

Towards Bridging the Performance Gaps of Joint Energy-based Models

Xiulong Yang, Qing Su, and Shihao Ji

Department of Computer Science
Georgia State University
Atlanta, USA



WED-PM-322

Summary of the paper

- Joint Energy-based Model (JEM) trains one single model for image classification and image generation.
- However, there remain two performance gaps
 - Classification accuracy gap
 - Image generation quality gap
- We introduce SADA-JEM to bridge both gaps
 - Extends Sharpness-Aware Minimization (SAM) to train JEM
 - Excludes data augmentation from the MLE pipeline of EBM
- Performance of SADA-JEM
 - Closed substantial performance gaps of JEM in **image classification** and **image generation**;
 - Outperforms JEM in **calibration**, **OOD detection** and **adversarial robustness** by a notable margin.

Outline

- Background
- Motivations
- Methodology
- Experimental Results

Background

- EBM stems from the observation that any pdf $p_{\theta}(\mathbf{x})$ can be expressed via a Boltzmann dist. as

$$p_{\theta}(\mathbf{x}) = \frac{\exp(-\boxed{E_{\theta}(\mathbf{x})})}{Z(\theta)}$$

energy function

- MLE training of parameter θ

$$\frac{\partial \log p_{\theta}(\mathbf{x})}{\partial \theta} = \mathbb{E}_{\underline{p_{\theta}(\mathbf{x})}} \left[\frac{\partial E_{\theta}(\mathbf{x})}{\partial \theta} \right] - \mathbb{E}_{\underline{p_d(\mathbf{x})}} \left[\frac{\partial E_{\theta}(\mathbf{x})}{\partial \theta} \right]$$

Samples from $p_{\theta}(\mathbf{x})$ Training set

- SGLD sampling

$$\mathbf{x}^0 \sim p_0(\mathbf{x}), \quad \text{Image generator}$$
$$\mathbf{x}^{t+1} = \mathbf{x}^t - \frac{\alpha}{2} \frac{\partial E_{\theta}(\mathbf{x}^t)}{\partial \mathbf{x}^t} + \alpha \epsilon^t, \quad \epsilon^t \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$$

Background

- JEM [Grathwohl et al. 2019] reinterpreted the standard softmax classifier as an EBM.

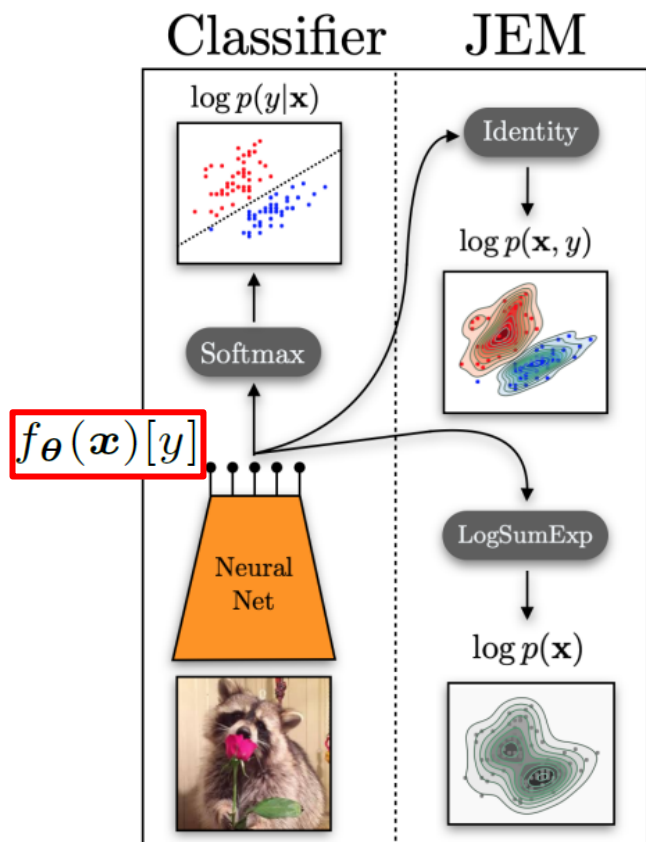


Image from [Grathwohl et al. 2019]

- Maximizes the log of joint density function

$$\log p_{\theta}(\mathbf{x}, y) = \underbrace{\log p_{\theta}(y|\mathbf{x})}_{\text{Cross-entropy for classification}} + \underbrace{\log p_{\theta}(\mathbf{x})}_{\text{MLE training of EBM}}$$

Cross-entropy
for classification

MLE training
of EBM

$$E_{\theta}(\mathbf{x}) = -\log \sum_y e^{f_{\theta}(\mathbf{x})[y]} = -\text{LSE}(f_{\theta}(\mathbf{x}))$$

$$p_{\theta}(\mathbf{x}) = \frac{\exp(-E_{\theta}(\mathbf{x}))}{Z(\theta)}$$

Motivations

- Two performance gaps of JEM [Grathwohl et al. 2019, Yang et al. 2021]
 - Classification accuracy gap
 - Image generation quality gap
- Hypothesis
 - Both performance gaps are the symptoms of lack of generalization of JEM trained models.

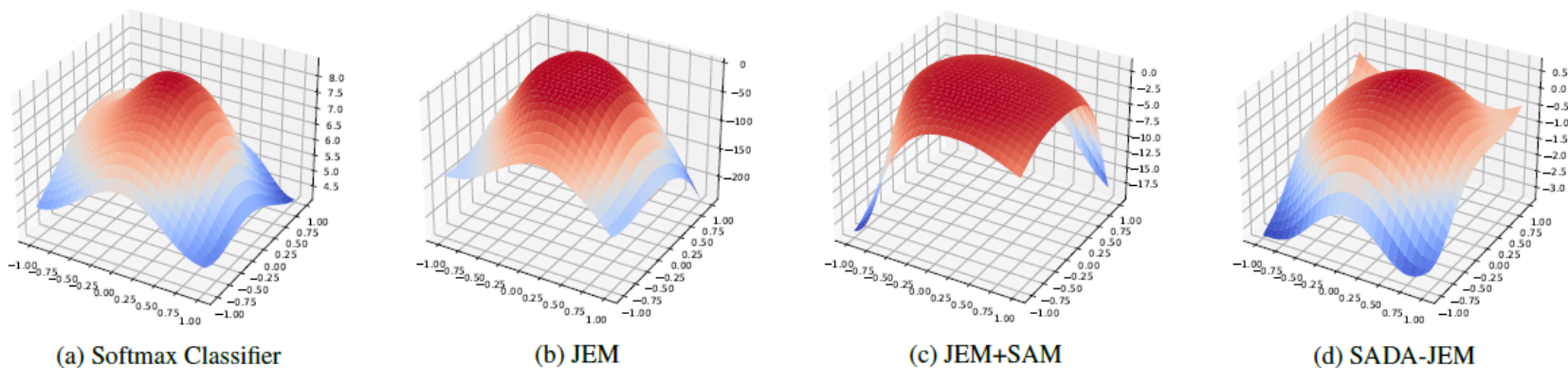


Figure 1. Visualizing the energy landscapes [34] of different models trained on CIFAR10. Note the dramatic scale differences of the y-axes, indicating SADA-JEM identifies the smoothest local optimum among all the methods considered.

Method: SADA-JEM

- Sharpness-Aware Minimization (SAM) [Foret et al. 2021]
 - Searches for model parameters θ whose entire neighborhoods have uniformly low loss values

$$\min_{\theta} \max_{\|\epsilon\|_2 \leq \rho} L_{train}(\theta + \epsilon) + \lambda \|\theta\|_2^2.$$

- Extension to optimize JEM


$$\max_{\theta} \min_{\|\epsilon\|_2 \leq \rho} \log p_{(\theta+\epsilon)}(\mathbf{x}, y) + \lambda \|\theta\|_2^2$$

Method: SADA-JEM

□ Image Generation w/o Data Augmentation

- The actual objective function of JEM with Data Augmentation

$$\log p_{\theta}(\mathbf{x}, y) = \log p_{\theta}(y|T(\mathbf{x})) + \log p_{\theta}(T(\mathbf{x}))$$


$$\log p_{\theta}(\mathbf{x}, y) = \log p_{\theta}(y|T(\mathbf{x})) + \log p_{\theta}(\mathbf{x})$$

This can be implemented efficiently by using two data loaders.

Experiments

□ Hybrid Modeling

Table 1. Results on CIFAR10

Model	Acc % \uparrow	IS \uparrow	FID \downarrow
SADA-JEM (K=5)	95.5	8.77	9.41
SADA-JEM (K=10)	<u>96.0</u>	8.63	<u>11.4</u>
SADA-JEM (K=20)	96.1	8.40	13.1
Single Hybrid Model			
IGEBM (K=60) [10]	49.1	8.30	37.9
JEM (K=20)* [17]	92.9	8.76	38.4
JEM++ (M=5)* [48]	91.1	7.81	37.9
JEM++ (M=10) [48]	<u>93.5</u>	8.29	<u>37.1</u>
JEM++ (M=20) [48]	94.1	8.11	38.0
JEAT [51]	85.2	8.80	38.2
Other EBMs			
CF-EBM (K=50) [50]	-	-	16.7
ImCD (K=40) [9]	-	7.85	25.1
DiffuRecov (K=30) [13]	-	8.31	9.58
VAEBM (K=6) [47]	-	8.43	12.2
VERA [18]	93.2	8.11	30.5
Other Models			
Softmax	<u>96.2</u>	-	-
Softmax + SAM	<u>97.2</u>	-	-
SNGAN [37]	-	8.59	21.7
StyleGAN2-ADA [28]	-	9.74	<u>2.92</u>

* The training is unstable and regularly diverged.

Table 2. Results on CIFAR100

Model	Acc % \uparrow	IS \uparrow	FID \downarrow
SADA-JEM (K=5)	75.0	11.63	14.4
SADA-JEM (K=10)	<u>76.4</u>	10.95	<u>15.1</u>
SADA-JEM (K=20)	77.3	10.78	19.9
Single Hybrid Model			
JEM (K=20)* [17]	72.2	10.22	38.1
JEM++ (M=5)* [48]	72.1	8.05	38.9
JEM++ (M=10)* [48]	<u>74.2</u>	9.97	<u>34.5</u>
JEM++ (M=20)* [48]	75.9	10.07	33.7
VERA ($\alpha=100$)* [18]	72.2	8.25	29.5
VERA ($\alpha=1$)* [18]	48.7	7.84	25.1
Softmax	<u>81.3</u>	-	-
Softmax + SAM	83.4	-	-
SNGAN [37]	-	9.30	15.6
BigGAN [4]	-	11.0	11.7

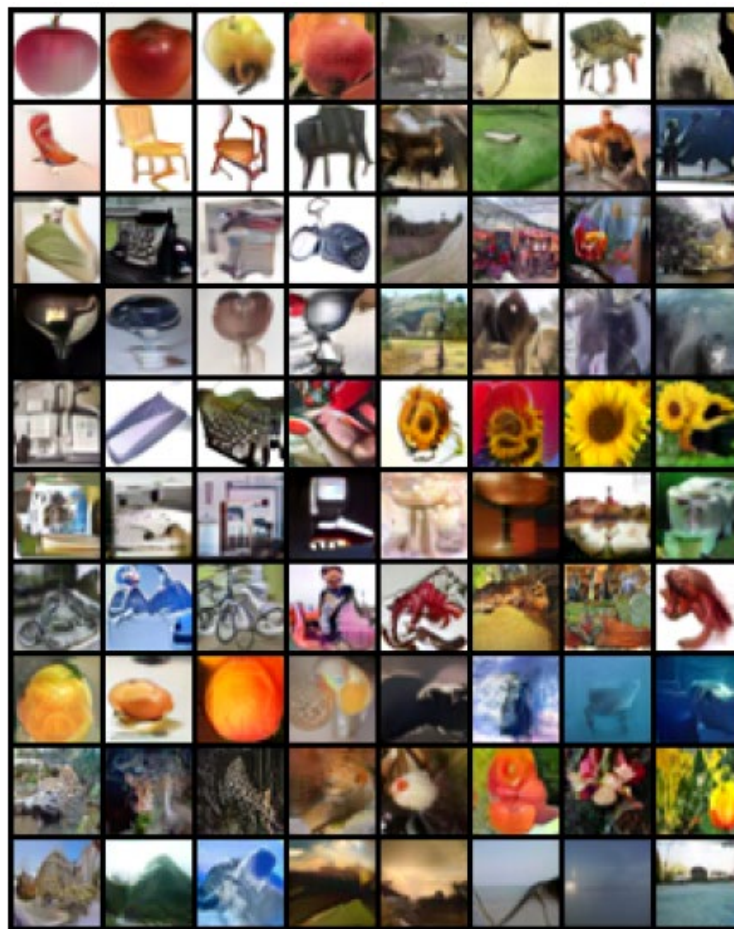
* No official IS and FID scores are reported. We run the official code with the default settings and report the results.

Experiments

- Generated samples from SADA-JEM



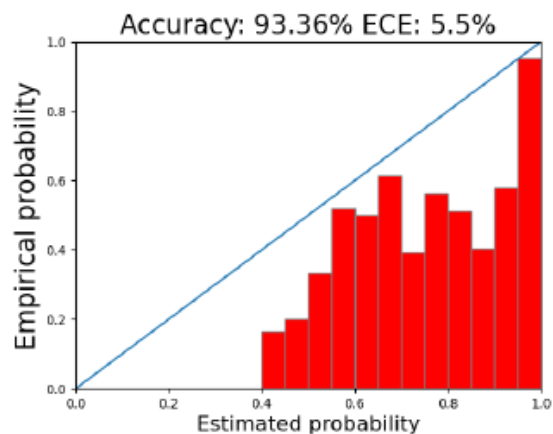
(a) CIFAR10



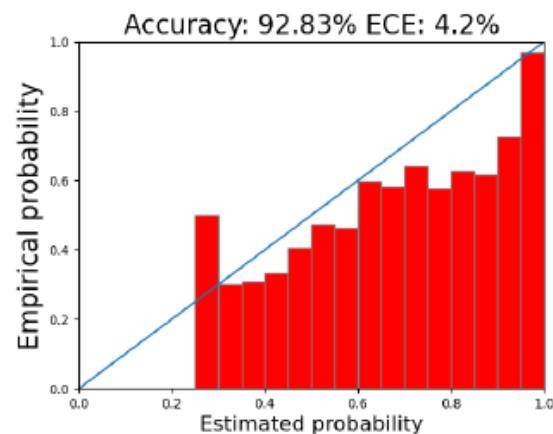
(b) CIFAR100

Experiments

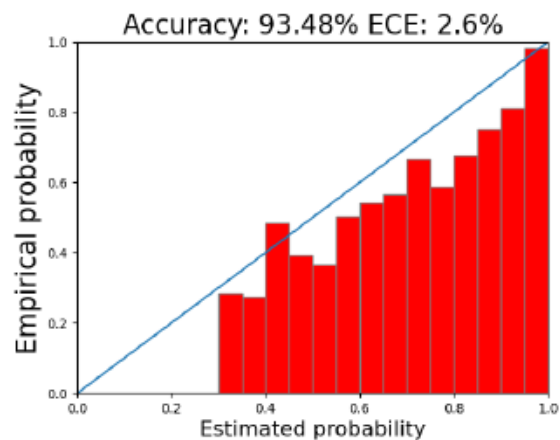
□ Calibration



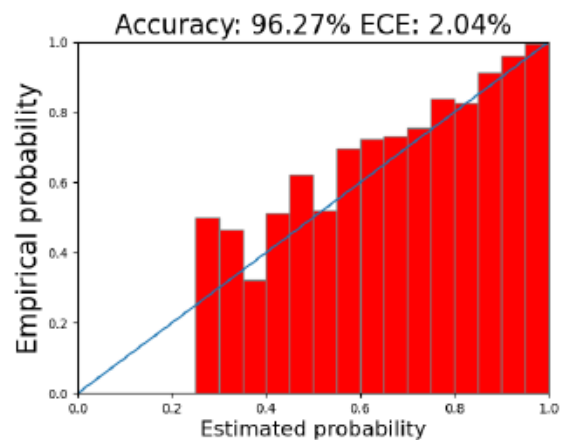
(a) Softmax (w/o BN)



(b) JEM (K=20)



(c) JEM++ (M=10)



(d) SADA-JEM (K=10)

Experiments

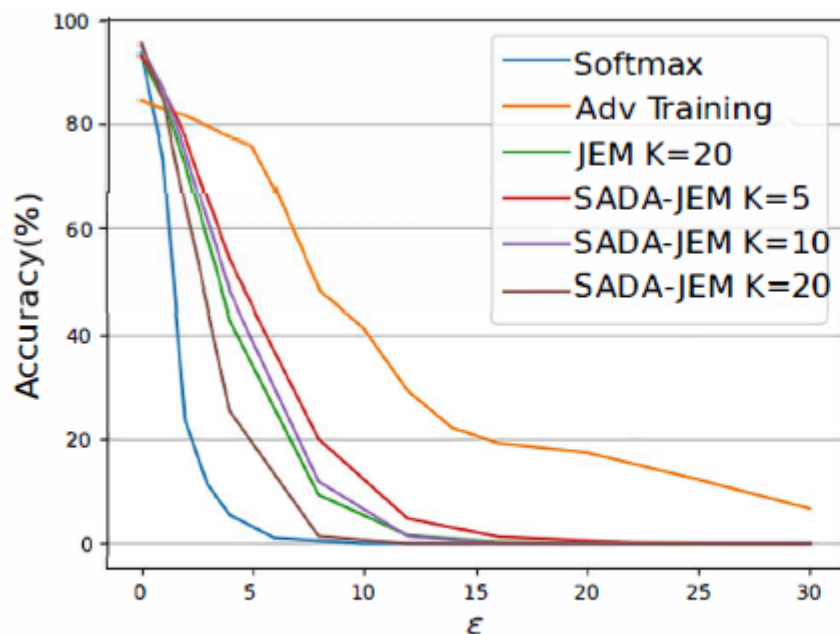
□ OOD Detection

Table 3. OOD detection results. Models are trained on CIFAR10. Values are AUROC.

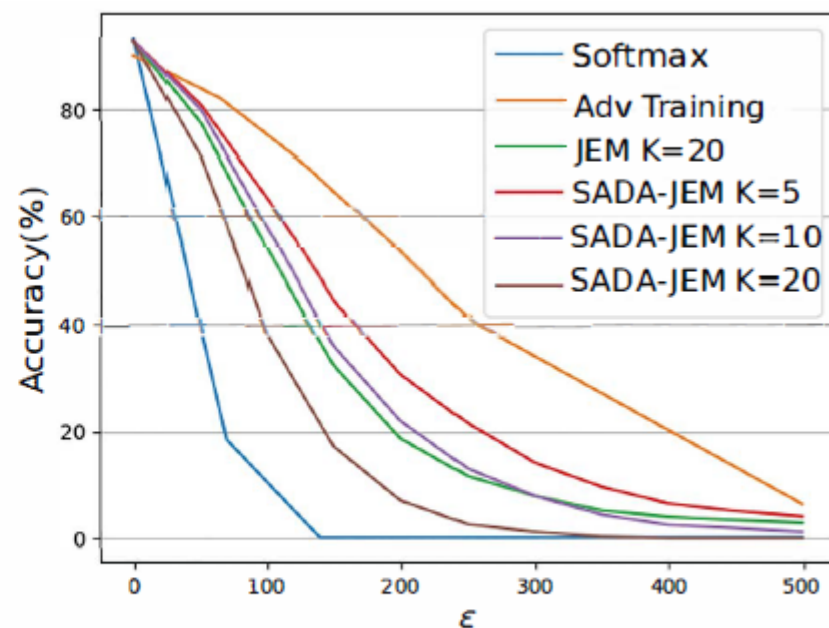
$s_{\theta}(x)$	Model	SVHN	CIFAR10 Interp	CIFAR100	CelebA
$\log p_{\theta}(x)$	WideResNet [35]	.91	-	.87	.78
	IGE BM [10]	.63	.70	.50	.70
	JEM (K=20) [17]	.67	.65	.67	.75
	JEM++ (M=20) [48]	.85	.57	.68	.80
	VERA [18]	.83	.86	.73	.33
	ImCD [9]	.91	.65	.83	-
	SADA-JEM (K=5)	.91	.79	.90	.82
	SADA-JEM (K=10)	.95	.81	.90	.88
	SADA-JEM (K=20)	.98	.83	.92	.95
$\max_y p_{\theta}(y x)$	WideResNet	.93	.77	.85	.62
	IGE BM [10]	.43	.69	.54	.69
	JEM (K=20) [17]	.89	.75	.87	.79
	JEM++ (M=20) [48]	.94	.77	.88	.90
	SADA-JEM (K=5)	.92	.77	.88	.81
	SADA-JEM (K=10)	.93	.78	.89	.78
	SADA-JEM (K=20)	.96	.80	.91	.84

Experiments

- Adversarial Robustness under PGD attack



(a) L_∞ Robustness



(b) L_2 Robustness

Experiments

□ Ablation Study

Ablation	Acc% \uparrow	FID \downarrow
JEM	89.5	36.2
JEM +SAM	90.1	35.0
JEM++	93.5	37.1
JEM++ +SAM	94.1	36.6
JEM++ w/o DA	93.6	12.9
JEM++ w/o DA + L_2^*	93.4	-
SADA-JEM	96.0	11.4

Conclusion

- We introduce SADA-JEM to bridge the classification accuracy gap and the generation quality gap of JEM.
- By incorporating the framework of SAM to JEM and excluding the undesirable data augmentation from the training pipeline of JEM, SADA-JEM promotes the energy landscape smoothness and hence the generalization of trained models.
- Our experiments verify the effectiveness of these techniques and demonstrate the state-of-the-art results in most of the tasks of image classification, generation, calibration, OOD detection and adversarial robustness.
- Future works
 - Computation bottleneck is not SAM (2x) but SGLD (Kx)
 - EBM for large-scale benchmarks with high resolution images, such as ImageNet

Thank You!



<https://github.com/sndnyang/sadajem>

Poster: WED-PM-322