoint. BERT uses a WordPiece (WP) tokenizer [49, 64], and all others use the SentencePiece (SP) tokenizer [34], a component which we also ablate with the ViT model.

Table 4 shows the results of LiT using an AugReg-ViT-B/32 on YFCC100M-CLIP and our dataset using the *base* sized variant of these text models. We sweep over various learning-rates and weight-decays separately for each

| | Model | Tok | INet 0shot | I→T | T→I |
|---|---|---|---|---|---|
| YFCC-CLIP | ViT | SP | 57.2 | 29.7 | 16.9 |
| | T5 | SP | 57.8 (+1.4) | 29.4 (+1.6) | 17.2 (+1.2) |
| | mT5 | SP | 58.1 (+1.2) | 28.3 (+0.4) | 16.4 (+1.0) |
| | BERT | WP | **58.8** (+0.7) | **35.2** (+1.1) | **20.0** (+0.7) |
| | ViT | WP | 56.4 | 28.2 | 17.3 |
| Ours | ViT | SP | 68.8 | 43.6 | 28.5 |
| | ViT | WP | 68.8 | 45.4 | 29.7 |
| | BERT | WP | 65.8 | 43.8 | 28.6 |

Table 4. The effect of different text encoders on zero-shot performance. The main numbers show performance achieved when the text tower is randomly initialised; the numbers in brackets are the further improvement achieved when the text tower is initialized with a pre-trained language model. The *Tok* column indicates whether a SentencePiece or WordPiece tokenizer was used.

combination to avoid being misled. Our observations differ slightly between the *relatively* small YFCC100m-CLIP dataset, and our much larger dataset, we first discuss the former. First, we see a small but consistent improvement by initializing the text model with pre-trained weights. Second and somewhat unexpectedly, we find that the BERT model performs significantly better than others, especially for retrieval. In order to disentangle the contribution of the architecture from the tokenizer, we further apply LiT using a ViT text encoder paired with BERT's WordPiece tokenizer and see no improvement. We believe that small differences in the architecture, such as initialization and LayerNorm placement, are responsible for the slightly better generalization of BERT that we observe. However, we also found the BERT model to be less stable to train. For the large-scale experiments on our dataset, we do not observe this improvement anymore, and favor sticking with the more stable ViT SentencePiece combination.

**What about model capacity?** Previous works used relatively low-capacity text models. We show in Appendix B that increasing the text tower's capacity consistently improves performance. The same is true, and more pronounced, for the image tower.

### 5.5. Do duplicate examples matter for LiT?

One relevant question in the context of large-scale training is the role of duplicate examples between upstream datasets and downstream datasets. We answer this question by performing experiments on three different upstream de-duplication setups: (1) no de-duplication; (2) de-duplicate against downstream test splits only; (3) de-duplicate against downstream train and test splits. We conduct experiments using the `Lu` setup on our dataset. We use a B/32 image model pre-trained on the JFT-3B dataset [68], which has been de-duplicated against downstream train and test splits.

| Dedup | #tune | #eval | ImgNet | I→T | T→I |
|---|---|---|---|---|---|
| - | 0 | 0 | 70.2 | 43.6 | 28.4 |
| test | 2.6M | 76K | 70.2 | 43.3 | 28.3 |
| train+test | 3.6M | 220K | 69.9 | 43.7 | 28.4 |

Table 5. Results on various de-duplication setups. #tune images are removed from the LiT dataset due to #eval images in the evaluation datasets. We report results averaged across three runs.

In Table 5, we show the number of duplicate samples found between upstream datasets and downstream datasets during de-duplication. In the de-duplication process, a downstream image may have multiple upstream duplicate examples, e.g. due to image copies on the web. As a result, the number of duplicate examples on the upstream dataset is significantly larger than the number on the downstream datasets. The downstream number indicates how many downstream images had a duplicate detected, while the upstream number indicates how many images are removed from the image-text dataset.

We apply LiT on the three setups, and the zero-shot transfer results vary little. More results with larger backbone can be found in Appendix K, with consistent conclusions. It indicates that the duplication of examples here *does not* influence the results strongly. This observation is also consistent with previous conclusions [32,45]. A possible interpretation is that with a large upstream dataset, the model may not memorize those duplicate examples.

Throughout this paper, we report results using the strictest setup (3) with proper de-duplication against downstream train splits and test splits, to avoid data leakage.

### 5.6. Technical advantages of locked image models

Besides potential modelling advantages previously explored, using a locked image tower has several more benefits. First, the training is significantly sped-up and memory use reduced as no gradients are computed for the image tower. Second, if no augmentations are used, such as in our large-data experiment, the image model's embeddings can be precomputed once, further reducing computation time and memory requirements. Appendix G shows concrete measurements. Taken together, these implementation features unlock the use of enormous models at very large batch-sizes.

### 5.7. Preliminary multilingual experiments

It is currently common practice [30, 45] to filter image-text datasets to English language data only. We believe that removing this restriction has the potential to benefit a larger part of the world's population. Concurrent work [29] has relied on additional translated text pairs for training the text

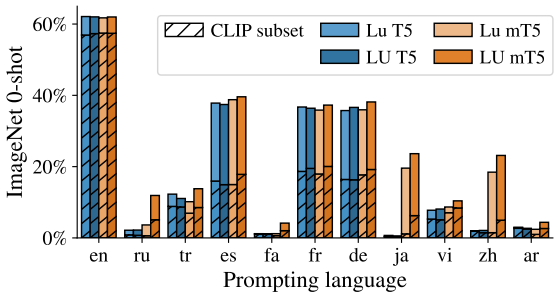


Figure 5. Including non-English data unlocks multilingual zero-shot models without hurting English performance. In such a regime, multilingual text pre-training can be more useful for low-resource languages.

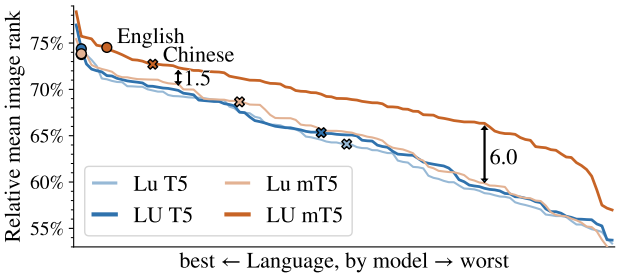


Figure 6. Image retrieval performance over 100 languages reveals that unfiltered data and a multilingually pre-trained text model can significantly increase long-tail performance.

encoder. In contrast, we do not require any translations and purely rely on the pre-trained, locked image model to bridge the language barrier. In this section, we report preliminary experiments that show the promise of LiT for multilingual image-text models.

We apply LiT on an AugReg-i21k ViT-B/32 with the T5 [46] and mT5 [66] base encoders, both with and without the pre-trained checkpoints. We do this on both the full YFCC100m dataset, and the reduced English-only CLIP subset, and we use all available text as supervision signal (See Appendix E). We evaluate the resulting model's multilingualism in two ways, both of which have limitations discussed in Appendix J. First, we translate the ImageNet prompts into the most common languages using an online translation service and perform zero-shot classification in each of them; this evaluation is shown in Figure 5. Second, we use the Wikipedia based Image Text (WIT) dataset [53] to perform T $\rightarrow$ I retrieval across more than a hundred languages. Figure 6 gives a summary of this evaluation; a more detailed variant is provided in Appendix J.

The high-level conclusions are consistent across both evaluations: training on the full dataset improves performance on non-English languages much more than on English, using a multilingual tokenizer (as in mT5) significantly helps languages that do not use the Latin script, and starting from a pre-trained multilingual text model can further help. The combination of all three improvements barely has any effect when evaluated in English, but significantly improves performance on the long tail of languages. This is a promising result for unlocking multimodal models for low-resource languages.

## 6. Discussion

**Limitations.** This work explores only classification and retrieval as zero-shot transfer tasks. We leave evaluating zero-shot transfer to a broader set of tasks such as detection, segmentation, visual question answering, and image captioning as future work in order to limit our scope.

On cross-modal retrieval tasks, we have not observed as clear a benefit of the `Lu` setup compared to `Uu` or `UU` (Figure 3). For very long tuning schedules, `Uu` or `UU` sometimes overtake `Lu` on these tasks. Our results suggest that the proposed `Lu` setup can still save computational cost within a fixed budget, but with a large enough budget, it may be useful to also consider the `Uu` setup if zero-shot classification is not the primary end goal.

**Societal impact.** This work shows how one can easily add a text-tower to a pre-trained image model. While there are many useful applications, like most research, it is a double-edged sword: the technique also makes it simpler to create malicious, offensive, or obscene text tower pendants to existing image models. Further research is needed on how to best equip open-world image-text models with the behaviour we desire.

## 7. Conclusion

We present a simple method named contrastive-tuning that allows transferring any pre-trained vision model in a zero-shot fashion. More specifically, the proposed LiT setup leads to substantial quality improvements on zero-shot transfer tasks. It halves the gap between the from-scratch contrastive learning setup, and the per-task supervised fine-tuning setup. LiT makes it possible to turn publicly available models into zero-shot classifiers using publicly available data, and rival the performance of previous works which rely on more, proprietary data.

We hope that this work motivates future research on how to smartly re-use and adapt already pre-trained models for different research problems.

## References

[1] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset

for pushing the limits of object recognition models. In *NeurIPS*, 2019. 2, 4

[2] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *CoRR*, abs/2006.07159, 2020. 4

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 6, 26

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

[6] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, 2021. 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020. 2

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 16

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *CoRR*, abs/2104.02057, 2021. 2, 6, 26

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In *ECCV*, 2020. 2

[12] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. CoAtNet: Marrying convolution and attention for all data sizes. In *NeurIPS*, 2021. 1, 2, 4

[13] Mostafa Dehghani, Anurag Arnab, Lucas Beyer, Ashish Vaswani, and Yi Tay. The efficiency misnomer. *CoRR*, abs/2110.12894, 2021. 12

[14] 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. 3

[15] Karan Desai and Justin Johnson. VirTex: Learning visual representations from textual annotations. In *CVPR*, 2021. 2

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 6, 12, 13, 26

[17] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, Sylvain Gelly, Neil Houlsby, Xiaohua Zhai, and Mario Lucic. On robustness and transferability of convolutional neural networks. In *CVPR*, 2021. 2

[18] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 2

[19] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014. 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 6, 12, 26

[21] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. 2

[22] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. DeViSE: A deep visual-semantic embedding model. In *NeurIPS*, 2013. 2

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 2

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[26] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *CoRR*, abs/2006.16241, 2020. 4

[27] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 4

[28] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. Zenodo, 2021. 4, 5

[29] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. MURAL: Multimodal, multitask representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3449–3463, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 7

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation

learning with noisy text supervision. In *ICML*, 2021. 1, 2, 3, 4, 5, 7

[31] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[32] 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, 2, 7, 12, 26

[33] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *CVPR*, 2019. 1, 2

[34] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*, 2018. 6

[35] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2

[36] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, 2008. 1, 2

[37] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. 2

[38] Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognit.*, 102:107256, 2020. 6

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 12, 26

[40] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2

[41] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 2

[42] Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 1

[43] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2

[44] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10), 2010. 1, 2

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 4, 5, 6, 7, 12, 13, 15

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. 6, 8, 12, 26

[47] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *CVPR*, 2014. 2

[48] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019. 4

[49] Mike Schuster and Kaisuke Nakajima. Japanese and Korean voice search. In *ICASSP*, 2012. 6

[50] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *ACL*, 2016. 17

[51] 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. 4

[52] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *ICML*, 2018. 12, 26

[53] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: wikipedia-based image text dataset for multimodal multilingual machine learning. *CoRR*, abs/2103.01913, 2021. 8

[54] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, augmentation, and regularization in vision transformers. *CoRR*, abs/2106.10270, 2021. 5, 6, 12, 13, 26

[55] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 2

[56] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017. 3

[57] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 12

[58] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *ICANN*, 2018. 2

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

[60] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-Mixer: An all-MLP architecture for vision. In *NeurIPS*, 2021. 2, 12, 26

[61] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 2

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 6

[63] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*

*2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10506–10518, 2019. 4

[64] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. 6

[65] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI*, 41(9):2251–2265, 2019. 1

[66] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *NAACL-HLT*, 2021. 6, 8, 12, 26

[67] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 2

[68] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *CoRR*, abs/2106.04560, 2021. 1, 2, 4, 7, 12, 17, 26

[69] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. The visual task adaptation benchmark. *CoRR*, abs/1910.04867, 2019. 4, 15, 26

[70] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *CoRR*, abs/2010.00747, 2020. 2, 3