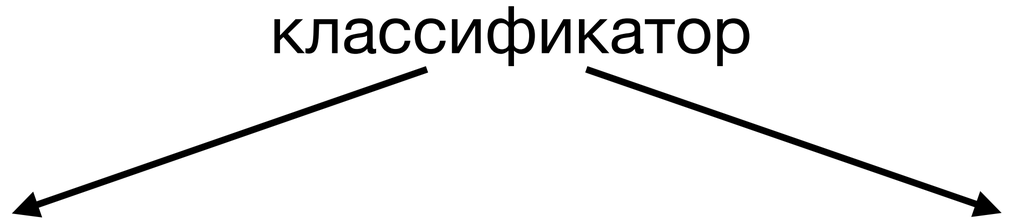
Your Classifier is Secretly an Energy Based model and You Should Treat it Like One

Иванов Семен, 182

TL DR



generative

- выучивает распределение данных
- $P_\theta(X, Y)$

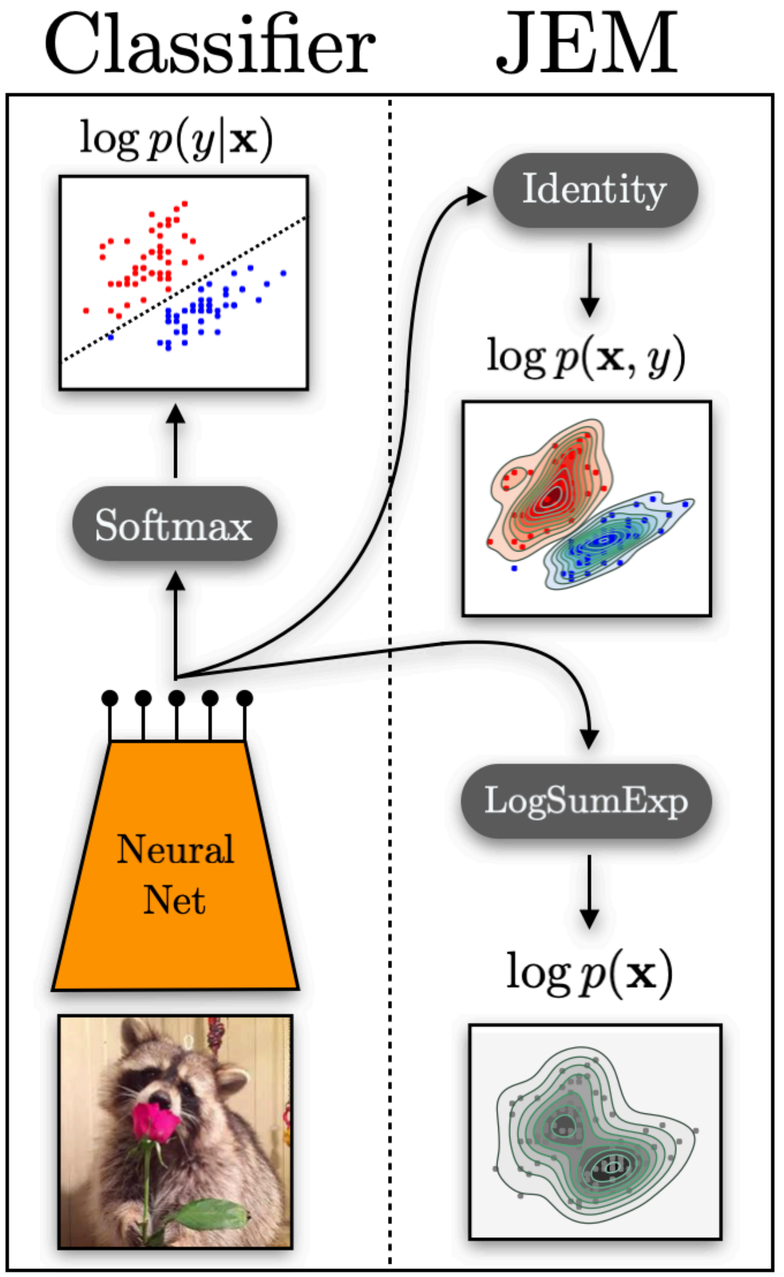
discriminative

- просто решает задачу классификации
- $\max_\theta P_\theta\left(Y \mid X\right)$

- на практике discriminative лучше классифицирует
- но у generative есть много полезных свойств: есть распределение; устойчивость; OOD; semisupervised; calibration

TL DR

Способ превратить discriminative модель в generative, просто используя логиты



Energy Based Model

Способ выучить $P_\theta(X)$

- Любую pdf можно записать в виде: $p_\theta(x) = \frac{\exp\left(-E_\theta(x)\right)}{Z(\theta)}$
- $Z(\theta) = \int_x \exp\left(-E_\theta(x)\right)$
- $E_\theta(x) : \mathbb{R}^D \to \mathbb{R}$ – energy function, чем меньше, тем больше pdf

Energy Based Model

Оптимизация

- MLE
- $$\frac{\partial \log p_\theta(x)}{\partial \theta} = E_{p_\theta(x')}\left[\frac{\partial E_\theta(x')}{\partial \theta}\right] - \frac{\partial E_\theta(x)}{\partial \theta}$$
- чтобы оценивать матожидание будем сэмплировать при помощи SGLD:
- $$x_0 \sim p_0(x), \qquad x_{i+1} = x_i - \frac{\alpha}{2}\frac{\partial E_\theta(x_i)}{\partial x_i} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0,\alpha)$$

What your classifier is hiding

Joint Energy Model

- Классификатор возвращает логиты $f_\theta : \mathbb{R}^D \to \mathbb{R}^K$, где K – число классов
- $p_\theta(y \mid x) = \dfrac{\exp\left(f_\theta(x)[y]\right)}{\sum_{y'} \exp\left(f_\theta(x)[y']\right)}$
- $p_\theta(x, y) = \dfrac{\exp\left(f_\theta(x)[y]\right)}{Z(\theta)}$, $E_\theta(x, y) = -f_\theta(x)[y]$
- $p_\theta(x) = \sum_y p_\theta(x, y) = \dfrac{\sum_y \exp\left(f_\theta(x)[y]\right)}{Z(\theta)}$, $E_\theta(x) = -\log \sum_y \exp\left(f_\theta(x)[y]\right)$

Joint Energy Model

Train

- MLE
- $\log p_\theta(x, y) = \log p_\theta(x) + \log p_\theta(y \mid x)$
- Тренируем по отдельности, так как иначе можем получить biased модель $p_\theta(y \mid x)$
- $\log p_\theta(y \mid x)$ оптимизируем как обычную discriminative модель
- $\log p_\theta(x)$ оптимизируем как EBM

Joint Energy Model

Train

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

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

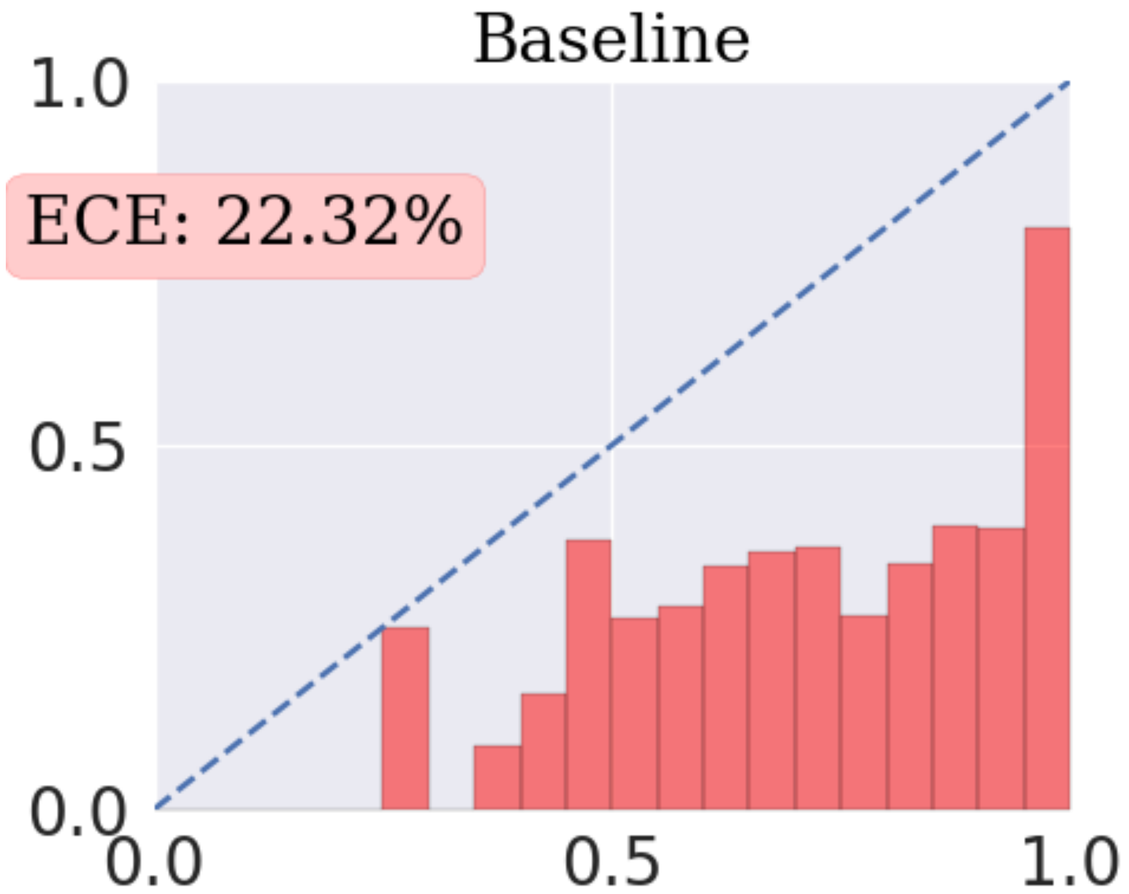
Joint Energy Model

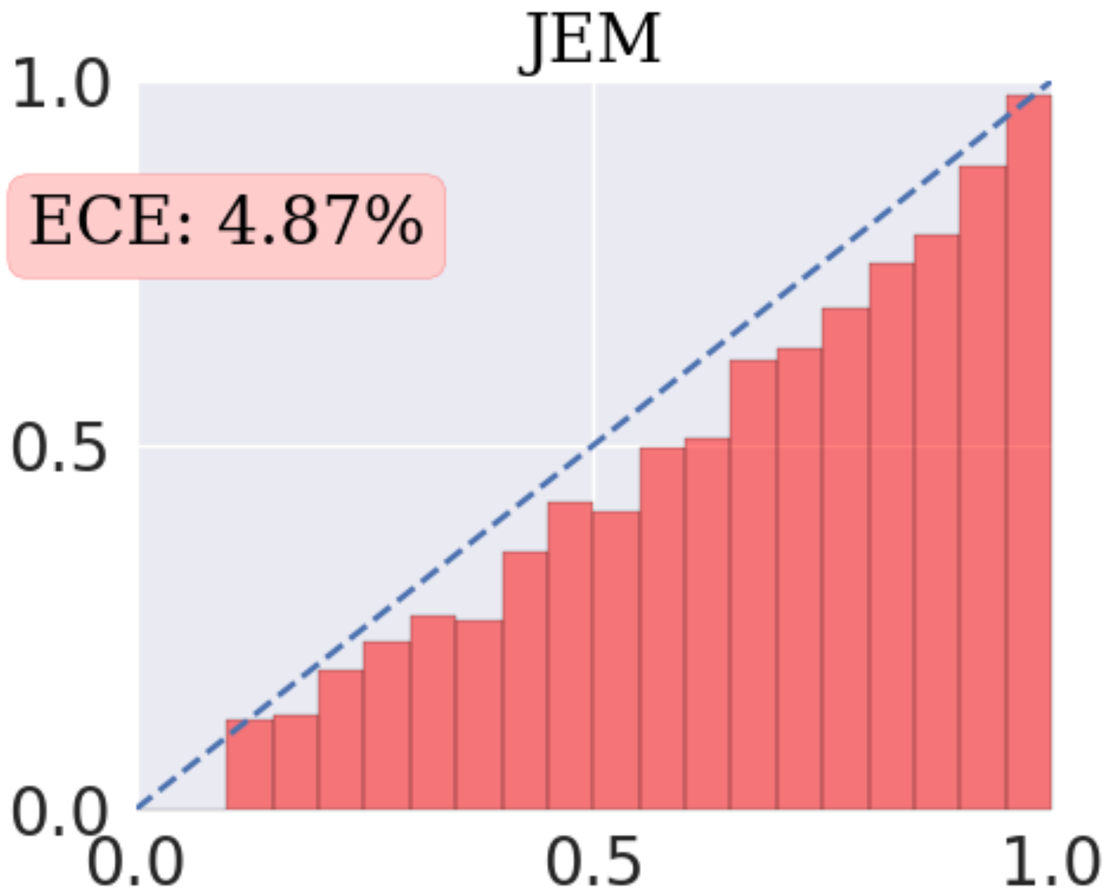
Результаты

- Обучили wide resnet на CIFAR-10
- Получилась сравнимая с discriminative моделями точность
- Плюс модель с хорошими генеративными качествами (IS - Inseption Score, то, насколько уверенно Inception v3 классифицирует сгенерированные картинки)
- Семплы генерировались по аналогии с оптимизацией (SGLD)

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

Calibration



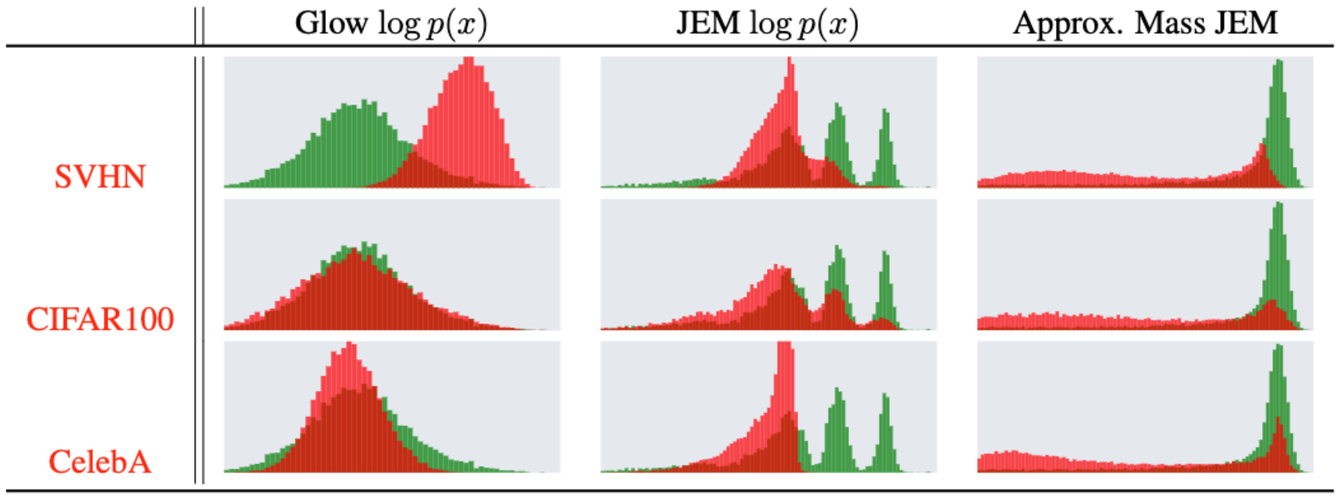


Calibration

- Скоры получились почти идеально откалиброванными
- Почему это может быть полезным?
- Устойчивость. Проще подбирать пороги при дисбалансе классов, меньше шума
- Пороги получаются более интерпретируемыми
- Можно что-то поверх этого считать и легче дальше использовать

Out-Of-Distribution Detection

- По скорам модели $p_\theta(x)$ понять, x вообще похож на данные обучения?
- Mass JEM: $s_\theta(x) = -\left|\frac{\partial \log p_\theta(x)}{\partial x}\right|_2$
- Можно улавливать изменения в данных



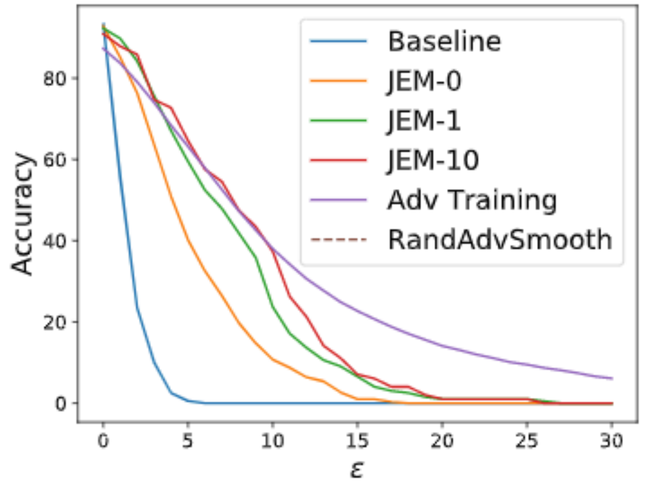
Out-Of-Distribution Detection

- По скорам модели $p_\theta(x)$ понять, x вообще похож на данные обучения?
- Mass JEM: $s_\theta(x) = -\left|\frac{\partial \log p_\theta(x)}{\partial x}\right|_2$
- Можно улавливать изменения в данных
- AUC-и:

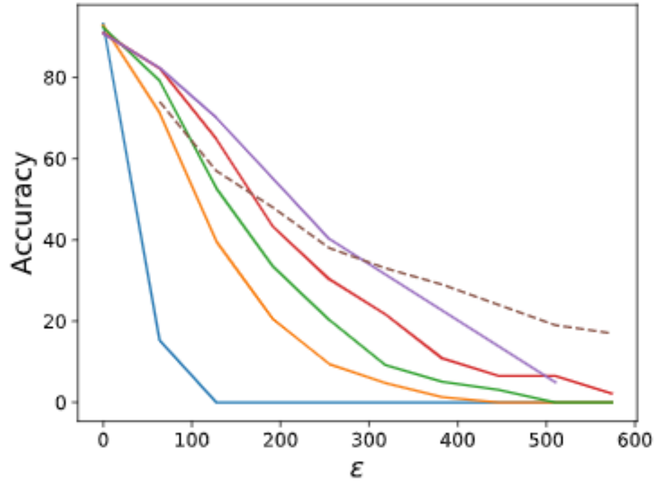
$s_\theta(\mathbf{x})$	Model	SVHN	CIFAR10 Interp	CIFAR100	CelebA
$\log p(\mathbf{x})$	Unconditional Glow	.05	.51	.55	.57
	Class-Conditional Glow	.07	.45	.51	.53
	IGEBM	.63	**.70**	.50	.70
	JEM (Ours)	**.67**	.65	**.67**	**.75**
$\max_y p(y\|\mathbf{x})$	Wide-ResNet	**.93**	**.77**	.85	.62
	Class-Conditional Glow	.64	.61	.65	.54
	IGEBM	.43	.69	.54	.69
	JEM (Ours)	.89	.75	**.87**	**.79**
$\left\|\left\|\frac{\partial \log p(\mathbf{x})}{\partial \mathbf{x}}\right\|\right\|$	Unconditional Glow	**.95**	.27	.46	.29
	Class-Conditional Glow	.47	.01	.52	.59
	IGEBM	.84	.65	.55	.66
	JEM (Ours)	.83	**.78**	**.82**	**.79**

Robustness

- $\tilde{x} = x + \delta,\ ||\tilde{x} - x||_p < \varepsilon$ (можно давать доступ к градиенту)
- RandAdvSmooth – SOTA в задаче robustness
- Adv Training – аугментации в обучении
- JEM-0 (0 шагов SGLD)
- Больше шагов семплирования -> более устойчивая модель



(a) L_∞ Robustness



(b) L_2 Robustness

Robustness

- Выдает высокий скор не совсем шуму

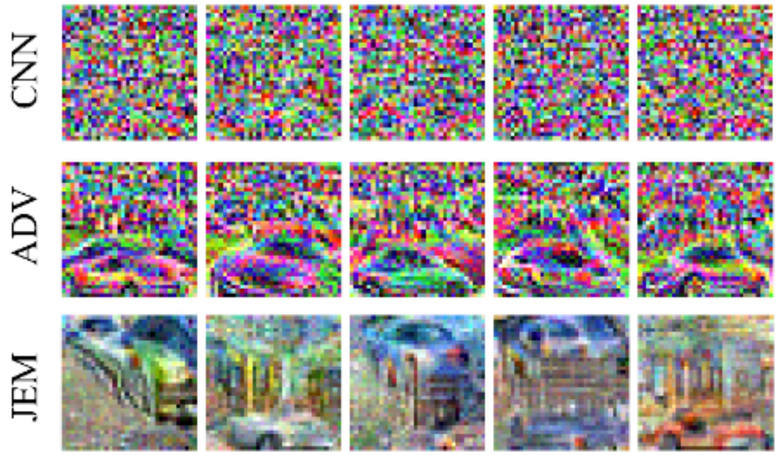


Figure 6: **Distal Adversarials.** Confidently classified images generated from noise, such that: $p(y = \text{“car”}|\mathbf{x}) > .9$.

Минусы

- Из-за того $Z(\theta)$ не посчитать, сложно понять, что модель вообще обучается
- Неусточивые оценки градиентов
- Модель постоянно расходилась и авторам приходилось перезапускать с меньшим lr, добавлять регуляризации

Итог

- Можно получить модель с хорошим качеством распознавания + она была бы генеративной и более устойчивой
- Но такие модели пока довольно сложно обучать

Рецензия

Содержание:

- Рассматривается способ интерпретации логитов стандартного нейросетевого классификатора p(y|x) для определения energy based модели на p(x, y)
- Обученная таким образом модель показывает хорошие результаты в задачах классификации и генерации
- Классификатор обретает ряд бонусов (adversarial устойчивость, лучшая откалиброванность)
- Модель можно применять для OOD detection

Рецензия

Плюсы:

- Важный вклад в область применения energy based моделей
- Предложенный подход хорошо и подробно описан, статья отлично структурирована и легко читается
- Проведены обширные и хорошо поставленные эксперименты
- Код обучения и экспериментов выложен на гитхаб, что облегчает воспроизводимость
- Подход имеет большой потенциал для практического применения

Рецензия

Минусы:

- Утверждения о том, что подход может тягаться с SOTA для генерации или классификации слабо обоснованы
- Авторы отмечают сложность и нестабильность процедуры обучения как одну из важнейших проблем, при не уделяют ей и способам борьбы с ней в статье почти никакого внимания
- В псевдокоде процедуры обучения есть ошибка
- Не очень понятно, зачем было выкидывать BatchNorm из архитектуры WideResNet

Рецензия

NIPS-like mark: 8/10

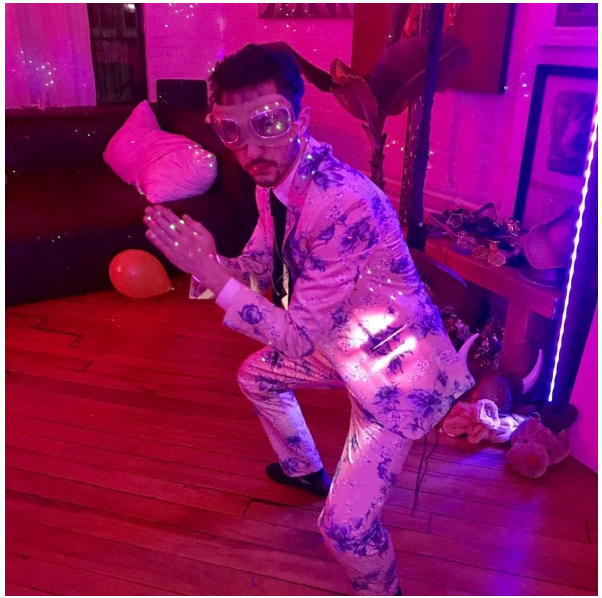
NIPS-like confidence mark: 5/5

Статья получила Oral на ICLR 2020

Авторы

Will Grathwohl, University of Toronto & Vector Institute, Google Research

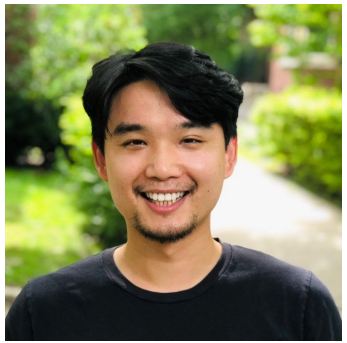
- DeepMind, Research Scientist, Oct 2021 – Present
- На момент написания статьи получал PhD в University of Toronto
- Новая статья на ICML 2021 получила Outstanding Paper Award Honorable Mention
- Все статьи автора после данной продолжают развитие темы, связанной с energy-based моделями



Авторы

Kuan-Chieh Wang & Jorn-Henrik Jacobsen, University of Toronto & Vector Institute

- Currently post-doctoral research fellow at Stanford CS
- PhD student at University of Toronto во время написания статьи
- В основном является соавтором в статьях, связанных с invertible neural networks



- Senior Research Scientist at Apple
- Was a postdoc at University of Toronto во время написания статьи
- В основном является соавтором в статьях, связанных с Generative models

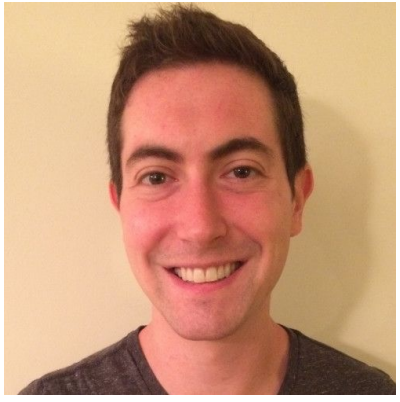
Авторы

David Duvenaud, University of Toronto & Vector Institute

- Google Brain Toronto March 2020 – present, Visiting Researcher (part time)
- University of Toronto July 2016 – present, Assistant Professor, Computer Science and Statistical Sciences, Canada Research Chair in Generative Models
- Является автором статей по различным темам, среди которых например Neural ODEs, Automatic chemical design using generative models и проч
- В том числе является соавтором и в новых статьях первого автора.

Авторы

Kevin Swersky & Mohammad Norouzi, Google Research



- Research scientist, Google Brain
- Тоже появляется соавтором первого автора в более свежих статьях.
- Помните, RMSProp был впервые сформулирован на курсе Coursera? Kevin Swersky был соведущим на этом курсе в этот момент.
- Тоже различные интересы: Bayesian optimization, Normalising flows и тд.

- Research scientist at Google Brain
- Научные интересы: self-supervised, semi-supervised learning, generative models

Most influential papers

LeCun, et al. A tutorial on energy-based learning.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics.

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models.

Related prior work

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet.

Yunfu Song and Zhijian Ou. Learning neural random fields with inclusive auxiliary generators.

Кто цитирует эту статью?

Есть ли у нее продолжение?

В основном развитием Energy-based модели занимается первый автор Will Grathwohl. У него есть 5 новых статей посвященных этому. Одна из статей получила награду на ICML 2021.

Есть популярная статья под названием Your GAN is Secretly an Energy-based Model and You Should use Discriminator Driven Latent Sampling

В основном эту работу цитируют в статьях по темам

- Out-of-distribution detection
- Continual learning. Изучается наличие “памяти” у модели, с помощью которой она долго помнит что учила при дообучении ее на новых данных или задаче. Например Supermasks in Superposition, Energy-Based Models for Continual Learning
- Применение Energy-based модели в сценариях обычной классификации: в задачах CV, в GANах и тд

Предложения по исследованию

Energy based модели по разному применяют в разных задачах: CV, GANs. Кроме того, что модель надо адаптировать под свой случай, интересно посмотреть на то, как влияет выбор марковской цепи в каждой конкретной задаче. Ведь у всех есть разные свойства, а распределения могут быть произвольно сложными (поэтому интереснее это изучать в GANax).

Вроде бы автор как раз этот вопрос и исследует. Однако он предлагает совсем отказаться от семплирования с помощью марковских цепей из-за того, что у них есть проблемы с переходом по модам. (Название статьи: No MCMC for me: Amortized sampling for fast and stable training of energy-based models и тд.)

Предложения по применению

В самой статье предложено много применений: Hybrid modeling, out-of-distribution detection и тд.

На данный момент EBMs нашли множество применений: text generation, molecule generation, anomaly detection, trajectory prediction, semi-supervised learning, etc.

For comprehensive review see CVPR 2021 Tutorial on EBMs: https://energy-based-models.github.io/