Impact of Language Guidance: A Reproducibility Study

Anonymous authors
Paper under double-blind review

Abstract

Modern deep-learning architectures need large amounts of data to produce state-of-the-art results. Annotating such huge datasets is time-consuming, expensive, and prone to human error. Recent advances in self-supervised learning allow us to train huge models without explicit annotation. Contrastive learning is a popular paradigm in self-supervised learning. Methods like SimCLR Chen et al. (2020) rely on image augmentations or directly minimizing cross-modal loss between image and text. El Banani et al. El Banani et al. (2021) propose to use language guidance to sample view pairs. They claim that language enables better conceptual similarity, eliminating the effects of visual variability. We reproduce their experiments to verify their claims. We find that their dataset, RedCaps Desai et al. (2021), sourced from Reddit, contains low-quality captions. We use an off-the-shelf image captioning model, BLIP-2 Li et al. (2022), to replace the captions and analyze the effect on model performance. We also use interpretability methods to demonstrate that the model learns semantic information well.

1 Introduction

Deep learning thrives on large datasets and compute-intensive training. In the age of the internet, unlabeled data is abundant. Supervised learning algorithms require annotated data. Annotation of huge datasets is prohibitively expensive, labor-intensive, and prone to human error. Self-supervised learning (SSL) enables the model to learn rich and transferable representations from unlabeled data Devlin et al. (2018); He et al. (2016). This has unlocked new possibilities and transformed both computer vision Chen et al. (2020); Caron et al. (2021) and natural language processing Devlin et al. (2018).

Contrastive learning is a self-supervised learning technique in which a model is trained to produce similar representations for similar images while ensuring distinct representations for dissimilar images. SimCLR Chen et al. (2020) uses image augmentations like random crop, Gaussian blur, and random flipping to generate a positive pair while treating other images as negative samples. Other methods Caron et al. (2018); Wu et al. (2018) use clustering algorithms or nearest neighbor operations to find positive samples. These methods use only visual similarity to determine similar images. Two objects might be visually similar, while objects of the same class might be visually dissimilar, as shown in Fig. A. In contrast to this, conceptually similar images are more often described similarly. This suggests that leveraging language modality can improve contrastive learning.

CLIP Radford et al. (2021) learns a joint embedding space for images and their captions. This yields highly generalizable and accurate representations. However, El Banani et al. El Banani et al. (2021) suggest that combining embedding spaces of different modalities might not give optimal results. They propose a new sampling procedure where image pairs are sampled using caption similarity for contrastive learning.

El Banani et al. El Banani et al. (2021) retrain existing self-supervised visual learning architectures Chen et al. (2020); Caron et al. (2018); Wu et al. (2018) with the proposed sampling strategy. Their experiments show that the newly proposed method outperforms all baselines on varying downstream tasks across multiple datasets. This substantiates the claim that language is a better proxy for conceptual similarity.

We aim to rigorously evaluate these claims by closely replicating the experimental setup and results reported in the original paper. We identify poor caption quality in the dataset Desai et al. (2021). We evaluate the

method using better captions from an off-the-shelf caption generator Li et al. (2022) and analyze the results. We also demonstrate that the model learns semantic information using interpretability methods.

2 Scope for Reproducibility

The main contribution of the original paper is a new language-based sampling strategy, and their claim is that this strategy improves the underlying self-supervision framework. Stu

In an effort to reproduce the paper and gain a deeper understanding, we discovered several key limitations:

- **High Computational Requirements:** The results reported utilize ResNet-50 with a batch size of 512 trained from scratch, demanding computational resources beyond typical academic settings.
- **Inefficient Captions:** The method's effectiveness is heavily dependent on image captions. Our analysis reveals that the captions scraped from Reddit are often noisy, vague, and frequently provide inaccurate image descriptions, potentially hampering the model training process significantly.
- **RedCaps Dependency:** The method achieves optimal performance when pairs are subsampled from specific subreddit subsets, introducing weak supervision and dataset-specific constraints.

2.1 Our Contributions

To address these limitations and extend the work, we make the following contributions:

- **Reproducibility:** We provide a detailed replication of the training pipeline and hyperparameters from the original paper, adapted for reduced computational environments.
- **Visual Backbone Optimization:** We investigate the generalization capabilities of language-guided SSL to smaller, more efficient architectures such as MobileNet, making the approach more accessible within academic resource constraints.
- **Caption Quality Enhancement:** We develop a curated set of refined captions for the existing dataset, enabling direct analysis of textual guidance impact on SSL frameworks.
- **Captioning Model Integration:** We incorporate a captioning model into the existing architecture, reducing RedCaps dataset dependency and enabling generalization to diverse datasets.
- **Weak Supervision Analysis:** We conduct a comprehensive study on the impact of subreddit sub-class weak supervision on model performance.
- **New metric:** We generate saliency maps, which are used to create a new metric for evaluating SSL-trained convnets.

3 Related Work

Visual Representation Learning involves learning an embedding space from data that captures visual information effectively. Unlike typical machine learning tasks like classification or segmentation we cannot optimize loss directly on embeddings. This is due to a lack of ground truth labels. There are two main approaches that have been explored for this task: generative and discriminative. Generative approaches [cite all] involve learning a model that can capture image distribution. Discriminative approaches [cite all] involve different downstream tasks like classification, dimensionality reduction and metric learning. The hypothesis is that the features internally learned for these tasks capture semantic information well.

Image only contrastive learning is used for visual representation learning. [cite Wu et al.] propose generating positive pairs for each image via augmentations and treating all others as negative samples for contrastive learning. While augmentation-based contrastive learning performs well, its utility has been

questioned [cite]. Some papers [cite] propose using objectness and saliency to correct for these issues [cite]. Later work [cite] uses nearest neighbour search or clustering to find positive pairs.

Using language for contrastive learning [cite clip etc] learn a joint vision language embedding space by minimizing cross modal loss. While this approach is effective have same space for vision and language means distance in vision space should map directly to language space. [cite original paper] propose to use language guidance to sample similar image pairs and do image-image contrastive learning.

4 Methodology

Traditional SSL methods, especially those based on instance discrimination, rely on image augmentations (e.g., cropping, color jitter) to generate positive pairs, assuming that visual similarity mirrors semantic similarity. However, these methods are limited to learning invariances specific to the applied augmentations, potentially missing higher-level semantic cues.

Our reproducibility study examines a language-guided sampling strategy that uses textual captions to identify semantically similar images. The hypothesis is that similar captions capture shared conceptual content beyond what visual augmentations can provide.

The original paper uses RedCaps, a dataset scrapped from Reddit. It is a set of images and the metadata originating along with it on Reddit, they created image caption pairs from this where caption are the user written captions.

4.1 Pair Sampling

For sampling image pairs, we need to find the most similar captions in our dataset we found their selection of SBERT with cosine similarity to be well-justified. Metrics like BLEU and CIDER are also generally used for finding caption similarity, but these n-gram based approaches would have been too sensitive to variations in phrasing and sentence structure. Even SPICE, which uses parse trees and handles structural variations better, is limited in dealing with different word choices for the same concept. From our reproduction perspective, this methodological choice was foundational to their framework's success. SBERT effectively identifies semantically similar caption pairs while being robust to surface-level text variations. The use of cosine similarity simplifies the implementation while maintaining reliable semantic matching capabilities. It is observes that the method is still agnostic to the sentence encoder chosen. Using the FAISS algorithm(cite this), nearest neighbors are calculated in the language embedding space for a caption, and the image corresponding to that caption is chosen as the positive sample when fed into various SSL frameworks. It took xxxxxx mins for complete similarity search over our dataset.

4.2 Improving Dataset

We identified that the quality of Reddit-sourced captions could be a significant limiting factor. The RedCaps dataset, while extensive, contains captions that are often vague, noisy, and inconsistent in their descriptive quality. To test this hypothesis and potentially improve the method, we introduced BLIPv2 as an alternative caption generation approach. BLIPv2 generates concise, descriptive captions that maintain consistent quality across the dataset. Our modification serves two key purposes: first, it allows us to evaluate whether higher-quality captions improve the performance of language-guided SSL, and second, it removes the dependency on pre-existing captions altogether. This latter point also enables the framework to be extended to any image dataset, regardless of whether it contains associated text descriptions. We adopt a filtering strategy similar to CapFilt proposed by (Cite BLIPv2 yaha), where we generate new captions using BLIPv2 and evaluate their quality using CLIP scores. Specifically, we compute the CLIP similarity scores between the original and newly generated captions with the corresponding image. The caption with the higher CLIP score is retained for sampling, ensuring that only the most semantically relevant descriptions guide the contrastive learning process.

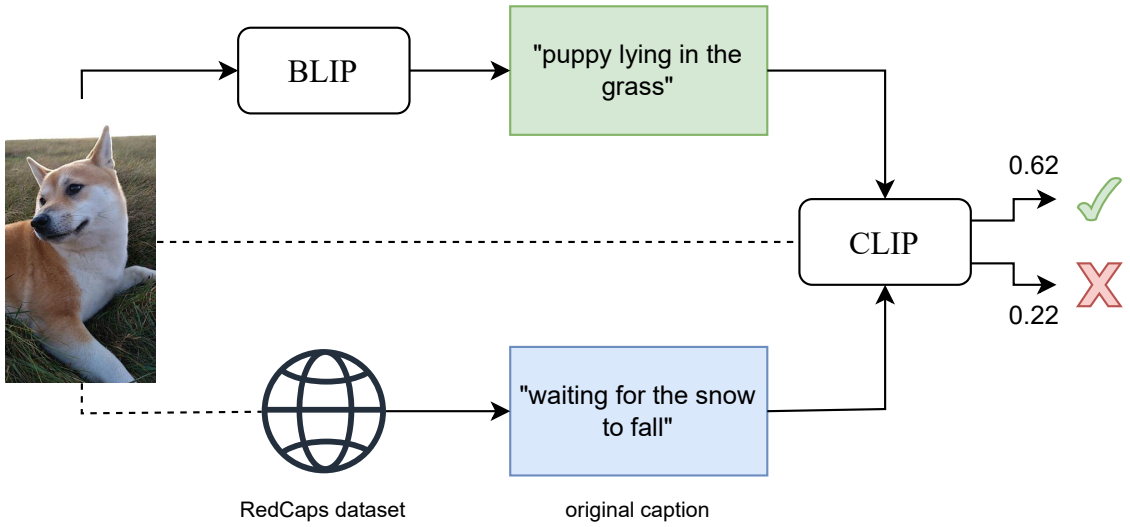

Figure 1: An overview of our caption improvement method.

4.3 Learning visual representations

Every SSL framework use some kind of visual backbone either Convnets or ViTs(Cite Dino) for learning visual representations. Orignal paper uses ResNet50 as the visual backbone but due to the high computational demands of ResNet50 with larger batch sizes, we opted for ResNet34. This choice not only reduced resource requirements, but also allowed us to evaluate the transferability of our method to smaller models and examine the effects of a reduced feature embed size (512 for ResNet34 versus 2048 for ResNet50).

4.4 Visualizing learned representations

We use self-supervised learning methods to learn visual representations. The original paper does not evaluate learned semantic information explicitly. [SirGur et al cite] propose a new unsupervised procedure that combines gradient based methods and attribution methods of visualization. We train a linear probe for classification on SimCLR and LGSimCLR. We use train split of ImageNet-S50 for this training. We then use the algorithm proposed by [cite Shirgur et al.] to generate saliency maps.

5 Experimental Setup

We primarily base our experiments on the code provided by the authors (ADD FOOTNOTE REFERENCE). Our experimental evaluation focuses on thoroughly validating the impact of improved captions on self-supervised learning frameworks. We explore multiple frameworks while maintaining consistent training conditions across all experiments to ensure fair comparisons.

5.1 Frameworks

To comprehensively evaluate the effect of enhanced captions, we conduct experiments across a diverse set of self-supervised learning frameworks. Our study includes SimCLR, LGSimCLR, SimSiam, SwAV, and NNCLR. This selection enables us to verify whether the performance improvements from better captions generalize across different architectural approaches to self-supervised learning, as demonstrated in Table **??**.

5.2 Training Details

We implement our experiments using a ResNet-34 backbone, chosen for its balance of computational efficiency and representational capacity. For optimization, we employ the AdamW optimizer with the originally used hyperparameters: a learning rate of 0.001 and weight decay of 0.01. The learning schedule follows a cosine decay pattern with 5000 warm-up steps, which helps stabilize early training.

To ensure meaningful comparisons across different experimental conditions, we maintain consistent training parameters throughout our studies. Each model processes data in batches of 512 images, leveraging efficient GPU utilization while staying within memory constraints. Training continues for a fixed number of steps across all experiments. This training and nearest neighbour search is done on NVIDIA V100 where each epoch of training took approximately 1.5hrs. We also observed that while training the bottleneck was not the training time on the GPU but the data loading step which varied according to the number of workers per GPU.

5.3 Datasets and Caption Sources

Our primary image source is RedCaps-2020, a subset of the RedCaps dataset comprising 2.8 million image-text pairs uploaded on Reddit in the year 2020. This dataset serves as our foundation for comparing caption quality effects. We explore two distinct caption sources in our experiments. First, we establish baseline performance using the original RedCaps captions. Then, we generate enhanced captions using pre-trained BLIPv2, allowing us to directly measure the impact of caption quality on model performance.

5.4 Evaluation Protocol

We run different downstream tasks on the frozen features for each model across multiple datasets to evaluate their performance. Similar to the original authors, we evaluate the model on linear-probe classification **?** and few-shot classification **? ?** . We were able to reproduce results for all datasets mentioned in the original paper except Sun397, Cars, Caltech-101 and Oxford Flowers. The Sun397 dataset has several corrupted images; Cars dataset has been removed from the host site; and the authors' code implementation to download Caltech-101 and Oxford Flowers is not working. Additionally, we report results using a new approach to evaluate self-supervised models using saliency maps.

Saliency Map Evaluation We generate saliency maps using method described in subsection 4.4. We evaluate the maps using segmentation metrics - IoU and mAP. We use validation split of ImageNet-S50 for evaluation. Results are reported in TableS

6 Results and Discussion

We report the results in **??** and **??**. The models trained with language guidance outperforms their corresponding baseline in most cases. However, our experiments suggest that the impact of language guidance is not as profound as indicated in the original paper. The performance disparity between ResNet34 and ResNet50 may be largely attributed to differences in the size of their feature embedding spaces rather than their overall parameter counts as the number of parameters in both models are comparable(21.79M and 25.5M respectively). While both architectures have a comparable number of parameters, ResNet50 produces a 2048-dimensional feature representation, compared to only 512 dimensions in ResNet34. This larger embedding space in ResNet50 likely allows the network to capture a richer and more nuanced set of features. Conversely, the reduced capacity of a 512-dimensional embedding may limit the model's ability to fully exploit the semantic cues provided by the language guidance, resulting in a less pronounced improvement in performance. Neglection of this factor lead to over-estimation of generalizability of this method to other models

Additionally, our caption improvement approach strengthens language guidance across all SSL frameworks by providing clearer, more consistent captions. This reduces ambiguity and allows models to better align visual features with concepts, resulting in improved performance.

standard linear probing approach(table x) and few-shot evaluation (table x)

Models	**Food101**	**CIFAR10**	**CIFAR100**	**Cub2011**	**Aircraft**	**DTD**	**Pets**	**STL10**	**Eurosat**	**Resisc45**	**Avg**
SimCLR	61.2	77.5	53.0	27.3	36.0	61.9	66.4	85.1	94.2	78.7	
VisSimSiam	56.0	72.0	46.1	21.1	27.9	58.6	55.9	85.2	92.3	72.4	
VisNNCLR	52.7	69.9	45.4	17.3	25.7	59.2	56.8	83.6	91.8	71.8	
SWAV											
Banani et al.											
LGSimCLR											
LGSimSiam											
Ours											
LGSimCLR	59.3	72.4	48.3	24.1	26	56.4	64.1	86.7	90.9	73.2	
LGSimSiam											

Table 1: Linear Probe Evaluation

Models	**Food101**	**CIFAR10**	**CIFAR100**	**Cub2011**	**Aircraft**	**DTD**	**Pets**	**STL10**	**Eurosat**	**Resisc45**	**Avg**
SimCLR	67.8	52.9	59.5	54.7	41.5	74.6	73.6	73.9	82.0	77.5	
VisSimSiam	61.2	51.4	56.6	43.6	33.5	72.1	62.9	73.1	75.2	68.6	
VisNNCLR	64.0	50.9	56.4	45.1	33.8	70.7	71.0	74.8	75.2	69.3	
SWAV											
Banani et al.											
LGSimCLR											
LGSimSiam											
Ours											
LGSimCLR	77.9	57.5	65.9	64.2	39.2	71.3	78.2	80.6	80.5	76.3	
LGSimSiam											

Table 2: Few Shot Evaluation

References

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT)*, 2018.

Mohamed El Banani, Luya Gao, and Justin Johnson. Unsupervisedr&r: Unsupervised point cloud registration via differentiable rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7129–7139, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.