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. Recent works like SimCLR and CLIP rely on image augmentations or directly minimizing cross-modal loss between image and text. Banani et al. (2023) 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, contains low-quality captions. We use an off-the-shelf image captioning model, BLIP-2, to replace the captions and improve performance. We also devise a new metric to evaluate the semantic capabilities of self-supervised models based on interpretability methods.

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, labour-intensive, and error-prone. Self-supervised learning (SSL) enables the model to learn rich and transferable representations from unlabeled data. This has unlocked new possibilities and transformed both computer vision (Chen et al., 2020a; 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., 2020a) uses image augmentations such as 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 neighbour operations to find positive samples. These methods only use visual similarity to find similar images. Two objects might be visually similar, while objects of the same class might be visually dissimilar. In contrast to this, conceptually similar images are more often described similarly. This suggests that leveraging language modality can improve contrastive learning.

Radford et al. (2021) propose learning a joint embedding space for images and their captions. This yields highly generalizable and accurate representations. However, Banani et al. (2023) suggest that combining embedding spaces of different modalities might lead to sub-optimal results. They propose a new sampling procedure for contrastive learning where image pairs are sampled using caption similarity based on embeddings generated using a pre-trained language encoder.

Banani et al. (2023) retrain existing self-supervised visual learning architectures (Chen et al., 2020a; 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 good 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 the poor caption quality of the dataset(Desai et al., 2021) used by the

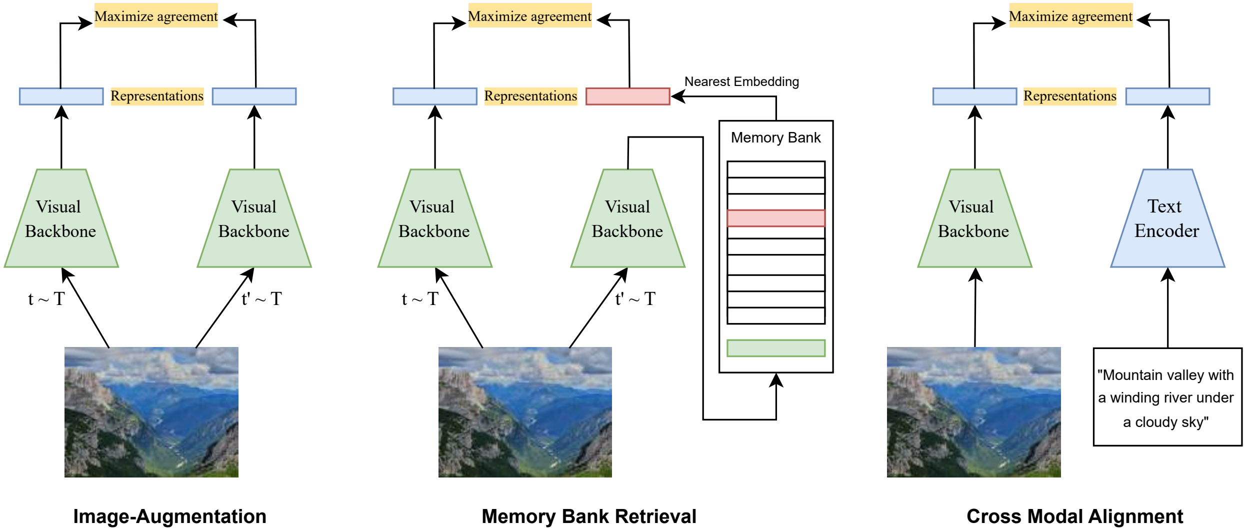


Figure 1: **Refining Representation Learning**: While early contrastive learning methods relied on simple image transformations, newer structured retrieval techniques have emerged to refine the learned embeddings beyond instance-level. These employ clustering, memory banks, or language-driven sampling to introduce structure into training signals and subsequently improve visual representation learning.

original authors and generate better captions from an off-the-shelf caption generator (Li et al., 2023) and analyze the performance improvement. We also demonstrate that the model learns semantic information using interpretability methods (Selvaraju et al., 2017).

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 (Banani et al., 2023).

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

- **Inefficient Captions:** The method's efficacy is heavily dependent on image captions. Manual inspection of the dataset, RedCaps, reveals that the captions scraped from Reddit are often noisy, vague and inaccurate potentially hampering the model training process.

- **RedCaps Dependency:** The method achieves optimal performance when pairs are sampled from specific subreddit subsets, introducing weak supervision (Zheng et al., 2021). The dataset-specific constraints reduce the generalizability of the method.

- **High Computational Requirements:** The reported results utilized ResNet-50 (He et al., 2016) with a batch size of 512 trained from scratch, which requires substantial computational resources that may not be readily available in many settings.

2.1 Our Contributions

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

- **Reproducibility:** We provide an exhaustive replication of most experiments in 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 ResNet34 (Wightman et al., 2021) and MobileNetV3 (Howard et al., 2019), making the approach accessible within resource constraints.
- **Caption Quality Enhancement:** We develop a curated set of refined captions for the existing dataset (Desai et al., 2021), improving the impact of language guidance and reducing dataset dependency and enabling generalization to diverse datasets.
- **New Metric:** We generate saliency maps (Selvaraju et al., 2017), which are used to create a new metric for evaluating SSL-trained ConvNets.

3 Background

In this section, we provide an overview of the foundational works and existing research that have contributed to advancements in this field. We discuss relevant literature, methodologies, and key developments that form the basis of our study, highlighting their significance in the context of our work.

Visual Representation Learning involves learning to encode visual information in an embedding space that preserves its semantics well. Unlike typical machine learning tasks like classification or segmentation, we cannot manually annotate ground truth labels for this task. So, we cannot directly optimize loss in embedding space. Two main approaches have been explored for this task: generative and discriminative. Generative approaches (Doersch et al., 2015; Gidaris et al., 2018; Oord et al., 2018; Vincent et al., 2008; Zhang et al., 2016) involve learning a model that can capture image distribution well. Such models are hypothesized to learn semantically relevant features. Discriminative approaches involve learning a model that can differentiate between images. Understanding semantically relevant features is essential for excelling in tasks like metric learning (Chopra et al., 2005), dimensionality reduction (Hadsell et al., 2006) and classification (Sharif Razavian et al., 2014). Self-supervised learning has recently gained popularity as a visual representation learning method. They relieve the need for human annotation and allow learning from large, unlabelled data sources. Various contrastive (Wu et al., 2018; Chen et al., 2020a;b;c; He et al., 2020), and non-contrastive (Chen & He, 2021; Grill et al., 2020) approaches have been proposed recently. The paper we reproduce (Banani et al., 2023) proposes a sampling strategy for contrastive visual representation learning.

Image-Image Contrastive Learning. Contrastive learning involves learning an embedding space in which similar images are close and dissimilar images are far away. Sampling positive and negative pairs effectively is an essential task for the effectiveness of contrastive learning. Chen et al. (2020a) propose a framework called SimCLR for contrastive learning. It involved using data augmentation to generate positive samples for each instance while treating all other images in the batch as negative samples. SimSiam (Chen & He, 2021) uses a similar approach but does not use negative pairs during training. It also implements a stop gradient operation in one branch of the Siamese network. It claims that the stop gradient operation is essential to prevent model collapse. Caron et al. (2020) propose a method called SwAV, which reduces computational complexity as it does not need to calculate explicit pairwise comparisons. It relies on an online clustering approach. We compute the feature of an image and then compute its code by matching it with a set of k prototype vectors. Then, we predict the code from an augmented view of the image. These approaches mitigate the demand for a memory bank and reduce computational complexity, ignoring similarities between different instances. NNCLR, proposed by Dwibedi et al. (2021), utilizes similarity between different instances along with transformations to account for more semantic variation. It uses a memory bank similar to He et al. (2020) and samples the nearest neighbour of the image in latent space. It then minimizes the loss between the nearest neighbour and random augmentation of the original image.

Using language for contrastive learning Past work has aimed to learn joint vision-language representations for tasks like Visual Question Answering (Antol et al., 2015; Goyal et al., 2017; Hudson & Manning, 2019; Zhu et al., 2016), Visual Reasoning (Kazemzadeh et al., 2014; Suhr et al., 2019; Zellers et al., 2019) and Retrieval (Park et al., 2022; Young et al., 2014). Radford et al. (2021) introduced CLIP, which learned a joint vision-language embedding space by directly minimizing cross-modal loss. This approach was widely adopted due to its generalization and few shot capabilities. Other work improved upon this by adding addi-

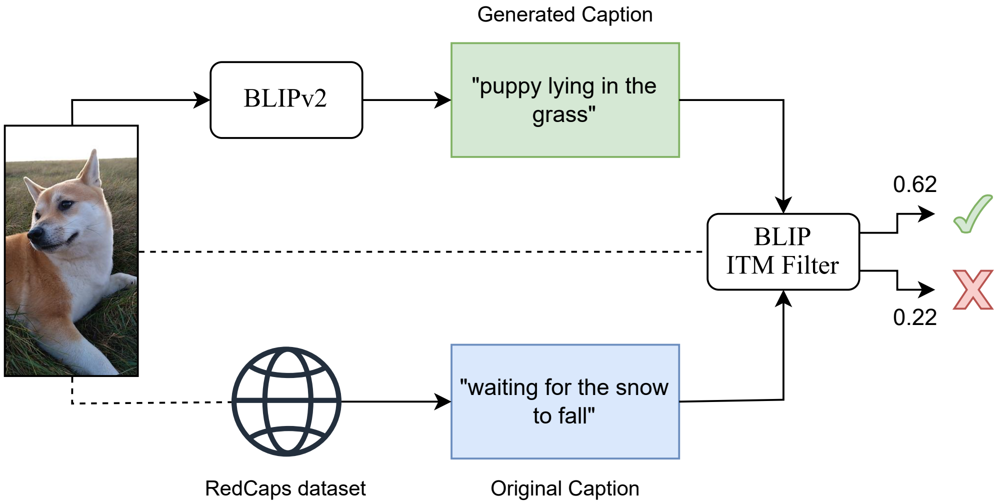


Figure 2: **Improving Captions via Contrastive Filtering**: Our caption improvement method leverages BLIPv2 to generate candidate captions that better describe an image. Since dataset-provided captions can be less relevant or inaccurate, we introduce an Image-Text Matching (ITM) Filter to evaluate, assign a relevance score and select the most appropriate caption between of the two. This ensures better semantic and visual alignment of a caption with its corresponding image.

tional losses and other improvements (Jia et al., 2021; Xu et al., 2022; Yao et al., 2022; Cui et al., 2022; Lee et al., 2022; Li et al., 2022; Mu et al., 2022).

Banani et al. (2023) aims to achieve good results in image-image contrastive learning with the help of language to sample positive pairs.

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(Desai et al., 2021), a dataset scraped 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 captions are user-written.

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 (Reimers & Gurevych, 2019) with cosine similarity to be well-justified. Metrics like BLEU Papineni et al. (2002) and CIDER Vedantam et al. (2015) 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 Anderson et al. (2016), 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 observed that the method is still agnostic to the sentence encoder chosen. Using the FAISS algorithm Johnson et al. (2019), 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 21 minutes on our system 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 where we generate new captions using BLIPv2 and evaluate their quality using Image-Text Matching (ITM) scores. The ITM score measures the alignment between an image and its caption by predicting whether they are a good match. Higher ITM scores indicate captions that are more semantically relevant to the image. Captions with higher ITM scores are retained to ensure better language guidance for contrastive learning, as shown in Figure 2.

4.3 Background

Every SSL framework use some kind of visual backbone either ConvNets or ViTs(Caron et al. (2021)) 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 embedding size as well (512 for ResNet34 versus 2048 for ResNet50).

4.4 Visualizing learned representations

We use self-supervised learning methods to learn visual representations. These representations are used for downstream tasks and evaluated on it. However, it is essential to inspect if the model is focussing on right regions of the image. We train a linear probe on the trained ResNet-34 backbone. We train of train set of ImageNetS-50. We apply the method proposed by Selvaraju et al. (2017), namely GradCAM. We apply GradCAM on second convolutional layer 4 of the backbone with true class of the image.

5 Experimental Setup

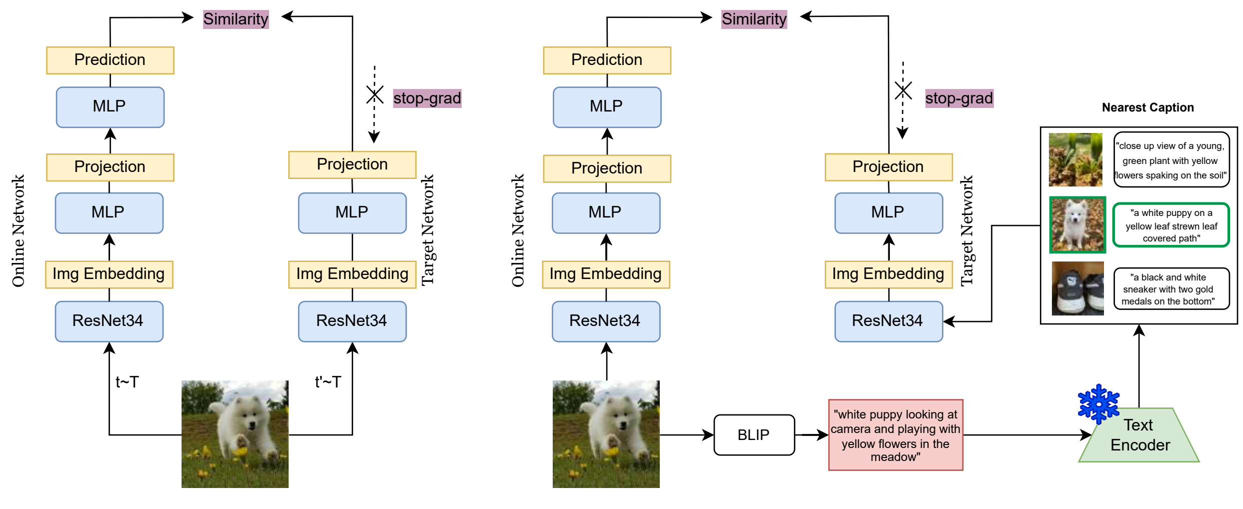
We primarily base our experiments on the code provided by the authors [1]. Our experimental evaluation focuses on thoroughly validating the impact of improved captions on self-supervised learning frameworks. We explore multiple frameworks while maintaining consistent training conditions across all experiments 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 2.

---

[1] https://github.com/mbanani/lgssl


Similarity
Prediction
MLP
Projection
MLP
Img Embedding
ResNet34
stop-grad
Online Network
Target Network
t~T
t'~T
Similarity
Prediction
MLP
Projection
MLP
Img Embedding
ResNet34
stop-grad
Nearest Caption
"close up view of a young, green plant with yellow flowers spaking on the soil"
"a white puppy on a yellow leaf strewn leaf covered path"
"a black and white sneaker with two gold medals on the bottom"
BLIP
"white puppy looking at camera and playing with yellow flowers in the meadow"
Text Encoder

SimSiam

Language Guided SimSiam (Ours)

Figure 3: A schematic comparison of SimSiam and Language Guided SimSiam trained using our pipeline.

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 (Loshchilov & Hutter, 2016) 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 1.5 hours.

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 (Kornblith et al., 2019) and few-shot classification (Wang et al., 2019). 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 saliency maps using Area Under Precision Recall Curve (AUC-PR) and Area Under ROC Curve (AUC-ROC) (Cong et al., 2018). These are calculated by treating ground truth segmentation map as ground truth labels for pixel level classification. We use validation split of ImageNet-S50 for evaluation.

Results are reported in Table 1. We can see that the metrics are not signficantly different. So we can conclude that language guidance has insignificant impact on quality of saliency maps.

Metric	SimCLR	LGSimCLR	LGSimCLR (*Ours*)
AUC-ROC	0.5411	0.5418	0.5427
AUC-PR	0.3419	0.3382	0.3425

Table 1: Saliency Map Evaluation. We evaluate across three models i.e. SimCLR, LGSimCLR on original captions and LGSimCLR on new captions (Ours) across two metrics. All the models were retrained with ResNet-34 backbone.

6 Results and Discussion

We report the results in Table 2 and Table 3. 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.

Models	**Food101**	**CIFAR10**	**CIFAR100**	**CUB**	**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	64.13
VisSimSiam	56.0	72.0	46.1	21.1	27.9	58.6	55.9	85.2	92.3	72.4	58.75
VisNNCLR	52.7	69.9	45.4	17.3	25.7	59.2	56.8	83.6	91.8	71.8	57.42
SWAV	50.7	69.0	43.0	15.3	23.2	57.6	54.1	83.2	90.4	68.6	55.51
Banani et al.											
LGSimCLR	73.1	81.3	58.7	46.5	42.2	63.1	73.7	88.4	92.1	80.3	69.94
LGSimSiam	68.2	74.6	52.1	39.5	36.0	54.6	63.4	90.1	92.5	76.0	55.49
Ours											
LGSimCLR	59.3	72.4	48.3	24.1	26.0	56.4	64.1	86.7	90.9	73.2	67.45
LGSimSiam	57.2	74.5	48.9	23.2	30.1	59.5	56.8	85.4	92.4	73.4	60.14

Table 2: We report performance of a linear probe using frozen features on 10 downstream tasks. LGSimCLR-*Ours* outperforms previous approaches on most datsets. *Ours* refers to the models trained on new captions. We retrain all models with ResNet-34 backbone.

7 Limitations

Our experiments, when compared to the base experiments in , support the observation that incorporating a captioning model improved upon the original results. However, the noisy and vague nature of the captions generated by the model limited the generality of our results. This underscores the need for a more accurate captioning model, which could potentially perform better in language-guided sampling tasks. Complete reproducibility of the original experiments could not be confirmed for three primary reasons. Firstly, computational constraints prevented us from replicating the experiments in the original paper that utilized the ResNet-50 architecture with a batch size of 512. Secondly, the original study conducted experiments on data

Models	**Food101**	**CIFAR10**	**CIFAR100**	**CUB**	**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	65.80
VisSimSiam	61.2	51.4	56.6	43.6	33.5	72.1	62.9	73.1	75.2	68.6	59.82
VisNNCLR	64.0	50.9	56.4	45.1	33.8	70.7	71.0	74.8	75.2	69.3	61.12
SWAV	64.5	51.6	55.8	44.0	33.9	71.2	69.6	74.3	72.3	68.5	60.57
Banani et al.											
LGSimCLR	89.2	72.3	59.1	63.8	81.2	74.5	86.3	84.5	83.1	79.0	77.3
LGSimSiam	77.2	68.8	63.4	71.3	35.6	74.2	74.5	86.8	76.4	74.2	70.24
Ours											
LGSimCLR	77.9	57.5	65.9	64.2	39.2	71.3	78.2	80.6	80.5	76.3	69.16
LGSimSiam	77.4	70.5	66.1	72.7	41.3	75.2	76.9	88.4	77.9	77.1	72.35

Table 3: We report performance of 5 way, 5 shot classification using frozen features on 10 downstream tasks. LGSimCLR-*Ours* outperforms previous approaches on most datsets. *Ours* refers to the models trained on new captions. We retrain all models with ResNet-34 backbone.

subsampled from different subreddits. The code provided by the original authors corresponding to these experiments lacked proper invocation of modules or definitions of essential modules themselves. Attempts to resolve these roadblocks by raising issues in the author's GitHub repository were in vain, as the issues remained unaddressed. Lastly, as mentioned in Section 5.4, certain datasets have could not be included in our evaluative experiments. Additionally, when we substituted the ResNet backbone with a MobileNet variant, the performance notably worsened, highlighting the backbone dependency present in the original codebase.

8 Conclusion

In conclusion, our reproducibility study confirms that incorporating language guidance in self-supervised learning can enhance visual representation, especially when supported by improved caption quality. By replacing noisy, user-generated captions with refined BLIP-2 descriptions, we observed consistent performance gains across multiple frameworks and architectures. Importantly, our results highlight that the magnitude of these improvements is sensitive to the feature embedding capacity, emphasizing a need for careful model selection in resource-constrained settings. Overall, our work not only validates the promise of language-guided sampling but also offers practical insights for its broader application.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015.

Mahdi Banani, Karan Desai, Justin Johnson, and Bernard Ghanem. Learning visual representations via language-guided sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

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)*, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. Review of visual saliency detection with comprehensive information. *IEEE Transactions on circuits and Systems for Video Technology*, 29(10):2941–2959, 2018.

Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A CLIP benchmark of data, model, and supervision. In *ICML Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward*, 2022.

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.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9588–9597, October 2021.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2019.

Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

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

Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2002.

Yookoon Park, Mahmoud Azab, Bo Xiong, Seungwhan Moon, Florian Metze, Gourab Kundu, and Kirmani Ahmed. Normalized contrastive learning for text-video retrieval. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 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.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPRW*, 2014.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2019.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

Yan Wang, Wei-Lun Chao, Kilian Q. Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019.

Ross Wightman, Hugo Touvron, and Herve Jegou. Resnet strikes back: An improved training procedure in timm. In *NeurIPS Workshop on ImageNet: Past, Present, and Future*, 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.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2021.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.