Variational Inference with Latent Space Quantization for Adversarial Resilience

Vinay Kyatham   Deepak Mishra   Prathosh AP

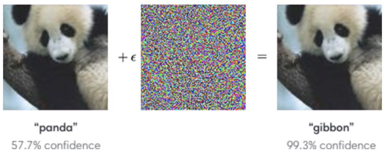
Abstract

Deep Neural Networks suffer from the threat of adversarial attacks - Existence of perceptually valid input-like samples obtained through careful perturbation that lead to degradation in the performance of the underlying model.

In this paper, we propose a generalized defense mechanism,

An adversarial filter, devoid of access to classifier and adversaries, which makes it usable in tandem with any classifier.

The basic idea is to learn a Lipschitz constrained mapping from the data manifold, incorporating adversarial perturbations, to a quantized latent space and re-map it to the true data manifold



An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon.

Introduction

Deep neural networks have shown tremendous success in omputer vision tasks. One of the primary factors contributing to their success is the availability of abundant data. Incomplete exploration of data space with the available training set, which in-turn results in loopholes in the data manifold . Adversarial attacks exploit these gaps in the data manifold, unexplored by the classifier.

Attack Methods:

FGSM (Fast Gradient Sign Method) that performs a one step gradient update along the direction of sign of gradient of loss at each pixel

DeepFool is another iterative attack which computes adversarial perturbations through an orthogonal projection of the sample on the decision boundary.

CW (Carlini-Wagner) attack is an optimization based attack

Defense Mechanisms:

Adversarial retraining: retrain the classifier using adversaries.

Adversarial filtering:
These defense mechanisms pre-process the adversarial examples to make them non-adversarial either by manifold projection or by using generative models.

MagNet trains a collection of detector networks to push adversarial examples close to the data manifold.

Defense-GAN trains a GAN only on legitimate examples and uses it for adversarial resilience - During inference, given an input image (adversarial or otherwise), a sample `close' to the input is generated by an iterative inference procedure in the latent space.

Methodology

Proposed Defense Method LQ-VAE is an Adversarial Filter,

LQ-VAE uses a Lipschitz constrained latent encoder which preserves the distances under a metric space on the latent and the data manifolds. This makes sure that real image and its its adversarial image are very close in the latent space,

LQ-VAE uses the Quantized latent space,For Training, LQ-VAE uses dual decoders, one works on real latent vector and other on the quantized.

complete overview of proposed Method LQ-VAE.

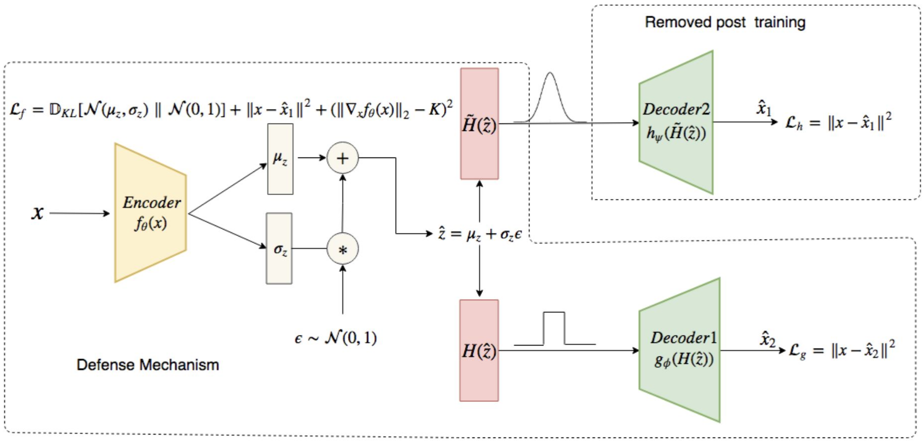

Removed post training
$\mathcal{L}_f = \mathbb{D}_{KL}[\mathcal{N}(\mu_z,\sigma_z) \parallel \mathcal{N}(0,1)] + \|x-\hat{x}_1\|^2 + (\|\nabla_x f_\theta(x)\|_2 - K)^2$
x
Encoder
$f_\theta(x)$
μ_z
σ_z
+
*
$\epsilon \sim \mathcal{N}(0,1)$
$\hat{z} = \mu_z + \sigma_z\epsilon$
$\tilde{H}(\hat{z})$
$H(\hat{z})$
Decoder2
$h_\psi(\tilde{H}(\hat{z}))$
$\hat{x}_1$
$\mathcal{L}_h = \|x-\hat{x}_1\|^2$
Decoder1
$g_\phi(H(\hat{z}))$
$\hat{x}_2$
$\mathcal{L}_g = \|x-\hat{x}_2\|^2$
Defense Mechanism

Fig. 1. **Proposed LQ-VAE - A Lipschitz constrained encoder (L_f) encodes the input image into a latent space quantized by the function H which is explored through a stochastic perturbation (ϵ). During inference, Decoder1 maps the quantized latent codes generated by the adversarial images back to the image space. Training is done only on real data samples by using an approximate differentiable version of Decoder1 (i.e. Decoder2).**

Algorithm 1 gives the details of the LQ-VAE training procedure

```
Algorithm 1 LQ-VAE algorithm
    Input: Dataset $\mathcal{D}$, Batchsize $B$, Encoder $f_\theta$, Decoder1
        $g_\phi$, Learning rate $\eta$, Quantization functions $H$, $\tilde{H}$
    Output Parameters $\theta^*$, $\phi^*$
Make a copy $h_\psi$ of decoder $g_\phi$
repeat
    Sample $\{\mathbf{x}^{(1)} \cdots \mathbf{x}^{(B)}\}$ from dataset $\mathcal{D}$
    $\mu_{\mathbf{z}}^{(i)}, \sigma_{\mathbf{z}}^{(i)} \leftarrow f_\theta(\mathbf{x}^{(i)})$
    Sample $\hat{\mathbf{z}}^{(i)}$ from $\mathcal{N}(\mu_{\mathbf{z}}^{(i)}, {\sigma_{\mathbf{z}}^{(i)}}^2)$
    $\hat{\mathbf{x}}_1^{(i)} \leftarrow h_\psi(\tilde{H}(\hat{\mathbf{z}}^{(i)}))$
    $\hat{\mathbf{x}}_2^{(i)} \leftarrow g_\phi(H(\hat{\mathbf{z}}^{(i)}))$
    $\mathcal{L}_h \leftarrow \sum_{i=1}^{B} \left\|\mathbf{x}^{(i)} - \hat{\mathbf{x}}_1^{(i)}\right\|_2^2$
    $\mathcal{L}_g \leftarrow \sum_{i=1}^{B} \left\|\mathbf{x}^{(i)} - \hat{\mathbf{x}}_2^{(i)}\right\|_2^2$
    $\mathcal{L}_f \leftarrow \mathcal{L}_h + \sum_{i=1}^{B} D_{KL}\left[\mathcal{N}(\mu_{\mathbf{z}}^{(i)}, {\sigma_{\mathbf{z}}^{(i)}}^2)||\mathcal{N}(0,1)\right] +$
        $\sum_{i=1}^{B}\left[\left\|\nabla_x f_\theta(x^{(i)})\right\|_2 - K\right]^2$
    $\theta \leftarrow \theta + \eta\nabla_\theta\mathcal{L}_f$
    $\phi \leftarrow \phi + \eta\nabla_\phi\mathcal{L}_g$
    $\psi \leftarrow \psi + \eta\nabla_\psi\mathcal{L}_h$
until convergence of $\theta$, $\phi$
```

Variational Inference with Latent Space Quantization for Adversarial Resilience

Vinay Kyatham   Deepak Mishra   Prathosh AP

Results

Classification accuracy of the MNIST classifiers on white box attacks with various defense strategies.

Attack	Model	No Attack	No Defense	LQ-VAE	Defense-GAN	Madry	Adv Tr
FGSM	A	99.40	20.16	89.17	90.43	96.85	67.95
	B	99.41	13.17	86.70	88.52	96.20	49.49
	C	98.37	5.66	83.02	86.7	84.71	80.75
DeepFool	A	99.40	7.38	97.60	95.41	67.82	3.10
	B	99.41	5.88	97.74	93.03	66.35	5.75
	C	98.37	48.24	97.42	92.32	62.38	10.97
CW	A	99.40	8.85	97.66	94.37	69.15	1.20
	B	99.41	5.07	97.20	90.56	71.35	1.45
	C	98.37	8.44	97.36	92.5	58.65	2.15
Average		99.06	13.65	**93.76**	91.54	74.83	24.76

Classification accuracy of the CelebA classifiers on white box attacks with various defense strategies.

Attack	Model	No Attack	No Defense	LQ-VAE	Defense-GAN	Madry	Adv Tr
FGSM	A	96.34	3.65	81.04	74.13	62.35	4.53
	B	96.60	3.40	64.74	67.06	71.42	72.88
	C	95.02	28.62	61.48	53.76	61.35	42.55
DeepFool	A	96.34	3.56	85.89	83.87	52.86	6.26
	B	96.60	2.43	83.81	83.65	49.39	14.17
	C	95.02	10.92	62.79	78.56	42.37	38.45
CW	A	96.34	6.98	85.90	84.64	58.62	11.88
	B	96.60	6.88	86.29	86.01	60.33	12.91
	C	95.02	10.92	79.20	78.56	45.02	38.45
Iter FGSM	A	96.34	3.12	85.44	81.00	82.34	3.50
	B	96.60	3.55	72.29	72.05	72.19	9.16
	C	95.02	11.92	52.12	42.13	90.87	19.47
Madry	A	96.34	2.84	85.11	81.43	76.35	3.52
	B	96.60	3.12	70.01	74.01	70.32	8.52
	C	95.02	8.57	54.00	45.11	84.09	18.59
Average		95.99	7.37	**74.01**	72.40	65.32	20.32

Classification accuracy of the FMNIST classifier on DeepFool black box attack images generated using substitute model.

Classifier/ Substitute	No Attack	No Defense	LQ-VAE	Defense-GAN	Madry	Adv Tr
A/B	92.76	29.14	77.74	74.41	60.11	48.27
A/C	92.76	35.44	77.11	74.11	62.58	57.53
B/A	91.17	67.82	81.33	77.97	80.71	76.61
B/C	91.17	45.55	78.83	74.5	69.19	64.05
C/A	89.06	79.11	82.12	78.82	80.99	81.84
C/B	89.06	47.26	80.76	76.6	67.46	59.64
Average	91.00	50.72	**79.65**	76.07	70.17	64.66

Classification accuracy of the CelebA classifier on CW black box attack images generated using substitute model

Classifier/ Substitute	No Attack	No Defense	LQ-VAE	Defense-GAN	Madry	Adv Tr
A/B	96.34	39.53	86.01	84.70	85.41	94.13
A/C	96.34	37.59	80.10	78.11	64.72	54.22
B/A	96.60	49.21	85.67	86.19	82.55	68.11
B/C	96.60	52.52	79.98	79.91	76.31	62.53
C/A	95.02	82.87	85.90	86.29	89.79	86.75
C/B	95.02	83.26	85.20	86.17	89.91	88.40
Average	95.99	57.50	**83.81**	83.56	81.45	75.69

Classification accuracy of end-to-end whitebox attack on LQ-VAE - classifier combination using BPDA.

Dataset	LQ-VAE	DGAN
MNIST	83.70	55.17
FMNIST	57.41	39.41
CELEBA	82.12	23.52

Results

Below Figure is a 2D t-SNE plot of the latent encodings (from the Lipschitz constrained encoder) of the true and the CW l_2 attacked adversarial samples from the MNIST data is shown. It can be seen that embeddings of the adversaries are extremely close to those of the true samples.

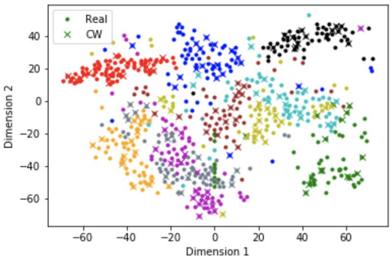
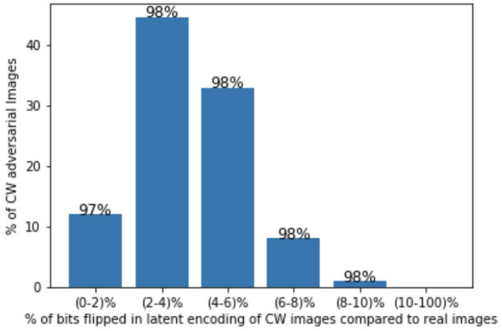


Fig. 2. t-SNE plot of the latent encodings of mnist real images and CW adversarial images. It can seen that the embeddings of real and adversarial data overlap.

Below Figure depicts the distribution of the bit-flippings in the latent codes of the CW adversaries on the MNIST data - It can be seen that about 90 % of the total adversaries undergo less than 6% of bits being flipped resulting in high classification accuracy (seen on top of the bars).



Conclusions

LQ-VAE offers several advantages over Defense-GAN such as -

(i) LQ-VAE does not involve a run-time search on the latent space unlike Defense-GAN which makes it orders of magnitude faster.

(ii) Training a VAE is known to be easier and faster, yielding a better data likelihood than a GAN

(iii) Defense-GAN can be attacked too by a method called the Backward Pass Differentiable Approximation (BPDA), in which case its defense on MNIST is reported at 55% . When the same technique is used to attack LQ-VAE, we get much better accuracy of 83% on the same task which can be ascribed to the use of latent space constraining and quantization.

In principle, LQ-VAE can be re-trained using adversaries too, For instance, it is observed that if one retrains LQ-VAE using Madry adversaries, its performance is enhanced by 5-10% on FGSM attacks.