

Wasserstein GAN

Каглинская Мария

Ноябрь 2017

Содержание

1. Постановка задачи
2. Различные метрики
3. Теоретические гарантии на EMD
4. Новый вид метрики и алгоритм на ее основе
5. Эмпирические результаты

Задача генерации

- ▶ Имеем некоторое распределение \mathbb{P}_r
- ▶ Хотим уметь из него сэмплировать
- ▶ Логичный подход: взять некоторое параметрическое семейство $(\mathbb{P}_\theta)_{\theta \in R^d}$ и максимизировать правдоподобие данных для получения распределения

$$\max_{\theta \in R^d} \frac{1}{m} \sum_{i=1}^m \log P_\theta(x_i)$$

- ▶ У такого подхода много проблем: он плохо работает с низкоразмерными многообразиями, может быть дорого сэмплировать

Текущее решение: GAN

Плюсы:

- ▶ Позволяет иметь дело с низкоразмерными многообразиями
- ▶ Дает возможность сэмплировать из интересующего распределения, а не только оценивать плотность

Минусы:

- ▶ Нестабильность обучения
- ▶ Отсутствие интерпретируемой метрики качества
- ▶ Затухание градентов

Немного обозначений:

- ▶ \mathcal{X} - компактное метрическое пространство
(например, $[0; 1]^d$)
- ▶ Σ - множество всех Борелевских* подмножеств в \mathcal{X}
- ▶ $Prob(\mathcal{X})$ - множество всех вероятностных мер на \mathcal{X}
- ▶ Пусть $\mathbb{P}_r, \mathbb{P}_\theta \in Prob(\mathcal{X})$

*полученных из открытых множеств с помощью объединения, пересечения или разности конечного числа множеств

Метрики

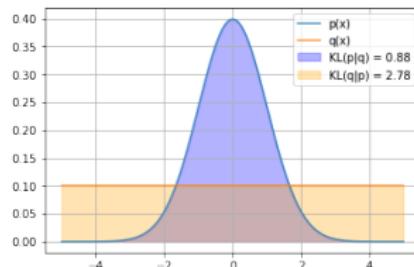
- ▶ *Total Variation*

$$\delta(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{A \in \Sigma} |\mathbb{P}_r(A) - \mathbb{P}_\theta(A)|$$

- ▶ *Kullback-Leibner*

$$KL(\mathbb{P}_r || \mathbb{P}_\theta) = \int_X \log \left(\frac{P_r(x)}{P_\theta(x)} \right) P_r(x) dx$$

Не симметрична, может быть бесконечна.



Интуитивно: $KL(p||q)$ - число бит, которое будет потеряно, если закодируем p через q

Метрики

- ▶ Jensen-Shannon

$$JS(\mathbb{P}_r, \mathbb{P}_\theta) = KL(\mathbb{P}_r || \mathbb{P}_m) + KL(\mathbb{P}_\theta || \mathbb{P}_m)$$

$$\mathbb{P}_m = (\mathbb{P}_r + \mathbb{P}_\theta)/2$$

Симметрична и везде определена.

Интуитивно: если $X = \frac{P+Q}{2}$ - смесь распределений, Z - скрытая переменная: $Z = I[\text{точка пришла из } Q]$, тогда $JS(P||Q)$ - взаимная информация X и Z ($I(A; B) = H(A) - H(A|B)$, где H - энтропия)

Earth Mover distance (Wassernstein - 1)

Если $\prod(\mathbb{P}_r, \mathbb{P}_\theta)$ - множество совместных распределений $\gamma(x, y)$:
их маргинальными являются $\mathbb{P}_r, \mathbb{P}_\theta$

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \inf_{\gamma \in \prod(\mathbb{P}_r, \mathbb{P}_\theta)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|] = \inf_{\gamma \in \prod(\mathbb{P}_r, \mathbb{P}_\theta)} \sum_{x,y} \|x - y\| \gamma(x, y)$$

Интуитивно это: минимальная стоимость перевода одного распределения в другое, где $\gamma(x, y)$ задает, сколько необходимо перевести, а $\|x - y\|$ стоимость перевода единицы вероятности из x в y .



Пример на понимание:

Пусть $z \sim Unif[0, 1]$, \mathbb{P}_r - распределение точек $(0, z) \in \mathbb{R}^2$.

Пусть $g_\theta(z) = (\theta, z)$, где $\theta \in \mathbb{R}$.

Тогда:

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = |\theta|$$

$$KL(\mathbb{P}_r || \mathbb{P}_\theta) = \begin{cases} \infty & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$$

$$\delta(\mathbb{P}_r, \mathbb{P}_\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$$

$$JS(\mathbb{P}_r, \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$$

Это означает, что при $\theta_t \rightarrow 0$, $(\mathbb{P}_{\theta_n})_{n \in N} \rightarrow \mathbb{P}_r$ только для $W(\mathbb{P}_r, \mathbb{P}_\theta)$, что ещё раз показывает осмыслинность её выбора.

Гарантии непрерывности и дифференцируемости

Теорема

Если Z - случайная величина на пространстве \mathcal{Z} . $g_\theta(z)$ - функция: $\mathcal{Z} \times \mathbb{R}^d \rightarrow \mathcal{X}$. \mathbb{P}_θ - распределение $g_\theta(Z)$.

1. Если g непрерывна по θ , то $W(\mathbb{P}_r, \mathbb{P}_\theta)$ также непрерывна
2. Если 1 верно и g локально Липшицева, то $W(\mathbb{P}_r, \mathbb{P}_\theta)$ непрерывна всюду и дифференцируема почти всюду
3. 1 и 2 неверны для $KL(\mathbb{P}_r, \mathbb{P}_\theta)$, $KL(\mathbb{P}_\theta, \mathbb{P}_r)$, $JS(\mathbb{P}_r, \mathbb{P}_\theta)$

Гарантии непрерывности и дифференцируемости

Следствие

Если

- ▶ g_θ - feed-forward нейронная сеть* (параметризованная θ и Липшицева)
- ▶ $p(z)$ - априорное распределение на z : $\mathbb{E}_{z \sim p(z)}[|z|] < \infty$

Тогда условия теоремы выполнены и $\Rightarrow W(\mathbb{P}_r, \mathbb{P}_\theta)$ непрерывна
всюду и дифференцируема почти всюду

* Под нейронной сетью понимается функция являющаяся композицией афинных преобразований и поточечных нелинейностей: они являются гладкими липшицевыми функциями

Отношение метрик

Теорема

Если \mathbb{P} распределение на компактном множестве \mathcal{X} и $(\mathbb{P}_n)_{n \in \mathbb{N}}$ последовательность распределений на \mathcal{X} ,
то при $n \rightarrow \infty$:

- ▶ $W(\mathbb{P}_n, \mathbb{P}) \rightarrow 0 \Leftrightarrow \mathbb{P}_n \xrightarrow{D} \mathbb{P}$
- ▶ Сходимость метрик соотносится таким образом:

$$KL(\mathbb{P}_n, \mathbb{P}) \Rightarrow (\delta(\mathbb{P}_n, \mathbb{P}) \Leftrightarrow JS(\mathbb{P}_n, \mathbb{P})) \Rightarrow W(\mathbb{P}_n, \mathbb{P})$$

И что нам это даёт?

У нас имеются теоретические гарантии на выбранную метрику:

- ▶ Для липшицевой нейронной сети наша метрика всюду непрерывна и почти всюду дифференцируема (что при этом неверно для J_S)
- ▶ Сходимость выбранной метрики к 0 эквивалентна сходимости по распределению
- ▶ Выбранная метрика имеет наиболее "слабую" сходимость из рассмотренных

Новый вид метрики

Рассмотренная метрика имеет хорошие свойства, но её трудно вычислить и оптимизировать по ней.

Поэтому будем решать двойственную к ней задачу:

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$$

Где $\|f\|_L \leq 1$ - все 1- Липшицевы функции.

Очевидно, для случая оптимизации по $\|f\|_L \leq K$ мы найдем значение $W(\mathbb{P}_r, \mathbb{P}_\theta)$ с точностью до умножения на константу K .

Новый вид метрики

Если $(f_w)_{w \in W}$ параметрическое семейство К-Липшицевых функций, можем теперь оценить снизу $W(\mathbb{P}_r, \mathbb{P}_\theta)$.

$$\sup_{\|f\|_L \leq K} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)] = K * W(\mathbb{P}_r, \mathbb{P}_\theta) \geq \max_{w \in W} \mathbb{E}_{x \sim \mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z \sim p(z)}[f_w(g_\theta(z))]$$

Стоит заметить, что размер K нас не интересует, только то, что она не изменяется во время обучения.

Как оптимизировать?

1. Сначала для фиксированного θ найдем f_w при которой достигается $\max_{w \in W} \mathbb{E}_{x \sim \mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z \sim p(z)}[f_w(g_\theta(z))]$.
Можно сделать это обучив нейронную сеть f_w максимизировать это выражение.
Это также даст нам оценку на значение $W(\mathbb{P}_r, \mathbb{P}_\theta)$.
2. С помощью f_w можем оценить

$$\nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta) = -\mathbb{E}_{z \sim p(z)}[\nabla_\theta f_w(g_\theta(z))]$$

и улучшить нашу оценку θ .

3. Повторять эти шаги до сходимости θ

Алгоритм

Require: : α , the learning rate. c , the clipping parameter. m , the batch size.
 n_{critic} , the number of iterations of the critic per generator iteration.

Require: : w_0 , initial critic parameters. θ_0 , initial generator's parameters.

1: **while** θ has not converged **do**

2: **for** $t = 0, \dots, n_{\text{critic}}$ **do**

3: Sample $\{x^{(i)}\}_{i=1}^m \sim \mathbb{P}_r$ a batch from the real data.

4: Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of prior samples.

5: $g_w \leftarrow \nabla_w \left[\frac{1}{m} \sum_{i=1}^m f_w(x^{(i)}) - \frac{1}{m} \sum_{i=1}^m f_w(g_\theta(z^{(i)})) \right]$

6: $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$

7: $w \leftarrow \text{clip}(w, -c, c)$

8: **end for**

9: Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of prior samples.

10: $g_\theta \leftarrow -\nabla_\theta \frac{1}{m} \sum_{i=1}^m f_w(g_\theta(z^{(i)}))$

11: $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$

12: **end while**

Обучение
“критика” f_w

+ вычисляем
 $W(\mathbb{P}_r, \mathbb{P}_\theta)$

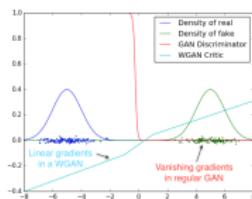
Обучение
генератора g_θ

* Для достижения липшицевости просто обрезаем веса w

Чем лучше обычного GAN?

Для каждого фиксированного θ (генератора) мы обучаем "критика" до оптимальности \Rightarrow

- ▶ Получаем более релевантные значения градиента θ
- ▶ Отсутствует mode collapse
- ▶ Посмотрим, что будет, если обучить дискриминатор до оптимальности в обычном GAN:

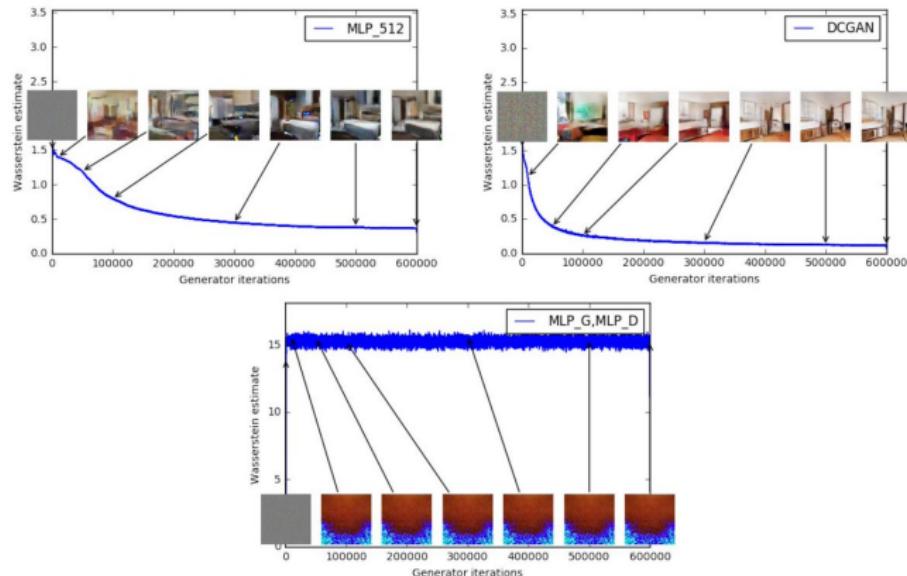


В случае JS градиент равен 0 почти везде \Rightarrow затухание
В случае W такого не происходит, так как каждый раз
после обучения критика мы улучшаем оценку θ

Эксперимент: WGAN лосс и качество

- ▶ Обучим WGAN с различными архитектурами генератора/критика и посмотрим на то, какое качество картинок получается и как оно соотносится со значениями метрики $W(\mathbb{P}_r, \mathbb{P}_\theta)$
- ▶ В качестве архитектур будем рассматривать DCGAN (сверточная сеть с несколькими модификациями) и многослойный персепtron
- ▶ На двух верхних графиках критик DCGAN, а дискриминатор отмечен на графике
- ▶ На нижнем графике и генератор и критик это многослойный персепtron

Эксперимент: WGAN лосс и качество

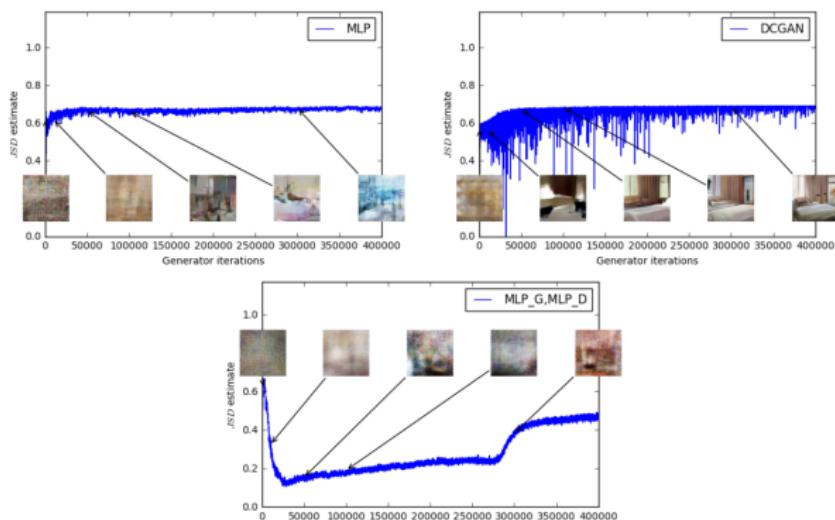


Видим четкую кореляцию между качеством картинок и выбранной метрикой!

ВАЖНО: все еще не можем сравнивать модели между собой.

Эксперимент: GAN лосс и качество

Аналогично предыдущему эксперименту обучим GAN-ы с такими же архитектурами и построим значения метрики $JS(\mathbb{P}_r, \mathbb{P}_\theta)$



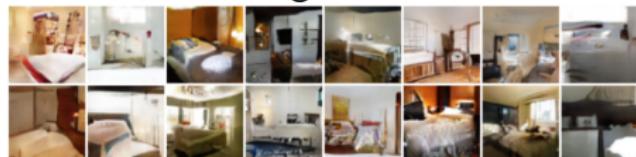
Видим, что нет корреляции качества изображений и значения метрики, а сама метрика не уменьшается.

Multiple Columns

WGAN

GAN

DCGAN generator



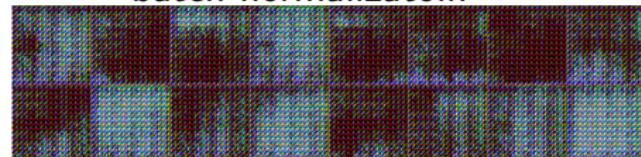
Training without



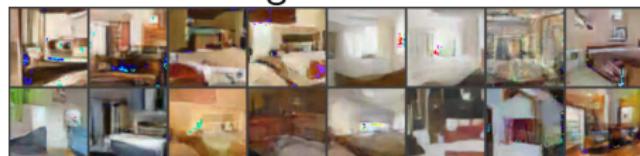
DCGAN generator



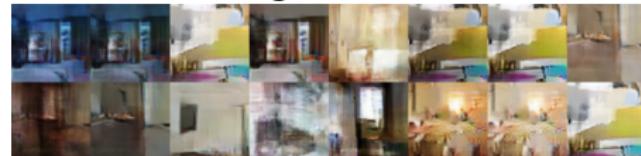
batch normalizatoin



MLP generator



MLP generator



Видим, что WGAN более устойчик к изменениям и у него отсутствует mode collapse.

Итоги:

- ▶ Была предложена метрика для измерения близости распределений, доказаны теоретические гарантии на неё
- ▶ Была предложена удобная для оптимизации оценка на эту метрику и алгоритм обучения WGAN на её основе
- ▶ Эксперименты показали, что предложенный алгоритм обучения имеет преимущества над **обычными GAN**:
 1. Стабильность обучения
 2. Метрика, кореллирующая с качеством получаемых изображений
 3. Отсутствие mode collapse

Источники

1. "Wasserstein GAN" Martin Arjovsky, Soumith Chintala, Leon Bottou (2017)
2. "Generative Adversarial Nets", Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio (2014)
3. Simple explanation of this paper
4. Some details about metric
5. Explanation of Kullback Leibner divergence