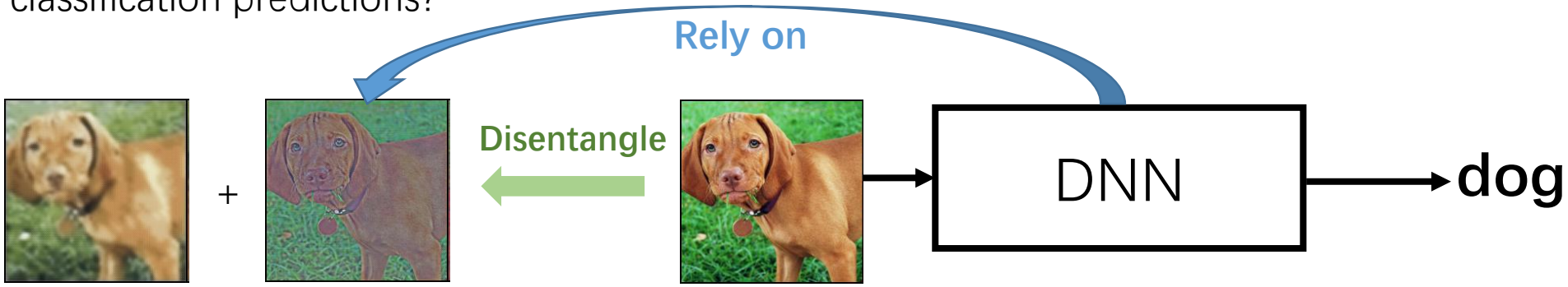
Class-Disentanglement and Applications in Adversarial Detection and Defense

Kaiwen Yang[1], Tianyi Zhou[2], Yonggang Zhang[1], Xinmei Tian[1], Dacheng Tao[3]

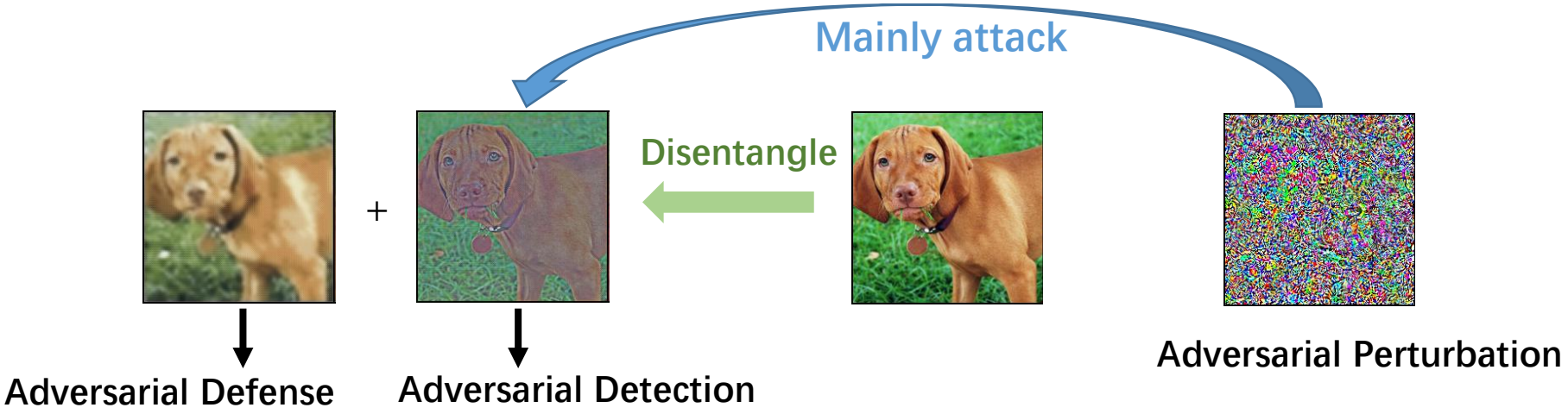
1. University of Science and Technology of China
2. University of Washington, University of Maryland
3. JD Explore Academy

Two Mysteries in Deep Neural Network:

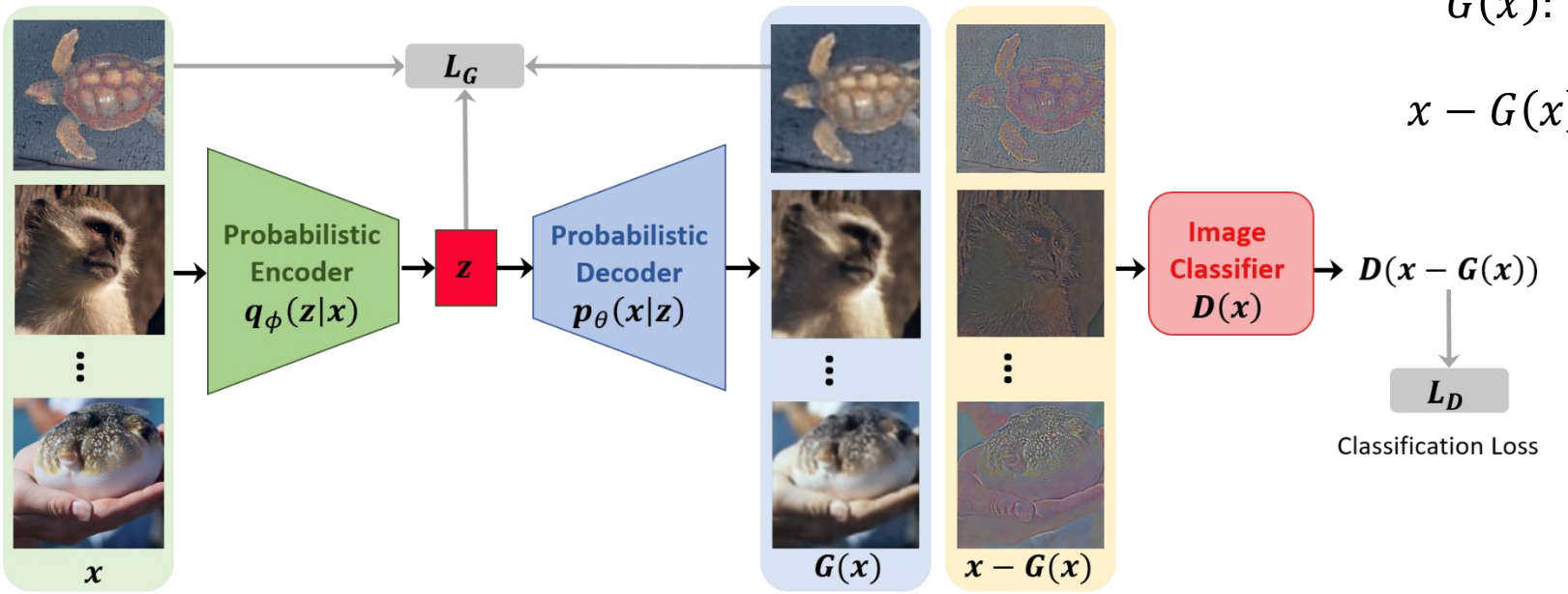
1. What essential (minimum necessary) information in the **input** do DNNs mainly rely on to make its classification predictions?



2. Disentangle the adversarial perturbation in **input** space for adversarial detection and defense.



Class Disentangled Variational Auto-Encoder (CD-VAE)



$G(x)$: Class-**redundant** part

$x - G(x)$: Class-**essential** part

$R(x) \triangleq x - G(x)$

$$\min_{\phi,\theta,\omega} \mathbb{E}_{(x,y)\sim p_{data}(x,y)} \left[L_G(\phi,\theta) + \gamma L_D(\omega) \right] \tag{1}$$

$$L_G(\phi,\theta) = -\mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) + \beta D_{KL}(q_\phi(z|x)\|p(z)), \tag{2}$$

$$L_D(\omega) = -\log D(x - G(x);\omega)[y], \tag{3}$$

Class Disentangle Results on Clean Data:

Training \ Test	x	$R(x)$	$G(x)$
x	96.01(99.84)	92.68(99.65)	18.86(67.93)
$R(x)$	95.81(99.81)	96.20(99.82)	18.12(66.30)
$G(x)$	51.84(86.52)	25.67(68.98)	75.25(97.39)

Tab 1: Training on one part of CD-VAE and test on another part: Top-1 (Top-5). $R(x) \triangleq x - G(x)$

- The classifier trained on $R(x)$ and x share similar important class information.
- $G(x)$ also contain some class (**redundant**) information.
- The classifiers trained on $R(x)$ and $G(x)$ uses different class information.

Class Disentangle Results on Clean Image $\boldsymbol{x}$ vs. Adversarial Image $\boldsymbol{x'}$:

$$\delta \triangleq x - x', \quad \delta_G \triangleq G(x) - G(x'), \quad \delta_R \triangleq R(x) - R(x')$$

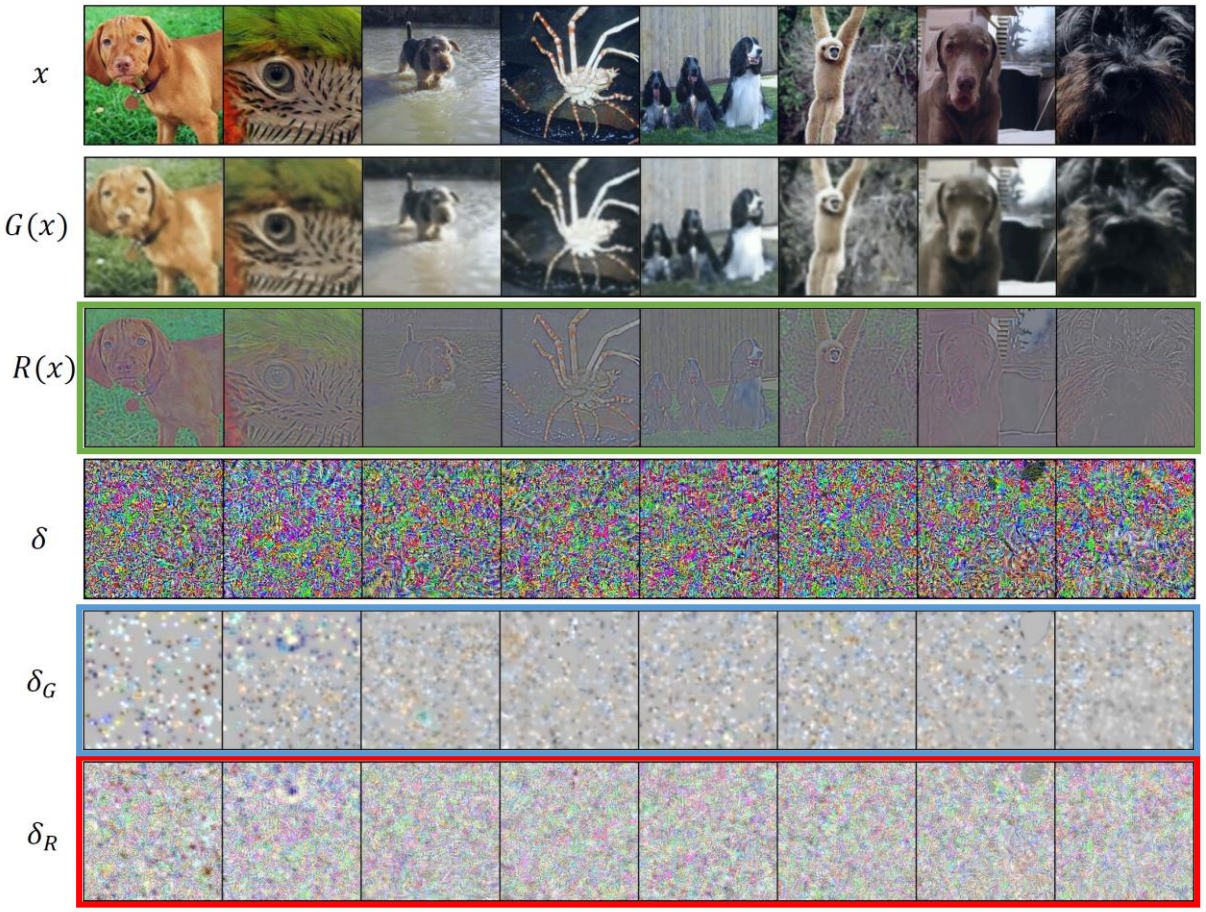


Fig. The visualization of each disentangled part.

- The adversarial perturbation **mainly lies** in the class-essential part $R(x)$.
- $G(x)$ is **not heavily distorted** by adversarial attack.
- $R(x)$ only captures **sparse and critical** regions of each image.

	$\ell_1 \times 10^{-3}$	ℓ_2	ℓ_∞
δ	13.65 ± 0.88	39.09 ± 1.40	0.14 ± 0.00
δ_G	4.15 ± 0.65	16.56 ± 2.47	0.39 ± 0.17
δ_R	13.78 ± 0.92	40.97 ± 1.87	0.48 ± 0.16

Tab. The l_p norm of each disentangled part.

Applications in Adversarial Detection and Defense

Adversarial Detection using $\boldsymbol{R(x')}$

- We found adversarial perturbation **mainly lies in** $\boldsymbol{R(x')}$.
- The **sparse** regions captured by $\boldsymbol{R(x')}$ largely narrow the search range for the attacked regions.
- Existing detection methods **use** $\boldsymbol{x}$ to detect adversarial examples against natural examples.

Adversarial Defense using $\boldsymbol{G(x')}$

- $\boldsymbol{G(x)}$ also contains some **redundant** class information.
- $\boldsymbol{G(x')}$ which is **not distorted** by adversarial attack.
- We can defend adversarial adversarial defense by using $\boldsymbol{G(x')}$ for classification.

Adversarial Detection Performance

Method	FGSM		BIM		C&W		PGD-l_∞		PGD-l_2	
	TNR	AUC	TNR	AUC	TNR	AUC	TNR	AUC	TNR	AUC
KD	42.38	85.74	74.54	94.82	73.33	94.75	73.12	94.59	70.62	93.62
KD ($R(x)$)	57.10	89.69	96.79	99.27	94.67	98.73	96.56	99.30	97.04	99.32
LID	69.05	93.60	77.73	95.20	74.98	94.32	71.52	93.19	72.57	93.46
LID ($R(x)$)	92.60	98.59	86.42	97.29	76.42	95.10	87.54	97.57	87.63	97.38
MD	94.91	98.69	88.33	97.66	86.30	97.36	77.23	95.38	76.70	95.33
MD ($R(x)$)	99.68	99.36	98.92	99.74	98.94	99.68	99.13	99.79	99.13	99.77

Table 4: TNR and AUC (%) of adversarial detection on x vs. $R(x)$ (ours) against 5 attacks (CIFAR-10)

CD-VAE can **generally improve** existing methods, simply by **replacing $\boldsymbol{x}$ with $\boldsymbol{R(x)}$**.

Adversarial Defense Performance

Dataset	Defense	Attack					
		Clean	PGD-ℓ_∞	PGD-ℓ_2	C&W-ℓ_∞	C&W-ℓ_2	StAdv
CIFAR10	Normal	96.01	0.0	0.0	0.0	0.0	0.0
	AT PGD-ℓ_∞	86.8	51.7	24.3	52.0	26.0	4.8
	TRADES ℓ_∞	84.9	55.1	28.0	53.8	28.3	9.2
	AT PGD-ℓ_2	85.0	41.9	50.1	43.4	50.6	7.8
	AT StAdv	86.2	0.1	0.3	0.2	0.5	53.9
	HGD	80.75	75.93	75.44	75.84	77.15	23.04
	APE-GAN	90.93	59.28	65.17	59.23	65.30	7.28
	Ours	86.81	77.05	78.02	77.04	78.29	19.41
ImageNet	Normal	82.53	0.0	0.0	0.0	0.0	0.0
	AT PGD-ℓ_2	69.89	10.93	60.95	9.49	60.07	0.31
	Ours	65.26	52.48	63.12	52.95	64.98	4.75

CD-VAE outperforms both adversarial training based methods and other preprocessing based methods (HGD, APE-GAN)

Towards White-box Defense: Modified Adversarial Training

$$\min_{\phi,\theta,\omega,\omega_G} \quad \mathbb{E}_{\{(x',y):D_G(G(x'))[y]-\max_{y'\neq y} D_G(G(x'))[y']\leq c\}} \left[L_G(\phi,\theta) + \gamma L_D(\omega,\omega_G)\right] \tag{8}$$

$$L_G(\phi,\theta) = -\mathbb{E}_{q_\phi(z|x')} \log p_\theta(x'|z) + \beta D_{KL}(q_\phi(z|x')\|p(z)), \tag{9}$$

$$L_D(\omega,\omega_G) = -\log D_G(G(x');\omega_G)[y] - \log D(R(x');\omega)[\operatorname*{argmax}_{y'\neq y} D_G(G(x'))[y']], \tag{10}$$

- Slighly modify L_D in the previous objective.
- Train $G(x')$ to predict the right class and $R(x')$ to predict the attacked class.
- Enforce the class-essential information mainly distorted by the attack to move to $R(x')$ instead of $G(x')$.

Robustness against White-Box Attack

Dataset	Defense	Clean	Unseen Attacks (mean)	Attack				
				ℓ_∞	ℓ_2	JPEG	ReColor	StAdv
CIFAR10	Normal	96.0	0.1	0.0	0.0	0.0	0.4	0.0
	AT PGD-ℓ_∞	86.8	27.2	49.0	19.2	30.2	54.5	4.8
	TRADES ℓ_∞	84.9	31.0	52.5	23.3	-	60.6	9.2
	AT PGD-ℓ_2	85.0	40.3	39.5	47.8	60.3	53.5	7.8
	AT ReColorAdv	93.4	7.9	8.5	3.9	19.2	65.0	0.0
	AT StAdv	86.2	1.8	0.1	0.2	1.9	5.1	53.9
	HGD	80.8	0.1	0.0	0.0	0.0	0.4	0.0
	APE-GAN	90.9	0.2	0.0	0.0	0.0	1.1	0.0
	Ours-ℓ_∞	81.2	51.4	40.5	43.1	62.1	73.1	27.4
	Ours-ℓ_2	81.0	50.4	39.4	42.4	61.6	72.2	28.4

Table 7: Defense accuracy (%) of our strategy and baselines against white-box attacks. AT-adversarial training. "Unseen Attacks (mean)" reports the defense accuracy averaged over all the attacks that are not used for adversarial training of the defense model. Ours-ℓ_∞ and Ours-ℓ_2 is trained using adversarial examples generated by C&W attacks [7] within ℓ_∞-ball of ℓ_2-ball respectively.

CD-VAE is robust to white-box attack and it can generalize well to unseen white-box attacks (the attacks not used for adversarial training). It achieves the highest unseen attacks (mean) accuracy.

Thanks!

Welcome to our poster session.

Contact me if you have any questions:

kwyang@mail.ustc.edu.cn