SAGA: SELF-AUGMENTATION WITH GUIDED ATTENTION FOR REPRESENTATION LEARNING

Chun-Hsiao Yeh[1,2*], *Cheng-Yao Hong*[1*], *Yen-Chi Hsu*[1*], *and Tyng-Luh Liu*[1]

[1]Institute of Information Science, Academia Sinica, [2]UC Berkeley

ABSTRACT

Self-supervised training that elegantly couples contrastive learning with a wide spectrum of data augmentation techniques has been shown to be a successful paradigm for representation learning. However, current methods implicitly maximize the agreement between differently augmented *views* of the same sample, which may perform poorly in certain situations. For example, considering an image comprising a boat on the sea, one augmented view is cropped solely from the boat and the other from the sea, whereas linking these two to form a positive pair could be misleading. To resolve this issue, we introduce a Self-Augmentation with Guided Attention (SAGA) strategy, which augments input data based on predictive attention to learn representations rather than simply applying off-the-shelf augmentation schemes. As a result, the proposed self-augmentation framework enables feature learning to enhance the robustness of representation.

Index Terms— Self-supervised learning, Augmentation, Representation learning

1. INTRODUCTION

Unsupervised representation learning has been notably successful in practical usages because it exploits massive raw data without any annotated supervision. Early relevant work has focused on *pretext* tasks, which are addressed by generating pseudo-labels to the unlabeled data through different transformations [1, 2, 3]. However, the performance of the resulting representations has been fallen behind supervised learning. Recent contrastive learning-based methods [4, 5, 6, 7] have attracted more interest. It aims to learn an effective feature representation by maximizing the mutual information between different augmented *views* of an underlying example. All views are generated from stochastic augmentations, such as random cropping, color jittering and rotation.

Observe that mainstream contrastive learning approaches heavily rely on stochastic augmentations, which can improve model generalization but occasionally hinder representation learning. For example, as illustrated in Figure 1, using random cropping to augment the sample image yields two *different* views, `fish` and `a man`, respectively. The objective of contrastive learning refers to `fish` and `a man` as a positive pair in that the two views are from the same image, and therefore is designed to bring them closer, which could substantially confuse the representation learning.

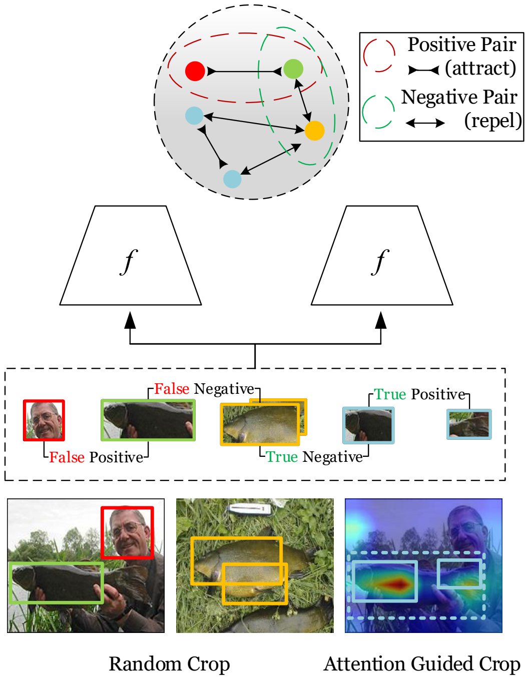


Fig. 1: While being visually dissimilar, the red and green views (from the same image) are referred to as a *positive* pair and are constrained to stay nearby in the embedding space. On the other hand, two cropped regions from different images would be treated as a *negative* pair, even when they are of the same class. To resolve this problem, the attention-guided self-augmentation generates views from the *salient* regions.

We introduce a novel framework for representation learning, which augments training data with guided attention induced by a learned embedding instead of applying hand-picked augmented schemes as in prior work. Its core idea is to prevent stochastic augmentations from degrading the representation learning by cropping views only in the attention-guided salient regions. More specifically, we develop a self-training strategy that employs an attention generator (see

∗ These authors contributed equally.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

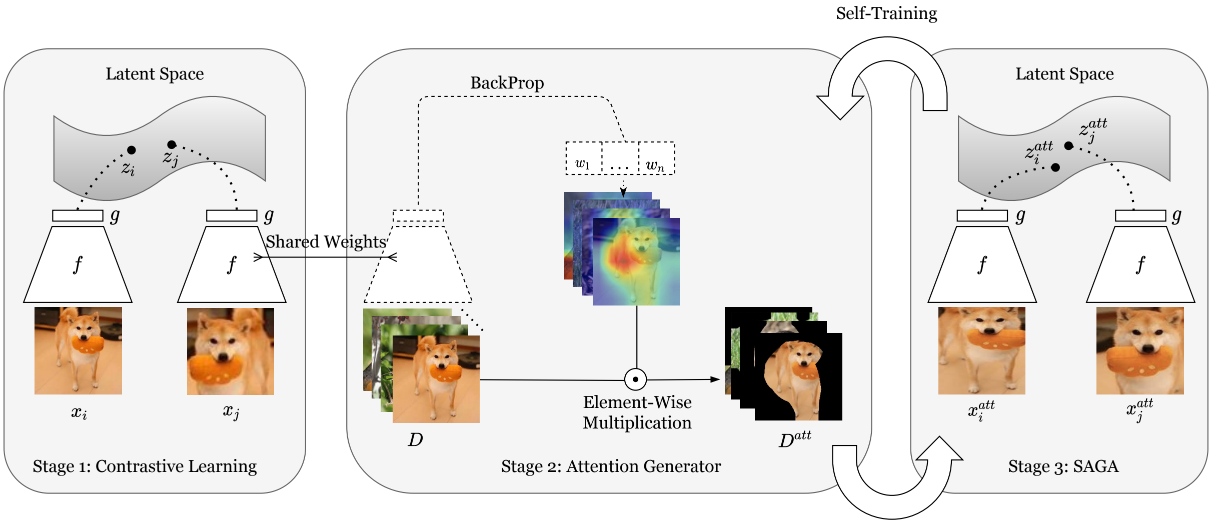


Fig. 2: The framework conceptually includes three stages, namely, contrastive learning, attention generator, and SAGA. Assume that a conventional contrastive learning-based model, which relies on hand-crafted augmentations, is given as the initial network. We establish an *attention generator* by applying Grad-CAM [8] to yield attention heatmaps and masked data D^{att}, focusing on salient regions of each image. The proposed SAGA then performs self-augmentations to generate more meaningful views from D^{att} as new input for representation learning. Over the process of iteratively carrying out stage 2 and stage 3, the mutual improvements of attention maps and the SAGA mechanism would effectively lead to a robust representation learning.

Figure 2) to obtain augmented images with guided attention. We then take the augmented images as additional input to refine the network training. Starting with a given unsupervised pretrained model, the proposed method iterates this learning process by incrementally improving the quality of augmented images (as new input) and thus achieving better representations.

Extensive experiments support that the proposed method can be potentially combined with several self-supervised approaches, SimCLR [7], CMC [5], MoCo [6], InsDis [9], PIRL [10], to outperform their fine-tuning version on established benchmarks: ImageNet [11] and ImageNet-100 [5]. The proposed representation learning achieves extra performance gains with increasing iterations of conducting self-augmentation training. Further analysis regarding why the proposed method should work better is addressed comprehensively in the experiment section.

2. THE PROPOSED METHOD

Concerning unsupervised representation learning, we propose to advance current contrastive methods by exploring two main ideas: 1) guided attention and 2) self-augmentation training in Figure 2. In this section, we present the proposed self-augmentation with guided attention (SAGA) in detail.

2.1. Contrastive Learning

Contrastive learning typically uses an encoder f and its projector g to contrast visual representations $\mathbf{z} = g(f(\mathbf{x})) \in \mathbb{R}^d$ for the input $\mathbf{x}$, where d is the dimension of the learned embedding z and in the setting $d = 128$ in all experiments. We follow [7, 6] to sample two random *views* $(\mathbf{x}_i = t(I), \mathbf{x}_j = t'(I))$ of the same image I under random data augmentations (t, t'), where $t, t' \sim \mathcal{T}$, and $\mathcal{T}$ is a set of data augmentations, passing through f and g to yield a positive pair $(\mathbf{z}_i, \mathbf{z}_j)$. Alternatively, negative pairs, $(\mathbf{z}_i, \mathbf{z}_k)$, are obtained by sampling views from different images. The InfoNCE loss [4] prefers the similarity of positive pairs to be higher than those of negatives and is defined for each pair of examples (i, j) as

$$L(\mathbf{z}_i, \mathbf{z}_j) = -\log \frac{\exp\left(s(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(s(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)} \quad (1)$$

where $s(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^\top \mathbf{z}_j / \|\mathbf{z}_i\| \|\mathbf{z}_j\|$ denotes the metric for measuring the similarity between two different views, and τ is the temperature hyper-parameter.

2.2. Attention-based Contrastive View

To improve the quality of augmented views and seek their meaningful paired relations, we draw on Grad-CAM [8] to explore guided attention. As the label information is not available in the proposed unsupervised setting, we provide a pseudo label for each sample as follows. We pass each view $\mathbf{x}$ through the current network to obtain $g(f(\mathbf{x})) = \mathbf{z} = (z^\ell)$

Algorithm 1: Attention Generator.

```
input : a set of images $D$, data size $S$, pretrained
        encoder $(f(\cdot), g(\cdot))$.
Function $G_{\text{Attention}}$ $(D, f(\cdot), g(\cdot))$ :
    for $s = 1$ to $\mathcal{S}$ do
        Extract features $z$ on set $D$ from $(f(\cdot), g(\cdot))$;
        Generate GradCAM heatmaps by (2) and (3);
        Obtain a subset $I^s_{\text{G-CAM}}$ by (4);
        $D^{att} \leftarrow D^{att}.append(I^s_{\text{G-CAM}})$
    end
    return $D^{att}$;
    End Function
output: a new set of images $D^{att}$
```

and define the pseudo label to be $c = \arg\max \mathbf{z}$. We can now obtain the discriminative localization map $L^c_{\text{G-CAM}} \in \mathbb{R}^{u \times v}$ of width u and height v by computing the the gradient of the score for the pseudo label c with respect to activation maps $\{A^k\}$ of the last convolutional layer, where k is the channel index. To this end, we compute the weight α^c_k of A^k by

$$\alpha^c_k = \frac{1}{Z} \sum\nolimits_i \sum\nolimits_j \frac{\partial z^c}{\partial A^k_{ij}} \tag{2}$$

where Z is the pixel numbers of feature map A^k. Then the attention heatmap is a weighted combination of $\{A^k\}$:

$$L_{\text{G-CAM}}(\mathbf{x}) = \text{ReLU}(\sum\nolimits_k \alpha^c_k A^k). \tag{3}$$

Notice that the coarse heatmap from (3) is of the same size as that of the convolutional feature maps. We apply bicubic interpolation to upsample it to the original image size. Denote the corresponding mask as $M_{\text{G-CAM}}(\mathbf{x})$ that is used to highlight the salient regions of $\mathbf{x}$. A pixel of $M_{\text{G-CAM}}(\mathbf{x})$ is set to 1 if the heatmap value is greater than δ_{grad}, and otherwise 0. We have δ_{grad} = 0.7 in all the experiments.

It is convenient to reduce I to its salient version I by computing a 4-tuple vector of the form $[x_{\min}, y_{\min}, w, h]$, where $(x_{\min}, y_{\min})$ are the coordinates of the top-left corner of the smallest rectangle enclosing the salient regions, and w and h are the respective width and height. Note that $x_{\min}, y_{\min}, w$, and h are normalized within $[0, 1]$. We have

$$I_{\text{G-CAM}} \leftarrow \text{imcrop}(I, [x_{\min}, y_{\min}, w, h]) \tag{4}$$

where self-augmentation to yield attention-guided augmented views can now be performed on $I_{\text{G-CAM}}$.

2.3. Self-augmentation with Guided Attention

The key improvements of the proposed SAGA lie in adding attention augmentation generated from the model itself and applying a self-training strategy to learn better representations. The self-training mechanism does not launch to learn a representation. We only use the objective of contrastive learning and the hand-crafted augmentations until the representation is reliable. We then conduct the SAGA by borrowing the given learned model as a self-training framework to augment the training data D via the function $G_{\text{attention}}$ illustrated in Algorithm 1. The augmented training data D^{att} is fed into the next step as the new input for representation learning. To conduct self-augmentation training with the self-training framework, we iterate this process by putting back the current network for the next step. The further learning updated the augmented data via attention generator as new input to learn better representations.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

We conduct ablation studies of the proposed framework on ImageNet-100 [5], a randomly selected 100 classes of ImageNet [11]. And training in standard large-scale image classification benchmark: ImageNet [11], which has around 1.28 million images in 1000 classes with the hyperparameter searched above. We train the embedding network with the unsupervised contrastive loss on a certain dataset. After pretraining process, replace the network's projection head with a new randomly initialized linear layer. This linear layer is trained with standard cross-entropy while the parameters of the embedding network are frozen.

3.2. SAGA on ImageNet-100

As we start to verify the effectiveness of each component that we proposed, adding guided attention to augment images as the new input and increasing the frequency of conducting SAGA by from every 40 to 10 epochs give us around 2% accuracy gain, which is shown in Table 2.

3.3. SAGA on ImageNet

For the sake of achieving higher efficiency in the experiments, we adopt the pretrained model of baselines: InsDis [9], PIRL [10], MoCo [6], CMC [5], and MoCo v2 [12] listed in Table 1 and closely follow most of their parameter settings. We then resume the pretrained models with 200 epochs and train for another 60 epochs with the fine-tuning and the proposed SAGA to each baseline methods. We conduct linear evaluations for the models after fine-tuning and SAGA by training a linear classifier while the embedding parameters are frozen. To evaluate the proposed work on an established large-scale benchmark, ImageNet, we compare SAGA with the fine-tuning of baselines listed in Table 1. The performance gains are reported to show that SAGA outperforms the fine-tuning of each baseline.

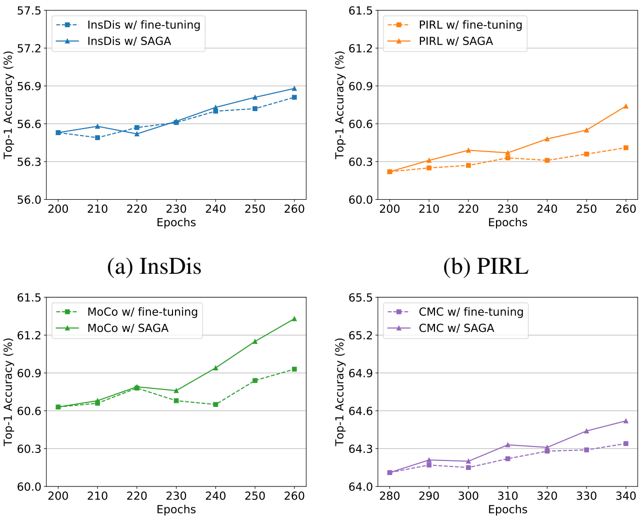
Figure 3 shows the linear top-1 performance (%) comparison between SAGA and fine-tuning more explicitly. Specif-

Methods	Architecture (#params)	#Epochs	Top-1	Fine-tuning Top-1 §	SAGA Top-1 §
InsDis [9]	R50 (24M)	200	56.5	56.8 (+0.3)	**56.9** (+0.4)
PIRL [10]	R50 (24M)	200	60.2	60.4 (+0.2)	**60.7** (+0.5)
MoCo [6]	R50 (24M)	200	60.6	60.9 (+0.3)	**61.3** (+0.7)
CMC [5]	$\mathrm{R50}_{\mathrm{L+ab}}$ (47M)	280	64.1	64.3 (+0.2)	**64.5** (+0.4)
MoCo v2 [12]	R50-MLP (28M)	200	67.5	67.7 (+0.2)	**68.0** (+0.5)

Table 1: Linear classifier top-1 accuracy (%) comparison of **self-supervised learning** methods on ImageNet between conducting SAGA and the fine-turning. Note that the experiments on ImageNet are conducted with 8 V-100 GPUs.

Accuracy (w/ kNN)	SimCLR [7]	Inv. Spread [13]
Pretraining (160 epochs)	50.3 ± 0.4	56.5 ± 0.2
Baseline: Fine-Tuning	51.1 ± 0.4	56.7 ± 0.3
SAGA	**53.2 ± 0.3**	**58.6 ± 0.2**
SAGA (supervised att.)	57.1 ± 0.3	62.3 ± 0.4

Table 2: SAGA outperforms the baseline: fine-tuning of SimCLR and Inv. Spread with another 80 epochs on ImageNet-100. The average top-1 kNN accuracy of 5 runs is reported.



(a) InsDis

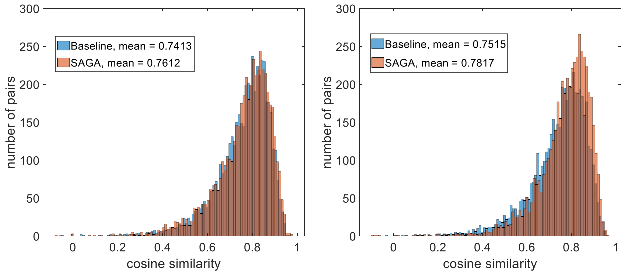
(b) PIRL

(c) MoCo

(d) CMC

Fig. 3: The linear top-1 performance (%) comparisons between SAGA and fine-tuning of each baseline with different epochs on ImageNet data.

ically, we observed that the SAGA performs almost the same as the fine-tuning in the early training steps, i.e., 200 $\sim$ 230 epochs, but achieves more performance gains when it gets more training steps. We think it might be the case that SAGA needs more training steps to learn a better augmentation with guided attention, which can directly improve the representation learning. Interestingly, it is observed that the proposed method does not have that much performance boost than the fine-tuning of InsDis in Figure 3 (c). It might be the case that cropping two differently augmented views in a sample image



(a) Epoch 200

(b) Epoch 240

Fig. 4: Based on the given SimCLR embedding at training epoch 160, SAGA is better than the baseline: fine-tuning at having higher similar instances in positive pairs while learning more epochs. The histograms of cosine similarities for positive pairs $s(z_i, z_j)$ evaluate on ImageNet-100 data at training epoch 200 and 240, respectively.

is not applied in InsDis. In Figure 4, we show the comparison of similarities among positive pairs between SAGA and the baseline at (a) 200 epochs and (b) 240 epochs. It's observed that SAGA has better learning scheme than the baseline at having more gains of the similarity mean from 200 to 240 epochs. That means we have addressed the problem illustrated in Figure 1.

4. CONCLUSION

Current contrastive learning methods rely on hand-selected augmentations, which may damage the representation learning if "positive pairs" are randomly cropped from visually different views. We proposed a new framework that augments input data with guided attention based on the learned model itself. We also introduce self-augmentation training by iterating the process of guided attention and representation learning. The proposed framework addresses the problem and outperforms the fine-tuning of each baseline on ImageNet and ImageNet-100 data. In the future work, we will investigate and tackle the problem of positive pairs as negatives that usually exist in current contrastive methods.

Acknowledgements. This work was supported in part by the MOST grants 110-2634-F-001-009, 110-2634-F-007-027 and 110-2221-E-001-017 of Taiwan. We are grateful to National Center for High-performance Computing for providing computational resources and facilities.

5. REFERENCES

[1] Mehdi Noroozi and Paolo Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–84.

[2] Richard Zhang, Phillip Isola, and Alexei A Efros, "Colorful image colorization," in *European conference on computer vision*. Springer, 2016, pp. 649–666.

[3] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

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

[5] Yonglong Tian, Dilip Krishnan, and Phillip Isola, "Contrastive multiview coding," *arXiv preprint arXiv:1906.05849*, 2019.

[6] 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*, 2020, pp. 9729–9738.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[8] 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 the IEEE international conference on computer vision*, 2017, pp. 618–626.

[9] 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*, 2018, pp. 3733–3742.

[10] Ishan Misra and Laurens van der Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6707–6717.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

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

[13] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2019, pp. 6210–6219.