

Your Classifier is Secretly an Energy Based model and You Should Treat it Like One

Сабина Даянова
Тимур Ваньков
Артем Цыганов
Никита Башаев

План

- Мотивация
- Energy Based Models
- Joint Energy-based Model
- Приложения

Модели

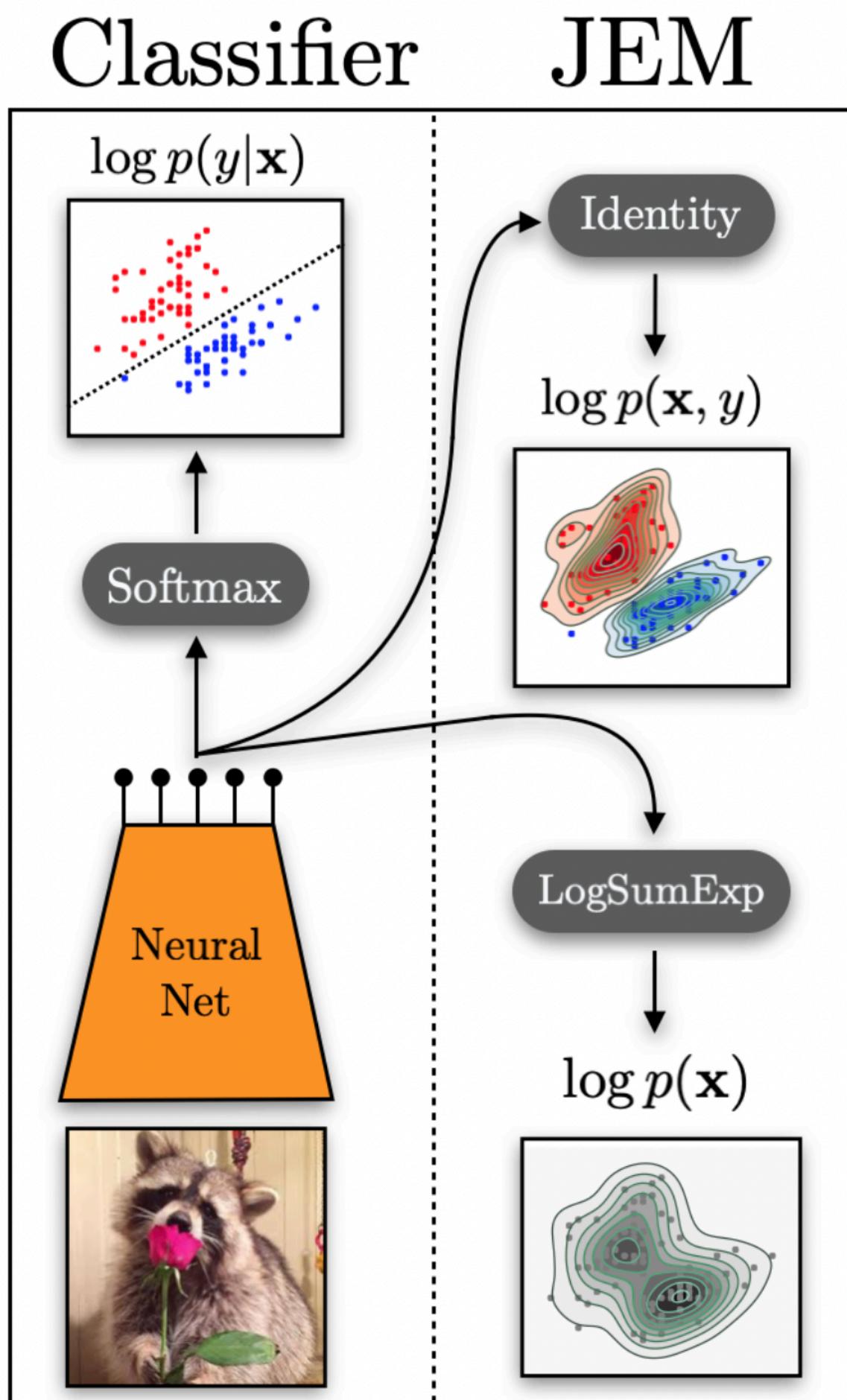
Генеративные

Дискриминативные

- оценивают $P_{\theta}(X, Y)$
- проигрывают дискриминаторам в классификации
- обладают полезными качествами: calibration, robustness

- оценивают $P_{\theta}(Y|X)$
- лучше генмоделей в классификации
- мало полезных качеств

Идея: сделать из
логитов $P(X, Y)$



Energy based models

cringe

- $E_\theta(x) : \mathbb{R}^D \rightarrow \mathbb{R}$ - функция энергии
- $p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z(\theta)}$ - связь плотности с энергией
- $Z(\theta) = \int_x \exp(-E_\theta(x))$ - нормализующая константа, невозможно посчитать!!!

What your classifier is hiding

- $p_{\theta}(y | x) = \frac{\exp(f_{\theta}(x)[y])}{\sum_{y'} \exp(f_{\theta}(x)[y'])}$ - softmax
- $p_{\theta}(x, y) = \frac{\exp(-E_{\theta}(x, y))}{Z(\theta)} \Rightarrow p_{\theta}(x, y) = \frac{\exp(f_{\theta}(x)[y])}{Z(\theta)}$
реинтерпретируем логиты, $E_{\theta}(x, y) = -f_{\theta}(x)[y]$
- $p_{\theta}(x) = \sum_y p_{\theta}(x, y) = \frac{\sum_y \exp(f_{\theta}(x)[y])}{Z(\theta)}, E_{\theta}(x) = -\text{LogSumExp}_y(f_{\theta}(x)[y])$
- $p_{\theta}(y | x) = \frac{p_{\theta}(x, y)}{p_{\theta}(x)}$
- Joint Energy-based Model

Optimization

- $p_{\theta}(x, y) = p_{\theta}(y | x) p_{\theta}(x)$
- $\log p_{\theta}(x, y) = \log p_{\theta}(y | x) + \log p_{\theta}(x)$ - хорошая факторизация
- $\log p_{\theta}(y | x)$ обучаем стандартно через кросс-энтропию
- что делать с $\log p_{\theta}(x)$?

- $\frac{\partial \log p_\theta(x)}{\partial \theta} = M_{p_\theta(x')} \left[\frac{\partial E_\theta(x')}{\partial \theta} \right] - \frac{\partial E_\theta(x)}{\partial \theta}$ - страшно
- contrastive divergence -

$$L(\theta) = \text{LogSumExp}_{y'}(f(x)[y']) - \text{LogSumExp}_{y'}(f(\hat{x}_t)[y']) = -E(x) + E(\hat{x}_t)$$
- $\hat{x}_0 \sim p_0(x), \hat{x}_{i+1} = \hat{x}_i - \frac{\alpha}{2} \frac{\partial E_\theta(\hat{x}_i)}{\partial \hat{x}_i} + \varepsilon, \varepsilon \sim \mathcal{N}(0, \alpha)$ - Stochastic Gradient Langevin Dynamics

Algorithm 1 JEM training: Given network f_θ , SGLD step-size α , SGLD noise σ , replay buffer B , SGLD steps η , reinitialization frequency ρ

- 1: **while** not converged **do**
- 2: Sample \mathbf{x} and y from dataset
- 3: $L_{\text{clf}}(\theta) = \text{xent}(f_\theta(\mathbf{x}), y)$
- 4: Sample $\hat{\mathbf{x}}_0 \sim B$ with probability $1 - \rho$, else $\hat{\mathbf{x}}_0 \sim \mathcal{U}(-1, 1)$ ▷ Initialize SGLD
- 5: **for** $t \in [1, 2, \dots, \eta]$ **do** ▷ SGLD
- 6: $\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_{t-1} + \alpha \cdot \frac{\partial \text{LogSumExp}_{y'}(f_\theta(\hat{\mathbf{x}}_{t-1})[y'])}{\partial \hat{\mathbf{x}}_{t-1}} + \sigma \cdot \mathcal{N}(0, I)$
- 7: **end for**
- 8: $L_{\text{gen}}(\theta) = \text{LogSumExp}_{y'}(f(\mathbf{x})[y']) - \text{LogSumExp}_{y'}(f(\hat{\mathbf{x}}_t)[y'])$ ▷ Surrogate for Eq 2
- 9: $L(\theta) = L_{\text{clf}}(\theta) + L_{\text{gen}}(\theta)$
- 10: Obtain gradients $\frac{\partial L(\theta)}{\partial \theta}$ for training
- 11: Add $\hat{\mathbf{x}}_t$ to B
- 12: **end while**

Applications

SVHN



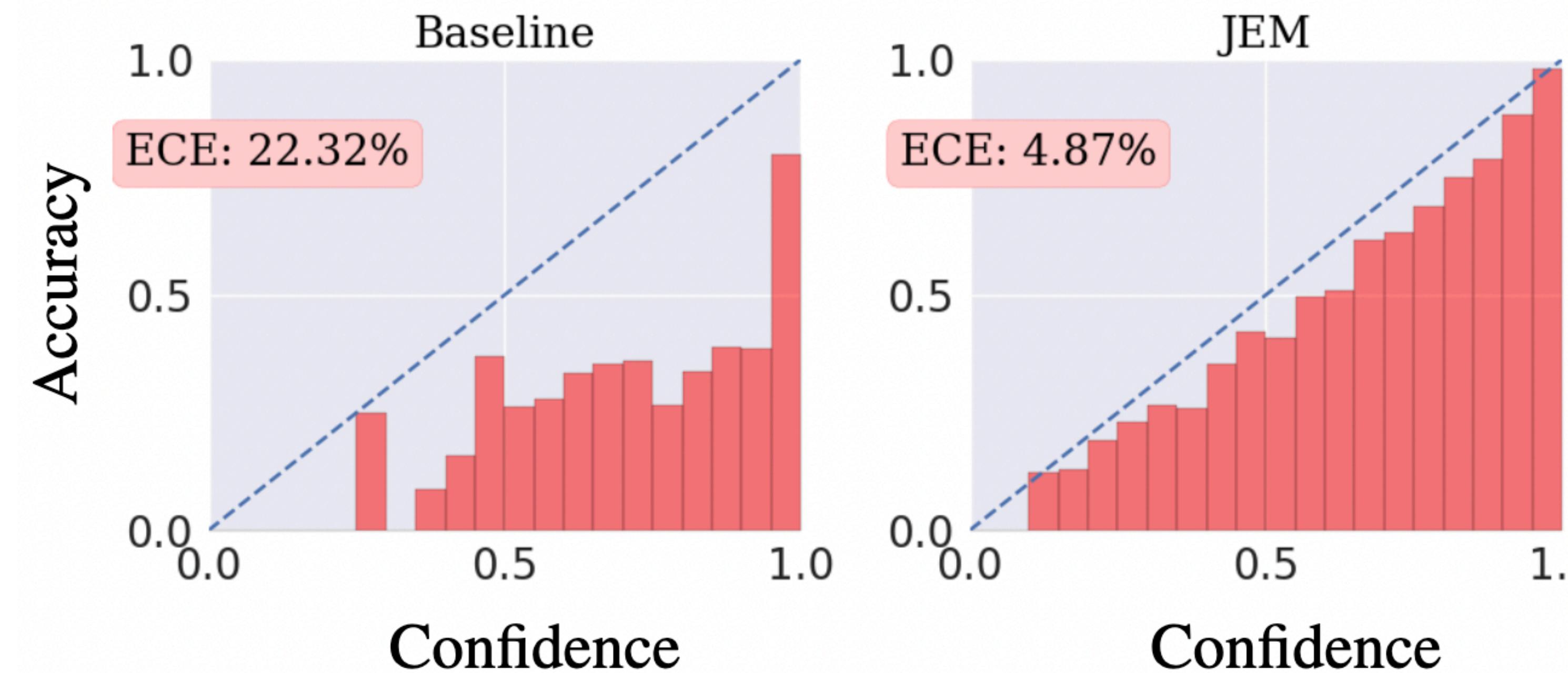
Class	Model	Accuracy % ↑	IS↑	FID↓
Hybrid	Residual Flow	70.3	3.6	46.4
	Glow	67.6	3.92	48.9
	IGEBM	49.1	8.3	37.9
	JEM $p(\mathbf{x} y)$ factored	30.1	6.36	61.8
	JEM (Ours)	92.9	8.76	38.4
Disc.	Wide-Resnet	95.8	N/A	N/A
Gen.	SNGAN	N/A	8.59	25.5
	NCSN	N/A	8.91	25.32

CIFAR100



Calibration

Калиброванность: уверенность модели в ответе близка к Accuracy



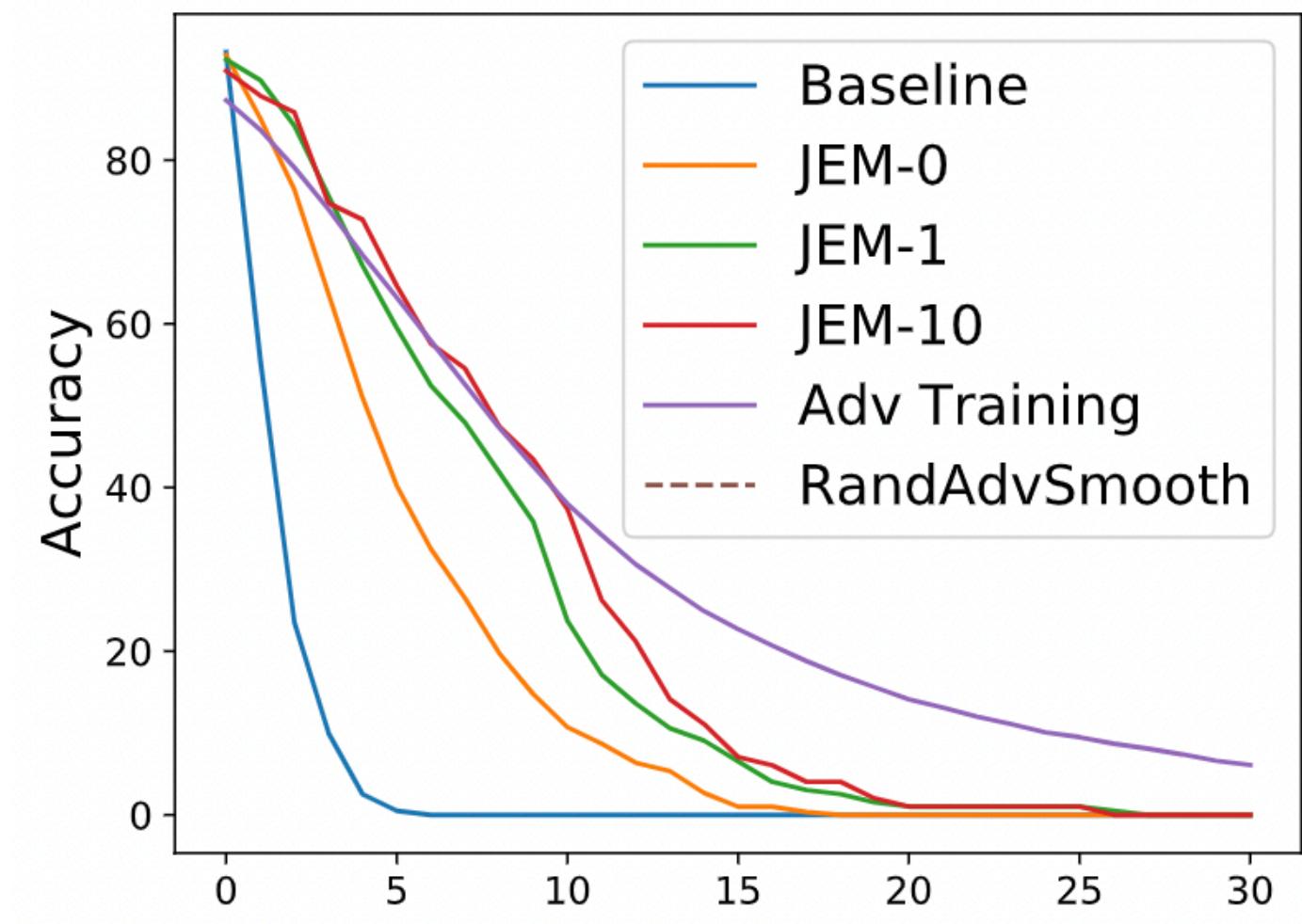
Robustness

- $\tilde{x} = x + \delta$, $||\tilde{x} - x||_p < \epsilon$
- JEM-K \Leftrightarrow делаем K итераций SGLD
- двигаем мало-вероятную точку к самой близкой более вероятной

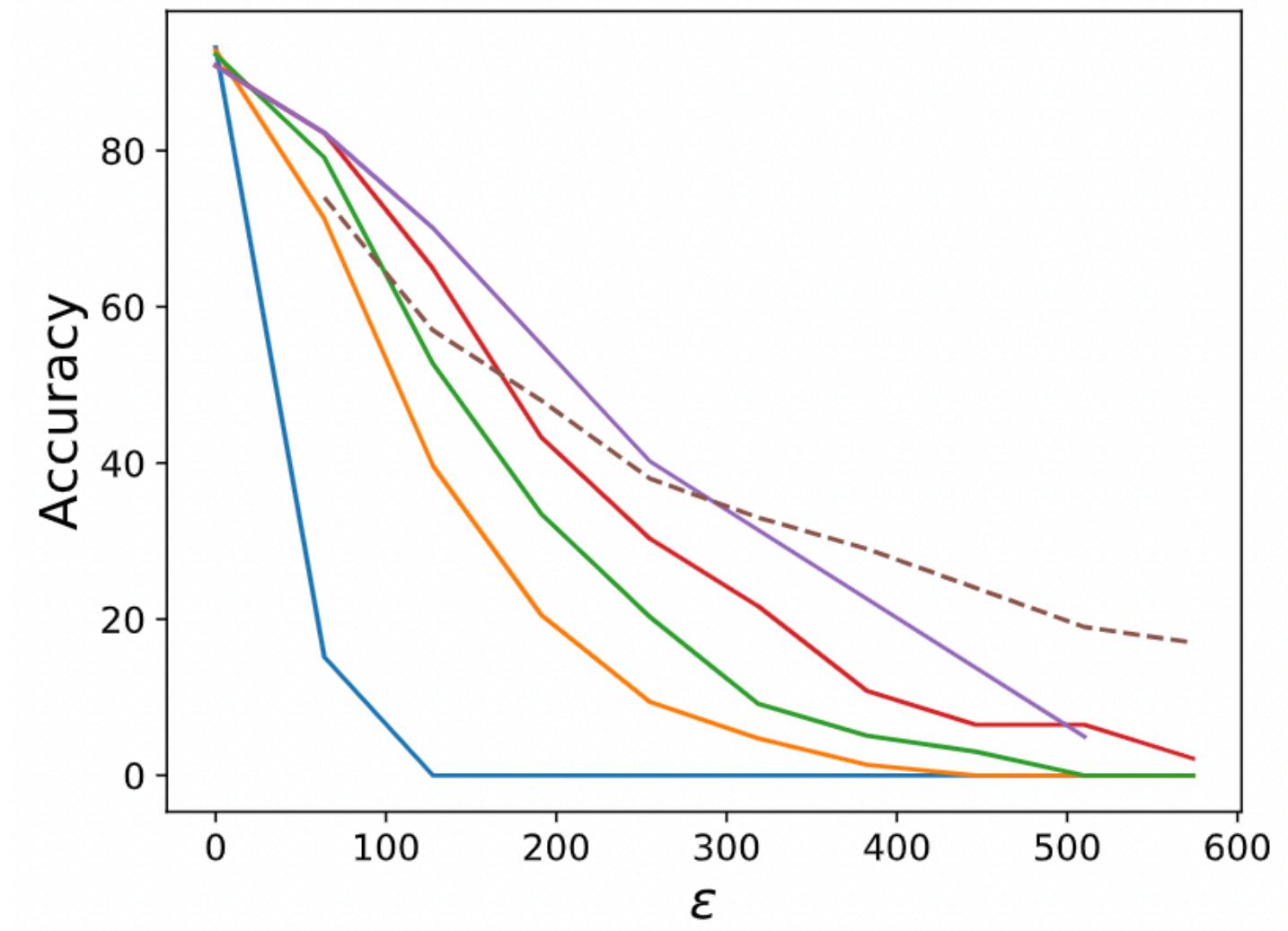
```

for  $t \in [1, 2, \dots, \eta]$  do
     $\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_{t-1} + \alpha \cdot \frac{\partial \text{LogSumExp}_{y'}(f_\theta(\hat{\mathbf{x}}_{t-1})[y'])}{\partial \hat{\mathbf{x}}_{t-1}} + \sigma \cdot \mathcal{N}(0, I)$ 
end for

```



(a) L_∞ Robustness



(b) L_2 Robustness

Conclusion

- создали модель, сохранившую сильные стороны дискриминатора, добавив плюшки от генеративных моделей
- модель откалибрована и робастна
- оценки градиента в обучении нестабильны
- тщательно подбирали гиперпараметры, модель разваливается при других

Рецензент

Сильные стороны:

1. Новизна идеи
2. Понятность статьи
3. Прикреплены понятные результаты экспериментов

Слабые стороны:

1. Отсутствие экспериментов на устойчивость моделей
2. Не описан алгоритм генерации изображений

Оценки по критериям НИПСа:

1. Оценка: 8 из 10
2. Уверенность: 3 из 5

Практик-исследователь

Кто авторы?

- Трое из авторов работают в Google Research
- Четверо - University of Toronto & Vector Institute
- Наибольшее число цитирований - Mohammad Narouzi, более 17 тысяч раз
- Как минимум у 4 авторов есть публикации по теме медицины

История публикаций

- Представлена на ICLR 2020 Conference в качестве Oral Paper
- Есть вторая версия статьи, опубликованная уже после конференции

Повлиявшие статьи

- Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. 2006 - идея Energy Based Models
- Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. 2019 - рассматривается схожая идея, но другое распределение. С данной статьей проводится большое количество сравнений, даже описываются некоторые недостатки

Future work

- 170 цитирований
- В одной из статей - эксперименты по устойчивости к разным типам атак
- В продолжение данной статьи, авторы опубликовали работу Learning the Stein Discrepancy for Training and Evaluating Energy-Based Models without Sampling - направлена на улучшение процесса обучения Energy Based моделей