Improved Training of Wasserstein GANs

И. Гаврилов[1]

[1]Факультет Компьютерных Наук
Высшая Школа Экономики

12 Февраля 2018 г.

План

1. Мотивация
2. Recap
   - GAN
   - Wasserstein GAN
3. Проблема WGAN
   - Обрезание весов
   - Взрыв и затухание градиента
4. Регуляризация градиента (WGAN-GP)
5. Эксперименты
   - Обучение случайных архитектур
   - Обучение архитектур на LSUN bedrooms
   - Улучшение скорости и качества обучения
   - Сравнение с другими архитектурами на CIFAR-10
   - Корреляция с качеством
6. Литература
7. Доп информация

Мотивация

- GAN являются мощными генеративными моделями, однако сильно страдают от нестабильности обучения.
- Wasserstein GAN (WGAN) решает проблему нестабильности, но порой генерирует плохие данные или не вообще не сходится. В статье рассмотрен метод, который призван решить эту проблему

План

Generative adversarial networks

Постановка задачи:

$$\min_G \max_D E_{x \sim P_r}[log(D(x))] + E_{x^* \sim P_g}[log(1 - D(x^*))]$$

- P_r - распределение исходных данных
- P_g - распределение генератора,
- $x^* = G(z)$, $z \sim p(z)$ (z -некоторый шум, подающийся на вход генератора)

План

Wasserstein GAN

- Проблема с обучением GAN'ов - минимизируемые функции потенциально не непрерывны.
- В WGAN используется Earth-Mover (Wasserstein-1) расстояние $W(p, q)$. При определенных условиях она непрерывна и дифференцируема.
- Формально задачу можно переписать в виде:

$$\min_G \max_{D \in \Delta} E_{x \sim P_r}[D(x)] - E_{x^* \sim P_g}[D(x^*)]$$

- Δ - семейство 1-липшицевых функций.

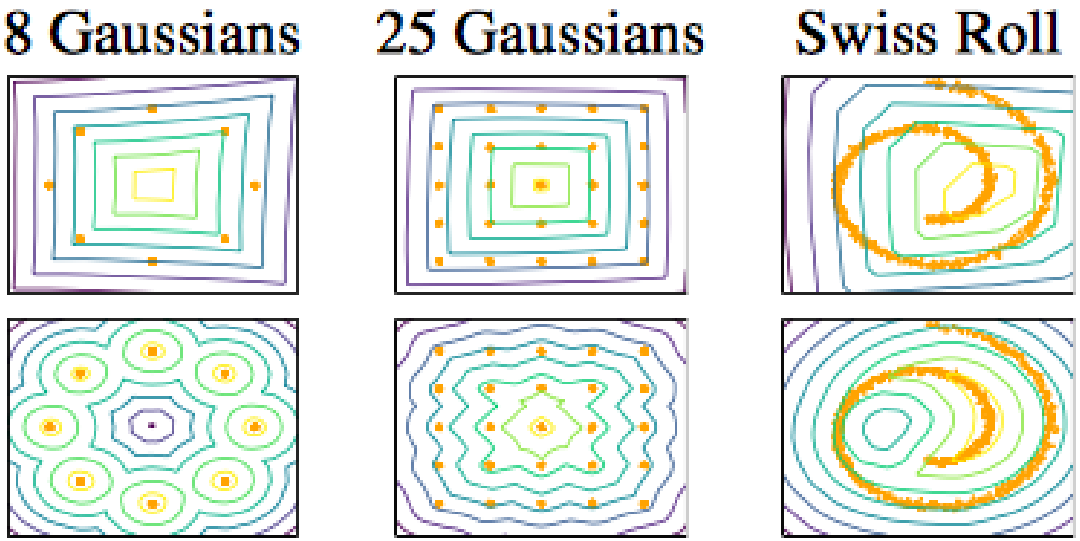
Wasserstein GANs

- Есть гарантии на непрерывность и дифференцируемость.
- Лучшая сходимость.
- Корреляция метрики и качества генерации.
- Требуется липшицевость, которая достигается за счет закрепления весов в пределах $[-c, c]$.
- Норма градиента оптимального дискриминатора WGAN равна 1 почти везде.

План

Ограничение весов

Закрепление весов некотором промежутке приводит дискриминатор к слишком простым функциям.

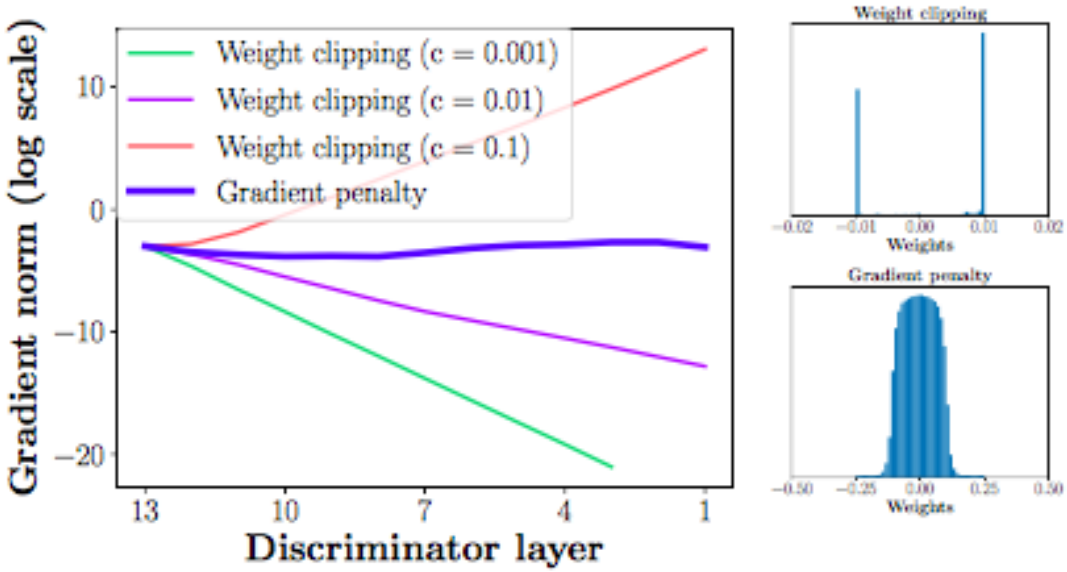


(a) Value surfaces of WGAN critics trained to optimality on toy datasets using (top) weight clipping and (bottom) gradient penalty. Critics trained with weight clipping fail to capture higher moments of the data distribution. The 'generator' is held fixed at the real data plus Gaussian noise.

План

Взрыв и затухание градиента

Без тщательного подбора ограничений на веса, градиенты дискриминатора взрываются или затухают



(b) (left) Gradient norms of deep WGAN critics during training on the Swiss Roll dataset either explode or vanish when using weight clipping, but not when using a gradient penalty. (right) Weight clipping (top) pushes weights towards two values (the extremes of the clipping range), unlike gradient penalty (bottom).

Регуляризация градиента

Поскольку дифференцируемая функция 1-липшицева тогда и только тогда, когда норма градиента не превосходит 1 почти всюду, то рассмотрим прямое ограничение нормы градиента дискриминатора:

$$L = E_{x^* \sim P_g}[D(x^*)] - E_{x \sim P_r}[D(x)] + \lambda E_{x^* \sim P_{x^*}}[(||\Delta_{x^*} D(x^*)||_2 - 1)^2]$$

- Распределение P_{x^*} - равномерное распределение вдоль прямых между парами точек, сэмплированых из распределений P_r и P_g
- Во всех экспериментах $\lambda = 10$
- В дискриминаторе не используется batch normalization
- Штрафуем, если норма градиента больше или меньше 1.

План

Обучение случайных архитектур

- Score: Inception score
- Данные: ImageNet, 32×32

Table 1: We evaluate WGAN-GP's ability to train the architectures in this set.

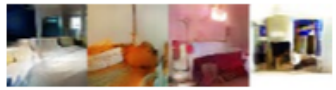
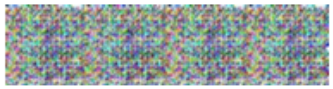
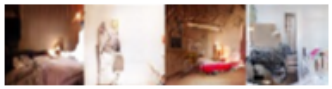
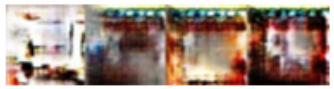
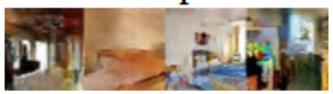
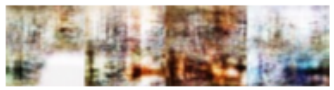
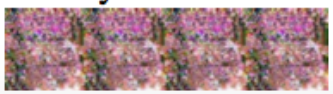
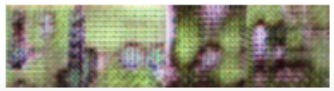
Nonlinearity (G)	[ReLU, LeakyReLU, $\frac{\text{softplus}(2x+2)}{2} - 1$, tanh]
Nonlinearity (D)	[ReLU, LeakyReLU, $\frac{\text{softplus}(2x+2)}{2} - 1$, tanh]
Depth (G)	[4, 8, 12, 20]
Depth (D)	[4, 8, 12, 20]
Batch norm (G)	[True, False]
Batch norm (D; layer norm for WGAN-GP)	[True, False]
Base filter count (G)	[32, 64, 128]
Base filter count (D)	[32, 64, 128]

Table 2: Outcomes of training 200 random architectures, for different success thresholds. For comparison, our standard DCGAN scored 7.24.

Min. score	**Only GAN**	**Only WGAN-GP**	**Both succeeded**	**Both failed**
1.0	0	8	192	0
3.0	1	88	110	1
5.0	0	147	42	11
7.0	1	104	5	90
9.0	0	0	0	200

План

Обучение архитектур на LSUN bedrooms

DCGAN	LSGAN	WGAN (clipping)	WGAN-GP (ours)
Baseline (G: DCGAN, D: DCGAN)			
G: No BN and a constant number of filters, D: DCGAN			
G: 4-layer 512-dim ReLU MLP, D: DCGAN			
No normalization in either G or D			
Gated multiplicative nonlinearities everywhere in G and D			
tanh nonlinearities everywhere in G and D			
101-layer ResNet G and D			

План

Улучшение скорости и качества обучения

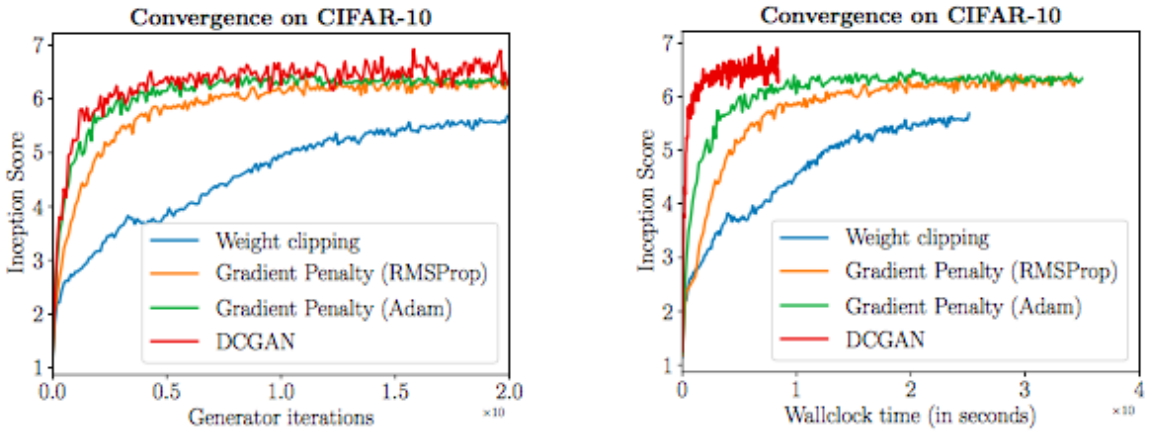


Figure 3: CIFAR-10 Inception score over generator iterations (left) or wall-clock time (right) for four models: WGAN with weight clipping, WGAN-GP with RMSProp and Adam (to control for the optimizer), and DCGAN. WGAN-GP significantly outperforms weight clipping and performs comparably to DCGAN.

План

Сравнение с другими архитектурами на CIFAR-10

Table 3: Inception scores on CIFAR-10. Our unsupervised model achieves state-of-the-art performance, and our conditional model outperforms all others except SGAN.

Unsupervised		Supervised	
Method	Score	Method	Score
ALI [8] (in [27])	$5.34 \pm .05$	SteinGAN [26]	6.35
BEGAN [4]	5.62	DCGAN (with labels, in [26])	6.58
DCGAN [22] (in [11])	$6.16 \pm .07$	Improved GAN [23]	$8.09 \pm .07$
Improved GAN (-L+HA) [23]	$6.86 \pm .06$	AC-GAN [20]	$8.25 \pm .07$
EGAN-Ent-VI [7]	$7.07 \pm .10$	SGAN-no-joint [11]	$8.37 \pm .08$
DFM [27]	$7.72 \pm .13$	WGAN-GP ResNet (ours)	$8.42 \pm .10$
WGAN-GP ResNet (ours)	$7.86 \pm .07$	**SGAN** [11]	$8.59 \pm .12$

План

Корреляция с качеством

Важное преимущество WGAN - корреляция с качеством генерируемой выборки, чтобы показать, что это свойство сохраняется, был обучен WGAN-GP на LSUM badrooms

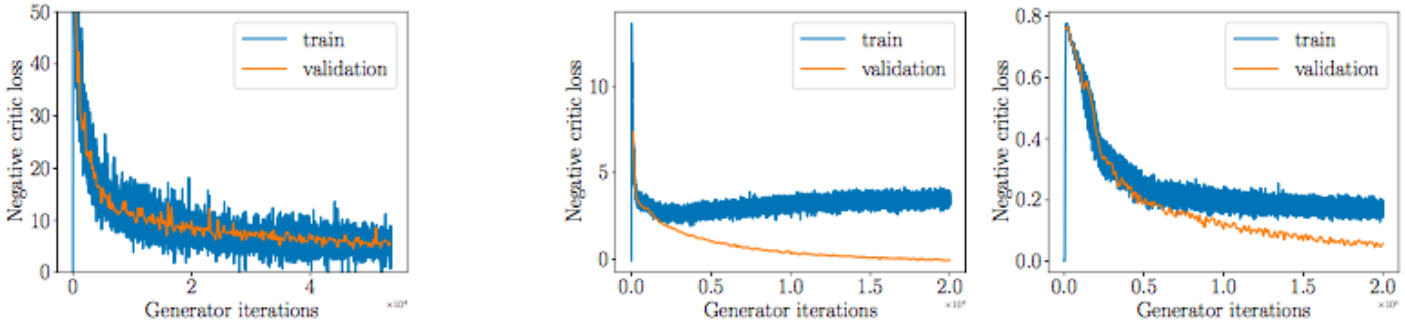


Figure 5: (a) The negative critic loss of our model on LSUN bedrooms converges toward a minimum as the network trains. (b) WGAN training and validation losses on a random 1000-digit subset of MNIST show overfitting when using either our method (left) or weight clipping (right). In particular, with our method, the critic overfits faster than the generator, causing the training loss to increase gradually over time even as the validation loss drops.

Литература

- Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville. Improved Training of Wasserstein GANs. *arXiv* : 1704.00028, 2017
- Shane Barratt, Rishi Sharma. A Note on the Inception Score. *arXiv* : 1801.01973, 2018
- Augustus Odena, Christopher Olah, Jonathon Shlens. Conditional Image Synthesis with Auxiliary Classifier GANs. *arXiv* : 1801.01973, 2017

Properties

Proposition 1. *Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be two distributions in $\mathcal{X}$, a compact metric space. Then, there is a 1-Lipschitz function f^* which is the optimal solution of $\max_{\|f\|_L \leq 1} \mathbb{E}_{y \sim \mathbb{P}_r}[f(y)] - \mathbb{E}_{x \sim \mathbb{P}_g}[f(x)]$. Let π be the optimal coupling between $\mathbb{P}_r$ and $\mathbb{P}_g$, defined as the minimizer of: $W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\pi \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y) \sim \pi}[\|x - y\|]$ where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ is the set of joint distributions $\pi(x, y)$ whose marginals are $\mathbb{P}_r$ and $\mathbb{P}_g$, respectively. Then, if f^* is differentiable[‡], $\pi(x = y) = 0$[§], and $x_t = tx + (1-t)y$ with $0 \leq t \leq 1$, it holds that $\mathbb{P}_{(x,y) \sim \pi}\left[\nabla f^*(x_t) = \frac{y - x_t}{\|y - x_t\|}\right] = 1$.*

Corollary 1. *f^* has gradient norm 1 almost everywhere under $\mathbb{P}_r$ and $\mathbb{P}_g$.*

Supervised GAN

3. AC-GANs

We propose a variant of the GAN architecture which we call an auxiliary classifier GAN (or AC-GAN). In the AC-GAN, every generated sample has a corresponding class label, $c \sim p_c$ in addition to the noise z. G uses both to generate images $X_{fake} = G(c, z)$. The discriminator gives both a probability distribution over sources and a probability distribution over the class labels, $P(S \mid X)$, $P(C \mid X) = D(X)$. The objective function has two parts: the log-likelihood of the correct source, L_S, and the log-likelihood of the correct class, L_C.

$$L_S = E[\log P(S = real \mid X_{real})] + E[\log P(S = fake \mid X_{fake})] \quad (2)$$

$$L_C = E[\log P(C = c \mid X_{real})] + E[\log P(C = c \mid X_{fake})] \quad (3)$$

D is trained to maximize $L_S + L_C$ while G is trained to maximize $L_C - L_S$. AC-GANs learn a representation for z that is independent of class label (e.g. (Kingma et al., 2014)).

Inception score

The Inception Score is a metric for automatically evaluating the quality of image generative models [Salimans et al., 2016]. This metric was shown to correlate well with human scoring of the realism of generated images from the CIFAR-10 dataset. The IS uses an Inception v3 Network pre-trained on ImageNet and calculates a statistic of the network's outputs when applied to generated images.

$$\mathrm{IS}(G) = \exp\left(\mathbb{E}_{\mathbf{x} \sim p_g} D_{KL}(p(y|\mathbf{x}) \parallel p(y)) \right), \tag{1}$$

where $\mathbf{x} \sim p_g$ indicates that $\mathbf{x}$ is an image sampled from p_g, $D_{KL}(p\|q)$ is the KL-divergence between the distributions p and q, $p(y|\mathbf{x})$ is the conditional class distribution, and $p(y) = \int_{\mathbf{x}} p(y|\mathbf{x}) p_g(\mathbf{x})$ is the marginal class distribution. The exp in the expression is there to make the values easier to compare, so it will be ignored and we will use $\ln(\mathrm{IS}(G))$ without loss of generality.

The authors who proposed the IS aimed to codify two desirable qualities of a generative model into a metric:

1. The images generated should contain clear objects (i.e. the images are sharp rather than blurry), or $p(y|\mathbf{x})$ should be low entropy. In other words, the Inception Network should be highly confident there is a single object in the image.
2. The generative algorithm should output a high diversity of images from all the different classes in ImageNet, or $p(y)$ should be high entropy.

If both of these traits are satisfied by a generative model, then we expect a large KL-divergence between the distributions $p(y)$ and $p(y|x)$, resulting in a large IS.

Качество на LSUN bedrooms

Сгенерированные изображения на LSUN bedrooms:

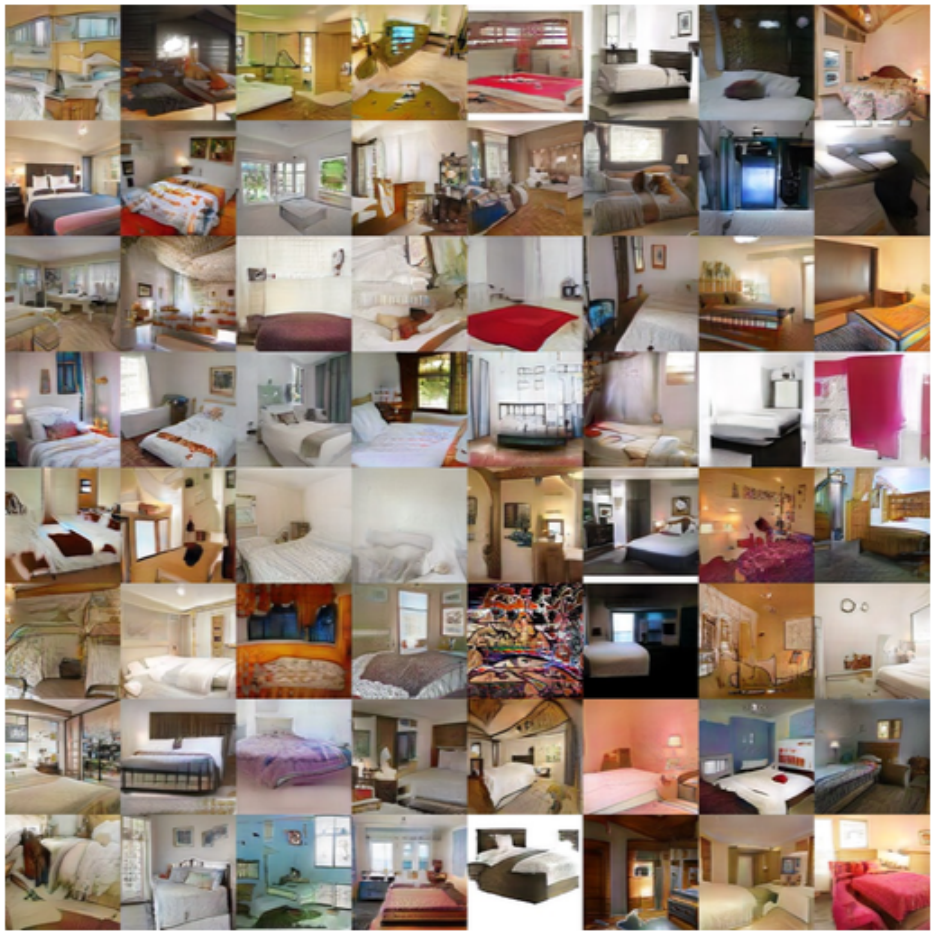


Figure 4: Samples of 128×128 LSUN bedrooms. We believe these samples are at least comparable to the best published results so far.