Wen LI, Hengyou WANG, Lianzhi HUO, Qiang HE, Linlin CHEN, Zhiquan HE, Wing W. Y. Ng, 2024. Low-rank matrix recovery with total generalized variation for defending adversarial examples. *Frontiers of Information Technology & Electronic Engineering*, 25(3):432-445. https://doi.org/10.1631/FITEE.2300017

Low-rank matrix recovery with total generalized variation for defending adversarial examples

Key words: Total generalized variation; Low-rank matrix; Alternating direction method of multipliers; Adversarial example

Corresponding author: Hengyou WANG
E-mail: wanghengyou@bucea.edu.cn
ORCID: https://orcid.org/0000-0001-6693-0161

Motivation

- With the development of deep neural network (DNN), it has been found that DNNs are easily disturbed by adversarial noise. If a subtle perturbation is added to the input image, the given DNN could be misclassified with high confidence. How to enhance the robustness of DNNs is crucial.
- Recently, defense methods combining the low-rank matrix theory and total variation (TV) regularization have been proposed to eliminate adversarial perturbation. However, the above defense methods usually perform over-smoothing operations on the global image, causing significant damage to the non-attacked images. Therefore, these defense methods reduce the classification accuracy of the original (or non-attacked) images.

Main idea

- A new method of defending adversarial examples based on low-rank matrix recovery with total generalized variation (TGV) is proposed. It not only removes the adversarial perturbation but also guarantees restoration of the edges and detailed information.
- To deal with the challenging optimal model, an algorithm based on the alternating direction method of multipliers (ADMM) is designed. It divides the multi-variable optimization problem into several single-variable optimization sub-problems.

Framework and method

A DNN can be easily destroyed by adversarial noise. To eliminate the adversarial noise, we followed three steps:

- Firstly, the pixels of images in the dataset are randomly masked with probability *p* to obtain the masked images. Pixel values of the masked part are set to zero.
- Then, the LRTGV regularization algorithm is applied to obtain the reconstructed image. The edges and local detail information can be better restored.
- Finally, the reconstructed images can be classified correctly with higher probability by the trained DNN.

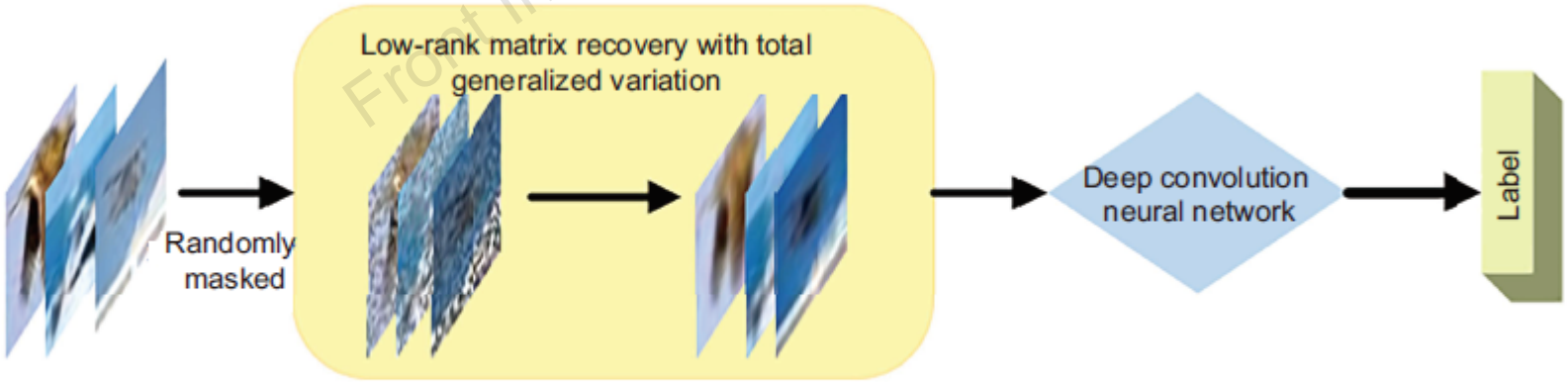

Fig. 1 Training process of low-rank matrix recovery with total generalized variation (LRTGV)

Framework and method

- TGV (i.e., high-order TV regularization) can better preserve image details. Inspired by TGV, we present the LRTGV to defend the adversarial samples. The proposed method can effectively remove adversarial noise while preserving the global structure and local detail information of the image. This method solves the over-smoothing problem of first-order TV regularization and improves the classification accuracy of the original images. The LRTGV regularization model can be formulated as follows:

$$\min_{X,E} \left(\sum_{j=1}^{n} w_j^X \sigma_j + \lambda \left\| W^E \odot \hat{M} \odot E \right\|_1 + \mathrm{TGV}_\alpha^2(X) \right)$$
$$\text{s.t. } X^{\mathrm{adv}} = X + E,$$

where $\hat{M}$ is the binary mask matrix of the noise location and λ is the balance parameter for the three items in the objective function. If there is a noisy pixel at position (i, j), $\hat{M}(i, j)$ is set as 0; $\hat{M}(i, j) = 1$ otherwise.

Framework and method

- We adopt the ADMM to solve the optimization problem. To solve the problem easily, the auxiliary variable A is introduced and the augmented Lagrangian function can be reformulated as follows:

$$
\begin{aligned}
& f(X, A, E, Y_1, Y_2) \\
= & \sum_{j=1}^{m} w_j^X \sigma_j + \|W^E \odot \hat{M} \odot E\|_1 + \mathrm{TGV}_\alpha^2(A) \\
& + \langle Y_1, X^{\mathrm{adv}} - X - E \rangle + \langle Y_2, A - X \rangle \\
& + \frac{\mu}{2} \left\| X^{\mathrm{adv}} - X - E \right\|_{\mathrm{F}}^2 + \frac{\mu}{2} \|A - X\|_{\mathrm{F}}^2 ,
\end{aligned}
\tag{13}
$$

- We divide problem (13) into several sub-problems. In our problem, there are five main variables X, A, E, Y_1, Y_2, which should be updated in each iteration. The updated iteration formulas of variables and their optimal solutions are given.

Major results

- We evaluated the LRTGV algorithm's capability to restore images from normal noise.

Table 1 PSNR comparison of PCP, reweighted L_1, NSVT, ROUTE, SRLRMR, and our LRTGV algorithms

Noise density	Image No.	PSNR (dB)						ΔPSNR (dB)
		PCP	Reweighted L_1	NSVT	ROUTE	SRLRMR	LRTGV	LRTGV−SRLRMR
$p = 0.2$	1	26.75	23.84	32.15	28.50	32.47	**34.42**	+1.95
	2	24.23	22.33	26.28	23.44	26.36	**28.87**	+2.51
	3	28.41	26.75	31.99	29.22	**33.71**	33.32	−0.39
	4	25.90	22.79	28.10	25.52	**29.64**	26.08	−3.56
	5	29.77	26.43	34.63	29.56	**34.88**	32.61	−2.27
	6	29.98	27.96	34.01	28.79	33.73	**34.30**	+0.57
	7	25.71	23.49	28.78	26.18	29.31	**31.12**	+1.81
	8	21.72	20.43	22.12	20.00	23.71	**24.40**	+0.69
	9	25.22	23.01	27.65	25.31	28.35	**32.03**	+3.68
	10	18.89	18.93	**19.33**	16.94	19.13	19.23	+0.10
	11	26.34	26.29	28.28	27.60	29.88	**31.26**	+1.38
	12	24.13	24.09	25.20	23.70	25.55	**28.27**	+2.72
	13	24.09	24.02	26.20	25.04	27.86	**29.04**	+1.18
	14	24.05	24.00	26.50	25.77	27.99	**32.71**	+4.72
	15	22.67	22.53	23.10	19.95	23.58	**23.63**	+0.05
	Average	25.19	23.79	27.62	25.03	28.41	**29.42**	+1.01
$p = 0.3$	1	21.27	22.17	23.84	23.52	30.67	**33.45**	+2.78
	2	19.10	20.85	20.77	20.43	24.35	**25.72**	+1.37
	3	23.00	25.42	25.56	26.07	31.15	**32.20**	+1.05
	4	22.52	23.04	26.61	23.26	27.36	**27.53**	+0.17
	5	25.20	26.69	28.73	27.41	33.31	**33.65**	+0.34
	6	27.29	26.57	30.57	28.31	33.01	**33.99**	+0.98
	7	20.92	22.01	22.93	22.85	27.60	**29.39**	+1.79
	8	18.09	19.21	18.12	18.00	21.90	**22.96**	+1.06
	9	21.25	21.62	22.01	21.94	27.66	**29.98**	+2.32
	10	16.43	16.40	17.56	14.82	**18.12**	17.61	−0.51
	11	21.41	21.43	22.00	23.92	25.31	**31.83**	+6.52
	12	19.61	19.14	21.34	20.82	22.25	**26.79**	+4.54
	13	19.22	18.82	22.14	21.94	23.22	**28.41**	+5.19
	14	20.11	19.91	20.54	21.76	26.83	**29.14**	+2.31
	15	18.23	18.11	20.43	17.73	21.34	**22.49**	+1.15
	Average	20.91	21.43	22.88	22.19	26.27	**28.34**	+2.07

Best results are in bold. PSNR: peak signal-to-noise ratio

Major results

Fig. 5 Reconstruction results of the six algorithms for images with 20% salt and pepper noise: (a) original; (b) noisy; (c) PCP; (d) reweighted L_1; (e) NSVT; (f) ROUTE; (g) SRLRMR; (h) LRTGV

Major results

- We compared our proposed LRTGV adversarial defense algorithm with state-of-the-art defense method.

Table 2 Black-box defense on CIFAR-10

Method	Classification accuracy (%)			
	Clean	PGD ($\varepsilon = 8/255$ / $16/255$)	Momentum ($\varepsilon = 8/255$ / $16/255$)	Diverse-input ($\varepsilon = 8/255$ / $16/255$)
No defense	87.06	34.18 / 21.30	24.36 / 12.16	20.81 / 10.39
JPEG	**85.22**	48.03 / 30.05	34.75 / 15.78	30.14 / 13.49
Bit-depth reduction	85.02	44.50 / 28.38	32.54 / 15.43	28.28 / 13.52
TV minimization	82.71	50.47 / 32.57	36.63 / 16.15	32.36 / 14.07
Image quilting	77.49	64.76 / 52.86	53.61 / 34.12	51.25 / 30.51
PixelDefend	78.88	46.9 / 29.12	34.86 / 16.00	31.09 / 13.78
LCHP	77.60	62.20 / 56.32	53.86 / 45.36	51.02 / 44.46
LRTGV	83.85	**71.45 / 62.06**	**67.17 / 52.19**	**61.39 / 50.84**

Best results are in bold. ε represents the maximum perturbation magnitude of the attack methods. No defence indicates the reference classification trained on ResNet-18

Table 3 Black-box defense on SVHN

Method	Classification accuracy (%)			
	Clean	PGD ($\varepsilon = 8/255$ / $16/255$)	Momentum ($\varepsilon = 8/255$ / $16/255$)	Diverse-input ($\varepsilon = 8/255$ / $16/255$)
No defense	96.82	30.18 /13.62	29.96 / 9.99	28.79 / 9.29
JPEG	93.75	30.76 / 13.98	30.35 / 10.24	29.43 / 9.58
Bit-depth reduction	92.66	33.19 / 15.51	32.63 /11.36	31.58 / 10.68
TV minimization	92.59	34.27 / 15.32	33.61 / 11.06	32.71 /10.36
Image quilting	93.30	36.04 / 16.68	35.14 / 11.6	34.71 / 11.12
PixelDefend	92.54	30.90 /13.56	30.63 / 9.89	29.38 / 9.28
LCHP	91.56	57.87 / 29.81	**60.56** / 25.17	**60.39** / 24.23
LRTGV	**95.99**	**58.89 / 31.35**	60.17 / **25.90**	59.82 / **25.12**

Best results are in bold. ε represents the maximum perturbation magnitude of the attack methods. No defence indicates the reference classification trained on ResNet-18

Conclusions

- In this paper, to overcome the disadvantage of the first-order TV regularization denoising method, we integrated TGV regularization into the reweighted low-rank matrix decomposition model to remove the adversarial noise. It can be used to defend against different types of adversarial attacks.
- To address the proposed optimization problem, an iterative solution based on the alternating direction method of multipliers was designed, which can be applied to effectively eliminate adversarial noise. Experimental results showed that our proposed model consistently outperforms state-of-the-art baselines in image restoration and defense attacks, and improves the overall robustness under various adversarial attacks.