Adversarial Data Augmentation Improves Unsupervised Machine Learning

Chia-Yi Hsu[1], Songtao Lu[2], Pin-Yu Chen[2], Sijia Liu[3] and Chia-Mu Yu[1]

[1]National Yang Ming Chiao Tung University, Taiwan
[2]AI Foundations Learning Group, IBM Thomas J. Watson Research Center, USA
[3]Michigan State University, USA

ABSTRACT

- 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 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**.
- The unsupervised attack formulation is as follows:

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

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

- Here we use an auto-encoder $\Phi(.)$ 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 there construction loss of $\Phi(x+\delta)$ is no greater than $\Phi(x)$(i.e., the criterion of successful attack for data reconstruction).

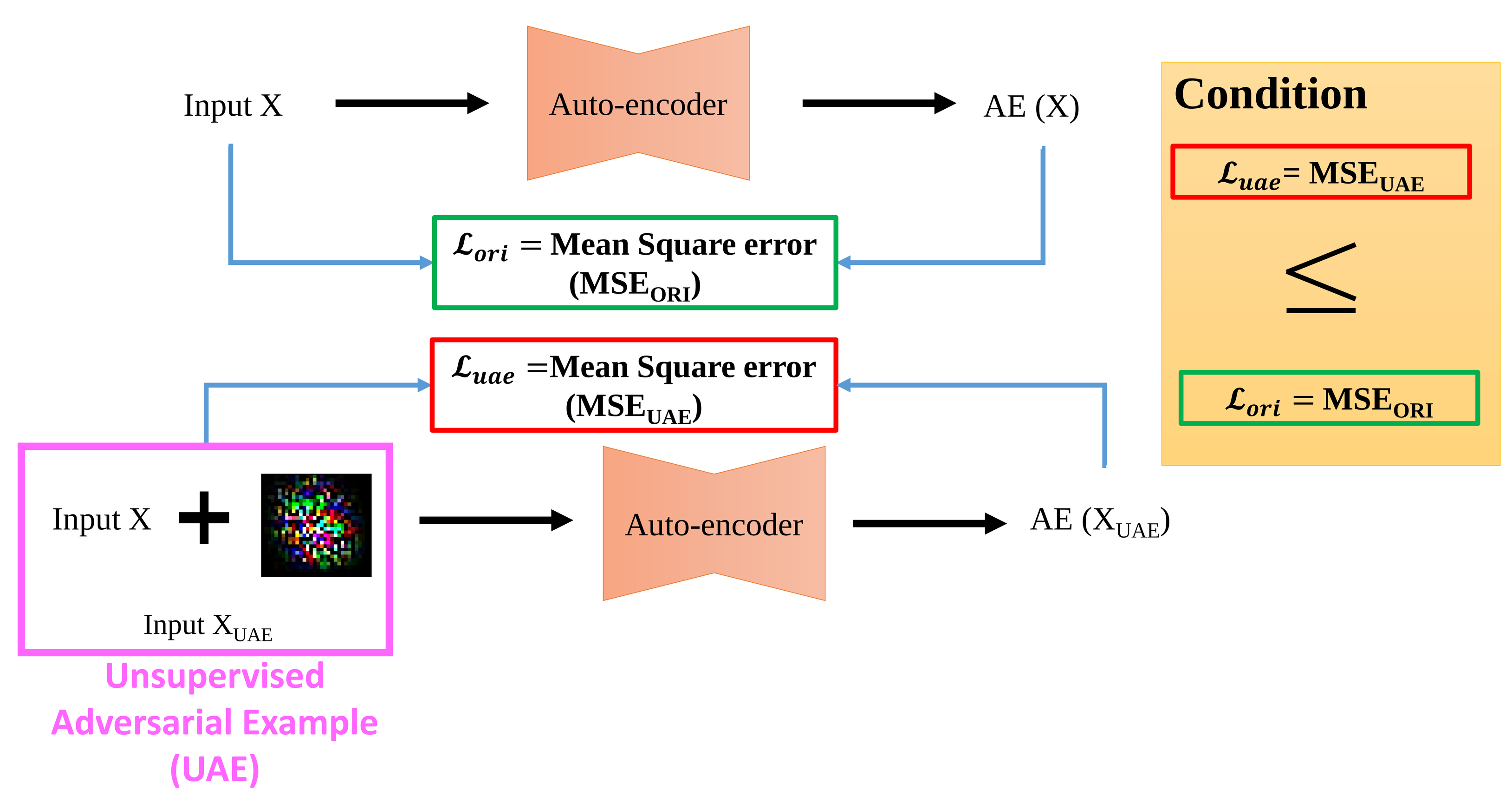

Figure 1: Generation of unsupervised adversarial examples (UAEs)

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

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

The outer minimization problem finds the best perturbation δ with data and perturbation feasibility constraints $x+\delta \in [0,1]^d$ and $\delta \in [\epsilon, -\epsilon]^d$. The inner maximization associates a variable $c \ge 0$ with the original attack criterion $f_x(x+\delta) \le 0$.

Experiments

UAE Improves Data Reconstruction.

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 ($\sigma=10^{-3}$)
	Reconstruction Error (test set)					ASR (training set)			
MNIST									
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%

Table 1: 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.

UAE Improves Representation Learning.

The concrete autoencoder proposed in Balın et al. [1] 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.

Dataset	Original	MINE-UAE	Original	MINE-UAE	MINE-UAE
	Reconstruction Error (test set)		Accuracy (test set)		ASR
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%

Table 2: Performance evaluation of representation learning by the concrete autoencoder and the resulting classification accuracy.

UAE Improves Contrastive Learning.

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

Table 3: Comparison of contrastive loss and the resulting accuracy on CIFAR-10 using SimCLR Chen et al. [2]. 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.

CONCLUSION

- 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.

REFERENCE

[1] 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.
[2] 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.