A Unified Contrastive Energy-based Model for Understanding the Generative Ability of Adversarial Training

ICLR 2022

Yifei Wang, Yisen Wang, Jiansheng Yang, Zhouchen Lin

Peking University

Background

- Adversarial Training (AT)
  - by far the most effective defense against adversarial attack
  - Minimax training objective

$$\min_{\theta} \mathbb{E}_{p_d(x,y)} \left[\max_{\|\hat{x}-x\|_p \le \epsilon} \ell_{CE}(\hat{x}, y; \theta) \right],$$

  - Max (inner–loop): generate worst–case adversarial example $\hat{x}$ with maximal loss
  - Min (outer–loop): update parameters on adversarial pair $(\hat{x}, y)$
- The Unexpected Bonus of Adversarial Training
  - Tsipras et al. (2018) show that robust features (by AT) align well with human perception
  - Engstrom et al. (2019) show that we can reconstruct inputs from robust representations
  - Santurkar et al. (2019) show that we can generate high–quality images from noise by targeted attack

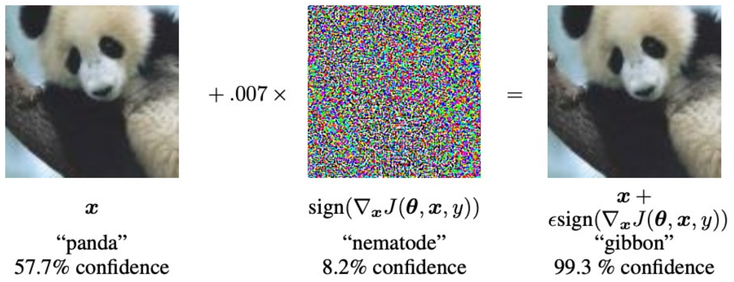

Figure from Goodfellow et al. (2015)

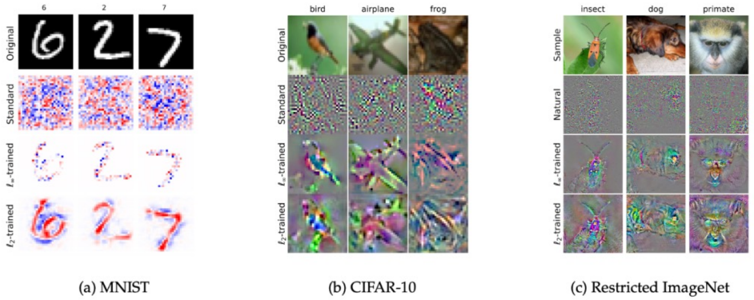

Figure from Tsipras et al. (2018)

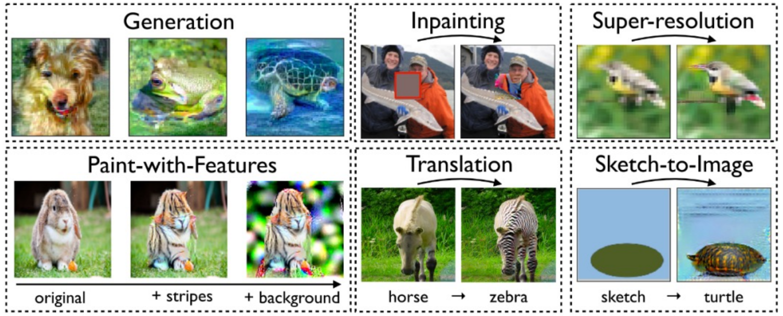

Figure from Santurkar et al. (2019)

Overview of CEM

PEKING UNIVERSITY 北京大学

- **Where does the generative ability of AT come from?**
  - There still lacks theoretical understandings of this phenomenon
- **Our work demystifies AT with a unified framework Contrastive Energy-based (CEM)**
  - General Form for two random variables (u, v)

$$p_{\boldsymbol{\theta}}(\mathbf{u},\mathbf{v}) = \frac{\exp(f_{\boldsymbol{\theta}}(\mathbf{u},\mathbf{v}))}{Z(\boldsymbol{\theta})},$$

    - where $f_\theta(u, v)$ measures the similarity between the two variables
  - Parametric CEM for supervised learning with input x and y

$$p_{\boldsymbol{\theta}}(\mathbf{x},y) = \frac{\exp(f_{\boldsymbol{\theta}}(\mathbf{x},y))}{Z(\boldsymbol{\theta})} = \frac{\exp(g_\theta(\mathbf{x})^\top \mathbf{w}_y)}{Z(\boldsymbol{\theta})},$$

    - where $g_\theta(x)$ is the encoder output, and w_k is a parameterized class cluster
    - Can be shown as a equivalence of JEM (Grathwohl et al. 2019)
  - Non–Parametric CEM for unsupervised learning with two inputs x and x'

$$p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{x}') = \frac{\exp(f_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{x}'))}{Z(\boldsymbol{\theta})} = \frac{\exp\left(g_{\boldsymbol{\theta}}(\mathbf{x})^\top g_{\boldsymbol{\theta}}(\mathbf{x}')\right)}{Z(\boldsymbol{\theta})},$$

Overview of CEM

- Where does the generative ability of AT come from?
  - There still lacks theoretical understandings of this phenomenon
- Our work demystifies AT with a unified framework **Contrastive Energy-based (CEM)**
  - General Form for two random variables (u, v)

$$p_{\boldsymbol{\theta}}(\mathbf{u}, \mathbf{v}) = \frac{\exp(f_{\boldsymbol{\theta}}(\mathbf{u}, \mathbf{v}))}{Z(\boldsymbol{\theta})},$$

    - where $f_\theta(u, v)$ measures the similarity between the two variables
  - Parametric CEM for supervised learning with input x and y

Contributions:
1. A Probabilistic Framework for Analyzing of AT
2. Unifying Supervised and Unsupervised AT as a whole, allow us to derive principled unsupervised AT
3. A principled framework for designing sampling algorithms for robust models

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \frac{\exp(f_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}'))}{Z(\boldsymbol{\theta})} = \frac{\exp\left(g_{\boldsymbol{\theta}}(\mathbf{x})^\top g_{\boldsymbol{\theta}}(\mathbf{x}')\right)}{Z(\boldsymbol{\theta})},$$

Supervised Case – Demystifying AT's generative Ability

PEKING UNIVERSITY 北京大学

- Maximization Process –
  - PGD

$$\hat{\mathbf{x}}_{n+1} = \hat{\mathbf{x}}_n + \alpha \nabla_{\hat{\mathbf{x}}_n} \ell(\hat{\mathbf{x}}_n, y; \boldsymbol{\theta}) = \hat{\mathbf{x}}_n - \alpha \nabla_{\hat{\mathbf{x}}_n} \log p_{\boldsymbol{\theta}}(y|\hat{\mathbf{x}}_n)$$
$$= \hat{\mathbf{x}}_n + \alpha \boxed{\nabla_{\hat{\mathbf{x}}_n} \left[\log \sum_{k=1}^{K} \exp(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_n, k)) \right]} - \alpha \nabla_{\hat{\mathbf{x}}_n} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_n, y),$$

  - Langevin Sampling

PGD as a (biased) sampling process

$$\hat{\mathbf{x}}_{n+1} = \hat{\mathbf{x}}_n + \alpha \nabla_{\hat{\mathbf{x}}} \log p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_n) + \sqrt{2\alpha} \cdot \boldsymbol{\varepsilon}$$
$$= \hat{\mathbf{x}}_n + \alpha \boxed{\nabla_{\hat{\mathbf{x}}_n} \left[\log \sum_{k=1}^{K} \exp(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_n, k)) \right]} + \sqrt{2\alpha} \cdot \boldsymbol{\varepsilon}.$$

- Minimization Process
  - Decomposing the likelihood gradient of CEM

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{p_d(\mathbf{x},y)} \log p_{\boldsymbol{\theta}}(\mathbf{x}, y) = \mathbb{E}_{p_d(\mathbf{x},y) \otimes p_{\boldsymbol{\theta}}(\hat{\mathbf{x}},\hat{y})} \left[\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{x}, y) - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}, \hat{y}) \right]$$
$$= \mathbb{E}_{p_d(\mathbf{x},y) \otimes p_{\boldsymbol{\theta}}(\hat{\mathbf{x}},\hat{y})} \Big[\underbrace{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{x}, y) - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}, y)}_{\text{consistency gradient}} + \boxed{\underbrace{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}, y) - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}, \hat{y})}_{\text{contrastive gradient}}} \Big].$$

Equivalent to AT loss!

Supervised Case – Demystifying AT's generative Ability

- Maximization Process –
  - PGD

$$\hat{\mathbf{x}}_{n+1} = \hat{\mathbf{x}}_n + \alpha \nabla_{\hat{\mathbf{x}}_n} \ell(\hat{\mathbf{x}}_n, y; \boldsymbol{\theta}) = \hat{\mathbf{x}}_n - \alpha \nabla_{\hat{\mathbf{x}}_n} \log p_{\boldsymbol{\theta}}(y|\hat{\mathbf{x}}_n)$$

$$= \hat{\mathbf{x}}_n + \boxed{\alpha \nabla_{\hat{\mathbf{x}}_n} \left[\log \sum_{k=1}^{K} \exp(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_n, k)) \right]} - \alpha \nabla_{\hat{\mathbf{x}}_n} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_n, y),$$

  - Langevin Sampling

$$\hat{\mathbf{x}}_{n+1} = \hat{\mathbf{x}}_n + \alpha \nabla_{\hat{\mathbf{x}}} \log p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_n) + \sqrt{2\alpha} \cdot \boldsymbol{\varepsilon}$$

$$= \hat{\mathbf{x}}_n + \boxed{\alpha \nabla_{\hat{\mathbf{x}}_n} \left[\log \sum_{k=1}^{K} \exp(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_n, k)) \right]} + \sqrt{2\alpha} \cdot \boldsymbol{\varepsilon}.$$

PGD as a (biased) sampling process

- Minimization Process
  - Decomposing the likelihood gradient of CEM

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{p_d(\mathbf{x},y)} \log p_{\boldsymbol{\theta}}(\mathbf{x}, y) = \mathbb{E}_{p_d(\mathbf{x},y) \otimes p_{\boldsymbol{\theta}}(\hat{\mathbf{x}},\hat{y})} \left[\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{x}, y) - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}, \hat{y}) \right]$$

$$= \mathbb{E}_{p_d(\mathbf{x},y) \otimes p_{\boldsymbol{\theta}}(\hat{\mathbf{x}},\hat{y})} \Big[\underbrace{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{x}, y) - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}, y)}_{\text{consistency gradient}} + \boxed{\underbrace{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}, y) - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}, \hat{y})}_{\text{contrastive gradient}}} \Big].$$

Equivalent to AT loss!

Supervised Case – Demystifying AT's generative Ability

北京大学 PEKING UNIVERSITY

- Maximization Process –
  - PGD

$$\hat{\mathbf{x}}_{n+1} = \hat{\mathbf{x}}_n + \alpha\nabla_{\hat{\mathbf{x}}_n}\ell(\hat{\mathbf{x}}_n, y; \boldsymbol{\theta}) = \hat{\mathbf{x}}_n - \alpha\nabla_{\hat{\mathbf{x}}_n}\log p_{\boldsymbol{\theta}}(y|\hat{\mathbf{x}}_n)$$

$$= \hat{\mathbf{x}}_n + \alpha\boxed{\nabla_{\hat{\mathbf{x}}_n}\left[\log\sum_{k=1}^{K}\exp(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_n, k))\right]} - \alpha\nabla_{\hat{\mathbf{x}}_n} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_n, y),$$

  - Langevin Sampling

$$\hat{\mathbf{x}}_{n+1} = \hat{\mathbf{x}}_n + \alpha\nabla_{\hat{\mathbf{x}}}\log p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_n) + \sqrt{2\alpha}\cdot\boldsymbol{\varepsilon}$$

PGD as a (biased) sampling process

AT ≈ MLE training of CEM, which explains the generative ability

$$= \mathbb{E}_{p_d(\mathbf{x},y)\otimes p_{\boldsymbol{\theta}}(\hat{\mathbf{x}},\hat{y})}\Big[\underbrace{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{x}, y) - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}, y)}_{\text{consistency gradient}} + \boxed{\underbrace{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}, y) - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}, \hat{y})}_{\text{contrastive gradient}}}\Big].$$

Equivalent to AT loss!

Image Generation Experiments

北京大学 PEKING UNIVERSITY

- Comparing different adversarial sampling algorithms
  - Our proposed algorithms achieve significant improvement over baselines
  - Comparable to state-of-the-art generative models
  - Unsupervised robust models can be almost equally good at sampling

Method	IS (↑)	FID (↓)
Auto-regressive		
PixelCNN++* (Salimans et al., 2017)	5.36	119.5
GAN-based		
DCGAN* (Radford et al., 2016)	6.69	35.6
WGAN-GP (Gulrajani et al., 2017)	7.86	36.4
PresGAN (Dieng et al., 2019)	-	52.2
StyleGAN2-ADA (Karras et al., 2020)	**10.02**	-
Score-based		
NCSN (Song & Ermon, 2019)	8.87	25.32
DDPM (Ho et al., 2020)	9.46	3.17
NCSN++ (Song et al., 2020)	9.89	**2.20**
EBM-based		
JEM (Grathwohl et al., 2019)	8.76	38.4
DRL (Gao et al., 2021)	8.58	9.60
AT-based		
TA (Santurkar et al., 2019) (w/ ResNet50)	7.5	-
Supervised CEM (w/ ResNet50)	**9.80**	55.91
Unsupervised CEM (w/ ResNet18) (ours)	8.68	**36.4**
Unsupervised CEM (w/ ResNet50) (ours)	9.61	40.25

Training	Sampling	Method	IS (↑)	FID (↓)
Supervised	Cond	TA	9.26	56.72
		Langevin	9.65	63.34
		CS	9.77	56.26
		RCS	**9.80**	**55.91**
Unsupervised (w/ ResNet18)	Uncond	PGD	5.35	74.27
		MaxEnt	**8.24**	**41.80**
	Cond	PGD	5.85	68.54
		MaxEnt	**8.68**	**36.44**
Unsupervised (w/ ResNet50)	Uncond	PGD	5.24	141.54
		MaxEnt	**9.57**	**44.86**
	Cond	PGD	5.37	137.68
		MaxEnt	**9.61**	**40.25**

Takeaways

- A probabilistic perspective of Adversarial Training helps explain its generative ability
- Unsupervised Adversarial Training loss can be developed via CEM in a principled way
- Adversarial Sampling from Robust Models can generate high-equality samples on par with state-of-the-art generative models

Thanks for Listening