Adversarial Data Augmentation Improves Unsupervised Machine Learning

Anonymous authors
Paper under double-blind review

Abstract

Adversarial examples causing evasive predictions are widely used to evaluate and improve the robustness of machine learning models. However, current studies focus on supervised learning tasks, relying on the ground-truth data label, a targeted objective, or supervision from a trained classifier. In this paper, we propose a framework of generating adversarial examples for **unsupervised** models and demonstrate novel applications to data augmentation. Our framework exploits a mutual information neural estimator as an information-theoretic similarity measure to generate adversarial examples without supervision. We propose a new MinMax algorithm with provable convergence guarantees for efficient generation of unsupervised adversarial examples. When using unsupervised adversarial examples as a simple plug-in data augmentation tool for model retraining, significant improvements are consistently observed across different unsupervised tasks and datasets, including data reconstruction, representation learning, and contrastive learning.

1 Introduction

Adversarial examples are known as prediction-evasive attacks on state-of-the-art machine learning models (e.g., deep neural networks), which are often generated by manipulating native data samples while maintaining high similarity measured by task-specific metrics such as L_p-norm bounded perturbations Goodfellow et al. (2015); Biggio & Roli (2018). Due to the implications and consequences on mission-critical and security-centric machine learning tasks, adversarial examples are widely used for robustness evaluation of a trained model and for robustness enhancement during training (i.e., adversarial training).

Despite of a plethora of adversarial attacking algorithms, the design principle of existing methods is primarily for *supervised* learning models — requiring either the true label or a targeted objective (e.g., a specific class label or a reference sample) for generating adversarial examples. Some recent works have extended to the *semi-supervised* setting, by leveraging supervision from a classifier (trained on labeled data) and using the predicted labels on unlabeled data for generating (semi-supervised) adversarial examples Miyato et al. (2018); Zhang et al. (2019); Stanforth et al. (2019); Carmon et al. (2019). On the other hand, recent advances in unsupervised and few-shot machine learning techniques show that task-invariant representations can be learned and contribute to downstream tasks with limited or even without supervision Ranzato et al. (2007); Zhu & Goldberg (2009); Zhai et al. (2019), which motivates this study regarding their robustness. Our goal is to provide efficient robustness evaluation and data augmentation techniques for unsupervised (and self-supervised) machine learning models through *unsupervised* adversarial examples (UAEs). Table 1 summarizes the fundamental difference between conventional supervised adversarial examples and

Table 1: Illustration of adversarial examples for supervised and unsupervised machine learning tasks. Both settings use a native data sample x as reference. For supervised setting, adversarial examples refer to *similar* samples of x causing inconsistent model predictions. For unsupervised setting, adversarial examples refer to *dissimilar* samples yielding smaller loss in reference to x, which can be interpreted as generalization errors on low-loss samples.

(I) *Mathematical notation*	
$M^{\text{sup}}/M^{\text{unsup}}$: trained supervised/unsupervised machine learning models x/x_{adv}: original/adversarial data sample $\ell_x^{\text{sup}}/\ell_x^{\text{unsup}}$: supervised/unsupervised loss function in reference to x	
(II) *Supervised tasks* (e.g. classification)	(III) *Unsupervised tasks* (our proposal) (e.g. data reconstruction, contrastive learning)
x_{adv} is similar to x but $M^{\text{sup}}(x_{\text{adv}}) \neq M^{\text{sup}}(x)$	x_{adv} is dissimilar to x but $\ell_x^{\text{unsup}}(x_{\text{adv}} \mid M^{\text{unsup}}) \leq \ell_x^{\text{unsup}}(x \mid M^{\text{unsup}})$

our UAEs. Notably, our UAE generation is supervision-free because it solely uses an information-theoretic similarity measure and the associated unsupervised learning objective function. It does not use any supervision such as label information or prediction from other supervised models.

In this paper, we aim to formalize the notion of UAE, establish an efficient framework for UAE generation, and demonstrate the advantage of UAEs for improving a variety of unsupervised machine learning tasks. We summarize our main contributions as follows.
• We propose a new per-sample based mutual information neural estimator (MINE) between a pair of original and modified data samples as an information-theoretic similarity measure and a supervision-free approach for generating UAE. For instance, see UAEs for data reconstruction in Figure 2 of supplementary material.
• We formulate the generation of adversarial examples with MINE as a constrained optimization problem, which applies to the unsupervised machine learning task. We then develop an efficient MinMax optimization algorithm (Algorithm 2).
• We show a novel application of UAEs as a simple plug-in data augmentation tool for several unsupervised machine learning tasks, including data reconstruction, representation learning, and contrastive learning on image and tabular datasets. Our extensive experimental results show outstanding performance gains (up to 73.5% performance improvement) by retraining the model with the generated UAEs.

2 Methodology

2.1 MINE of Single Data Sample and MINE-based MinMax attack Algorithm

Mutual information (MI) measures the mutual dependence between two random variables X and Z. For efficient computation of MI, the mutual information neural estimator (MINE) with consistency guarantees is proposed in Belghazi et al. (2018). However, the vanilla MINE is not applicable because it only applies to a batch of data samples (so that empirical data distributions can be used for computing MI estimates) but not to single data sample. To bridge this gap, we will propose two MINE-based sampling methods for single data sample denoted as the per-sample MINE $I_\Theta(x, x+\delta)$: (1) *random sampling* and (2) *convolution output*. (1) **Random Sampling**: we perform independent Gaussian sampling of a given data sample x to obtain a batch of K compressed samples $\{x_k, (x+\delta)_k\}_{k=1}^{K}$ for computing $I_\Theta(x, x+\delta)$ via MINE. (2) **Convolution output**: we propose to use the output of the first convolution layer of a data input, denoted by $conv(\cdot)$, to obtain K feature maps $\{conv(x)_k, conv(x+\delta)_k\}_{k=1}^{K}$ for computing $I_\Theta(x, x+\delta)$ based on the neural network model using a convolution layer to process the input data. We show the comparison between random sampling and convolution output and more detail of them in Appendix B.1 and B.2. In this paper we use convolution-based approach whenever applicable and otherwise use random sampling.

Many machine learning tasks such as data reconstruction and unsupervised representation learning do not use data labels. Here we use an autoencoder $\Phi(\cdot)$ for data reconstruction to illustrate the unsupervised attack formulation. The design principle can naturally extend to other unsupervised tasks. The autoencoder Φ takes a data sample x as an input and outputs a reconstructed data sample $\Phi(x)$. Different from the rationale of supervised attack, for unsupervised attack we propose to use MINE to find the *least similar* perturbed data sample $x+\delta$ with respect to x while ensuring the reconstruction loss of $\Phi(x+\delta)$ is no greater than $\Phi(x)$ (i.e., the criterion of successful attack for data reconstruction). The unsupervised attack formulation is as follows:

$$\underset{\delta}{\text{Minimize}} \quad I_\Theta(x, x+\delta)$$

$$\textit{such that } x+\delta \in [0,1]^d \text{ , } \delta \in [-\epsilon, \epsilon]^d \text{ and } f_x^{\text{unsup}}(x+\delta) \leq 0$$

which means the attack is considered successful (i.e., $f_x^{\text{unsup}}(x+\delta) \leq 0$) if the reconstruction loss of $x+\delta$ relative to the original sample x is smaller than the native reconstruction loss minus a nonnegative margin κ. That is, $\|x - \Phi(x+\delta)\|_2 \leq \|x - \Phi(x)\|_2 - \kappa$. In other words, our unsupervised attack formulation aims to find that most dissimilar perturbed sample $x+\delta$ to x measured by MINE while having smaller reconstruction loss (in reference to x) than the that of x. Such UAEs thus relates to generalization errors on low-loss samples.

Here we propose a unified MinMax algorithm for solving the aforementioned unsupervised attack formulation. For simplicity, we will use f_x to denote the attack criterion for f_x^{unsup}. We reformulate

Table 2: Comparison of data reconstruction by retraining the autoencoder on the UAE-augmented data. The reconstruction error is the average L_2 reconstruction loss of the test set. The improvement (in green/red) is with respect to the original model. The attack success rate (ASR) is the fraction of augmented training data having smaller reconstruction loss than the original loss (see Table 1 for definition).

MNIST									
	Reconstruction Error (test set)					ASR (training set)			
Autoencoder	Original	MINE-UAE	L_2-UAE	GA ($\sigma = 0.01$)	GA ($\sigma = 10^{-3}$)	MINE-UAE	L_2-UAE	GA ($\sigma = 0.01$)	GA ($\sigma = 10^{-3}$)
Sparse	0.00561	**0.00243** (↑ 56.7%)	0.00348 (↑ 38.0%)	0.00280±2.60e-05 (↑ 50.1%)	0.00280±3.71e-05 (↑ 50.1%)	100%	99.18%	54.10%	63.95%
Dense	0.00258	**0.00228** (↑ 11.6%)	0.00286 (↓ 6.0%)	0.00244±0.00014 (↑ 5.4%)	0.00238±0.00012 (↑ 7.8%)	92.99%	99.94%	48.53%	58.47%
Convolutional	0.00294	**0.00256** (↑ 12.9%)	0.00364 (↓ 23.8%)	0.00301±0.00011 (↓ 2.4%)	0.00304±0.00015 (↓ 3.4%)	99.86%	99.61%	68.71%	99.61%
Adversarial	0.04785	**0.04581** (↑ 4.3%)	0.06098 (↓ 27.4%)	0.05793±0.00501 (↓ 21%)	0.05544±0.00567 (↓ 15.86%)	98.46%	43.54%	99.79%	99.83%
SVHN									
Sparse	0.00887	**0.00235** (↑ 73.5%)	0.00315 (↑ 64.5%)	0.00301±0.00137 (↑ 66.1%)	0.00293±0.00078 (↑ 67.4%)	100%	72.16%	72.42%	79.92%
Dense	0.00659	**0.00421** (↑ 36.1%)	0.00550 (↑ 16.5%)	0.00858±0.00232 (↓ 30.2%)	0.00860±0.00190 (↓ 30.5%)	99.99%	82.65%	92.3%	93.92%
Convolutional	0.00128	**0.00095** (↑ 25.8%)	0.00121 (↑ 5.5%)	0.00098 ± 3.77e-05 (↑ 25.4%)	0.00104±7.41e-05 (↑ 18.8%)	100%	56%	96.40%	99.24%
Adversarial	0.00173	**0.00129** (↑ 25.4%)	0.00181 (↓ 27.4%)	0.00161±0.00061 (↑ 6.9%)	0.00130±0.00037 (↑ 24.9%)	94.82%	58.98%	97.31%	99.85%

the attack generation via MINE as the following MinMax optimization problem with simple convex set constraints:

$$\underset{\delta: x+\delta \in [0,1]^d,\ \delta \in [-\epsilon,\epsilon]^d}{\mathrm{Min}} \quad \underset{c \geq 0}{\mathrm{Max}} \quad F(\delta, c) \triangleq c \cdot f_x^+(x+\delta) - I_\Theta(x, x+\delta)$$

The outer minimization problem finds the best perturbation δ with data and perturbation feasibility constraints $x + \delta \in [0,1]^d$ and $\delta \in [-\epsilon, \epsilon]^d$, which are both convex sets with known analytical projection functions. The inner maximization associates a variable $c \geq 0$ with the original attack criterion $f_x(x+\delta) \leq 0$, where c is multiplied to the ReLU activation function of f_x, denoted as $f_x^+(x+\delta) = \text{ReLU}(f_x(x+\delta)) = \max\{f_x(x+\delta), 0\}$. The use of f_x^+ means when the attack criterion is not met (i.e., $f_x(x+\delta) > 0$), the loss term $c \cdot f_x(x+\delta)$ will appear in the objective function F. On the other hand, if the attack criterion is met (i.e., $f_x(x+\delta) \leq 0$), then $c \cdot f_x^+(x+\delta) = 0$ and the objective function F only contains the similarity loss term $-I_\Theta(x, x+\delta)$. Therefore, the design of f_x^+ balances the tradeoff between the two loss terms associated with attack success and MINE-based similarity. We propose to use alternative projected gradient descent between the inner and outer steps to solve the MinMax attack problem, which is summarized in Algorithm 2 (see Appendix B.3).

2.2 Data Augmentation using UAE

With the proposed MinMax attack algorithm and per-sample MINE for similarity evaluation, we can generate MINE-based unsupervised adversarial examples (UAEs). Section 3 will show novel applications that MINE-based UAEs can be used as a simple plug-in data augmentation tool to boost the model performance of several unsupervised machine learning tasks.

3 Results

In this section, we conduct extensive experiments on a variety of datasets and neural network models to demonstrate the performance of our proposed MINE-based MinMax adversarial attack algorithm and the utility of its generated UAEs for data augmentation. The detail of experiment setup and datasets sees in the supplementary material.

UAE Improves Data Reconstruction. Data reconstruction using an autoencoder $\Phi(\cdot)$ that learns to encode and decode the raw data through latent representations is a standard unsupervised learning task. Here we use the default implementation of the four different autoencoders to generate UAEs based on the training data samples of MNIST and SVHN for data augmentation, retrain the model from scratch on the augmented dataset, and report the resulting reconstruction error on the original test set. All autoencoders use the L_2 reconstruction loss defined as $\|x - \Phi(x)\|_2$. We provide more details about the model retraining on augmented data in the supplementary material.

We also compare the performance of our proposed MINE-based UAE (MINE-UAE) with two baselines: (i) L_2-*UAE* that replaces the objective of minimizing $I_\Theta(x, x+\delta)$ with maximizing the L_2 reconstruction loss $\|x - \Phi(x+\delta)\|_2$ in the MinMax attack algorithm while keeping the same attack success criterion. (ii) *Gaussian augmentation* (GA) that adds zero-mean Gaussian noise with a diagonal covariance matrix of the same constant σ^2 to the training data.

We observe significant and consistent performance improvement in data reconstruction (up to 73.5% improvement) in Table 2. Table 2 shows the reconstruction loss and the ASR. The improvement of reconstruction error is measured with respect to the reconstruction loss of the original model (i.e., without data augmentation). We find that MINE-UAE can attain much higher ASR than L_2-UAE and GA in most cases. More importantly, data augmentation using MINE-UAE achieves consistent and significant reconstruction performance improvement across all models and datasets (up to 56.7% on MNIST and up to 73.5% on SVHN), validating the importance and effectiveness of using MINE-UAE for data augmentation. On the other hand, in several cases L_2-UAE and GA lead to notable performance degradation. The results suggest that MINE-UAE can be an effective and plug-in data augmentation tool for unsupervised machine learning models, as it simply uses the training data and the original model to generate UAEs for model retraining. Moreover, we compare MINE-UAE to other data augmentation methods such as flipping and rotation in Appendix B.5. UAE can improve data reconstruction when the original model involves augmented training data.

UAE Improves Representation Learning. The concrete autoencoder proposed in Balın et al. (2019) is an unsupervised feature selection method which recognizes a subset of the most informative features through an additional *concrete select layer* with M nodes in the encoder for data reconstruction. We apply MINE-UAE for data augmentation on a variety of datasets and use the same post-hoc classification evaluation procedure as in Balın et al. (2019) for the learned representations, which passes the selected features to an extremely randomized tree classification model Geurts et al. (2006).

Table 3: Performance evaluation of representation learning by the concrete autoencoder and the resulting classification accuracy.

	Reconstruction Error (test set)		Accuracy (test set)		ASR
Dataset	Original	MINE-UAE	Original	MINE-UAE	MINE-UAE
MNIST	0.01170	**0.01142** (↑ 2.4%)	94.97%	95.41%	99.98%
Fashion MMIST	0.01307	**0.01254** (↑ 4.1%)	84.92%	85.24%	99.99%
Isolet	0.01200	**0.01159** (↑ 3.4%)	81.98%	82.93%	100%
Coil-20	**0.00693**	0.01374 (↓ 98.3%)	98.96%	96.88%	9.21%
Mice Protein	0.00651	**0.00611** (↑ 6.1%)	89.81%	91.2%	40.24%
Activity	0.00337	**0.00300** (↑ 11.0%)	83.38%	84.45%	96.52%

The six datasets and the resulting classification accuracy are reported in Table 3. We select $M = 50$ features for every dataset except for the Mice Protein dataset (we set $M = 10$) owing to its small data dimension. We find that MINE-UAE can attain up to 11% improvement for data reconstruction and up to 1.39% increase in accuracy among 5 out of 6 datasets, corroborating the utility of MINE-UAE in representation learning and feature selection.

UAE Improves Contrastive Learning. The SimCLR algorithm Chen et al. (2018) is a new framework for contrastive learning of visual representations. It uses self-supervised data modifications for efficient contrastive learning and is shown to improve several downstream image classification tasks.

We use the default implementation of SimCLR on CIFAR-10 and generate MINE-UAEs using the training data and the defined contrastive training loss for SimCLR. Table 4 shows the loss, ASR and the resulting classification accuracy using a linear classifier on the learned representations. We find that using MINE-UAE for additional data augmentation and model retraining can yield 7.8% improvement in contrastive loss and 1.58% increase in classification accuracy, suggesting advanced contrastive learning performance. Moreover, we find that MINE-UAE data augmentation also leads to a significant gain in adversarial robustness (see Appendix B.11).

Table 4: Comparison of contrastive loss and the resulting accuracy on CIFAR-10 using SimCLR Chen et al. (2018). The attack success rate (ASR) is the fraction of augmented training data having smaller contrastive loss than the original loss. The SimCLR model is ResNet-18 and the batch size is set to be 512.

	CIFAR-10		
Model	Loss (test set)	Accuracy (test set)	ASR
Original	0.29010	91.30%	-
MINE-UAE	**0.26755** (↑ 7.8%)	**92.88%**	100%

4 Conclusion

In this paper, we propose a novel framework for studying adversarial examples in unsupervised learning tasks, based on our developed per-sample mutual information neural estimator as an information-theoretic similarity measure. We also propose a new MinMax algorithm for efficient generation of MINE-based unsupervised adversarial examples. As a novel application, we show that MINE-based UAEs can be used as a simple yet effective plug-in data augmentation tool and achieve significant performance gains in data reconstruction, representation learning, and contrastive learning.

REFERENCES

Muhammed Fatih Balın, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *International Conference on Machine Learning*, pp. 444–453, 2019.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pp. 531–540, 2018.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *Neural Information Processing Systems*, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2018.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

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

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Marc'Aurelio Ranzato, Fu Jie Huang, Y-Lan Boureau, and Yann LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

Robert Stanforth, Alhussein Fawzi, Pushmeet Kohli, et al. Are labels required for improving adversarial robustness? *Neural Information Processing Systems*, 2019.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international Conference on Computer Vision*, pp. 1476–1485, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482, 2019.

Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.