Democratizing Contrastive Language-Image Pre-training: A CLIP Benchmark of Data, Model, and Supervision

Yufeng Cui [*1,2] **Lichen Zhao** [*1,2] **Feng Liang** [*3] **Yangguang Li** [2] **Jing Shao** [2]

Abstract

Contrastive Language-Image Pretraining (CLIP) has emerged as a novel paradigm to learn visual models from language supervision. While researchers continue to push the frontier of CLIP, reproducing these works remains challenging. This is because researchers do not choose consistent training recipes and even use different data, hampering the fair comparison between different methods. In this work, we propose CLIP-benchmark, a first attempt to evaluate, analyze, and benchmark CLIP and its variants. We conduct a comprehensive analysis of three key factors: data, supervision, and model architecture. We find considerable intuitive or counter-intuitive insights: (1). Data quality has a significant impact on performance. (2). Certain supervision has different effects for Convolutional Networks (ConvNets) and Vision Transformers (ViT). (3). Curtailing the text encoder reduces the training cost but not much affect the final performance. Moreover, we further combine DeCLIP (Li et al., 2021) with FILIP (Yao et al., 2021), bringing us the strongest variant DeFILIP. The CLIP-benchmark is released at: https://github.com/Sense-GVT/DeCLIP for future CLIP research.

1. Introduction

Over the past few years, supervised pre-training on well-labeled ImageNet (Deng et al., 2009) and then transferred to downstream tasks (Girshick et al., 2014; Long et al., 2015; Vinyals et al., 2015) has greatly transformed the computer vision (CV) community. However, supervised pre-training is hard to scale since we need arduous human labeling to specify new visual concepts. More Recently, Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) has emerged as a scalable pre-training paradigm via learning visual models from language supervision, or more specifically, image-text pairs. Basically, CLIP adopts the contrastive loss to push the embeddings of matched image-text pairs together while pushing those of non-matched pairs apart. Benefiting from abundant image-text pairs on the Internet, CLIP learns general visual features that could perform zero-shot recognition, *i.e.*, predict an image's category without seeing a single labeled example. CLIP's transferable features could also be well transferred to various downstream tasks.

Witnessing its great success, researchers continue to push the frontier of CLIP. For instance, SLIP (Mu et al., 2021), DeCLIP (Li et al., 2021) and FILIP (Yao et al., 2021) achieve considerable improvements via embracing different kinds of supervision within the image-text pairs. However, it remains challenging to make fair comparison between these methods. This is because they do not choose consistent training recipes and even use different data. Although CLIP (Radford et al., 2021), DeCLIP (Li et al., 2021) and SLIP (Mu et al., 2021) use the same amount of 15 million data from YFCC (Thomee et al., 2016), they adopt different filtering strategies. Moreover, methods (Radford et al., 2021; Jia et al., 2021; Li et al., 2021; Yao et al., 2021; Pham et al., 2021) crawl their datasets from the Internet, making the fair comparison more difficult.

This paper aims to democratize large-scale CLIP, *i.e.*, to build a fair and reproducible CLIP community. We propose CLIP-benchmark, a first attempt to evaluate, analyze, and benchmark CLIP and its variants. We do a comprehensive empirical study on three key factors: data, supervision, and model architecture. We find considerable intuitive or counter-intuitive insights:

Data: Mid-scale 15M data is a good balance of the training cost and performance. Thus, most methods use a 15M subset from YFCC (Thomee et al., 2016) to verify the effectiveness of their methods. We carefully compare the current two YFCC15M versions, V1 from CLIP (Radford et al., 2021) and V2 from DeCLIP (Li et al., 2021). Interestingly, we find

[*]Equal contribution [1]School of Software, Beihang University, China [2]SenseTime Research, Beijing, China [3]University of Texas at Austin, USA. Correspondence to: Yangguang Li <liyangguang@sensetime.com>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

that in terms of the zero-shot performance on ImageNet, V2 is much better than V1 (details in Tab. 2). We conjecture that V2 includes a more meticulous filtering strategy, making its data quality better than V1. This also helps us conclude that data quality is crucial in CLIP training.

Supervision: We first reproduce all the methods using a unified training recipe (details in Tab. 3). We find that fine-grained alignment supervision (Yao et al., 2021) could benefit ViT image encoder but hurts ConvNets. Intuitively, fine-grained alignment needs the image features to be non-overlapped, which is unachievable for ConvNets. For ViT image encoder, aggregating self-supervision (Mu et al., 2021; Li et al., 2021), multi-view supervision (Li et al., 2021), nearest-neighbor supervision (Li et al., 2021) and fine-grained alignment supervision (Yao et al., 2021) brings us the strongest variant DeFILIP.

Model: While most attention is paid to image encoders, little research is conducted on text encoders. Most literature follows the exact setting from CLIP, *i.e.*, a 12-layer transformer (Radford et al., 2019). We find that CLIP's text encoder is not necessary to be so much deep; a 3-layer transformer performs even better than the default 12-layer setting under the mid-scale data scenarios (details in Tab. 4). Therefore, pay attention to the text encoder when designing your CLIP models.

In a nutshell, this paper proposes the first CLIP-benchmark that includes the state-of-the-art methods. We benchmark these methods under the same training recipe using the same data. Our CLIP-benchmark also brings some insights about data, supervision and model. The CLIP-benchmark would be released to the public for future research.

2. Related Work

Concurrently to this work, many researchers continue to push the frontier of CLIP (Radford et al., 2021). SLIP (Mu et al., 2021) introduces self-supervision to Contrastive Language-Image Pretraining. DeCLIP (Li et al., 2021) utilizes widespread supervision among the image-text pairs. FILIP (Yao et al., 2021) leverages the finer-grained alignment between image patches and textual words. LiT (Zhai et al., 2021) adopt contrastive-tuning to tune the text tower using image-text data while using a pre-trained, strong image model as the image tower. OTTER (Wu et al., 2021) uses online entropic optimal transport to find a soft image-text match as labels for contrastive learning.

The representations learned by CLIP have shown excellent transferability over various tasks. CLIP2Video (Fang et al., 2021) and CLIP4Clip (Luo et al., 2021) apply CLIP to video retrieval task. ActionCLIP (Wang et al., 2021) utilizes CLIP for action recognition task. More works about improving image captioning with CLIP, *e.g.*CLIPCap (Mokady et al., 2021), CLIP4Caption (Tang et al., 2021). Interestingly CLIP can be even used in text-guided image generation task (StyleCLIP (Patashnik et al., 2021)) and Embodied AI (EmbCLIP (Khandelwal et al., 2021)). CLIP has also contributed to the development of general vision (Shao et al., 2021). Witnessing CLIP's active community and wide applications, we propose the first work to benchmark CLIP.

3. Methods

CLIP (Radford et al., 2021) and its variants (*e.g.*, DeCLIP (Li et al., 2021), SLIP (Mu et al., 2021), and FILIP (Yao et al., 2021)) follow a common high-level structure(see Fig. 1). The model consists of an image encoder (*e.g.*, ResNet (He et al., 2016) or ViT (Dosovitskiy et al., 2020)) and a text encoder(*e.g.*, transformer (Vaswani et al., 2017)), with a multimodal interaction at the top. Take the most straightforward CLIP as an example, the image encoder (the text encoder) extracts the image embedding (the text embedding) based on the input image-text pair. A contrastive objective is used to push the embeddings of matched image-text pairs together while pushing non-matched pairs apart. At the test phase, the learned text encoder synthesizes a zero-shot linear classifier by embedding the arbitrary categories of the test dataset. Because it is rare in the dataset that image caption is just a single word, CLIPs use prompts to make up the context of the category {`label`}, such as `"a photo of a {label}"`. As shown in the Fig.1, different variants further explore the widespread supervised signal of the image-text pair for better visual representations. This section will briefly introduce the above CLIP variants and bring the strongest variant DeFILIP.

3.1. CLIP

CLIP (Radford et al., 2021) only uses the original image-text supervsion. In a batch of N image-text pairs $\{(x_i^I, x_i^T)\}$, we denote x_i^I and x_i^T as image and text of the i_{th} pair. Let z_i^I and z_j^T be the normalized embedding of the i_{th} image and j_{th} text, respectively. CLIP uses InfoNCE loss (Van den Oord et al., 2018). The loss for the image encoder can be denoted as Eq. 1.

$$L_I = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathrm{sim}(z_i^I, z_i^T)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(z_i^I, z_j^T)/\tau)} \quad (1)$$

Here, the similarity function $\mathrm{sim}(,)$ is measured by dot product, and τ is a learnable temperature variable to scale the logits. We have a symmetrical loss for image and text encoder; thus, the overall loss function L_{CLIP} is the average of L_I and L_T.

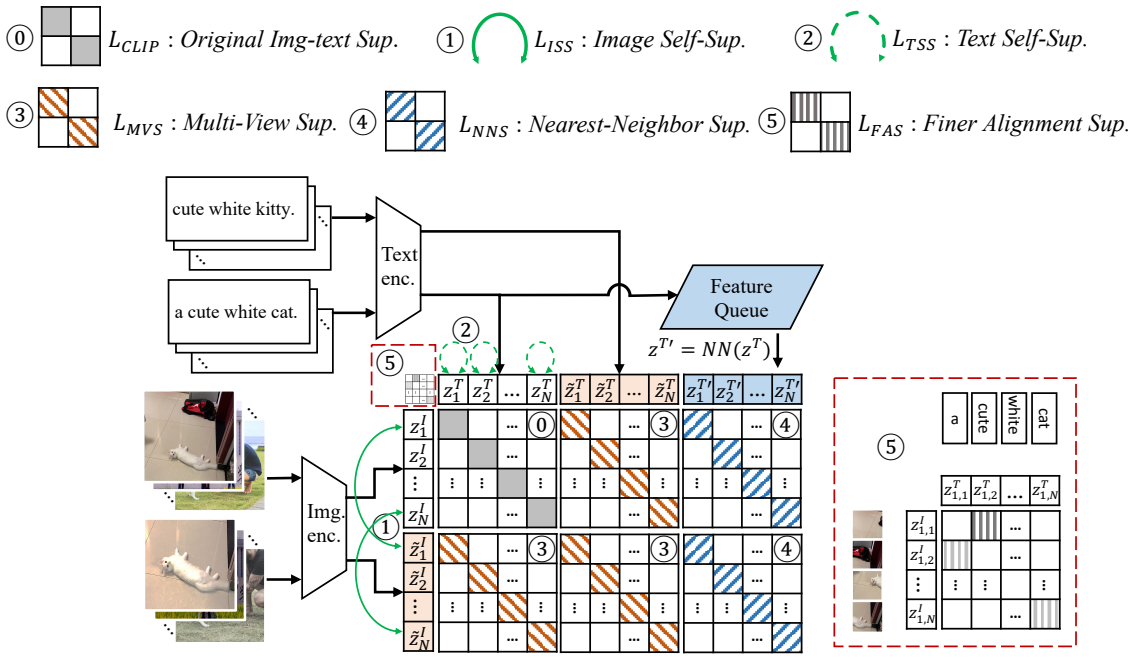

Figure 1. A unified framework of CLIP variants. Combining different supervision leads to different variants. CLIP: {⓪}, SLIP: {⓪,①}, FILIP: {⑤}, DeCLIP: {⓪,①,②,③,④}, DeFILIP: {⓪,①,②,③,④,⑤},

$$L_{CLIP} = (L_I + L_T)/2 \tag{2}$$

3.2. SLIP

SLIP (Mu et al., 2021) introduces self-supervision to CLIP for better visual representations. Built upon CLIP, SLIP gets two more strong augmented views for image self-supervised contrastive loss L_{ISS}. SLIP further compares different image self-supervised methods and finally selected SimCLR (Chen et al., 2020) for the final framework. The overall loss function of SLIP is shown in Eq. 3. α is the scale of self-supervision and is set to 1.

$$L_{SLIP} = L_{CLIP} + \alpha L_{ISS} \tag{3}$$

3.3. FILIP

FILIP (Yao et al., 2021) perform finer-grained alignment supervision on token level rather than image-text level. The similarity $\{sim(z_i^I, {}_i^T)\}$ of the i_{th} image and j_{th} text is improved to token-wise maximum similarity which is calculated as:

$$\begin{cases} \mathrm{sim}^{\mathrm{I}}(z_i^I, z_j^T) = \frac{1}{n_1}\sum_{k=1}^{n_1} z_{i,k}^I z_{j,m_k^I}^T \\ \mathrm{sim}^{\mathrm{T}}(z_i^I, z_j^T) = \frac{1}{n_2}\sum_{k=1}^{n_2} z_{i,m_k^T}^I z_{j,k}^T \end{cases} \tag{4}$$

Where $m_k^I = argmax_{0<r<n_2} z_{i,k}^I z_{j,r}^T$ and $m_k^T = argmax_{0<r<n_1} z_{i,r}^I z_{j,k}^T$. FILIP achieves finer-level alignment through a cross-modal late interaction mechanism, which uses a token-wise maximum similarity between visual and textual tokens to guide the contrastive objective. Though the late cross-modal interaction can capture finer-grained features, it relies on the token-wise representations of both modalities and can be inefficient in terms of communication, memory, and computation. To alleviate this problem, the authors carefully reduce the precision and embedding size of the model and further select the 25% tokens with the highest token-wise maximum similarity score among all texts (*resp*, images) in the same local worker before node communication. Denoting the loss of fine-grained alignment supervision as L_{FAS}, The overall loss function of FILIP is shown in Eq. 5.

$$L_{FILIP} = L_{FAS} \tag{5}$$

3.4. DeCLIP

DeCLIP (Li et al., 2021) utilizes widespread supervision among the image-text pairs, including Self-Supervision(SS), Multi-View Supervision(MVS), and Nearest-Neighbor Supervision(NNS). DeCLIP contains image SS and text SS: Image SS maximizes the similarity between two augmented views of the same instance while text SS leverages Masked Language Modeling(MLM) within a text sentence. For MVS, DeCLIP has two augmented views of both image and text, then contrasts the 2×2 image-text pairs. For NNS, De-

CLIP sample text NN in the embedding space as additional supervision.

In summary, DeCLIP denote L_{ISS} and L_{TSS} as the loss function of image SS and text SS, respectively. L_{MVS} is multi-view loss, and L_{NNS} is nearest-neighbor loss. The overall loss function of DeCLIP is shown in Eq. 6. α, β, γ are the loss scales and are both set to 0.2.

$$\begin{aligned} L_{DeCLIP} =& (1-\alpha-\beta-\gamma)L_{CLIP} \\ &+ \alpha(L_{ISS}+L_{TSS}) \\ &+ \beta L_{MVS} + \gamma L_{NNS} \end{aligned} \tag{6}$$

3.5. DeFILIP

By introducing the above methods, we can find a large number of possible supervision signals in the image-text pairs, which can improve the efficiency of training and generalization ability. In order to learn better visual representations and improve the data efficiency of the model, we further combine DeCLIP (Li et al., 2021) with FILIP (Yao et al., 2021), bringing us the strongest variant DeFILIP. The overall loss function of DeFILIP is shown in Eq. 7.

$$\begin{aligned} L_{DeFILIP} =& (1-\alpha-\beta-\gamma)L_{CLIP} \\ &+ \alpha(L_{ISS}+L_{TSS}) \\ &+ \beta L_{MVS} + \gamma L_{NNS} \\ &+ \lambda L_{FAS} \end{aligned} \tag{7}$$

L_{FAS} is applied to improve fine-grained learning of visual representations further. The loss weight λ is set to 0.2 in this work. As shown in fig. 1, our DeFILIP is a summary and development of the existing SOTA methods, which applies the existing supervision and achieves a new state-of-the-art performance.

4. CLIP-Benchmark

4.1. Setup

Evaluation Metric In this paper, we mainly evaluate zero-shot performance of different models on ImageNet (Deng et al., 2009), which is regarded as the main feature of CLIP methods. We perform prompt ensemble by averaging the caption embeddings for each class across the prompt templates. The prompts are the same as proposed in CLIP(Radford et al., 2021).

Implementation details The models in this work are trained and tested in the same codebase. Unless otherwise specified, all models are realized with Pytorch, and are trained with 32 NVIDIA A100 GPUs. When pretraining, we use an AdamW optimizer (Loshchilov & Hutter, 2017) with a total batch size of 4,096 (single GPU batch size 128 with 32 NVIDIA A100 GPUs). Starting with a learning rate (LR) of 0.0001, we linearly warm-yp the LR to 0.001 in one epoch, and then we use the cosine anneal LR decay strategy (Loshchilov & Hutter, 2016) to decrease the LR. The weight decay rate is set to 0.1. The input resolution of the image encoder is 224 $\times$ 224, and the maximum context length of the text encoder is 76. The learnable temperature parameter τ in Eq.1 is initialized to 0.07. All models are trained from scratch for 32 epochs.

Table 1. The basic statistics of the two versions of YFCC15M. V1 is filtered by CLIP. V2 is filtered by DeCLIP.

Dataset	Examples	Caption length	En-word ratio	Unique Tokens
V2	15,388,848	16.7±29.2	0.92	770,996
V1	14,747,529	26.1±69.6	0.72	8,262,556

Table 2. Zero-shot top1 accuracy on ImageNet. We train CLIP-ViT-B32 and our DeFILIP-ViT-B32 using different datasets

Method	Accuracy w/ V1	Accuracy w/ V2
CLIP	26.1	32.8
DeFILIP	36.4	45.0

4.2. Data

Data is a crucial part of CLIP. This section does a holistic study of two mid-scale YFCC15M versions. V1 from CLIP (Radford et al., 2021) and V2 from DeCLIP (Li et al., 2021).

Data statistics We present statistics of two versions YFCC15M on examples number, mean/std of caption length, mean English word ratio, and the vocabulary size (unique tokens) in Table 1. The V2 consists of 15.4M image-text pairs, 0.6M(3%) more than V1. V2 is generally shorter and more evenly distributed than V1 regarding the caption length. The English word ratio (*i.e.*, # of English words divided by # of all words) of V2 is about 0.92, which is significantly better than V1's 0.72. For the vocabulary size (unique tokens), V1 is one order larger than V2 mainly because V1 contains many non-English characters. We can infer from these statistics that V2 has better quality than V1 because V2 is more evenly distributed and has fewer non-English characters. We believe that V2 includes a more meticulous filtering strategy, making its data quality better than V1.

Performance over V1-V2. To further evaluate the quality of the two YFCC15M versions and explore the impact of data quality on CLIP, we perform a comparison with V1 (Radford et al., 2021) and V2 (Li et al., 2021), using the same methodology. As shown in Tab. 2, training with V2

Table 3. Zero-shot top1 accuracy on ImageNet. All models are trained with YFCC15M-V2 (Li et al., 2021). Δ denotes the improvement.

Method	Image encoder	Accuracy	Δ
CLIP	ResNet50	37.2	-
SLIP		28.5	-
FILIP		21.3	-
DeCLIP		**44.4**	+7.2
CLIP	ViT-B/32	32.8	-
SLIP		34.3	+1.5
FILIP		39.5	+6.7
DeCLIP		**43.2**	+10.4
DeFILIP	ViT-B/32	**45.0**	+12.2

leads to a better zero-shot performance than V1 under the same experimental setup. On the one hand, it proves that the data quality of V2 is better regarding final performance. On the other hand, it also proves that data quality significantly impacts the performance of CLIP methods.

4.3. Supervision

We perform a comprehensive comparison of our re-implemented pretraining methods (Radford et al., 2021; Li et al., 2021; Mu et al., 2021; Yao et al., 2021) to benchmark these methods under the same training recipe. We report the zero-shot top-1 accuracy on ImageNet in Tab. 3 When the image encoder is ViT, all supervision is proved to be effective. DeCLIP, which utilizes the maximum supervision, obtains the best results. Moreover, we further integrate the existing supervision to make the strongest variant, named DeFILIP. Our proposed DeFILIP reaches 45.0% accuracy, surpassing the CLIP baseline by a considerable 12.2% margin.

When we use ResNet as the image encoder, some methods seem cannot preserve the improvement. Worth mentioning, SLIP (Mu et al., 2021) and FILIP (Yao et al., 2021) do not report the results of ResNet models. We conjecture there are two reasons: 1) ResNet models might need more dedicated hyper parameter tuning. 2). Fine-grained alignment requires the image features to be non-overlapped, which is unachievable for ConvNets. However, DeCLIP can still brings 7.2% improvement over the CLIP baseline.

4.4. Model

While most attention is paid to image encoders, little research is conducted on text encoders. Most literature follows the exact setting from CLIP, *i.e.*, a 12-layer transformer. Therefore, we expect to study the role of the text encoder, and further explore whether the training efficiency can be improved by reducing the parameters of the text encoder without affecting the performance.

Table 4. Zero-shot top1 accuracy on ImageNet. All models are trained with YFCC15M-V2 (Li et al., 2021). The image encoder is ViT-B32, we vary the layer number of transformers in the text encoder.

Method	Layer number	Accuracy
CLIP	1	29.9
	3	34.2
	6	34.3
	12	32.8
DeFILIP	1	39.7
	3	44.1
	6	44.3
	12	45.0

As shown in Tab. 4, we try 1/3/6/12-layer transformer for CLIP-ViTB32 and DeFILIP-ViTB32. Surprisingly, we find that (1) For the primitive CLIP method, text encoders with 6 layers of transformers achieve the best results instead of the default 12 layers. A 3-layers transformer is enough to achieve high results. (2) For the DeCLIP, which applies more supervision, the text encoder is more critical. However, even if half the number of layers, it does not significantly affect the final performance. Such an exciting result shows that curtailing the text-encoder is an efficient approach to reducing training costs.

5. Conclusions

In this paper, we propose the first CLIP-benchmark that includes state-of-the-art methods. We benchmark these methods under the same training recipe using the same data. Our CLIP-benchmark also brings some insights about data, supervision, and model. Moreover, we further propose DeFILIP to make a stronger baseline for this task. The CLIP-benchmark would be released to the public for future research. We hope this technical report could avoid duplicate data cleaning efforts and provide a consistent benchmark to facilitate fair comparisons.

References

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.

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.

Fang, H., Xiong, P., Xu, L., and Chen, Y. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

Khandelwal, A., Weihs, L., Mottaghi, R., and Kembhavi, A. Simple but effective: Clip embeddings for embodied ai. *arXiv preprint arXiv:2111.09888*, 2021.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., and Li, T. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

Mokady, R., Hertz, A., and Bermano, A. H. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

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

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 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., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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. *arXiv preprint arXiv:2103.00020*, 2021.

Shao, J., Chen, S., Li, Y., Wang, K., Yin, Z., He, Y., Teng, J., Sun, Q., Gao, M., Liu, J., et al. Intern: A new learning paradigm towards general vision. *arXiv preprint arXiv:2111.08687*, 2021.

Tang, M., Wang, Z., Liu, Z., Rao, F., Li, D., and Li, X. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4858–4862, 2021.

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.

Van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv e-prints*, pp. arXiv–1807, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Wang, M., Xing, J., and Liu, Y. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

Wu, B., Cheng, R., Zhang, P., Vajda, P., and Gonzalez, J. E. Data efficient language-supervised zero-shot recognition with optimal transport distillation. *arXiv preprint arXiv:2112.09445*, 2021.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. *arXiv preprint arXiv:2111.07991*, 2021.