KL-GAN: метод обучения состязательных нейронных сетей посредством сопоставления распределений признаков реальных и сгенерированных данных

Е. Д. Сауткин
ООО «ИТ-ЛИДЕР»
evgenijsautkin29@gmail.com

А. Б. Цветков
Сибирский государственный индустриальный университет
atsvet@mail.ru

***Аннотация*. В данной работе предлагается новый метод KL-GAN, предназначенный для обучения генеративных состязательных нейронных сетей. В отличие от аналогов, метод построен на сопоставлении распределений признаков посредством дивергенции Кульбака–Лейблера. Это позволяет одновременно обновлять параметры генератора и дискриминатора за один шаг обратного распространения ошибки, что делает его альтернативой классическим методам. Эксперименты на датасете CelebA показывают, что KL-GAN конкурентоспособен по метрике FID, а в ряде случаев превосходит классические подходы.**

Ключевые слова: генеративные состязательные сети; KL-дивергенция; мини-батч дискриминация; нейронные сети; глубокое обучение

I. Введение

Генеративные состязательные сети (GAN) представляют собой класс методов, позволяющих моделировать сложные распределения данных путём сопоставления двух сетей — генератора и дискриминатора. Генератор стремится вырабатывать «реалистичные» сэмплы, а дискриминатор пытается отличить их от настоящих. Однако классическая постановка обучения, сформулированная в [1], сводится к бинарной классификации, в которой дискриминатор разделяет реальные данные и данные, созданные генератором, что часто приводит к ряду проблем: неустойчивости обучения, исчезающим градиентам и модовому коллапсу.

Многочисленные модификации, такие как LS-GAN [2], Hinge-GAN (известный также как Geometric GAN) [3], WGAN-GP [4], и R1-GAN [5], лишь частично решают эти проблемы, сохраняя базовую дихотомию в задаче дискриминатора. Несмотря на улучшения архитектур и функций потерь, необходим более принципиальный пересмотр самой формулировки задачи обучения GAN.

В данной работе предлагается KL-GAN — метод обучения, который принципиально отличается от классического. Его ключевая идея — отказаться от классификации отдельных образцов и вместо этого сопоставлять статистические характеристики распределений признаковых векторов для реальных и сгенерированных данных. В качестве меры расхождения этих распределений используется дивергенция Кульбака–Лейблера (KL) [6]. Эксперименты с датасетом CelebA (32 x 32) [7] показали, что предложенный подход как минимум не уступает ряду классических алгоритмов в метрике FID [8], а при применении техники Minibatch Discrimination [9] существенно опережает их. Результаты исследования могут внести вклад в разработку более стабильных генеративных моделей и преодоление фундаментальных ограничений существующих подходов к обучению GAN.

II. Обзор литературы

A. *Классические постановки GAN*

Появилось множество модификаций GAN, направленных на улучшение сходимости и качества генерируемых данных. Так, LS-GAN вводит квадратичную функцию потерь для смягчения жёсткой логарифмической функции, Hinge-GAN использует SVM-подобную функцию потерь (hinge loss), основанную на принципах метода опорных векторов, а WGAN-GP применяет приближение расстояния Васерштейна.

Для повышения устойчивости обучения дискриминатора используется спектральная нормализация [10], а также различные формы липшицевой регуляризации, например R1-GAN. Параллельно велась работа над улучшением архитектур (DCGAN [11], StyleGAN [12], BigGAN [13] и т. д.). Однако общим для большинства модификаций остаётся концепция бинарной классификации: дискриминатор определяет вероятность того, что образец принадлежит распределению реальных данных.

B. *Переосмысление процесса обучения*

В ряде исследований пытались уйти от бинарной классификации в сторону прямого сопоставления распределений. Так, некоторые работы предлагают минимизировать критерий MMD или другие f-дивергенции [14]. В частности, подходы, использующие KL-дивергенцию, часто рассматриваются в контексте вариационного вывода (VAE) или теоретических обобщений GAN [15]. Однако существующие варианты не фокусируются на прямом сравнении распределений признаковых векторов.

Предлагаемый метод KL-GAN смещает акцент на статистические характеристики признаков, извлекаемых дискриминатором. Дискриминатор в данном подходе не

классифицирует отдельные образцы, а формирует вектор признаков, статистические характеристики которого (средние, дисперсии) сопоставляются между реальными и сгенерированными данными. Такой подход обеспечивает более плавное выравнивание распределений без необходимости в дискретных решениях о принадлежности каждого образца, что потенциально смягчает проблемы «мертвых зон» и исчезающих градиентов при сильном доминировании дискриминатора.

III. Метод KL-GAN

A. *Основные обозначения*

Пусть в одном мини-батче собраны n реальных образцов и n сгенерированных:

$$\{x_i^{\text{real}}\}_{i=1}^n, \quad \{x_i^{\text{gen}} = G(z_i)\}_{i=1}^n, \quad z_i \sim p_z.$$

где x_i^{real} — реальные образцы из набора данных, x_i^{gen} — сгенерированные образцы, G — генератор, z_i — случайные векторы из распределения p_z.

Дискриминатор $D(x)$ сопоставляет входному изображению x вектор признаков $\mathbf{f} \in \mathbb{R}^k$, где k — размерность пространства признаков.

B. *Оценка признаковых распределений*

Для аналитического вычисления KL-дивергенции между распределениями признаков предполагается, что компоненты вектора f независимы и их распределения могут быть аппроксимированы нормальными. В статье предлагается вычислять векторы средних значений и дисперсий для реальных и сгенерированных данных по всем примерам в мини-батче:

$$\boldsymbol{\mu}_{\text{real}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i^{\text{real}}, \quad \boldsymbol{\mu}_{\text{gen}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i^{\text{gen}},$$

$$\boldsymbol{\sigma}^2_{\text{real}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{f}_i^{\text{real}} - \boldsymbol{\mu}_{\text{real}})^2 + \text{т}, \quad \boldsymbol{\sigma}^2_{\text{gen}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{f}_i^{\text{gen}} - \boldsymbol{\mu}_{\text{gen}})^2 + \text{т},$$

где $\boldsymbol{\mu}_{\text{real}}$, $\boldsymbol{\mu}_{\text{gen}}$ — векторы средних значений признаков для реальных и сгенерированных данных соответственно, $\mathbf{f}_i^{\text{real}}$, $\mathbf{f}_i^{\text{gen}}$ — векторы признаков для i-го реального и сгенерированного образца соответственно, $\boldsymbol{\sigma}^2_{\text{real}}$, $\boldsymbol{\sigma}^2_{\text{gen}}$ — векторы дисперсий признаков, а ϵ – малая константа для численной стабильности. Операции возведения в квадрат и суммирования выполняются покомпонентно.

C. *KL-дивергенция*

Для реальных и сгенерированных признаков вычисляются статистики, и предполагается, что признаки распределены согласно многомерным нормальным распределениям и с диагональными ковариационными матрицами:

$$P = \mathrm{N}(\boldsymbol{\mu}_{\text{real}}, \text{diag}(\boldsymbol{\sigma}^2_{\text{real}})), \quad Q = \mathrm{N}(\boldsymbol{\mu}_{\text{gen}}, \text{diag}(\boldsymbol{\sigma}^2_{\text{gen}})),$$

где P и Q — многомерные нормальные распределения с векторами средних $\boldsymbol{\mu}_{\text{real}}$ и $\boldsymbol{\mu}_{\text{gen}}$, $\text{diag}(\boldsymbol{\sigma}^2_{\text{real}})$ и $\text{diag}(\boldsymbol{\sigma}^2_{\text{gen}})$ — диагональные ковариационные матрицы соответственно.

Далее KL-дивергенция между такими распределениями вычисляется с помощью известной формулы [6] и для обеспечения устойчивости градиентов при обучении нейронной сети применяется логарифмирование:

$$\mathrm{KL}^*(P \parallel Q) = \ln(1 + \mathrm{KL}(P \parallel Q)),$$
$$\mathrm{KL}^*(Q \parallel P) = \ln(1 + \mathrm{KL}(Q \parallel P)).$$

где $\mathrm{KL}(P \parallel Q)$ и $\mathrm{KL}(Q \parallel P)$ — стандартные KL-дивергенции между распределениями P и Q в обоих направлениях, а $\mathrm{KL}^*(P \parallel Q)$ и $\mathrm{KL}^*(Q \parallel P)$ — их сглаженные версии.

Так как KL-дивергенция не является симметричной, то вычисляется среднее полученных KL-дивергенций и определяется величина для построения функции потерь:

$$\mathrm{SKL}(P, Q) = \frac{1}{2} \cdot \left[\mathrm{KL}^*(P \parallel Q) + \mathrm{KL}^*(Q \parallel P)\right].$$

Функция потерь для обучения имеет вид:

$$\mathrm{L}(\theta_G, \theta_D) = \mathrm{SKL}(P_{\theta_D}, Q_{\theta_G, \theta_D}), \qquad (1)$$

где P_{θ_D} и Q_{θ_G, θ_D} – распределения признаков реальных и сгенерированных данных, зависящие от параметров дискриминатора θ_D и генератора θ_G.

D. *Одновременное обновление*

Классическая постановка GAN предполагает отдельные шаги обновления для дискриминатора и генератора. В KL-GAN введён одношаговый (совместный) подход:

- Формируется мини-батч: реальные (x_i^{real}) и сгенерированные (x_i^{gen}) данные.
- Дискриминатор вычисляет признаки для реальных и сгенерированных примеров.
- Для каждого набора признаков вычисляются векторы средних значений и дисперсий.
- Моделируются распределения признаков P и Q с использованием полученных статистик.
- Вычисляются KL-дивергенции в обоих направлениях (P→Q и Q→P), к каждой из них применяется логарифмическое сглаживание, а затем вычисляется симметричная KL-дивергенция как среднее этих сглаженных значений.
- Определяется функция потерь $\mathrm{L}(\theta_G, \theta_D)$ согласно формуле (1).
- Оптимизационная задача формулируется как минимаксная игра:

$$\min_{\theta_D} \max_{\theta_G} -\mathrm{L}(\theta_G, \theta_D).$$

- Делается один шаг обратного распространения ошибки, после чего параметры дискриминатора сдвигаются в направлении максимизации $\mathrm{L}(\theta_G, \theta_D)$, а параметры генератора — в направлении её минимизации. Технически это достигается инвертированием знака градиента для параметров генератора:

$$\theta_D \leftarrow \theta_D - \eta \cdot \frac{\partial(-\mathrm{L})}{\partial \theta_D}; \quad \theta_G \leftarrow \theta_G + \eta \cdot \frac{\partial(-\mathrm{L})}{\partial \theta_G}.$$

Таким образом, генератор использует актуальные параметры дискриминатора при каждом обновлении, что облегчает реализацию и может положительно повлиять на процесс обучения.

IV. Эксперименты

A. *Настройки*

Для проверки эффективности KL-GAN использовался датасет CelebA, уменьшенный до разрешения 32 x 32. Обучение проводилось в течение 300 эпох, размер мини-батча равнялся 1024, латентный вектор имел размер 128. Сравнивались KL-GAN, LS-GAN, Hinge-GAN, WGAN-GP и R1-GAN. Метрикой качества генерации выступала Fréchet Inception Distance (FID), вычислявшаяся по 10000 образцам, сгенерированным каждой моделью.

B. *Техника Minibatch Discrimination*

Одной из ключевых техник, примененных в экспериментах, является Minibatch Discrimination. Эта техника помогает бороться с проблемой коллапса мод путем анализа близости образцов в пространстве признаков дискриминатора. Для каждого образца в мини-батче вычисляется вектор сходств с другими образцами того же мини-батча. Это позволяет дискриминатору отличать ситуации, когда генератор производит однообразные образцы, тем самым стимулируя разнообразие генерируемых данных и препятствуя схлопыванию распределения. В нашей реализации использовалась стандартная архитектура Minibatch Discrimination, описанная в оригинальной работе [9].

C. *Результаты*

В таблице I и на рис. 1 показаны усреднённые по нескольким запускам значения FID. Из анализа рис. 1 следует то, что KL-GAN дает лучшие результаты по FID по сравнению с другими методами. На рис. 2 приведены примеры изображений, сгенерированных каждым из методов.

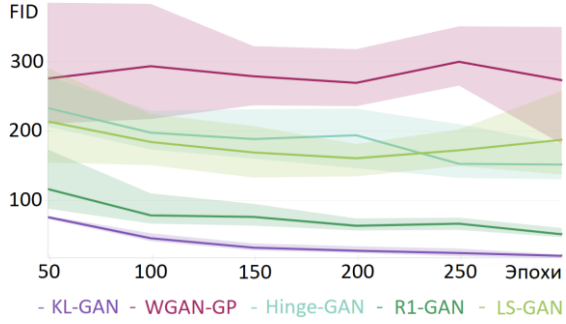

Рис. 1. Графики динамики FID при обучении

Рис. 2. Пример сравнения сгенерированных изображений предлагаемым методом KL-GAN с известными методами R1-GAN [5], Hinge-GAN [3], WGAN-GP [4] и LS-GAN [2]

ТАБЛИЦА I. СРЕДНИЕ FID (ПО 5 ЗАПУСКОВ) НА CELEBA (32 X 32, 300 ЭПОХ)

Метод	Без Minibatch Discrimination	С Minibatch Discrimination
R1-GAN	51.5 ± 5.4	92.1 ± 41.9
Hinge-GAN	341.7 ± 45.6	151.9 ± 20.0
WGAN-GP	351.3 ± 21.6	273.6 ± 71.8
LS-GAN	340.6 ± 37.8	187.6 ± 47.5
KL-GAN (наш)	237.4 ± 102.8	20.2 ± 2.0

- Без Minibatch Discrimination: R1-GAN дал лучший результат среди классических (примерно 51.5 ± 5.4), KL-GAN был на уровне порядка 237.4 ± 102.8, что хуже R1-GAN, но лучше, чем Hinge-GAN, WGAN-GP и LS-GAN (их FID превышал 300).
- С Minibatch Discrimination: KL-GAN значительно улучшил результат до ~20.2 (± 2.0), существенно превзойдя все указанные методы.

Таким образом, KL-GAN без специальных техник остаётся как минимум конкурентоспособным, однако с добавлением Minibatch Discrimination существенно опережает более «классические» подходы по метрике FID. Интересно отметить, что для R1-GAN добавление Minibatch Discrimination, наоборот, ухудшило результаты (с 51.5 до 92.1). Это наблюдение указывает на то, что техника градиентной регуляризации в R1-GAN может не согласовываться оптимально с дополнительными признаками, вводимыми через Minibatch Discrimination, либо гиперпараметры объединенной модели требуют

специфической настройки при таком сочетании регуляризационных подходов.

V. Заключение и направления для дальнейших исследований

В данной работе представлен KL-GAN — альтернативный метод обучения GAN, в который позволяет непосредственно сопоставлять статистические характеристики распределений признаков для реальных и сгенерированных данных. Этот подход реализуется через дивергенцию Кульбака–Лейблера и одношаговое обновление параметров генератора и дискриминатора. Эксперименты на CelebA показывают, что KL-GAN в некоторых случаях превосходит классические методы по метрике FID, особенно при использовании Minibatch Discrimination.

Дальнейшие исследования могут включать:

- Тестирование KL-GAN на более крупных разрешениях (например, 128 x 128 и выше) и современных архитектурах (StyleGAN, BigGAN).
- Исследование возможности интеграции прогрессивного роста [16] в архитектуру KL-GAN для улучшения стабильности генерации изображений высокого разрешения.
- Сравнение KL с другими метриками сходства распределений (MMD, α-дивергенции, JS-дивергенция) в контексте выравнивания признаков.
- Углублённый анализ сходимости, в том числе формальные эксперименты по изучению поведения алгоритма в различных режимах настройки гиперпараметров.

Таким образом, KL-GAN раскрывает потенциал переосмысления задачи обучения GAN через сопоставление статистических характеристик распределений признаков вместо классификации отдельных образцов. Полученные результаты подтверждают, что этот подход конкурентоспособен по качеству и заслуживает дальнейшей проработки.

Список литературы

[1] Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative Adversarial Nets // Advances in Neural Information Processing Systems, 2014, vol. 27, pp. 2672–2680.

[2] Mao X., Li Q., Xie H., Lau R. Y. K., Wang Z., Smolley S. P. Least Squares Generative Adversarial Networks // IEEE International Conference on Computer Vision, 2017, pp. 2813–2821.

[3] Lim J. H., Ye J. C. Geometric GAN // arXiv preprint arXiv:1705.02894, 2017.

[4] Gulrajani I., Ahmed F., Arjovsky M., Dumoulin V., Courville A. Improved Training of Wasserstein GANs // Advances in Neural Information Processing Systems, 2017, vol. 30, pp. 5767–5777.

[5] Mescheder L., Geiger A., Nowozin S. Which Training Methods for GANs Actually Converge? // International Conference on Machine Learning, 2018, pp. 3481–3490.

[6] Bishop C. M. Pattern Recognition and Machine Learning. – New York : Springer, 2006. – 738 p.

[7] Liu Z., Luo P., Wang X., Tang X. Deep Learning Face Attributes in the Wild // IEEE International Conference on Computer Vision, 2015, pp. 3730–3738.

[8] Heusel M., Ramsauer H., Unterthiner T., Nessler B., Hochreiter S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium // Advances in Neural Information Processing Systems, 2017, vol. 30, pp. 6626–6637.

[9] Salimans T., Goodfellow I. J., Zaremba W., Cheung V., Radford A., Chen X. Improved Techniques for Training GANs // Advances in Neural Information Processing Systems, 2016, vol. 29, pp. 2226–2234.

[10] Miyato T., Kataoka T., Koyama M., Yoshida Y. Spectral Normalization for Generative Adversarial Networks // International Conference on Learning Representations, 2018.

[11] Radford A., Metz L., Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks // International Conference on Learning Representations, 2016.

[12] Karras T., Laine S., Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks // IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4396–4405.

[13] Brock A., Donahue J., Simonyan K. Large Scale GAN Training for High Fidelity Natural Image Synthesis // International Conference on Learning Representations, 2019.

[14] Li C.-L., Chang W.-C., Cheng Y., Yang Y., Póczos B. MMD GAN: Towards Deeper Understanding of Moment Matching Network // Advances in Neural Information Processing Systems, 2017, vol. 30, pp. 2203–2213.

[15] Nowozin S., Cseke B., Tomioka R. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization // Advances in Neural Information Processing Systems, 2016, vol. 29, pp. 271–279.

[16] Karras T., Aila T., Laine S., Lehtinen J. Progressive Growing of GANs for Improved Quality, Stability, and Variation // International Conference on Learning Representations, 2018.