When Noises Help: Improve Text-Image Multimodal Contrastive Learning with Stochastic Label Augmentations

Anonymous ACL submission

Abstract

Contrastive learning (CL) has been widely used for self-supervised representation learning in text-image multimodal representation learning. However, there are two setbacks in the SOTA contrastive learning framework. One lies in the design of contrastive learning, where the model aims to pull together positive pairs and push away negative pairs. For one image, CL only considers one unique text as its positive sample, and treat all remaining text data as negative samples. Such design inevitably brings in learning bias towards overfitting into specific data pairs. Another setback comes from the web-crawled datasets that are commonly used in CL such as Conceptual Caption, YFCC and LAION. These datasets brings benefit due to its large size, yet contain significant noisy or vague labels. In this paper, we examine how augmenting the ground-truth labels with randomness can bring significant improvements in text-image multimodal contrastive learning. Through the simple addition of noise to ground-truth labels, we observe substantial improvements in model performance and robustness, requiring no additional computational overhead. We introduce three distinct stochastic label augmentation strategies and evaluate their effectiveness across various benchmarks, including zero-shot transfer, distribution shift, and linear probing tasks. Furthermore, we conduct comprehensive experiments involving different model architectures and noise rates, demonstrating the generalizability and substantial benefits of stochastic label augmentation across diverse tasks and models.

1 Introduction

Vision-language representation learning aims to learn generic representations from images and texts that could benefit multimodal downstream applications. One prominent technique that has garnered significant attention in this domain is contrastive learning (CL), which has emerged as a powerful paradigm for self-supervised representation learning in multimodal tasks. CL aims to learn robust representations by contrasting positive pairs, where similar instances are brought together, against negative pairs, where dissimilar instances are pushed apart. Recent works in text-image multimodal learning (Radford et al., 2021; Mokady et al., 2021; Shen et al., 2021; Jia et al., 2021; Li et al., 2021; Duan et al., 2022; Yang et al., 2022; Shukor et al., 2022; Kwon et al., 2022; Jiang et al., 2023) handle the image and text modality separately with modality-specific encoders and utilizes contrastive learning to align the modalities, achieving state-of-the-art performance on multiple downstream applications such as Zero-shot Classification, Image-Text Retrieval (Duan et al., 2022; Li et al., 2021) and Visual Question Answering (Jia et al., 2021; Goyal et al., 2017).

However, despite its effectiveness, SOTA CL frameworks face notable challenges that hinder their performance and generalizability. One of the primary setbacks in CL lies in its design, which often leads to biases towards specific data pairs. For instance, in text-image multimodal tasks, CL typically treats all but one text sample as negative pairs for a given image, potentially resulting in overfitting to particular associations. Additionally, the reliance on web-crawled datasets like Conceptual Caption, YFCC, and LAION introduces noise and ambiguity into the training data, undermining the quality of learned representations.

In light of these challenges, this paper explores novel approaches to address the limitations of current CL frameworks and enhance text-image multimodal representation learning. Specifically, we investigate the potential of augmenting ground-truth labels with randomness to mitigate biases and improve the robustness of learned representations. By introducing stochastic label augmentation strategies, we aim to enhance the performance and generalizability of CL models without imposing addi-

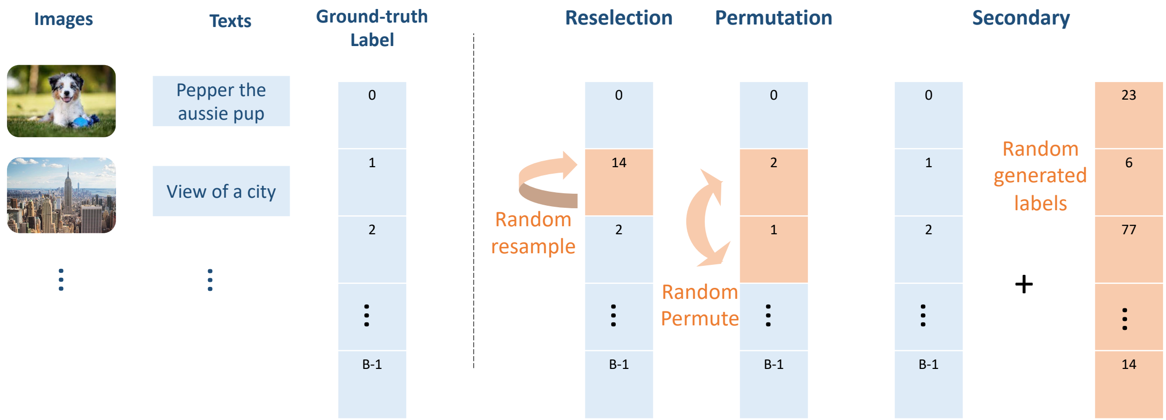

Figure 1: Label augmentation approaches illustration.

tional computational overhead. To summarize, our contributions are as follows:

- We address the inherent biases problem of contrastive learning framework by using stochastic strategies to mitigate overfitting to web-sourced dataset.
- We propose three simple yet effective randomized approaches to augment the groundtruth labels and enhance robustness in contrastive learning for text-image multimodal representation learning.
- We demonstrate the effectiveness and generalizability of our proposed approaches through comprehensive empirical evaluations across various benchmarks and model architectures.

2 Methods

In a conventional contrastive learning setup, data samples are divided into 'positive' and 'negative' categories based on ground-truth labels. During training, the model is encouraged to pull positive pairs closer in the embedded space while pushing the negative pairs farther apart. One key assumption here is that all negative samples are equally different from the positive sample.

Algorithm 1 Label Reselection

Require: noise rate $0 < \gamma < 1$
$B \leftarrow$ batch size
$\mathbf{y} = [0, 1, 2, , \cdots, B-1] \leftarrow$ Ground-truth
Random select a subset $\tilde{\mathbf{y}}$ of size γB from $\mathbf{y}$
for $y_i \in \tilde{\mathbf{y}}$ **do**
  $y_i =$ Random sample $\sim \{0, 1, \cdots, B-1\}$
end for

Algorithm 2 Label Permutation

Require: noise rate $0 < \gamma < 1$
$B \leftarrow$ batch size
$\mathbf{y} = [0, 1, 2, , \cdots, B-1] \leftarrow$ Ground-truth
Random select a subset $\tilde{\mathbf{y}}$ of size γB from $\mathbf{y}$
$\tilde{\mathbf{y}}$ = Random permute($\tilde{\mathbf{y}}$)

Algorithm 3 Secondary Random Label

Require: noise rate $0 < \gamma < 1$
$B \leftarrow$ batch size
$\mathbf{y} = [0, 1, 2, , \cdots, B-1] \leftarrow$ Ground-truth
Initialize $\tilde{\mathbf{y}} \in \{0, 1, \cdots, B-1\}^B$
for $y_i \in \tilde{\mathbf{y}}$ **do**
  $y_i =$ Random sample $\sim \{0, 1, \cdots, B-1\}$
end for
$\mathbf{y} = (1-\gamma)\mathbf{y} + \gamma\tilde{\mathbf{y}}$

This may not be true, especially in noisy web datasets. This "one-size-fits-all" treatment of negative samples limits the model's power to generalize.

In order to improve the generalizability of contrastive learning framework we propose to augment the ground-truth labels with random noises. The idea is that the webdatasets can be very noisy and the way contrastive learning treats all negative data samples equally can limit its power to generalize.

Our hypothesis is that by adding more noise to the label space, we prevent the contrastive learning trained model from overfitting on noisy datasets. By introducing random noise into the ground-truth labels, we hypothesize that the contrastive learning model will become more robust to outliers and label noise. The perturbed labels force the model to not overly rely on the exact boundary conditions

Method	Noise Rate γ	ResNet-50		ViT-B/16		ViT-B/32	
		Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑
CLIP	-	17.01	34.38	16.0	32.39	12.07	26.14
Re-selection	0.1	18.84	35.82	15.72	31.99	12.15	26.02
	0.3	18.52	37.66	14.88	31.07	11.46	25.57
Permutation	0.1	20.44	39.23	**18.01**	**34.98**	**14.1**	**29.17**
	0.3	20.39	39.90	16.09	33.02	12.55	27.03
Secondary	0.1	**21.17**	38.78	17.73	34.20	13.86	28.68
	0.3	21.14	**39.65**	17.31	34.48	12.07	26.14

Table 1: Zero-Shot Classifiction Accuray on ImageNet-1K (%).

Method	$\gamma = 0.1$		$\gamma = 0.3$		$\gamma = 0.5$		$\gamma = 0.7$		$\gamma = 0.9$	
	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑
Re-selection	18.84	35.82	18.52	37.66	15.49	32.23	0.1	0.5	0.1	0.5
Permutation	20.44	39.23	20.39	39.90	18.89	38.40	14.28	32.29	0.1	0.5
Secondary	21.17	38.78	21.14	39.65	20.24	39.17	18.31	36.92	0.1	0.5

Table 2: Zero-Shot Classifiction Accuray (%) on ImageNet-1K. The effect of different noise rate scale is studied. All reported numbers are based on ResNet-50. Complete results on different encoders are included in Appendix.

defined by the original labels, hence mitigating the risk of overfitting.

2.1 Label Augmentation by Random Reselection

We first use a fully randomized approach to augment the ground-truth labels by simply changing the ground-truth label with random resampling. As illustrated in Algorithm. 1, after choosing a noise rate between 0 and 1, for every batch, we randomly select samples that will have augmented labels based on the noise rate. Then for all the selected samples, randomly re-select its ground-truth within the same batch. This randomized re-selection could lead to a situation where multiple data points can have the same positive sample.

2.2 Label Augmentation by Random Permutation

The second approach slightly differ from the first one in the sense that we guarantee that every datapoints in the batch has its own positive sample, thus the one-to-one mapping nature of origin dataset is preserved. As illustrated in Algorithm. 2, after choosing a noise rate between 0 and 1, for every batch, we randomly select samples that will have augmented labels based on the noise ratio. Then for all the selected samples, we randomly switch their ground-truth.

2.3 Label Augmentation by Random Secondary Labels

The last approach differ from previous two by adding randomized secondary label to all the training data. In this way we are imposing randomness to all the training data. As shown in Algorithm. 3, for every batch, we randomly construct false ground-truth labels with random permutation. Compute the contrastive loss using the permutated labels and add it onto the original contrastive loss with the noise rate hyperparameter. Now that the contrastive loss composes of one true loss and one loss from random labels.

3 Experiments

We conduct experiments on image-text contrastive learning on CLIP model, where two separate encoders are trained to align features from the image and text modalities. **Setup:** Our CLIP model adopts ResNet-50 (He et al., 2016) and ViT (Dosovitskiy et al., 2021) as the image encoder and BERT (Devlin et al., 2018) as the text encoder. We adopt the official code from OpenCLIP to incorporate our approachs. Our reproduced CLIP results are consistent with the recent works (Mu et al., 2021; Gao et al., 2021). Note that all methods are under the same codebase and same hyper-parameter setting, thus the comparisons are fair. **Pre-training:** We follow the protocol of previous works to pre-train the model with the CC3M (Sharma et al., 2018) dataset, which contains 3M unique images and 4M image-text pairs. All models are pretrained with 8 Tesla V100 machines for 32 epochs.

3.1 Zero-Shot and Linear Probing Evaluation

We perform zero-shot transfer on standard image classification tasks, with ImageNet1K (Russakovsky et al., 2015) datasets and its distribution shift benchmarks (Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021b,a), and linear prob-

Method	ImageNetV2		ImageNetSketch		ImageNet-A		ImageNet-R	
	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑
CLIP	15.24	31.0	9.84	22.49	2.97	11.3	22.14	42.61
Re-selection	16.33	33.09	10.31	23.01	2.63	9.65	21.16	38.87
Permutation	17.92	**36.51**	12.28	26.58	**4.17**	**14.96**	25.82	47.30
Secondary	**18.50**	36.32	**12.77**	**27.10**	3.65	14.01	**26.0**	**47.8**

Table 3: Zero-Shot Natural Distribution Shift Classification Accuracy (%) using $\gamma = 0.1$ on ResNet-50.

	Caltech101	**SVHN**	**STL10**	**CIFAR10**	**CIFAR100**	**DTD**	**FGVCAircraft**	**OxfordPets**	**SST2**	**Food101**	**GTSRB**	**StanfordCars**	**Flowers102**	**ImageNet-1K**	**Average**
CLIP	80.67	48.62	88.55	77.91	56.54	56.97	24.54	61.68	**55.74**	58.23	72.91	19.57	80.09	51.58	59.54
Re-selection	78.47	43.77	89.86	76.70	54.46	**61.65**	24.06	61.68	54.20	57.37	68.90	19.09	77.52	50.72	58.46
Permutation	80.36	47.12	**89.64**	77.65	56.43	59.04	24.18	60.53	54.04	58.52	73.6	18.78	80.19	52.28	59.45
Secondary	**81.15**	**54.31**	89.09	**78.37**	**57.72**	59.52	**25.53**	**63.78**	55.24	**60.56**	76.41	**20.99**	**81.62**	**53.61**	**61.28**

Table 4: Linear Probing Top1 Classification Accuracy (%) on Vision Benchmarks using $\gamma = 0.1$ on ResNet-50.

ing on 14 vision benchmarks. We use the standard evaluation strategy of prompt engineering. For each dataset, we construct the text prompts using the name of the class, *e.g.* "a photo of the `[class name]`". For each class, we obtain the normalized class text embedding. During the evaluation, the class with the highest similarity score to the image embedding is predicted to be the label.

We show in Tab. 1 the perforamnce on ImageNet-1K, we can see that our label augmentations improves the performance by an average of 2-3%. We show in Tab. 2 that with changing noise rate, the model gradually changing from better performance to degraded performance then failed to train if the noise rate is extreme. In Tab. 3, the performance on distribution shift benchmark validates the robustness improvement with our methods.

We perform standard linear probing testing to evaluate the generalizability of learned models. We evaluate on 14 vision benchmarks with fixed encoders and fit a linear classifier for classification. We show in Tab.4 that secondary random label augmentation method has substantially improved the baseline performance.

4 Related Works and Limitations

Contrastive Learning: CLIP (Radford et al., 2021) introduced a unified model that learns to align visual and textual representations through contrastive learning, achieving impressive performance across various tasks. Other works (Li et al., 2020) extends contrastive learning principles to simultaneously pre-train image and text encoders, leading to state-of-the-art performance.

Vision-Language Pretraining: Most recent works on vision-language representation learning use separate encoders for images and texts (CLIPRadford et al. (2021); Mokady et al. (2021); Shen et al. (2021), ALIGNJia et al. (2021)), and rely on contrastive loss Oord et al. (2018); He et al. (2020); Chen et al. (2020) to align multiple modalities. These methods have been shown to achieve state-of-the-art (SOTA) performance on image-text tasks.

However, despite its efficacy, CL frameworks encounter significant challenges that impede their performance and ability to generalize. Hence we propose three randomized label augmentation methods to mitigate such issues. Yet our approach is limited to the paired image-text web datasets.

5 Conclusion

While contrastive learning (CL) is widely utilized for text-image multimodal tasks, existing frameworks face challenges stemming from biased design and noisy datasets. This paper proposes augmenting ground-truth labels with randomness to mitigate these issues. Significant improvements in model performance and robustness are achieved without additional computational overhead. We introduce three stochastic label augmentation strategies and demonstrate their effectiveness across various benchmarks, showcasing the generalizability and substantial benefits of this technique in enhancing multimodal representation learning.

References

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. 2022. Multi-modal alignment using representation codebook. In *Proc. CVPR*.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proc. CVPR*.

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

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proc. CVPR*.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proc. ICCV*.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Xiaodong Song. 2021b. Natural adversarial examples. In *Proc. CVPR*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. ICML*.

Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. 2023. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proc. CVPR*, pages 7661–7671.

Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefan 0 Soatto. 2022. Masked vision and language modeling for multi-modal representation learning. *ArXiv*, abs/2208.02131.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Proc. NeurIPS*.

Liunian Harold Li, Mark Yatskar, Da Yin Yin, Po-Sen Hsieh, Dragomir Chang, Jaemin Choi, Yanai Elazar, Yongxin Sung, and Minjoon Seo. 2020. Unified multimodal pre-training for image and text. In *NeurIPS*.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2021. Slip: Self-supervision meets language-image pre-training.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *Proc. ICML*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *IJCV*, 115.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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)*. Association for Computational Linguistics.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Mustafa Shukor, Guillaume Couairon, and Matthieu Cord. 2022. Efficient vision-language pretraining with visual concepts and hierarchical alignment. *ArXiv*, abs/2208.13628.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. In *Proc. NeurIPS*.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pretraining with triple contrastive learning. In *Proc. CVPR*.