How well do contrastively trained models transfer?

M. Moein Shariatnia [*1] **Rahim Entezari** [*2] **Mitchell Wortsman** [3] **Olga Saukh** [2] **Ludwig Schmidt** [3]

Abstract

There are two prevailing methods for pre-training on large datasets to learn transferable representations: 1) supervised pre-training on large but weakly-labeled datasets; 2) contrastive training on image only and on image-text pairs. While supervised pre-training learns good representations that can be transferred to a wide range of tasks, contrastively trained models such as CLIP have demonstrated unprecedented zero-shot transfer. In this work we compare the transferability of the two aforementioned methods to multiple downstream tasks. The pre-training distributions we consider include YFCC, Conceptual Captions, and ImageNet-21K while pre-training objectives range from supervised to SimCLR, CLIP, and SLIP. We observe that different pre-training methods with the same training source transfer similarly given their ImageNet accuracy.

1. Introduction

The last few years of computer vision have witnessed scaling in both dataset and model size. Supervised pre-training on large but weakly-labeled datasets such as Instagram images (Mahajan et al., 2018) and JFT (Sun et al., 2017; Zhai et al., 2021) learns good representations which can be transferred to a wide range of tasks. Abnar et al. (2021) study the effect of pre-training on JFT-300 (Sun et al., 2017) and ImageNet-21K (Deng et al., 2009) on transferring to multiple downstream tasks, across different architectures, upstream dataset and model sizes, concluding that learned representations could be transferred well to downstream tasks, however, such transfer performance is saturated.

In contrast to supervised pre-training, another promising direction to scale the required data for large models is self-supervised learning with contrastive loss. SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) are two examples which learn the representation in image modality, while CLIP (Radford et al., 2021), SLIP (Mu et al., 2021), ALIGN (Jia et al., 2021), and BASIC (Pham et al., 2021) leverage data from both image and text modalities, demonstrating unprecedented robustness to the challenging distribution shifts. However, such robustness improvements are at best in the zero-shot setting.

Ericsson et al. (2021) show that semi-supervised contrastive pre-training on ImageNet transfers well to downstream tasks related to object recognition in natural images. Cole et al. (2021) extend the ImageNet transfer study and show that the learned representations by supervised methods can transfer better than semi-supervised methods on non-ImageNet domains, *e.g.*, fine-grained classification. Radford et al. (2021) and Jia et al. (2021) try finetuning contrastive image and text models and show that the accuracy of their few-shot fine-tuned models on downstream tasks is lower than that of the original zero-shot model, and such performance drop is compensated with higher shots. Wortsman et al. (2021) also propose a finetuning strategy (WiSE-FT) for pre-trained CLIP models on downstream tasks to reduce this performance drop. Their proposed method linearly interpolates the weights of the pre-trained model and that of the fine-tuned model.

In this work, we investigate the transferability of learned representations by different pre-training objectives to ImageNet, CIFAR100, DTD, and CALTECH-101. We compare supervised pre-training on ImageNet-21K (Deng et al., 2009) with 14M images to contrastive image and contrastive image+text pre-training on YFCC-15m (Radford et al., 2021) and Conceptual Captions 12M (Changpinyo et al., 2021). We explore different impacting factors on transferability and make the following observations:

- In few-shot transfer, supervised pre-training on ImageNet-21K shows higher accuracy than contrastive pre-training on YFCC-15m.
- Among contrastive models pretrained on YFCC-15m, SLIP shows the best performance when transferring learned representations from YFCC-15m to the downstream tasks, followed by SimCLR and CLIP.
- Given the same ImageNet accuracy in few-shot transfer, supervised pre-training on ImageNet-21K shows lower

[*]Equal contribution [1]Tehran University of Medical Sciences [2]TU Graz / CSH Vienna [3]University of Washington. Correspondence to: Rahim Entezari <entezari@tugraz.at>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

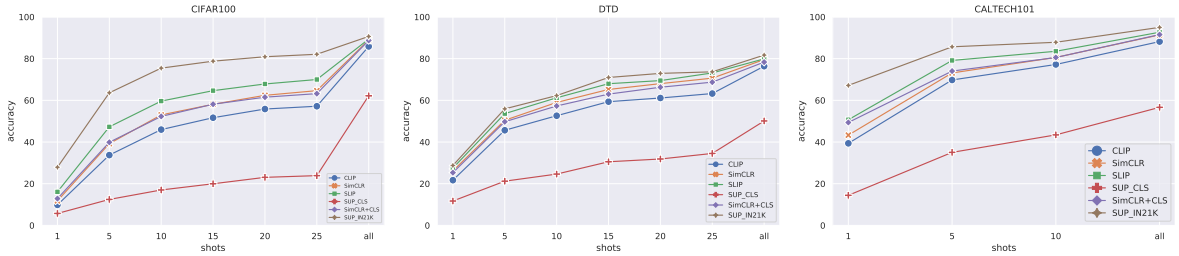


Figure 1: **Comparison of Supervised (cross-entropy) and Self-supervised (contrastive) pre-training with respect to transferability to downstream tasks.** When evaluated few-shot, supervised pre-training on ImageNet-21K shows higher accuracy than contrastive pre-training on YFCC-15m. The difference is large for CIFAR100 and smaller for DTD and CALTECH-101. The gap shrinks when we use all samples from the downstream task for finetuning. Supervised pre-training on a subset of YFCC (YFCC-15m-cls) performs worst. Among contrastive pre-training methods, SLIP shows the best performance when transferring the learned representations from YFCC-15m to the downstream tasks, followed by SimCLR and CLIP.

accuracy, while SimCLR shows higher accuracy than other methods.

- Different pre-training methods with the same training source transfer similarly given their ImageNet accuracy.

2. Background

In this section we first review pre-training methods used in this work and then discuss the details of the transfer learning framework which we used. Details of different datasets can be viewed in Appendix Section A.

2.1. Methods

SimCLR (Chen et al., 2020) applies random transformations to each image to get a pair of two augmented images. These images are then passed through an encoder. The resulting representations then get transformed and projected via non-linear layers to obtain the final representations for images and their augmented views. The similarity between two augmented versions of an image is calculated using cosine similarity. The idea is to pull together similar and push away dissimilar images.

SUP-IN21K pre-trains ViT-B/16 on ImageNet-21K with cross-entropy loss.

SUP_CLS is the supervised classification using cross-entropy loss and from scratch on YFCC-15m-cls dataset.

SimCLR+CLS is a SimCLR model pre-trained on YFCC-15m, and fine-tuned using cross-entropy on YFCC-15m-cls.

CLIP (Radford et al., 2021) is directly trained on images and their corresponding unstructured text from the web (400 million image and text pairs). Given a batch of M (image, text) pairs, CLIP learns a multi-modal embedding space, by jointly training an image-encoder and a text-encoder, such that the cosine similarity of the valid M (image, text) pairs is maximized. The resulting models achieve decent robustness even on a series of challenging distribution shifts such as ImageNetV2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-Adversarial (Hendrycks et al., 2021), and ObjectNet (Barbu et al., 2019).

SLIP (Mu et al., 2021) combines self-supervision of SimCLR with CLIP for better visual representations. The CLIP and self supervised objectives are computed on the relevant embedding and then accumulated into a single scalar loss.

2.2. Transfer Learning

In this work we investigate the transferability of the learned representations in a few-shot learning setup. This is motivated by the previous findings that as the number of downstream samples increases, the effect of transfer learning shrinks (Kornblith et al., 2019; Mensink et al., 2021; Zoph et al., 2020; Abnar et al., 2021). Therefore we mostly focus on comparison between different settings where transferability differs most *i.e.*, few-shot learning. All pre-trained models are ViT-B/16 (Dosovitskiy et al., 2020). For Supervised-ImageNet-21K we used the pre-trained model from Wightman (2019). SLIP, CLIP, and SimCLR checkpoints are from Mu et al. (2021). SimCLR+CLS and SUP_CLS checkpoints are from Fang et al. (2022). We use the training procedure from BeiT (Bao et al., 2021) for end-to-end finetuning on downstream tasks. This procedure takes advantage of significant regularization and data augmentation, as well as layer-wise learning rate decay. We finetune the pre-trained models for 100 epochs.

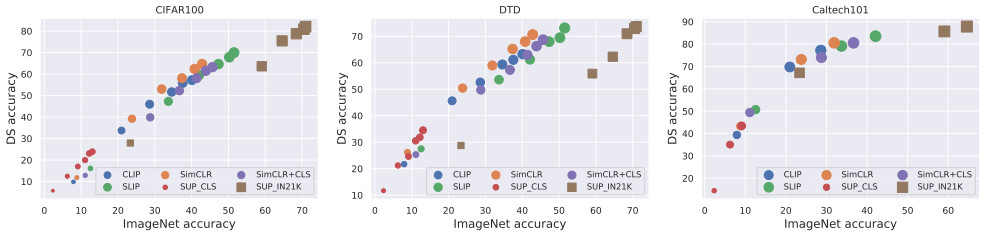


Figure 2: **ImageNet vs. downstream accuracy in few shot learning setting across different pre-training methods and datasets.** ImageNet accuracy has positive correlation with downstream accuracy. Removing the confounder of ImageNet accuracy from Figure 1 changes the picture *i.e.*, given the same ImageNet accuracy when evaluating few-shot, supervised pre-training on ImageNet-21K shows lower accuracy, while SimCLR shows higher accuracy than other methods. Different pre-training methods with the same training source transfer similarly given their ImageNet accuracy. Bigger points represent larger number of shots.

3. Experiments and Results

Figure 1 shows the accuracy of multiple methods after few-shot fine-tuning on three downstream tasks[1][2]. When evaluated in few-shot setting, supervised pre-training on ImageNet-21K shows higher accuracy than contrastive pre-training on YFCC-15m. The difference is large for CIFAR100 and smaller for DTD and CALTECH. As stated in Section 2.2, when we use all samples from the downstream task for finetuning, the gap between different models shrinks. Supervised pre-training on a subset of YFCC (YFCC-15m-cls) performs worst. Among contrastive pre-training methods, SLIP shows the best performance when transferring learned representations from YFCC-15m to the downstream tasks, followed by SimCLR and CLIP. Below we run multiple dedicated studies to investigate the effect of various impacting factor on transferability.

3.1. Does supervised finetuning help transferability?

To answer this question, we compare SimCLR pre-trained on YFCC-15m, with the same method but further finetuned on YFCC-15m-cls using cross-entropy loss, referred to as SimCLR and SimCLR+CLS in the figures, respectively. Both methods follow almost the same transfer accuracy across different shots and datasets implying that supervised finetuning on the upstream dataset has no effect on transferability.

3.2. Does language supervision help transferability?

We compare SimCLR pre-trained on YFCC-15m images with CLIP pre-trained on YFCC-15m with both images and text data. Figure 1 shows that when evaluated few-shot, SimCLR pre-training shows higher accuracy than CLIP. This finding is aligned with (Fang et al., 2022), in which they show that language supervision does not contribute to a model's robustness, but simplifies training on a diverse distribution of images by removing the need for consistent annotation with class labels.

3.3. What is the effect of pre-training distribution?

In order to answer this question, we compare CLIP pre-trained on YFCC-15m dataset with the same model but pre-trained on Conceptual Captions-12M. Figure 3(a) shows that different pre-training sources, keeping the pre-training objective fixed (CLIP), result in different few-shot transfer performance. Few-shot transfer of CC-12m learned representations shows higher accuracies. The fine-tuning performance gap for these two sources is at smallest for DTD and larger for CIFAR100 and CALTECH-101. Higher performance on Conceptual Captions could be attributed to the quality of the text captions (Hendricks et al., 2021; Hendricks & Nematzadeh, 2021) or the image distribution. This is worth of further experiments as a potential future direction. We also leave the investigations on the language similarity (measured by perplexity) between pre-training distribution and downstream tasks for future works.

3.4. What is the role of pre-training dataset size?

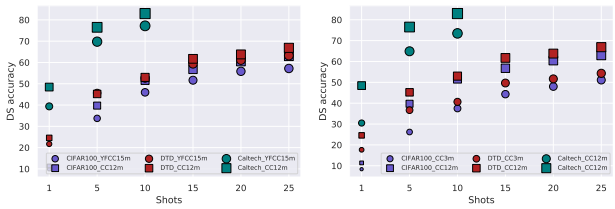
We compare CLIP models pre-trained on two different sizes of Conceptual Captions dataset: 3M and 12M. Figure 3(b) shows that the larger pre-training dataset results in higher few-shot transfer accuracy. For Conceptual Captions experiments, we used checkpoints from SLIP GitHub repository (Mu et al., 2021).

3.5. What is the role of pre-training objective in transfer learning?

Comparison between SLIP and CLIP pre-training objectives shows the effect of adding self-supervised contrastive learn-

[1]CLIP Zero-shot: 34.5(CIFAR100), 21,2(DTD), 60.9(CALTECH)

[2]SLIP Zero-shot: 45.2(CIFAR100), 26.1(DTD), 71.0(CALTECH)



(a) Effect of pre-training dataset (b) Effect of pre-training size

Figure 3: **Effect of pre-training dataset and pre-training dataset size. Left:** We compare CLIP pre-trained on YFCC-15m vs. CC-12m. Different pre-training sources, keeping the pre-training objective fixed (CLIP), result in different few-shot transfer performance. Few-shot transfer of CC-12m learned representation shows higher accuracies. **Right:** We fix the pre-training objective and dataset to CLIP and Conceptual Captions. Larger pre-training dataset results in higher few-shot transfer accuracies.

ing (SimCLR) objective. Figure 1 shows that SLIP performs better than CLIP in all shots and datasets. However, the gap between these two methods shrinks as the number of shots increases; in the full-finetuning setting, SLIP performs only slightly better than CLIP.

3.6. Distance to pre-training dataset

Table 1 compares CLIP and SLIP in linear and full-finetuning settings, across different datasets. The difference between Full and Linear finetuning can be used as proxy of the distance to the upstream dataset, here YFCC-15m. Intuitively, a small gap between full and linear finetuning would suggest that the upstream features can be transferred more easily. It is interesting to see that the ranking of differences between full and linear finetuning is the same for CLIP and SLIP. *i.e.*, CALTECH $<$ DTD $<$ ImageNet $<$ CIFAR100. It is worth noting that the image resolution of downstream tasks differs which might be a confounding factor in such a ranking.

Method	ImageNet	CIFAR100	DTD	CALTECH
CLIP	66.5	70.7	66.0	85.3
SLIP	72.1	71.6	73.9	89.7
CLIP	80.6 (+14.1)	85.8 (+15.1)	76.3 (+10.3)	88.1 (+2.8)
SLIP	82.6 (+10.5)	89.3 (+17.7)	79.8 (+5.9)	92.6 (+2.9)

Table 1: Comparison between linear (first block) and full-finetuning (second block). The difference between full and linear finetuning can be used as a proxy for distance to the upstream dataset *e.g.*, CALTECH $<$ DTD $<$ ImageNet $<$ CIFAR100.

3.7. Does better transfer to ImageNet correspond to better transfer to other downstream tasks?

Figure 2 shows the correlation between ImageNet few-shot evaluation and different downstream tasks. We can observe that models with the same pre-training source which achieve higher ImageNet accuracy, also perform better on the downstream task. This shows that ImageNet is a good proxy for transferability across different downstream tasks. In contrast to Figure 1, given the same ImageNet accuracy when evaluating few-shot, supervised pre-training on ImageNet-21K shows lower accuracy, while SimCLR shows higher accuracy than other methods. Different pre-training methods with the same training source transfer similarly given their ImageNet accuracy.

3.8. Can we predict the full-finetune accuracy with lower computation?

	Kendall's τ	p-value
CIFAR100	0.867	0.017
DTD	1.000	0.003
CALTECH	1.000	0.003
IMAGENET	1.000	0.083

Table 2: Kendall's τ coefficient to capture correlation between 1-shot and full-finetuning accuracy.

1-shot accuracy as a proxy. Table 2 shows the correlation between full-finetuning and 1 shot accuracy as a low computation proxy. As the Kendall's τ coefficients and their p-values indicate, 1 shot accuracy can be used as a proxy to choose the best pre-trained model to finetune on a downstream task without actually having to do so. For ImageNet, we used the 3 full finetuning accuracies reported in Mu et al. (2021) and another from Dosovitskiy et al. (2020). For other datasets, we employ all six methods described in Section 2.1. Exact values for full-finetuning accuracies can be seen in Table 4. Appendix Section C.1 also shows the correlation between 1-NN and full-finetuning accuracy.

4. Discussion

In this work we compare supervised and contrastive self-supervised pre-training on transferability of learned representations to multiple downstream tasks. We consider ImageNet-21K for supervised pre-training, YFCC-15m and Conceptual Captions-12m for contrastive image only and image+text pre-training. Pre-training objectives include supervised, SimCLR, CLIP, and SLIP. Our observations show that different pre-training methods with the same training source transfer similarly given their ImageNet accuracy. We leave further investigations on more pre-training sources and downstream tasks for future works.

References

Abnar, S., Dehghani, M., Neyshabur, B., and Sedghi, H. Exploring the limits of large scale pre-training. *arXiv preprint arXiv:2110.02095*, 2021.

Bao, H., Dong, L., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Cole, E., Yang, X., Wilber, K., Mac Aodha, O., and Belongie, S. When does contrastive visual representation learning work? *arXiv preprint arXiv:2105.05837*, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ericsson, L., Gouk, H., and Hospedales, T. M. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5414–5423, 2021.

Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., and Schmidt, L. Data determines distributional robustness in contrastive language image pre-training (clip). *arXiv preprint arXiv:2205.01397*, 2022.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hendricks, L. A. and Nematzadeh, A. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.

Hendricks, L. A., Mellor, J., Schneider, R., Alayrac, J.-B., and Nematzadeh, A. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Kendall, M. G. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-100 and cifar-10 (canadian institute for advanced research), 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html. MIT License.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and Van Der Maaten, L. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 181–196, 2018.

Mensink, T., Uijlings, J., Kuznetsova, A., Gygli, M., and Ferrari, V. Factors of influence for transfer learning across diverse appearance domains and task types. *arXiv preprint arXiv:2103.13318*, 2021.

Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

Pham, H., Dai, Z., Ghiasi, G., Liu, H., Yu, A. W., Luong, M.-T., Tan, M., and Le, Q. V. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

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

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Wortsman, M., Ilharco, G., Li, M., Kim, J. W., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. arxiv 2021. *arXiv preprint arXiv:2106.04560*, 2021.

Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E. D., and Le, Q. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33:3833–3845, 2020.

Appendix

A. Datasets

YFCC. We use the YFCC-15m (Radford et al., 2021) dataset, a subset of YFCC-100M (Thomee et al., 2016) filtered to only images with English titles or descriptions. The dataset contains 14,829,396 images with natural language captions associated with each image.

We also followed Fang et al. (2022) to convert YFCC to a classification dataset (**YFCC-15m-cls**) with class labels. They assign ImageNet labels to each image using a simple strategy: if the title or description contains the name of an ImageNet synset or synonym Miller (1995), they assign the corresponding synset label to the image. If an image contains no or multiple ImageNet synsets, they discard that image. This results in 1,694,125 images (11.4% of the full dataset) covering 953 ILSVRC classes.

ImageNet-21K. To have same number of images as YFCC-15m, we chose ImageNet-21K (Deng et al., 2009) for supervised pretraining, which consists of 14,197,122 images, each tagged in a single-label fashion by one of 21,841 possible classes.

Conceptual-Captions. To investigate the effect of pre-training dataset distribution and dataset size we also use pre-trained checkpoints on Conceptual Captions 3M (CC-3M) and Conceptual Captions 12M (CC-12M) (Sharma et al., 2018; Changpinyo et al., 2021). CC-3M represtns a wide variety of images and caption styles and is built via extraction and filtering of images and their associated texts from billions of web pages. To build CC-12M dataset, the authors take a step further and relax the data collection pipeline used to build the previous CC-3M version and obtain 12 million image and text pairs.

Downstream tasks. We measure the transferablity of YFCC learned representations using different methods on CIFAR100 (Krizhevsky et al., 2009), DTD (Cimpoi et al., 2014), and Caltech-101 (Fei-Fei et al., 2004)

Downstream Task	Train Size	Test Size	Classes	Resolution
CIFAR100	50000	10000	100	32× 32
DTD	3760	1880	47	300× 300 to 640× 640
CALTECH-101	3060	6084	102	300× 200

Table 3: Details of down stream datasets used in our experiments.

B. Training Hyperparameters

B.1. Pretraining on YFCC

We used CLIP, SLIP and SimCLR checkpoints from SLIP GitHub repository (Mu et al., 2021). For SUP-CLS and SimCLR+CLS we use the checkpoints trained in Fang et al. (2022).

B.2. Transfer to downstream tasks

We follow the procedure from BeiT (Bao et al., 2021) by using fine-tuning scripts from SLIP GitHub repository (Mu et al., 2021) for our full fine-tuning and few-shot training experiments. Mu et al. (2021) use DeiT training procedure for smaller datasets but we followed the BeiT training procedure for all of our experiments, including few-shot learning experiments.

We use a batch size of 800 for full tuning and 100 for few-shot training experiments (except for 1 shot experiments on DTD dataset where we use a batch size of 32 due to lower number of classes in this dataset). We use the learning rate of 4e-3 when using a batch size of 800 and adjust it proportional to batch size whenever it is changed; according to Goyal et al. (2017). Moreover, we employ layerwise learning rate decay to exponentially decrease the learning rate across layers in the model.

All the experiments in this study are trained for 100 epochs using AdamW optimizer with a weight decay of 0.05 and by employing different data augmentation methods and regularization techniques. We also set drop path to 0.1 and layer decay to 0.65 for all the experiments.

Method	ImageNet[1]	CIFAR100	DTD	CALTECH-101
SUP_CLS	-	62.17	50.10	56.51
SimCLR	82.5	88.94	79.57	91.83
SimCLR+CLS	-	88.74	78.35	91.35
CLIP	80.5	85.85	76.38	88.18
SLIP	82.6	89.36	79.89	92.69
SUP-IN21K	83.9	90.65	81.70	95.00

[1] Values from (Mu et al., 2021; Dosovitskiy et al., 2020)

Table 4: **Comparison between different methods pretrained on YFCC-15m and then full-finetuned on different downstream tasks.** ImageNet accuracy is a good proxy for the full fine-tuning performance on other downstream tasks.

C. Full fine-tuning performance on downstream tasks

C.1. KNN as a proxy for full-finetuning

While full-finetuning is computationally expensive, calculating KNN ($K = 1$) is cheap and if we can show that the 1-NN and full-finetuning accuracies are correlated, 1-NN can be used as a low computation proxy to select the best pre-training method without actual finetuning. Here we use K-NN with $K = 1$ as follows: For each pre-training method, we first obtain the final embeddings of all training and test images. Then, for each test sample, the predicted label is the label of the closest sample from the training set, *i.e.*, the sample with the highest cosine similarity to its embedding vector. We used Kendall's τ coefficient (Kendall, 1938) to calculate the ordinal association between ground truth full-finetuning results and the 1-NN results for each dataset.

Table 5 shows the correlation between full-finetuning and 1-NN accuracy. The Kendall's τ coefficient for correlation between ranking of different methods on CIFAR100, DTD, and CALTECH101 datasets are 0.60, 1.00, and 0.80, respectively. This indicates a good correlation between 1-NN and fine tuning, showing that 1-NN can be used to select the best pre-training method without actual finetuning.

	Kendall's τ	p-value
CIFAR100	0.60	0.23
DTD	1.00	0.02
CALTECH	0.80	0.08

Table 5: Kendall's τ coefficient to capture correlation between $1-$NN and full-finetuning.