Contrastive Learning with Adversarial Examples

Chih-Hui Ho, Nuno Vasconcelos

University of California, San Diego

Statistical Visual Computing Lab

UC San Diego

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

- Contrastive learning (CL) is one of the popular technique for self-supervised learning (SSL) of visual representations.
- CL treats instances as classes and aims to learn an invariant instance representation.
- This is implemented by generating a pair of examples per instance, and feeding them through an encoder, which is trained with a constrastive loss.
- The design of positive pairs is one of the research focuses of CL. For example, [1] shows that data augmentation is critical for the success of CL with different augmentation approaches having a different impact on SSL performance.
- While CL resembles metric learning approaches such as noise contrastive estimation [2] and N-pair [3] losses, the design of negative pairs has received less emphasis in the CL literature, unlike the plethora of positive pair selection proposals.
- In this work, we seek a general algorithm for the generation of diverse positive and challenging negative pairs for CL algorithms.
- This is framed as the search for instance augmentation sets that induce the largest optimization cost for CL with adversarial examples
- We show that it is possible to leverage the interpretation of CL as instance classification to produce a sensible generalization of classification attacks to the CL problem.
- The new attacks are then combined with recent techniques from the adversarial literature which treat adversarial training as multi-domain training.
- We show that the novel procedure Contrastive Learning with Adversarial Examples (CLAE) can boost the performance of several CL baselines across different datasets.

Contrastive learning

- Contrastive learning (CL) is formulated as

$$L_{cl}(x_i^{q_i}, x_i^{p_i}; \theta, \mathcal{T}) = -log\frac{e^{f_\theta(x_i^{q_i})^T f_\theta(x_i^{p_i})/\tau}}{\sum_{k=1}^{B} e^{f_\theta(x_i^{q_i})^T f_\theta(x_k^{p_k})/\tau}}\ , q_i\, , p_i \sim \mathcal{T} \qquad (1)$$

where f is an embedding parameterized by θ, τ is the temperature, B is the batch size and $x_i^{p_i}$, $x_i^{q_i}$ are augmentations of x_i under transformations q_i , p_i randomly sampled from some set of transformations $\mathcal{T}$.

- While [12] has shown that the choice of $\mathcal{T}$ has a critical role on SSL performance, most prior works do not give much consideration to the individual choice of q_i and p_i.
- In this work, we seek augmentations that maximize the risk defined by the loss of (1), i.e.

$$\{r_i^*, q_i^*\} = \underset{\{r_i, q_i\} \sim \mathcal{T}}{argmax} \sum_i L_{cl}(x_i^{r_i}, x_i^{q_i}; \theta, \mathcal{T}) \qquad (2)$$

- However, optimizing (2) is difficult. Instead, we proposed to fix the augmentation q_i and minimize

$$\{r_i^*\} = \underset{\{r_i\} \sim \mathcal{T}}{argmax} \sum_i L_{cl}(x_i^{r_i}, x_i^{q_i}; \theta, \mathcal{T}) \qquad (3)$$

Code available at
https://github.com/chihhuiho/CLAE

Proposed method

- To make the search more efficiently, we proposed to constrain the search as

$$x_i^{r_i} = \underset{x \in \mathcal{A}(x_i^{q_i})}{argmax} \sum_i L_{cl}(x, x_i^{q_i}; \theta, \mathcal{T}) \qquad (4)$$

where $\mathcal{A}(x_i^{q_i})$ is a set of adversarial perturbations of $x_i^{q_i}$, defined by

$$\mathcal{A}(x) = \{x' | x' = x + \delta, ||\delta||_p < \epsilon\} \qquad (5)$$

- To optimize (4), we reformulate the contrastive loss of (1) as the cross-entropy loss

$$L_{ce}(x_i^{q_i}, i; \{\frac{f_\theta(x_k^{p_k})}{\tau}\}, \theta) \qquad (6)$$

where $L_{ce}(x, y; W, \theta) = -log\frac{e^{w_y^T f_\theta(x)}}{\sum_k e^{w_k^T f_\theta(x)}}$.

- Then (4) becomes

$$\{x_k^{r_k}\} = \underset{\{x \in \mathcal{A}(x_i^{q_i})\}}{argmax} \sum_i L_{ce}(x_i^{q_i}, i; f_\theta(x_k)/\tau, \theta) \;\; q_i \sim \mathcal{T} \qquad (7)$$

- By substitute (5) into (7)

$$\{\delta_k^*\} = \underset{\{\delta_k\}}{argmax} \sum_i L_{ce}(x_i^{q_i}, i; f_\theta(x_i^{q_i} + \delta_k)/\tau, \theta) \text{ s.t. } ||\delta_k||_p < \epsilon, \quad q_i \sim \mathcal{T} \qquad (8)$$

- In this work, we rely on untargeted FGSM [] to solve (8) and compute $\{x_k^{r_k}\}$ as

$$x_k^{r_k} = x_k^{q_k} + \delta_k^* = x_k^{q_k} + \epsilon sign(\nabla_{x_k^{q_k}} \sum_i L_{ce}(x_i^{q_i}, i; f_\theta(x_k^{q_k})/\tau, \theta)\,),\; ||\delta_k||_2 < \epsilon \qquad (9)$$

- To perform SSL training, we adopt the training scheme of AdvProp [], which uses two separate batch normalization (BN) layers for clean and adversarial examples.
- The overall loss function contains CL losses computed with augmented examples and adversarial examples and is formulated as

$$\underset{\theta}{argmax} \sum_i L_{ce}(x_i^{q_i}, i; \{\frac{f_\theta(x_k^{p_k})}{\tau}\}, \theta) + \alpha \sum_i L_{ce}(x_i^{q_i}, i; \{\frac{f_\theta(x_k^{r_k})}{\tau}\}, \theta) \qquad (10)$$

- We refer this as Contrastive Learning with Adversarial Example (CLAE) and the procedure of CLAE is summarized in Algorithm 1.

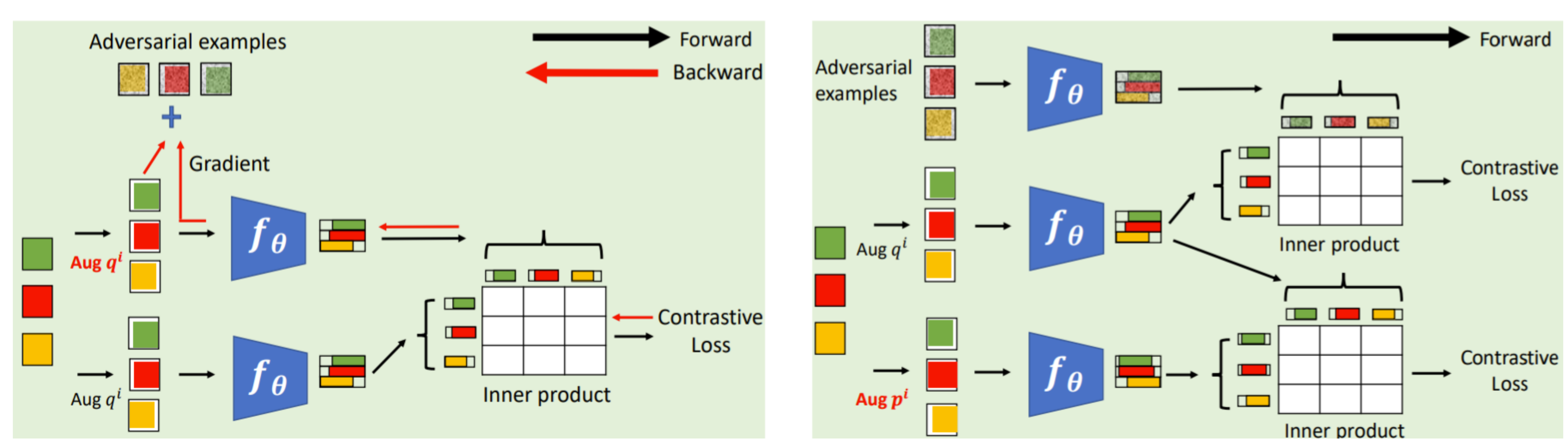

Figure 1. (Left) Generation of adversarial augmentations in step 4 of Algorithm 1 (Right) Adversarial training with contrastive loss in step 5 of Algorithm 1

Algorithm 1 Pseudocode of adversarial training with constrastive loss in a batch

1: **Input** $\mathcal{X} = \{x_i\}_{i=1}^B$, AUG:= Data Augmentation, Hyperparameter α
2: $\mathcal{W}^p = \{x_i^{p_i}\}_{i=1}^B = AUG(\mathcal{X})$
3: $\mathcal{W}^q = \{x_i^{q_i}\}_{i=1}^B = AUG(\mathcal{X})$
4: Compute (15) with $\{x_i^{q_i}\}_{i=1}^B$ and $\mathcal{W}^q$ to obtain $\mathcal{W}^*$
5: Compute L_{aug} of (16) with $\{x_i^{q_i}\}_{i=1}^B$ and $\mathcal{W}^p$, and L_{adv} of (16) with $\{x_i^{q_i}\}_{i=1}^B$ and $\mathcal{W}^*$
6: Minimize (16) with hyperparameter α

Experiments

Table 1: Downstream classification accuracy for three SSL methods, with and without (ϵ= 0) adversarial augmentation, on different datasets.

Method	ϵ	kNN Cifar10	kNN Cifar100	LR Cifar10	LR Cifar100	LR tinyImageNet
Plain	0	82.78±0.20	54.73±0.20	79.65±0.43	51.82±0.46	31.71±0.23
	0.03	**83.09±0.19**	**55.28±0.12**	**79.94±0.28**	52.04±0.32	**32.82±0.10**
	0.07	83.04±0.18	54.96±0.12	79.85±0.16	**52.14±0.21**	32.71±0.22
UEL [5]	0	83.63±0.14	55.23±0.28	80.63±0.18	52.99±0.25	32.32±0.30
	0.03	**84±0.15**	**55.96±0.06**	**80.94±0.13**	**54.27±0.40**	**33.72±0.30**
	0.07	83.72±0.19	55.36±0.22	80.82±0.12	53.90±0.11	33.16±0.36
SimCLR [1]	0	75.92±0.26	34.94±0.25	83.27±0.17	53.79±0.21	40.11±0.34
	0.03	76.45±0.32	**38.89±0.25**	**83.32±0.26**	**55.52±0.30**	**41.62±0.20**
	0.07	**76.70±0.36**	38.41±0.21	83.13±0.22	54.96±0.20	41.46±0.22

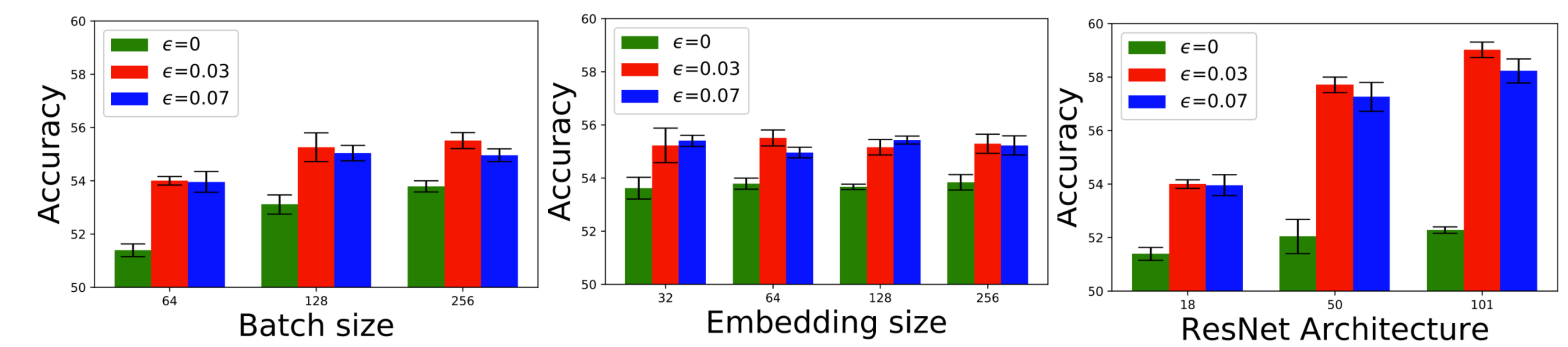

Figure2 Ablation study of (a) batch sizes ,(b) embedding dimensions and (c) ResNet architectures.

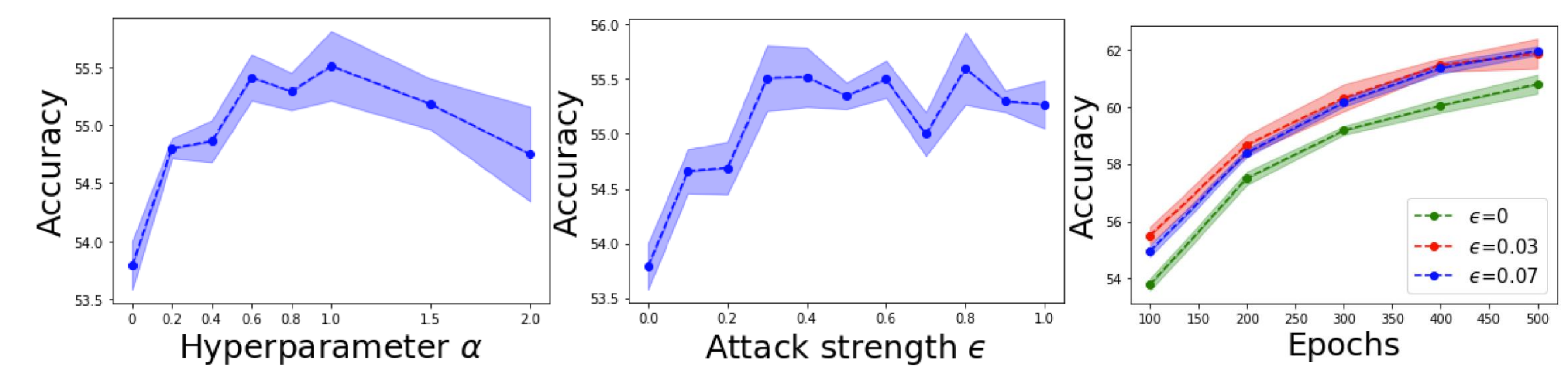

Figure 3: Ablation study for (a) hyperparameter α, (b) attack strength and (c) longer pretext training.

Conclusion

- Self-supervised learning approaches based on contrastive learning do not necessarily optimize on hard negative pairs.
- In this work, we have proposed a new algorithm (CLAE) that generates more challenging positive and hard negative pairs by leveraging adversarial examples.
- Adversarial training with the proposed adversarial augmentations was demonstrated to improve performance of several CL baselines.

Acknowledgement

This work was partially funded by NSF awards IIS-1637941, IIS-1924937, and NVIDIA GPU donations. We also acknowledge and thank the use of the Nautilus platform for some of the experiments discussed above.

References

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

[2] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington, editors, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, volume 9 of Proceedings of Machine Learning Research, pages 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[3] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 1857–1865. Curran Associates, Inc., 2016.

[4] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L. Yuille, and Quoc V. Le. Adversarial examples improve image recognition. ArXiv, abs/1911.09665, 2019.

[5] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. CoRR, abs/1904.03436, 2019