| | CIFAR10 | CIFAR100 | Cars | Aircraft |
|---|---|---|---|---|
| SimCLR | 72.8±0.3 | 45.3±0.1 | 12.25 ±0.0 | 14.8±0.9 |
| Ours | **73.6±0.3** | **47.0±0.2** | **12.60 ±0.1** | **16.2±0.6** |
| | DTD | Pets | Caltech-101 | Flowers |
| SimCLR | 50.6 ±1.2 | **44.4 ±0.4** | 68.2 ±0.3 | 46.0 ±0.3 |
| Ours | **51.5±0.5** | **44.4±0.0** | **69.1±0.3** | **47.1±0.8** |

Table 1: Comparison of transfer learning performance on 8 other datasets, using pretrained ResNet18 on ImageNet100.

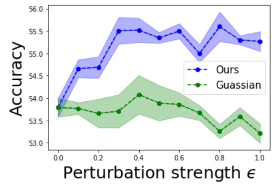


Figure 1: Comparison of adversarial and Guassian perturbations.

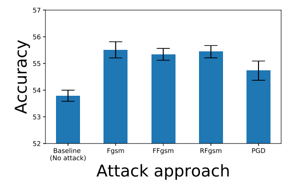


Figure 2: Comparison of different attack methods.

We thank the reviewers for the very thoughtful comments. Major issues are addressed here; minor suggestions are omitted (for space) and will be fixed as advised. **R1 Larger models**: For experiments with large ResNet models (i.e. 50 and 101) see Fig 5(c) and L259-265. **R1**&**R3 Larger datasets**: SSL is very computing intensive in general. SOTA results require very large datasets (ImageNet) and, more importantly, very large batch sizes (4,096). We are an academic group. Although in a large university, our clusters are not large enough to run extensive experiments with these settings, even for existing methods like SimCLR [12]. Nevertheless, to show that the proposed architecture works for large datasets, we compared to SimCLR on ImageNet100 (an ImageNet subset sampled by [55]), using a ResNet18 under the linear evaluation. This achieved 62.4±0.02 classification accuracy, outperforming [12] (61.7±0.02), which provides evidence that the proposed method scales up to big datasets. **R2 Theoretical justification**: We are working on a theoretical analysis of the benefits of adversarial learning for SSL. However, this is not ready and would require a paper of its own. We believe that an experimental showing of the benefits of adversarial examples is a first and important preliminary step, which will be of interest to the NeurIPS audience and motivate others to work on the topic, both experimentally and theoretically. **R3** & **R4 Related work**: [2] was released after the NeurIPS deadline. More importantly, it leverages SSL to increase robustness against adversarial attacks. This is a completely different goal form our work (which leverages adversarial attacks to increase downstream task performance of SSL models). Ref [1] of R3 is [7] in the submission. It uses a GAN-style manner for improving supervised learning problems in NLP, which is again different from this work. [3] is an orthogonal work to ours. It studies the relationships between the infomax criterion and minimization of the risk of (5), focusing on the impact of the choice of encoder and the tightness of mutual information estimator (See section 3 of [3]). It concludes that infomax is insufficient for SSL. This is unrelated to our work. It does highlight the importance of negative sampling (See section 4 and conclusion of [3]), which is one of the appealing features of the proposed approach. We will cite and discuss these works. **R3 Guassian noise**: Good suggestion! This was partially addressed in Fig. 3 of the submission (effect on loss of adversarial perturbations vs uniform noise). To really compare the classification performance, we extended Fig 6(b) of the submission to show the results obtained by adding guassian noise $\mathcal{N}(0,\epsilon)$ to the input image. As shown in Fig 1, this does not improve classification performance. Instead, accuracy degrades as the perturbation magnitude increases. **R3 Computation**: The proposed method requires an extra forward and backward pass per example during the pretraining stage. However, this is not what prevents us from doing the large scale experiments. We can't do them even for standard SimCLR. We note that pretraining cost is usually not seen as a major impediment in the literature, because the model is learned once and can be transferred to many tasks. This is the reason why SimCLR levels of computation are tolerated, even though few can afford to even perform the experiments at the scale needed to achieve SOTA results. **R3 Compare on various datasets**: Good suggestion! These datasets are used to measure transfer performance. We followed the linear evaluation protocol of [12], fixing the encoder pretrained on Imagenet100 and adding a linear layer, which is trained on each downstream dataset. This was done for the encoders learned by both SimCLR and the proposed method, with the results of Table1. The proposed approach outperformed [12] on 7 of the 8 datasets, indicating that the encoder trained with the proposed method generalizes better across downstream datasets. Since ImageNet100 does not contain any classes related to cars and airplanes, the performance on these 2 datasets is worse than on the others. In any case, our results beat [12] on several fine-grained datasets, such as Cars, Aircraft and Flowers. **R3 superior results** As shown above, we can show superior results for many SSL methods and downstream datasets. We cannot show SOTA results because we lack Google scale resources, namely the ability to train on ImageNet with batch sizes of 4,096. **R4 Other attack method**: Good question. While all experiments in the submission use FGSM [17] for simplicity, many untargeted attacks can be applied (See L182). This is now shown in Figure 2. Various attack methods (R-FGSM [57], F-FGSM [4], PGD [1]) are compatible with the proposed framework, all beating the baseline. Various attack methods lead to similar SSL performance. **R4 adversarial examples on** $x_i$: It is possible to compute the adversarial examples on $x_i$, in Algorithm 1 (by forcing $x_i^{q_i} = x_i$). However, it is a common practice in SSL [12, 68, 20, 64, 34] (See L31) to use one input example and one augmentation per pair. We simply follow this common practice.


[1] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[2] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *ArXiv*, abs/1907.13625, 2020.

[4] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.