Your Classifier is Secretly and Energy Based Model and You Should Treat it Like one

НИС «Машинное обучение и приложения»

Котов Егор, БПМИ172

Зачем нам нужны генеративные модели?

- discover latent structure in high-dimensional data
- can leverage unlabeled data
- 1 approach to solve many problems

Задачи генеративных моделей

task	method
• out-of-distribution detection	• reject low $p(x)$ inputs
• robust classification	• given x find $\hat{x}$ s.t. $p(\hat{x})$ high classify $p(y \mid \hat{x})$
• semi-supervised learning	• train $p(x, y)$, $p(x)$ when no label

Почему генеративные модели иногда работают плохо?

- current generative models are not flexible enough to fit data
- architectures for generative models have diverged considerably from discriminative models
- generative model architectures are worse at discriminative tasks

Энергетические модели (Energy based models)

- an energy-based model (EBM) parameterizes a density using its unnormalized log-density function

$$p_\theta(x) = \frac{e^{-E_\theta(x)}}{Z(\theta)} \qquad Z(\theta) = \int_x e^{-E_\theta(x)} dx$$

- where $E_\theta : \mathbf{R}^D \to \mathbf{R}$

Can be *almost* any function

easy to incorporate known structure

inctractable to compute or even estimate

cannot efficeintly compute likelihoods or draw samples

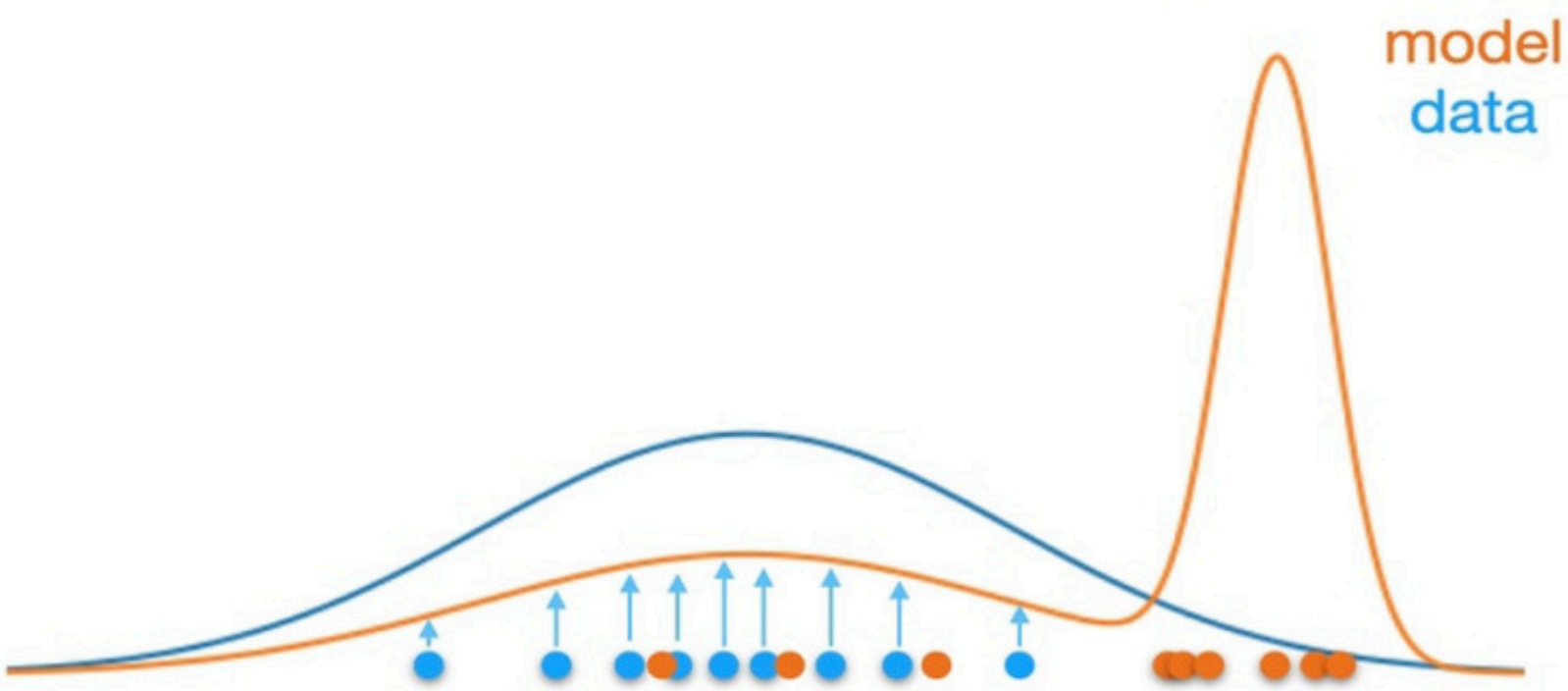
Обучение EBMs

- given we cannot compute likelihoods we cannot train to maximize likelihood
- we must be a bit more clever

Правдоподобие

- while $\log p_\theta(x)$ does not have a nice form, $\dfrac{\partial \log p_\theta(x)}{\partial \theta}$ can be written more simply as

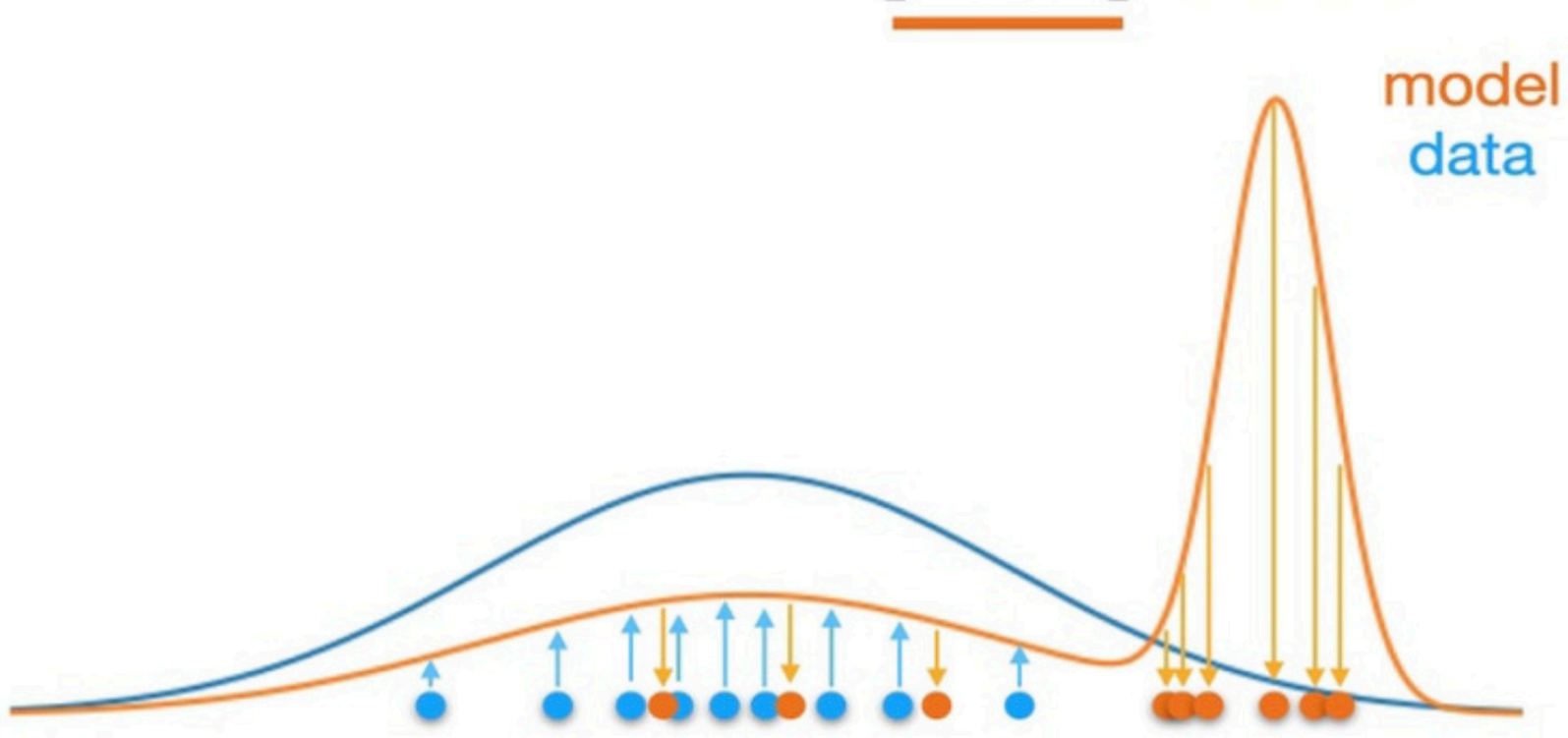
$$\frac{\partial \log p_\theta(x)}{\partial \theta} = \mathbf{E}_{p_\theta(x')}\left[\frac{\partial E_\theta(x)}{\partial \theta}\right] - \frac{\partial E_\theta(x)}{\partial \theta}$$



Правдоподобие

- while $\log p_\theta(x)$ does not have a nice form, $\dfrac{\partial \log p_\theta(x)}{\partial \theta}$ can be written more simply as

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

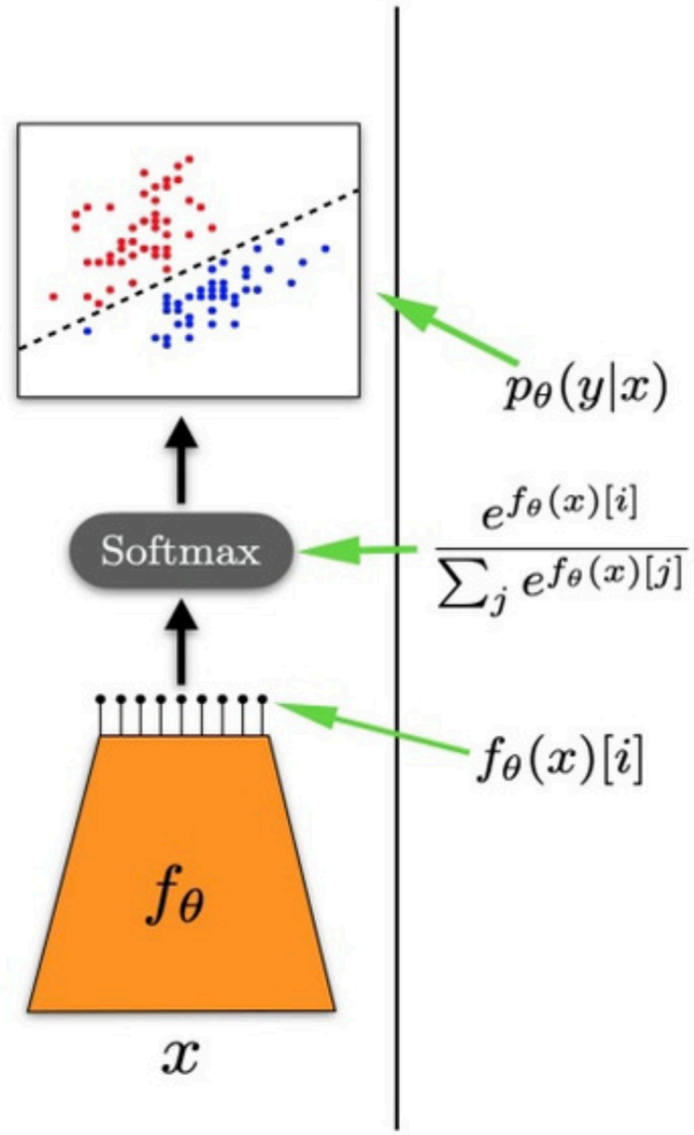


Архитектура в классификации

- classification tasks solved by modeling $p(y \mid x)$ and maximizing likelihood

- we do this with a function $f_\theta : \mathbf{R}^D \to \mathbf{R}^K$

- we pass these K outputs through a softmax function to obtain $p(y \mid x)$



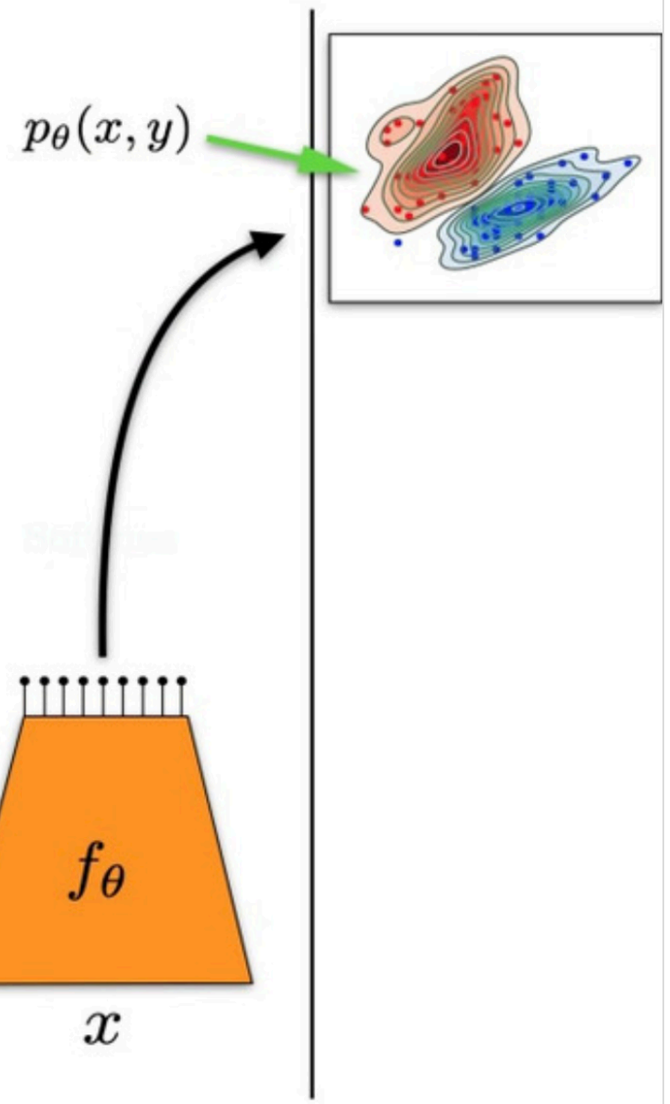
Архитектура в классификации

- lets redefine what these outputs mean

$$p_\theta(x, y) = \frac{e^{f_\theta(x)[y]}}{Z(\theta)}$$

- this is an EBM with energy

$$E_\theta(x, y) = -f_\theta(x)[y]$$



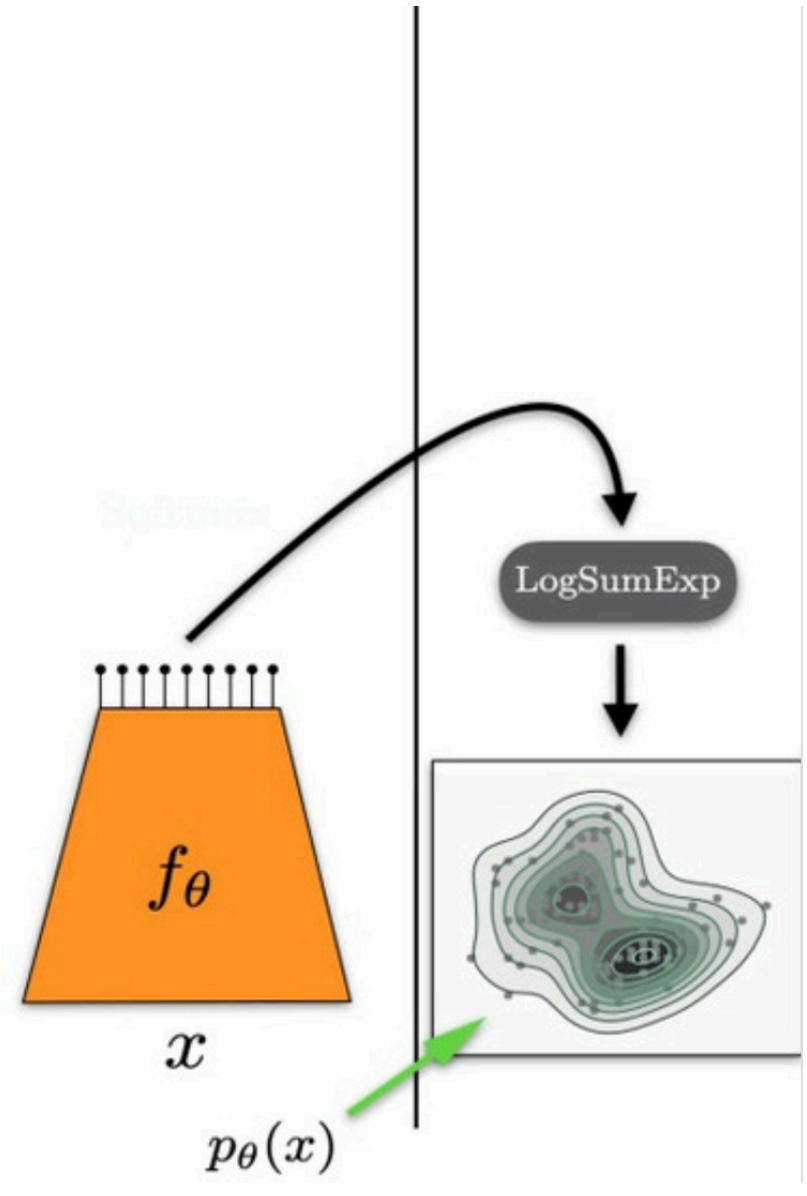
Архитектура в классификации

- summing out y we obtain

$$p_\theta(x) = \sum_y p_\theta(x, y) = \frac{\sum_y e^{f_\theta(x)[y]}}{Z(\theta)}$$

- which is an EBM with energy

$$E_\theta(x) = -\,\text{LogSumExp}_y(f_\theta(x)[y])$$

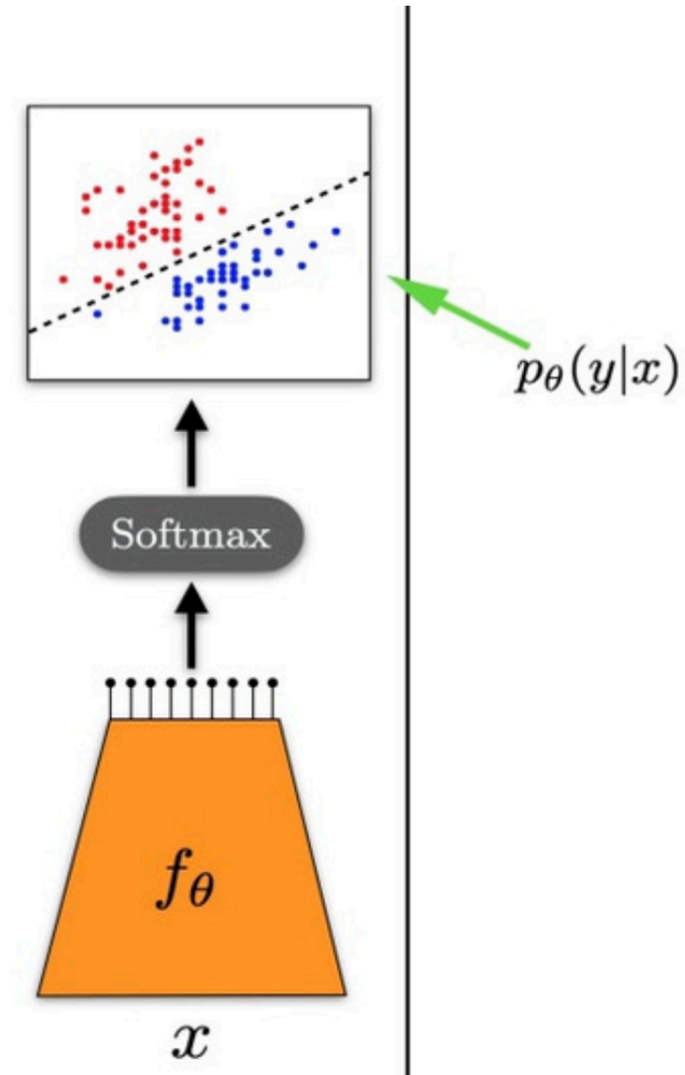


Архитектура в классификации

- when we compute $p_\theta(y \mid x) = \dfrac{p_\theta(x, y)}{p_\theta(x)}$ we get

$$p_\theta(y \mid x) = \frac{e^{f_\theta(x)[y]}}{\sum_{y'} e^{f_\theta(x)[y']}}$$

standard softmax!!



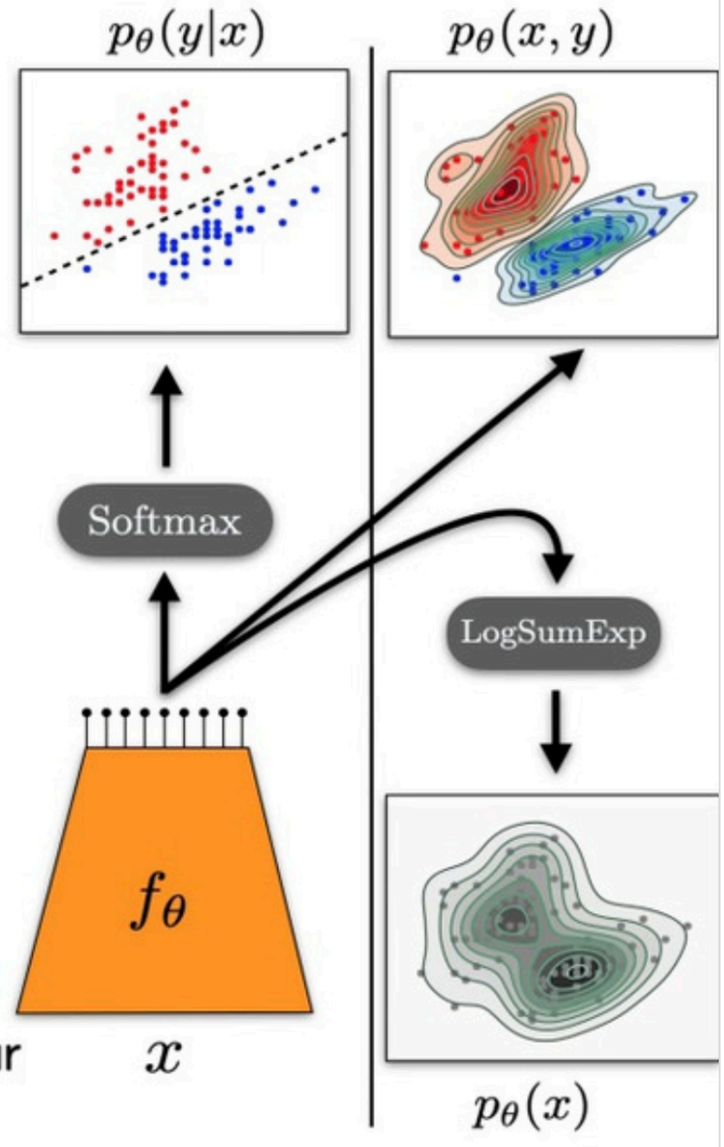
Архитектура в классификации

- without changing our architecture we have found models of

$$p_\theta(x, y) = \frac{e^{f_\theta(x)[y]}}{Z(\theta)}$$

$$p_\theta(x) = \frac{\sum_y e^{f_\theta(x)[y]}}{Z(\theta)}$$

$$p_\theta(y \mid x) = \frac{e^{f_\theta(x)[y]}}{\sum_{y'} e^{f_\theta(x)[y']}}$$



- we have found a Joint Energy Model inside of your classifier…a hidden JEM 😏

Обучение

- given x, y how should we train JEM?
- can maximize $p_\theta(x, y)$ using approximate likelihood gradient
- or can maximize $\log p_\theta(x) + \log p_\theta(y \mid x)$

using approximate gradient

standard x-entropy

Обучение

- given x, y how should we train JEM?
- can maximize $p_\theta(x, y)$ using approximate likelihood gradient
- or can maximize $\log p_\theta(x) + \log p_\theta(y \mid x)$

biased

biased

unbiased

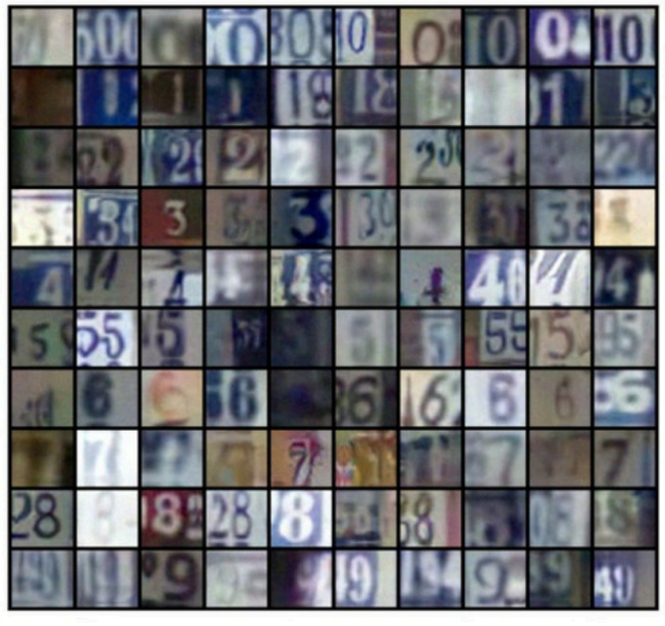
Результаты. Гибридные модели

- we obtain accuracy comparable to SOTA
- we obtain IS/FID comparable to SOTA
- accuracy decreases negligibly from classification only model

CIFAR10: Accuracy 92.9%

CIFAR100: Accuracy 72.2%

SVHN: Accuracy 96.7%



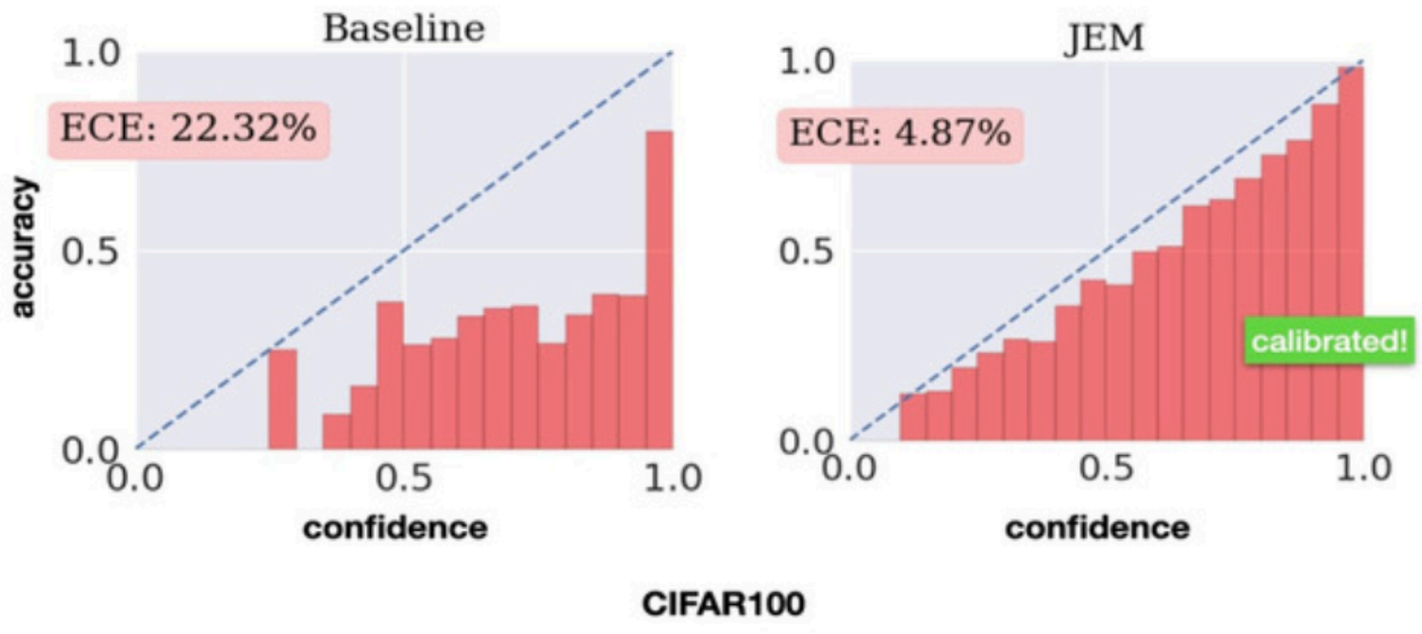
Результаты. Гибридные модели

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

CIFAR10 Quantitative Results

Результаты. Калибровка

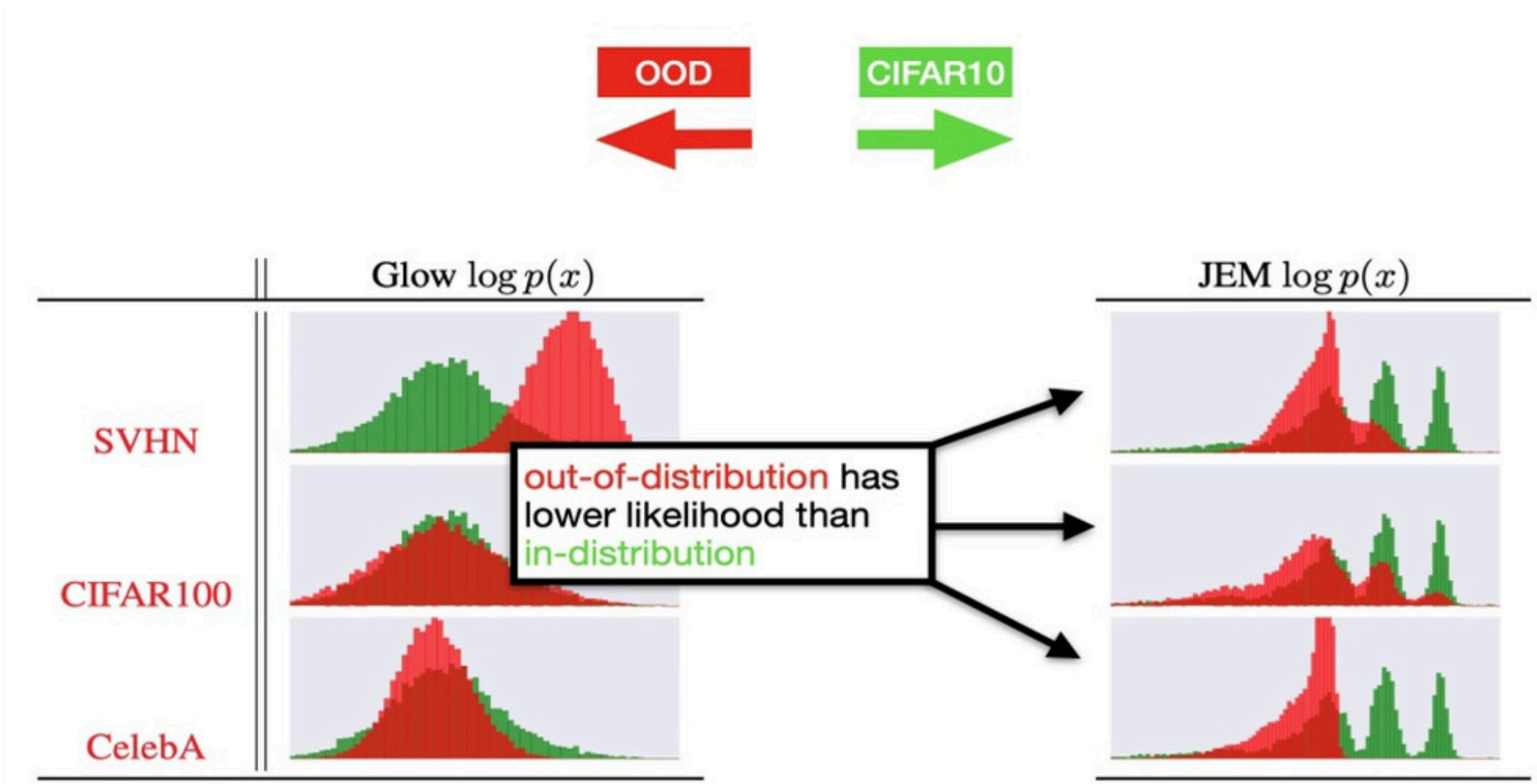
- deep nets are poorly calibrated
- predictive uncertainty meaningless
- problematic when deploying models in practice
- adding JEM training greatly improves calibration without significantly hurting accuracy
- requires no extra data



Результаты. Out-of-distribution detection

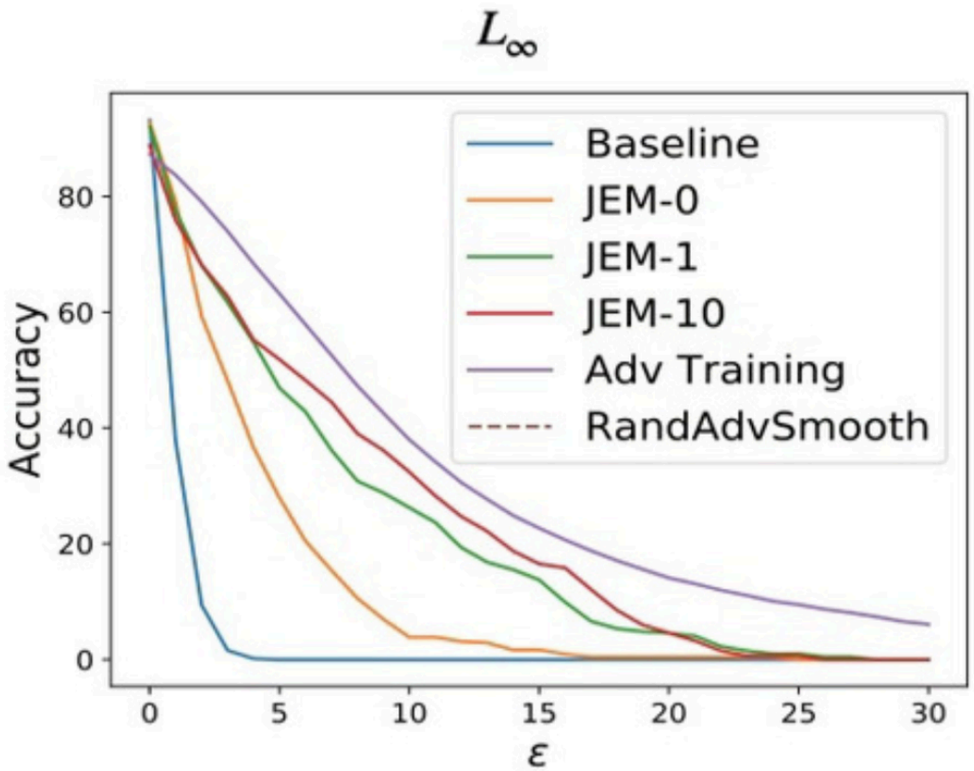
- we can do OOD detection with $p(x)$ model
- previous likelihood-based generative models have been shown to fail at this task
- most SOTA methods are based on reusing parts of classifier
- model is strong classifier, we can use both approaches

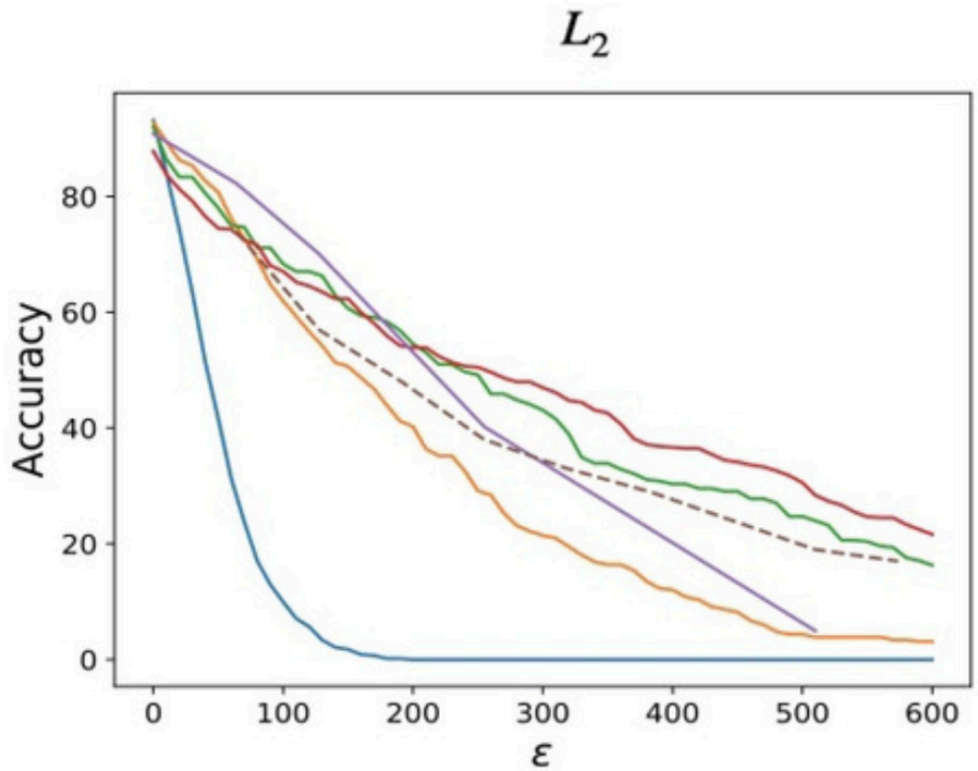
Результаты. Out-of-distribution detection



Результаты. Робастность

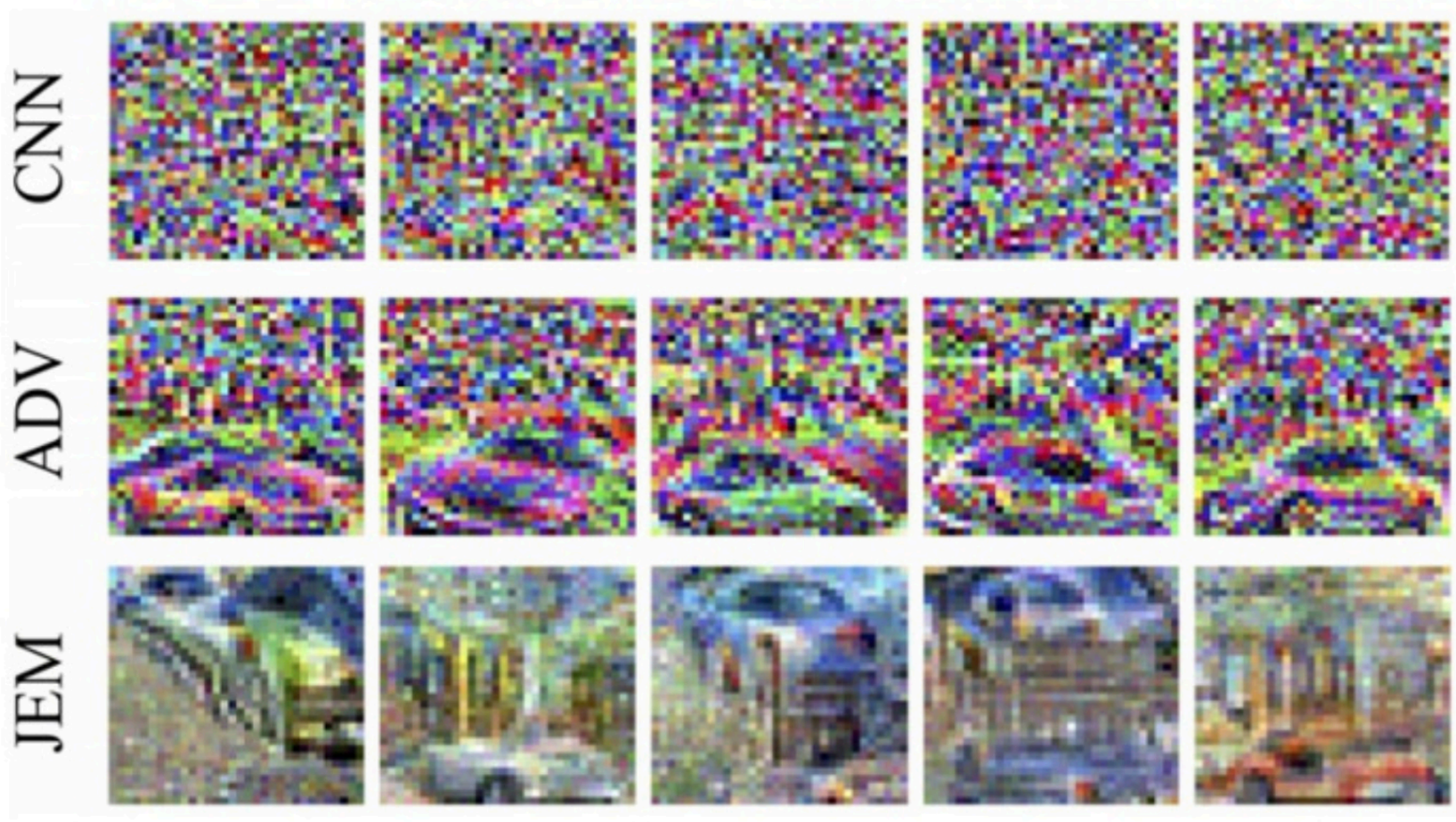
- does this make more robust models?





Результаты. Робастность

- does this make more robust models?



examples such that $p(y = \text{car} \mid x) > .9$

Основные проблемы

- hard to diagnose problems because no good metrics to track model fit
- EBM training can be unstable, is biased
- sampler parameters must be perfectly tuned
- relying on MCMC for training/eval complicates things

Вопросы

- Какой формулой описывается плотность вероятности в энергетической модели?
- Зачем нужна калибровка классификатора?
- Какие основные минусы энергетических моделей? скрытую структуру

Использованная литература

- https://iclr.cc/virtual_2020/poster_Hkxzx0NtDB.html

- https://openreview.net/pdf/df53e66f00cddbec2fc54bd79e0e5d84a31eaf9a.pdf

- Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397, 2017.

- Shane Barratt and Rishi Sharma. A note on the inception score. arXiv preprint arXiv:1801.01973, 2018.