Understanding Adversarial Robustness Through Loss Landscape Geometries

Joyce Xu
Stanford University
jexu@stanford.edu

Dian Ang Yap
Stanford University
dayap@stanford.edu

Vinay Uday Prabhu
UnifyID AI Labs
vinay@unify.id

Abstract

The pursuit of explaining and improving generalization in deep learning has elicited efforts both in regularization techniques as well as visualization techniques of the loss surface geometry. The latter is related to the intuition prevalent in the community that flatter local optima leads to lower generalization error. In this paper, we harness the state-of-the-art "filter normalization" technique of loss-surface visualization to qualitatively understand the consequences of using adversarial training data augmentation as the explicit regularization technique of choice. Much to our surprise, we discover that this oft deployed adversarial augmentation technique does not actually result in "flatter" loss-landscapes, which requires rethinking adversarial training generalization, and the relationship between generalization and loss landscapes geometries.

1. Introduction

State of the art deep learning models that exhibit low generalization error intriguingly exist in the regime where the number of trainable parameters in the model is often greater than the number of training data points. In order to address the susceptibility of such models to over-fitting, there exists a formidable body of literature of regularization techniques, which broadly fall into two categories. The first category is *implicit regularization* [8] that includes examples such as early stopping [2] during training or small-norm solution inducing Stochastic Gradient Descent (SGD) [12]. The second category is *explicit regularization* that includes techniques such as using $\ell_1 \backslash \ell_2$-norm weight penalty, training data augmentation, architectural changes (such as introducing skip connections) and dropouts [1]. Training data augmentation based explicit regularization in turn falls into two categories. The first entails usage of classical handcrafted transformations such as elastic transformation, random cropping, perturbations of image attributes such as brightness and contrast, and synthetic oversampling. The second entails harnessing adversarial perturbations to augment the dataset [6]. The work presented in this paper fits squarely into this second category, that is also termed *adversarial training*.

Co-temporal to this evolution of regularization techniques for deep learning, has been the quest to both qualitatively and quantitatively investigate the nexus between the geometry of the loss-landscape (flatness of the local optima) and the ability of regularized-model to generalize well [7]. Combining these two potent and hitherto unconnected set of ideas, in this paper, we seek to qualitatively address the following question:

> *"What happens to the loss surface of deep-nets*
> *when we use explicit regularization*
> *using adversarial data augmentation?"*

Guided by the pre-eminent merits of adversarial training, our *ansatz* at the beginning of this experimental work was that we would encounter mild to impressive *flattening* of the loss surfaces. Much to our surprise, our findings did not meet this expectation, which requires rethinking generalization [14] in the context of adversarial training. Through this dissemination, we'd like to call upon the community at large to further investigate this finding and have duly open sourced all our code in order to accelerate reproducibility and experimentation[1].

The rest of the paper is organized as follows: In section 2, we cover the methods and experiments used for performing the adversarial training and loss surface visualization. We present the results and visualizations in section 3 and conclude the paper in section 4.

2. Methods and Experiments

2.1. Adversarial Attacks

We experiment with adversarially augmented training data using three different common attacks, with all three attacks using the ℓ_∞ metric. The first is the Fast Gradient Sign Method (**FGSM**), which computes an adversarial

[1] https://github.com/joycex99/adversarial-training

sample x_{adv} from x [4] as:

$$x_{\text{adv}} = x + \epsilon \text{sgn}(\nabla_x L(\theta, x, y)). \tag{1}$$

Secondly, we consider projected gradient descent (PGD) as a *more powerful* adversary, which is considered a "universal" first-order technique and is essentially a multi-step variant FGSM^k with random starts [6]:

$$x_{\text{adv}}^{t+1} = \text{Clip}(x^t + \alpha \text{sgn}(\nabla_x L(\theta, x, y))). \tag{2}$$

Thirdly, we use the Spatially Transformed Adversarial Attack (**stAdv**) that crafts more perceptually realistic attacks by modifying image *geometry* [11]. Instead of directly the pixel values, these adversaries optimize the amount of displacement in each dimension, i.e. the flow vector $f_i = (\Delta u^{(i)}, \Delta v^{(i)})$. Each pixel value from input location $(u^{(i)}, v^{(i)})$ corresponds to a flow-displaced pixel in the adversarial image:

$$(u^{(i)}, v^{(i)}) = (u_{\text{adv}}^{(i)} + \Delta u^{(i)}, v_{\text{adv}}^{(i)} + \Delta v^{(i)}). \tag{3}$$

Since these locations can be fractional coordinates, each adversarial pixel $x_{\text{adv}}^{(i)}$ is calculated through bilinear interpolation of N neighboring pixels:

$$x_{\text{adv}}^{(i)} = \sum_{q \in N} x^{(q)} (1 - |u^{(i)} - u^{(q)}|)(1 - |v^{(i)} - v^{(q)}|). \tag{4}$$

2.2. The adversarial training procedure

We experiment with a standard ResNet-32 model architecture [5]. As a base model, we first train a clean model on the original CIFAR-10 dataset until convergence. Then, for each attack, we generate a single (untargeted) adversarial example per training point in the original dataset, such that the new augmented dataset has a one-to-one ratio of clean and adversarial samples. The base model is then fine-tuned for 200 epochs on the augmented dataset, with the learnt weights then used for loss landscape visualizations.

2.2.1 A caveat

We recognize and want to highlight that this choice of training procedure likely has an impact on the results we obtain. In particular, pre-training and then performing adversarial fine-tuning may result in a different loss landscape than performing adversarial training from scratch. We investigate this training procedure because we are interested in how much adversarial training based augmentation can *increase* robustness relative to existing trained models, potentially as part of a multi-step process to improve model generalization. However, we are also interested in and encourage future exploration of loss landscapes of models adversarially trained from scratch.

2.3. Loss Landscape Visualization

Traditional 2D contour plots visualize the change in loss moving from some center point θ^* (e.g. the optimization minimum) out along any two direction vectors δ and η:

$$f(\alpha, \beta) = L(\theta^* + \alpha\delta + \beta\eta) \tag{5}$$

Since naively adding these direction vectors to θ^* (which can vary greatly in magnitude model to model) sacrifices the scale invariance of network weights, we normalize the direction vectors d by the Frobenius norm of the filter:

$$d_{i,j} = \frac{d_{i,j}}{||d_{i,j}||} ||\theta_{i,j}|| \tag{6}$$

where $d_{i,j}$ is the jth filter of the ith layer of d [7].

With this modification, even if the adversarial models converge upon different magnitudes of weights, the *relative* curvatures and geometry of their loss landscapes are directly comparable.

3. Results and Visualizations

We evaluate original and augmented models on Top-1 accuracy on both the original test data and adversarial inputs. We use FGSM with the ℓ_∞ radius of $8/255$, PGD with radius $1/255$ for 10 iters, and stAdv with radius $0.3/64$. The relative robustness of our adversarially-trained models supports accuracies as reported by existing literature [11].

Data Augmented with attacks	Ground Acc.	Adversarial Acc.		
		FGSM	PGD	stAdv
None	**92.64**	16.88	0.0	0.0
FGSM	90.11	**75.23**	12.11	0.0
PGD	87.34	60.52	51.20	13.85
stAdv	91.05	55.84	**52.98**	**25.49**

Table 1. Top-1 Accuracies for Original and Adversarially-Augmented Data

Surprisingly, the loss landscape geometries do not follow intuition of common practices. Since adversarial training has been shown to improve generalization, and models with low generalization error tend to have smoother loss landscapes, we originally expected adversarially-trained nets to have smoother loss landscapes and fewer local minima. However, the clean model (trained on the original data) appears to have the smoothest/most convex loss landscape (Fig 1(a)). Meanwhile, the FGSM-augmented model has the sharpest/least convex geometry, followed by PGD and then the comparatively smoother stAdv (Fig 1(b), 1(c), 1(d)), which might allude that this form of adversarial-augmentation based fine-tuning might actually *hurt* generalization in the quest trade-offs between the *standard accuracy* and *adversarially robust accuracy* of a model [9].

DATA AUGMENTED WITH	1-SSIM
FGSM	0.049344
PGD	0.008206
STADV	0.002801

Table 2. Average SSIM Distance of Adversary from Original

3.1. Image similarity and loss surface smoothness

We also note that there is a positive correlation between the smoothness/convexity of the adversarial loss landscapes and the relative similarity between the adversarially perturbed images and their original counterparts. Table 2 shows the average difference of perturbed images from the original for each attack, measured by 1 - SSIM (the Structural SIMilarity index) [3][10]. Our results demonstrate that smaller 1-SSIM differences correspond to smoother loss landscapes. Specifically, models trained on stAdv perturbations - which intentionally modify image geometry as opposed to arbitrary pixel values - have the smoothest adversarial loss landscape. Meanwhile, training on FGSM-perturbed images that have lower structural similarity to the original results in a particularly sharp and chaotic landscape.

4. Conclusion

The loss landscape geometries of adversarially-trained neural networks reveal that overall model *generalization* does not necessarily improve with resilience to certain attacks. The sharper loss landscapes of adversarially-trained models are not only an unintuitive, surprising result for ongoing model interpretability efforts, but an important point of consideration for developing new adversarial training techniques.

Fine-tuning models on augmented data may also contribute to sharper and more chaotic local minima, since the optimization path from former local minima can differ from the optimization path from randomly initialized weights.

Based on our investigations, out of our three adversarial augmentations, stAdv shows the most promise in terms of accuracy and generalization abilities. Even so, under adversarial fine-tuning, we show models can default to memorizing perturbations and generalizing only *based on structural image similarity* [13][15]. It is possible that combining multiple attack methods might improve overall generalization and converge upon relatively flat loss landscapes, and is a potential avenue for future investigation.

References

[1] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 233–242. JMLR. org, 2017. 1

[2] R. Caruana, S. Lawrence, and C. L. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in neural information processing systems*, pages 402–408, 2001. 1

[3] J. R. Flynn, S. Ward, J. Abich, and D. Poole. Image quality assessment using the ssim and the just noticeable difference paradigm. In *International Conference on Engineering Psychology and Cognitive Ergonomics*, pages 23–30. Springer, 2013. 3

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[6] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 1, 2

[7] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2018. 1, 2

[8] B. Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017. 1

[9] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *stat*, 1050:11, 2018. 3

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

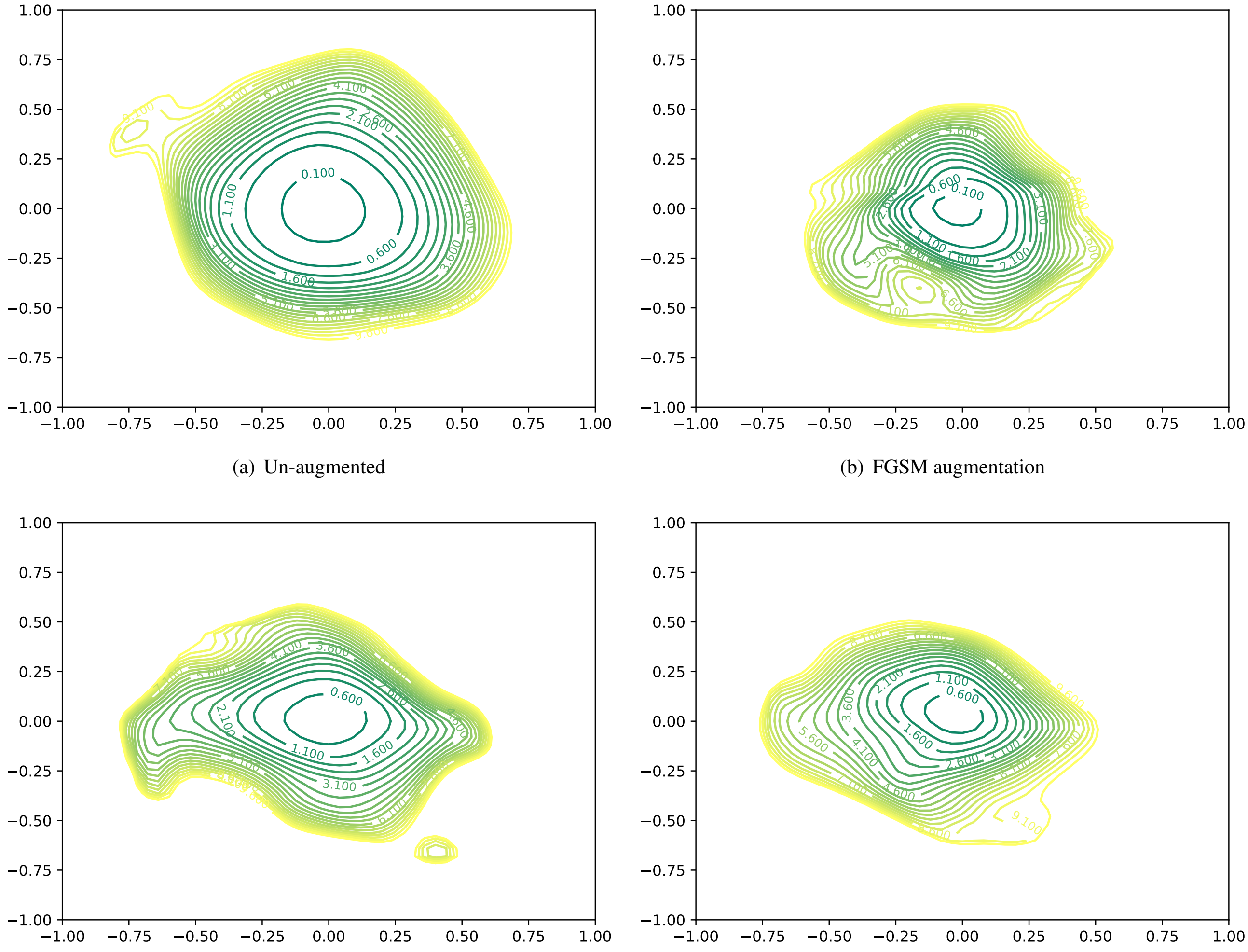
[11] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018. 2

[12] C. Xing, D. Arpit, C. Tsirigotis, and Y. Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018. 1

[13] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 2019. 3

[14] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016. 1

[15] J. Zhang and X. Jiang. Adversarial examples: Opportunities and challenges. *arXiv preprint arXiv:1809.04790*, 2018. 3



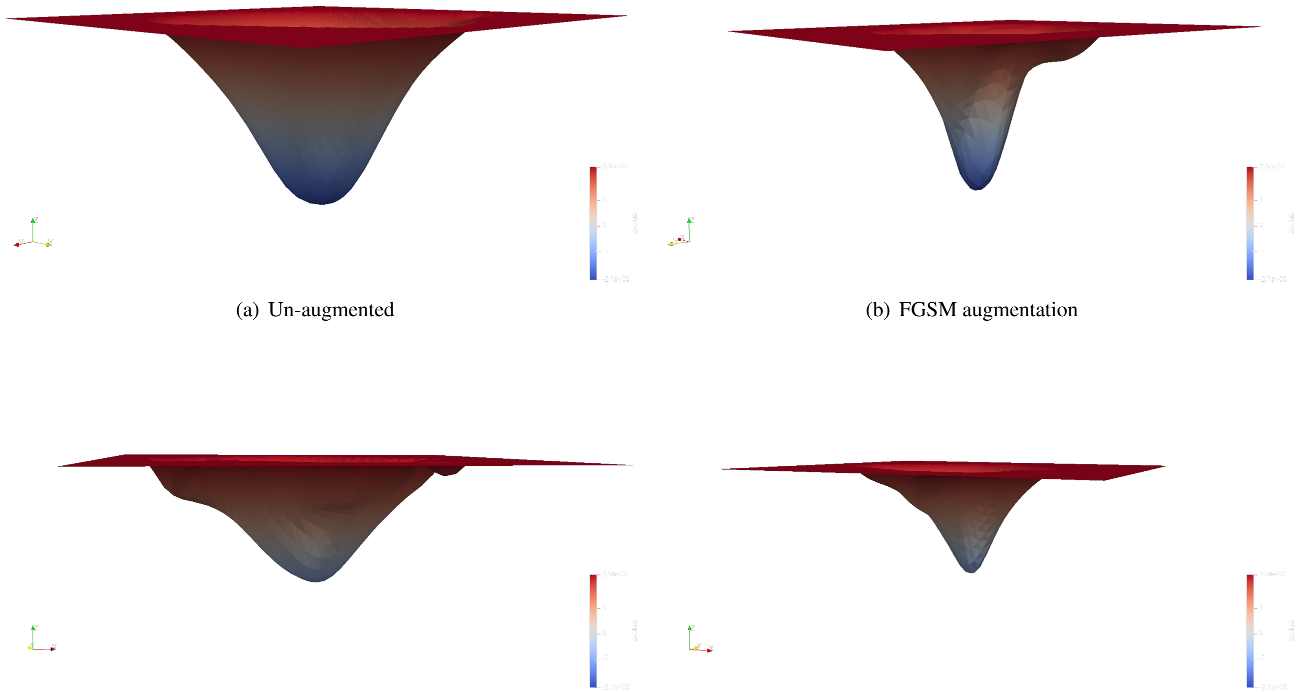
(a) Un-augmented

(b) FGSM augmentation

(c) PGD augmentation

(d) stAdv augmentation

Figure 1. Visualizations of loss landscapes of the same model. (b)-(d) are losses fine-tuned augmented data from different attacks.



(a) Un-augmented

(b) FGSM augmentation

(c) PGD augmentation

(d) stAdv augmentation

Figure 2. 3D Visualizations of ResNet-32 loss landscapes of the same model, trained on CIFAR-10 with and without adversarial training. (b)-(d) are losses fine-tuned augmented data from different attacks.