Adversarial Training Improves Joint Energy-Based Generative Modelling

Rostislav Korst [1] **Arip Asadulaev** [2 3]

Abstract

We propose the novel framework for generative modelling using hybrid energy-based models. In our method we combine the interpretable input gradients of the robust classifier and Langevin Dynamics for sampling. Using the adversarial training we improve not only the training stability, but robustness and generative modelling of the joint energy-based models.

1. Introduction

Adversarial examples tend classifier to make incorrect predictions. The classifier that is not prone to adversarial examples is called adversarially robust (Szegedy et al., 2014; Papernot et al., 2017; Yuan et al., 2019). The robustness can be measured based on the visual similarity of adversarial attack gradients to the real data (Shafahi et al., 2019). Moreover, when classifier is robust, adversarial attack adds robust features to the image, that is visually similar to the data (Ilyas et al., 2019). Based on this, robust classifiers can be used for images generation (Santurkar et al.).

In this paper we combine robust classifier abilities with the energy-based models for generative modelling. It was shown that classifier can be reinterpreted as an energy-based model for the joint distribution (Grathwohl et al., 2019). During image generation, (1) we sample the gaussian noise, (2) then turn samples into specific class using the adversarial attack, (3) and then minimize the energy using Stochastic Gradient Langevin Dynamics (SGLD).

Adversarial training can improve the out-of distribution detection of the energy based models (Grathwohl et al., 2019; Lee et al., 2020), but to best of our knowledge, the connection between adversarial training and energy-based inference for generative modelling was not previously studied. We tested our method on CIFAR-10 dataset and improved IS and FID in comparison to the standard robust classifier and joint energy-based model.

[1]MIPT, Moscow, Russia [2]ITMO, Saint-Petersburg, Russia [3]Artificial Intelligence Research Institute, Moscow, Russia. Correspondence to: Arip Asadulaev <aripasadulaev@itmo.ru>.

AdvML workshop of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

2. Background

Joint Energy-Based Model: Energy-based model (EBM) learns to assign low energy values to samples from training distribution and high values otherwise. The probability density $p_\theta(x)$ given by an EBM can be presented as: $p_\theta(x) = \frac{exp(-E_\theta(x))}{Z(\theta)}$, where $E_\theta(x)$ is an energy function that maps each input x to a scalar, and $Z(\theta)$ is an intractable normalizing constant. To generate samples by energy function, Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011) can be used:

$$x_{i+1} \leftarrow x_i - \frac{\alpha}{2}\frac{\partial E_\theta(x_i)}{\partial x} + \beta, \quad \beta \sim \mathcal{N}(0, \alpha) \quad (1)$$

Recently Joint Energy-based Model (JEM) was proposed (Grathwohl et al., 2019). JEM reinterprets a standard discriminative classifier of $p(y|x)$ as an energy-based model for the joint distribution $p(x, y)$. This model can classify images and generate realistic samples within one hybrid model. In our experiments we used JEM++ (Yang & Ji, 2021) that is more stable and fast version of JEM.

Adversarial Examples: Having the sample x, the real label y_{real}, model with parameters θ and loss function L, we can apply Projected Gradient Descent (PGD) (Madry et al., 2018) to generate the adversarial examples:

$$x_{i+1} = \text{Proj}_{(x,\varepsilon)}\left[x_i + \alpha\, \text{sign}\left(\nabla_x L\left(\theta, x_i, y_{real}\right)\right)\right] \quad (2)$$

Where, $\text{Proj}_{(x,\varepsilon)}$ is a projection operator onto the l_{inf} ball of radius ε around the original image x_i. After obtaining adversarial examples we can train the model on this examples to increase the robustness (Szegedy et al., 2014). The robust classifier can generate realistic images by minimizing cross-entropy loss on target class with PGD that is called target attack (Santurkar et al.).

3. Method

We propose to combine adversarial training of classifier and joint energy-based model. During training process we fed classifier with adversarial examples and we find that it improves training stability of JEM. Due to the use of adversarial training, the classifier contained in hybrid model got adversarially robust and obtained the ability to generate images with PGD attack (Santurkar et al.). Moreover,

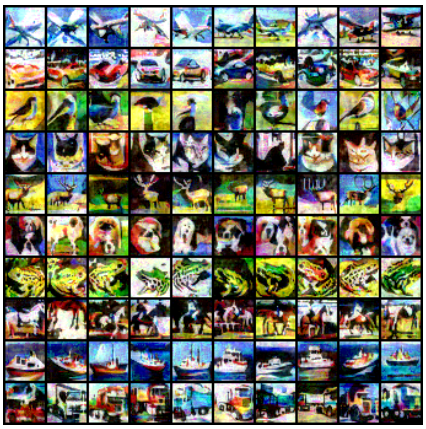

Figure 1: Single Robust Classifier (Santurkar et al.)

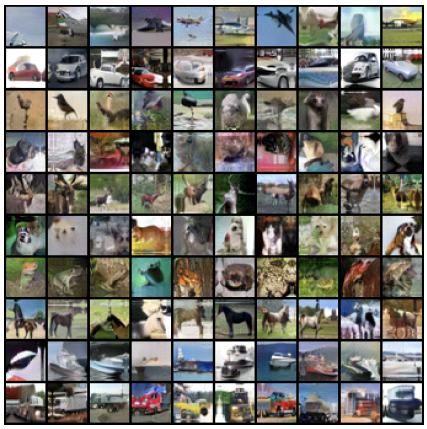

Figure 2: JEM++ (Yang & Ji, 2021)

Figure 3: Robust JEM++ **(Ours)**

our hybrid model can generate images by minimizing energy value similar to baseline JEM. We propose to combine both generation methods into **combined inference**. Energy-based SGLD steps can be treated as perturbation smoothing mechanism for prior obtained with target attack on noise.

To perform conditional generation we start from Gaussian mixture distribution estimated from the training dataset like in (Yang & Ji, 2021), then we implement target PGD attack with classifier contained in JEM. Finally we minimize joint energy function using Langevin dynamics given joint energy function $E_\theta(x, y) = -f_\theta(x)[y]$, proposed in (Grathwohl et al., 2019), where $f_\theta(x)$ - is a parametric function $\mathbb{R}^D \to \mathbb{R}^K$, which maps each data point $x \in \mathbb{R}^D$ to K real-valued numbers known as classifier logits and $f_\theta(x)[y]$ - the logit corresponding to the y^{th} class label. In other words **the initial sampling distribution for JEM is defined by adversarial attack to the classifier contained in JEM**.

4. Experiments

Dataset: We trained and evaluated our models on CIFAR-10 dataset. The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class.

Settings: We used JEM++ (Yang & Ji, 2021) for our experiments. All architectures used are based on Wide Residual Networks (Zagoruyko & Komodakis, 2016). JEM++ uses m outer loops and n inner loops for SGLD steps. For training JEM++ we use $m = 10$, $n = 5$ steps, and $n = 300$, $n = 5$ for unconditional inference. Same parameters were used for our model. For conditional inference we use $m = 50$, $n = 5$. Adversarial attack parameters for adversarial training was: $\varepsilon_{adv} = 0.1$, number of steps is 15, constraint is inf. For inference we use $\varepsilon_{adv} = 0.5$. Same parameters were used for training single robust classifier. In combined inference, we decrease contrast of adversarial prior, it allows pixels not to go beyond the range of $[-1, 1]$ after SGLD and not to clamp pixels. Empirically, this trick improves the performance.

Metrics: To evaluate the quality of generated images, we used the Inception Score (IS) (Salimans et al., 2016) and Frechet Inception Distance (FID) (Heusel et al., 2017).

Model	IS ↑	FID↓
Residual Flow (Chen et al., 2019)	3.69	46.40
Glow (Kingma & Dhariwal, 2018)	3.92	48.90
JEM++ (Yang & Ji, 2021)	7.54	48.25
Single Robust Classifier (Santurkar et al.)	7.05	85.12
Robust JEM-Energy (Ours)	8.71	**41.17**
Robust JEM (Ours)	**9.28**	47.92

Table 1: Hybrid models results on CIFAR-10. Robust JEM-Energy is sampling using only SGLD steps without adversarial attack. Samples were generated from scratch.

Results: We compared generative performance of our model (Figure 3) with JEM++(Figure 2) and Single Robust Classifier (Santurkar et al.),(Figure 1). Our model improved performance of energy-based inference of JEM++, see Table (1). The combined inference showed increase in IS rivaling the hybrid models state-of-the-art in generative learning.

5. Conclusion and Future Work

Thanks to the strong adversarial training, JEM improves its generative performance from noise. The study of visually interpretable gradients of robust networks is still unrepresented research area. In our opinion this property can find many applications in generative modelling. As a problem, we find that our model tend to generate less diverse images than standard energy based models and we are going to tackle this issue in the future work.

References

Chen, R. T., Behrmann, J., Duvenaud, D. K., and Jacobsen, J.-H. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., and Swersky, K. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *CoRR*, abs/1905.02175, 2019. URL http://arxiv.org/abs/1905.02175.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Lee, K., Yang, H., and Oh, S.-Y. Adversarial training on joint energy based model for robust classification and out-of-distribution detection. In *2020 20th International Conference on Control, Automation and Systems (ICCAS)*, pp. 17–21. IEEE, 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Papernot, N., McDaniel, P. D., Goodfellow, I. J., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017*, pp. 506–519, 2017. doi: 10.1145/3052973.3053009. URL https://doi.org/10.1145/3052973.3053009.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Santurkar, S., Tsipras, D., Tran, B., Ilyas, A., Engstrom, L., and Madry, A. Image synthesis with a single (robust) classifier.

Shafahi, A., Saadatpanah, P., Zhu, C., Ghiasi, A., Studer, C., Jacobs, D. W., and Goldstein, T. Adversarially robust transfer learning. *CoRR*, abs/1905.08232, 2019. URL http://arxiv.org/abs/1905.08232.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

Yang, X. and Ji, S. Jem++: Improved techniques for training jem. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6494–6503, 2021.

Yuan, X., He, P., Zhu, Q., and Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learning Syst.*, 30(9):2805–2824, 2019. doi: 10.1109/TNNLS.2018.2886017. URL https://doi.org/10.1109/TNNLS.2018.2886017.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.