*Кожин А.Р.*

*Вологодский государственный университет*

*Kozhin A.R.*

*Vologda state university*

**АНАЛИЗ КОЛИЧЕСТВЕННЫХ МЕТРИК ОЦЕНКИ ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНЫХ СЕТЕЙ**

**ANALYSIS OF QUANTITATIVE METRICS FOR EVALUATING GENERATIVE-ADVERSARIAL NETWORKS**

Аннотация

В этой статье мы рассматриваем 5 количественных метрик для оценки генеративно-состязательных моделей, углубляясь в идеи, лежащие в основе метрик и в их математические формулы. Провели анализ сильных и слабых сторон метрик и вывели ряд гипотез по корреляцию метрик. В конце статьи представили выводы о подходе к оценке генеративно-созтязательных моделей.

Ключевые Слова: Генеративно-состязательные сети, генеративные модели, оценки, глубокое обучение, нейронные сети

In this article, we look at 5 quantitative metrics for evaluating generative-adversarial models, delving into the ideas behind the metrics and their mathematical formulas. We analyzed the strengths and weaknesses of metrics and derived a number of hypotheses on the correlation of metrics. At the end of the article, conclusions were presented on the approach to evaluating generative-adversarial models.

Keywords: Generative-adversarial networks, generative models, estimates, deep learning, neural networks

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. **Введение**

Оценка генеративно-состязательных сетей – нетривиальная задача, которая может быть рассмотрена с разных сторон. Самый простой способ – подвергнуть сгенерированную информацию критике экспертов и посмотреть, насколько их оценка удовлетворяет потребностям. Однако, здесь имеет место субъективный взгляд. В попытках формализовать оценку генеративных сетей, были придуманы количественные оценки, отражающие качество модели через какое-то число. Дальше мы рассматриваем некоторые количественные методы оценки моделей.

1. ***Метрики оценки генеративно-состязательных сетей***

**Inception Score (IS).** Предложенная Salimans и др. [1], возможно, одна из широко используемых методов оценки генеративно-состязательных сетей (GAN) [2]. IS использует обученную нейронную сеть Inception Net [3], обученную на ImageNet [4]. Основная идея, лежащая в основе IS – мы хотим понять, насколько разнообразные данные может генерировать наша модель, а также насколько “конкретные” вещи генерируются. Следовательно, в идеальной ситуации мы хотим получить выполнение следующих пунктов:

1) Все классы хорошо представлены – разнообразие. Если мы сгенерировали много объектов, то мы хотим, чтобы сумма распределений каждого объекта выборки была примерно равномерным распределением.

2) Каждый отдельный объект имеет четкий класс – “конкретность”. Распределение вероятности для отдельного примера должно быть близко к δ-функции.

Выражая пункты 1 и 2 в математическом эквиваленте, получим:

1) , q – сгенерированные данные;

2) ;

Объединяя эту пару условий и максимизируя её получаем итоговую формулу для вычисления IS:

,

где – дивергенция Кульбека-Лейбнера. Её можно представить в виде суммы энтропии и кросс-энтропии: .

Чем больше IS(q), тем более разнообразные и более “конкретные” наши предсказания. Если хотя бы одно из условий не выполняется, тогда значение IS(q) будет маленьким. Тем самым Inception score демонстрирует разумную корреляцию между разнообразием и “конкретностью” сгенерированных изображений. Так же из определения КL дивергенции можно определить множество значений, которые принимает Inception Score: . Важно заметить, что IS использует только сгенерированные данные и никак не использует реальные данные. Однако, вычислив IS над реальными данными можно найти верхнюю границу для IS над сгенерированными.

Несмотря на очевидные плюсы, Inception score имеет свои ограничения и недостатки:

1) Не может ловить переобучение (Очевидно, что генерирующая модель, которая запоминает соответствующее подмножество обучающей выборки, будет работать чрезвычайно хорошо с точки зрения IS [5]).

2) Не замечает проблем с mode collapse (к примеру модель будет генерировать одно и тоже изображение для каждого класса, IS покажет хорошую оценку, но в идеале GAN должен уметь генерировать несколько разных изображений для одного класса).

3) Никак не использует в оценке сравнение реального распределения данных с сгенерированным.

4) Так как IS использует модель Inception, которая обучена на ImageNet с большим множеством классов, то модели, генерирующие понятные Inception изображения, а не реалистичные изображения с точки зрения человека.

5) KL дивергенция – ассиметричная мера

6) IS чувствительна к разрешению изображения. (см. рис. 1)

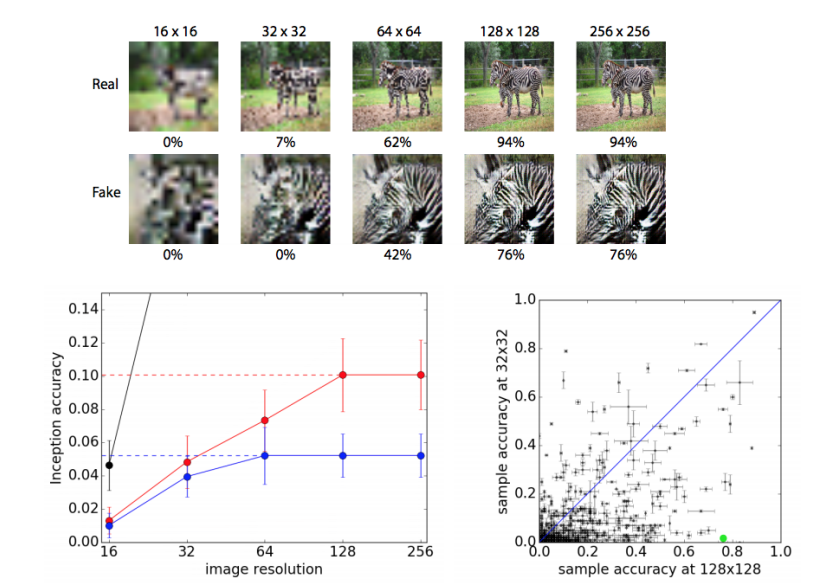


Рисунок 1: Чувствительность IS к разрешению изображения. Вверху: Размер реальных данных и сгенерированных изображений из класса зебры был изменен до 16 × 16 , а затем искусственно изменен до исходного разрешения (128 × 128 для красной и черной линий; 64 × 64 для синей линии). Слева внизу: оценка по различным пространственным разрешениям для обучающих данных и выборок изображений из моделей 64 × 64 и 128 × 128. Столбцы ошибок показывают стандартное отклонение по 10 подмножествам изображений. Пунктирные линии указывают точность при выходном разрешении модели. Внизу справа: Сравнение показателей точности при разрешении 128 × 128 и 32 × 32. Каждая точка соответствует классу ImageNet. 84,4% классов расположены ниже диагонали. Зеленая точка соответствует классу зебры. Рисунок из [2].

**Fréchet Inception Distance (FID).** Представлена Heusel и др. [6]. В отличии от IS, в основе FID лежит идея сравнения распределения сгенерированных и реальных данных. Так как мы не можем знать реальное распределение данных, то будем рассматривать его эмпирическое приближение. В основном рассматривается расстояние между двумя нормальными распределениями, и используется соответствующая формула для вычисления расстояния Фреше:

.

Для того, чтобы связать FID с формулой вычисления расстояния Фреше рассмотрим Inception Net (или любую другую CNN). И вместо выходов нейросети в FID будем использовать выходы с предпоследнего слоя. Они кодируют высокоуровневую семантику изображений, потому что по этим признакам модель предсказывает вероятность принадлежности изображения к тому или иному классу. При этом в них остаётся довольно много информации об исходном изображении и свойств локальных признаков, которые могут помочь распознать артефакты. Таким образом метрика FID сравнивает два распределения высокоуровневых признаков для реальных и сгенерированных изображений, используя в качестве их приближения многомерные гауссианы[13]. Для реальных и сгенерированных данных вычисляется среднее и матрица ковариаций глубоких признаков. Подставляя в формулу вычисления расстояния Фреше получаем:

Чем меньше значение FID, тем распределения находятся ближе друг к другу. Так же из можно определить множество значений, которые принимает FID: . Внушительным плюсом, в отличии от IS, является хорошая корреляция с человеческой оценкой [6], из-за чего FID ­– самая популярная метрика качества генерации. На практике метрика используется как для измерения реализма изображения, так и для обнаружения mode collapse, чего не может обнаружить IS. Было доказано, что FID более устойчив к шуму, чем IS [6]. При создании только одного изображения для каждого класса может быть получен высокий балл IS, а FID будет принимать большие значения. Кроме того, в отличие от IS, значение FID увеличивается по мере добавления к изображениям различных типов артефактов (см. рис. 2).

Недостатки и ограничения FID:

1) FID чувствителен к размеру тестовой выборки. Небольшие наборы данных могут привести к нестабильным оценкам

2) Предполагается, что распределения реальных и сгенерированных данных можно аппроксимировать многомерными гауссовыми распределениями. Не всегда можно гарантировать выполнение этого условия.

3) Хотя FID и хорошо коррелирует с человеческой оценкой, есть случаи, когда это не выполняется [7].

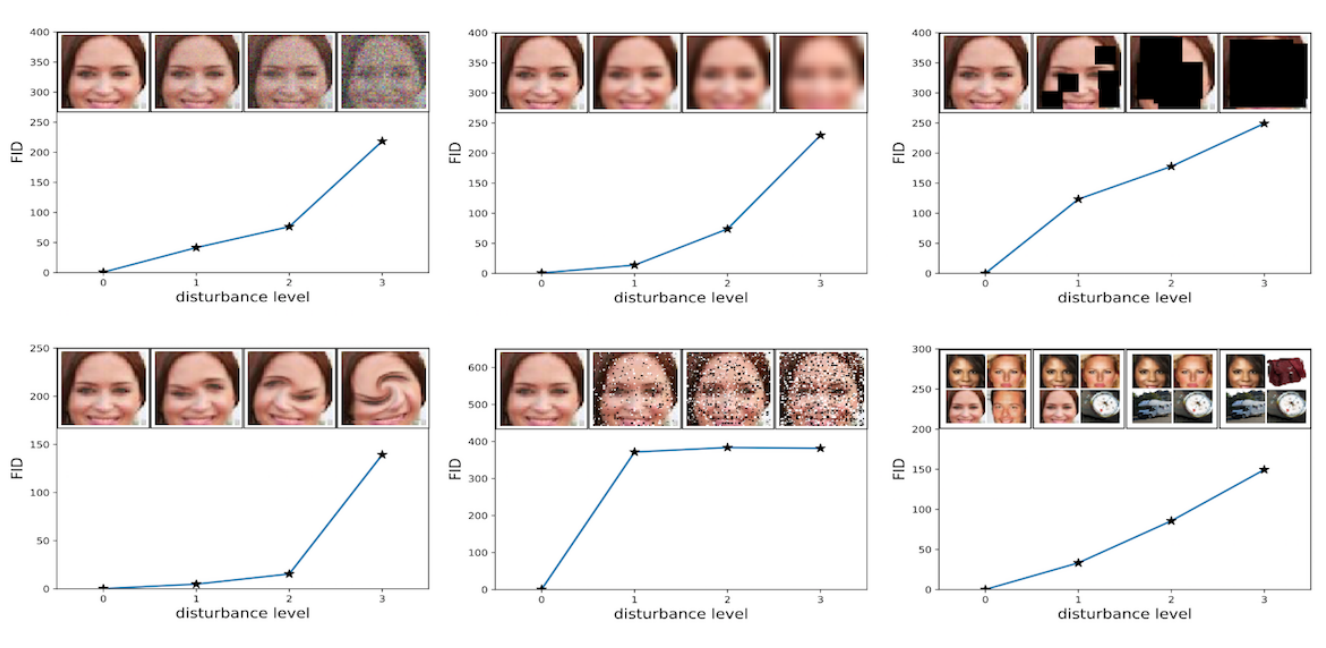
4)Не может ловить переобучение.

Рисунок 2: FID чувствителен к искажениям изображения. Слева направо: гауссовский шум, размытие по Гауссу, имплантированные черные прямоугольники, закрученные изображения, salt and pepper шум, а также набор данных CelebA, загрязненный изображениями ImageNet. Рисунок из [6].

**Kernel** **Maximum Mean Discrepancy (MMD).** Эта метрика вычисляет различие между двумя вероятностными распределениями и . MMD определяет норму разницы между их средними значениями в Гильбертовом пространстве. Меньшее значение MMD означает, что ближе к . Можно рассматривать Kernel MMD [7] как квадрат MMD между и с некоторой функций ядра k:

На практике для оценки расстояния MMD используются дискретные распределения. Сложность выборки и вычислительная сложность метрики крайне мала [8]. MMD может использоваться с различными ядрами, что позволяет адаптировать метрику под конкретные задачи и данные. Метрика напрямую работает с исходными данными, что позволяет избежать проблем, связанных с выбором модели для извлечения признаков, в отличии от предыдущих метрик. Как и FID, MMD чувствительна к различиям между распределениями данных. Значения MMD принадлежат тому же множеству, что и FID.

Недостатки и ограничения MMD:

1) Невозможно ловить переобучение.

2) Как и FID, MMD может быть чувствителен к размеру выборки, что может привести к нестабильным оценкам при малом количестве данных.

3) Эффективность MMD сильно зависит от выбора ядра.

**Wasserstein Critic** [9] обеспечивает аппроксимацию расстояния Вассерштейна между распределением реальных данных и распределению сгенерированных данных . В основе этой метрике лежит идея нахождения минимальной работы, требующейся для получения из одного распределения другое. Эта задача так же известна как нахождение Earth Mover’s Distance (EMD). Исходя из Kantorovich-Rubinstein duality [10] получаем:

,

где f – Липшицева функция с параметром 1 (Пусть , , если ).

Чем меньше значение , тем меньше расстояние между распределениями данных, следовательно, GAN генерирует хорошие изображения. принимает значения из того же множества, что и FID. The Wasserstein Critic направлена на обнаружение переобучения и mode collapse. Если генератор запоминает обучающую выборку, то метрика, обученная на реальных данных, может различать сгенерированные выборки и данные. Если происходит mode collapse, метрике будет легко отличить данные из реального распределения от выборки из сгенерированного распределения.

Ключевым недостатком является высокая выборочная и временная сложность (), что делает метрику Вассерштейна менее привлекательным в качестве практической оценки по сравнению с другими метриками [11].

**Number of Statistically-Different Bins (NDB).** Чтобы измерить разнообразие генерируемых данных и обнаруживать mode collapse, Richardson и Weiss [12] предложили метод оценки, основанный на следующей идее: Распределение данных разделяются на K ячеек, пусть даны два набора данных из одного и того же распределения, тогда количество данных, попадающих в данную ячейку, должно быть одинаковым с точностью до шума выборки.

Пусть – индикатор принадлежности ячейке B, p – реальное распределение, q – сгенерированное распределение. Пусть - множество данных из p, – количество элементов в этом множестве, - множество данных из q, – количество элементов в множестве. Тогда, если p совпадает с q, то ожидается:

ц

Пусть – количество данных из объединения множеств, попавших в ячейку B, деленное на количество элементов в ячейке, и стандартная ошибка вычисляется по следующей формулe: .

Тогда для статистической оценки используется , где - количество элементов из множества, попавших в ячейку B, деленное на количество элементов в B. Если меньше, чем какое-то пороговое значение (обычно рассматривается значение ), то ячейка будет *статически различной*. Реальные данные разделяются на K ячеек, сгенерированные данные связываются с ближайшей ячейкой по расстоянию в пространстве. Просчитывая для каждой ячейки, взяв их сумму и разделив на количество ячеек, получим Number of Statistically-Different Bins (NDB).

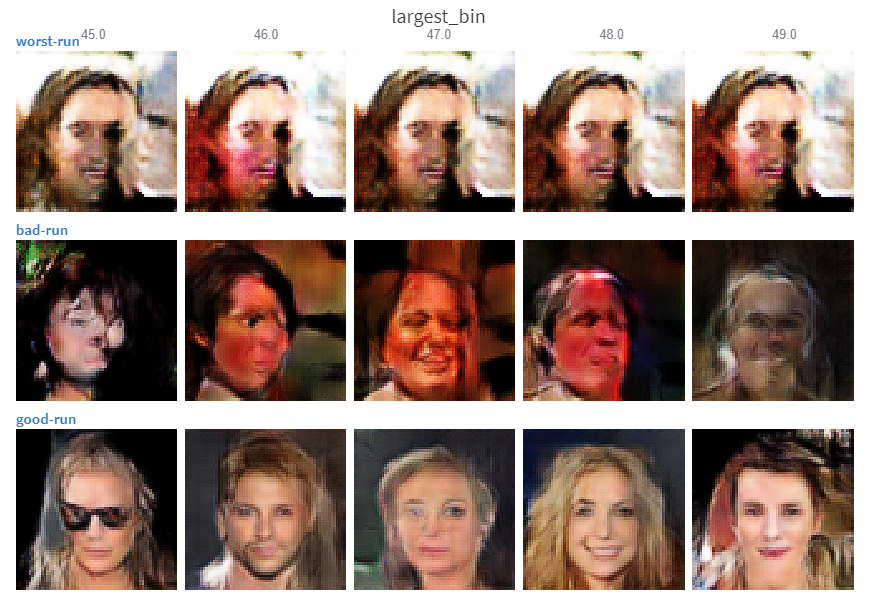
Маленько значение NDB интерпретируется хорошей работой генеративной сети, иначе – плохой. Принимает значение на отрезке . NDB чувствительна к различным артефактам изображения. Направлена на определение разнообразия изображений, в отличии от других метрик. 

Рисунок 3: слева сгенерированные изображения в разных случаях: сверху – mode collapse, по центру – частичный mode collapse, снизу -–без mode collapse. Справа показан график значений метрики NDB и их сходимость во всем случаях. [12]

Недостатки NDB:

1. Результаты NDB могут сильно зависеть от выбранных параметров, таких как количество ячеек и размер выборки. Неправильный выбор этих параметров может привести к искажению результатов.
2. При малом объеме данных NDB может давать нестабильные результаты, что затрудняет оценку качества генерации.
3. В зависимости от реализации, вычисление NDB может требовать значительных вычислительных ресурсов.

Стоит отметить, что метрика не отображает качества изображения, NDB рассчитана на обнаружения mode collapse. Из-за этого оценка может не соответствовать тому, как люди воспринимают качество изображений, что делает её менее надежной. Идеальный вариант будет использовать эту метрику вместе с другими метриками, оценивающими качество изображения.

1. **Сравнительный анализ метрик оценки GAN.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Метрика | Множество значений | Обнаружение переобучения | Обнаружение mode collapse | Чувствительность к артефактам | Вычислительная эффективность | Визуальная оценка |
| Inception Score [1] |  | - | - | - | Высокая | Высокая |
| FID [6] |  | - |  | + | Высокая | Высокая |
| Kernel MMD [7] |  | - |  | + | Очень высокая | Высокая |
| Wasserstein Critic [9] |  | - | + | - | Низкая | - |
| NDB [12] |  | - | + | - | Высокая | - |

Таблица 1: Общие сведения метрик оценки GAN. “-” – неизвестные данные. – возможно определение при визуальном анализе генерируемых изображений. Для ознакомления с полной таблицей см. [14]

Рассмотрев метрики оценки, можно попробовать выдвинуть гипотезы по поводу корреляции метрик, основываясь на идеях, лежащих в основе, и их математических интерпретаций.

- Исходя из того, что

1) метрики FID, Kernel MMD и Wasserstein Critic направлены на то, чтобы вычислить различия между реальным и сгенерированным распределением данных.

2) множество значений, принимаемое метриками, совпадает.

3) В идеальной ситуации метрики сходятся к нулю.

можно предположить, что эти три метрики взаимно коррелируют между собой.

- Метрика IS в идеальной ситуации стремиться к , в отличии от FID, Kernel MMD и Wasserstein Critic. Поэтому можно предположить, что IS имеет отрицательную корреляцию между этими метриками.

- Метрика NDB сильно отличается от других представленных метриках, однако, из того, что метрика в какой-то степени отображает разнообразие сгенерированных данных, можно предложить о положительной корреляции с IS. Аналогично, из возможности определения mode collapse, можно предположить о положительной корреляции между Wasserstein Critic. С остальными метриками ожидается отсутствие корреляции или её очень малое значение, из-за того, что они отражают не только разнообразие генерируемых изображений, но и их качество.

1. **Вывод.**

В этой работе мы провели анализ 5 метрик оценки генеративно-состязательных сетей, разобрали их сильные и слабые стороны, выдвинули гипотезы об корреляции представленных метрик. Исходя из анализа, получили, что существуют различные подходы к оценке работы генеративно-состязательных сетей: одни направлены на связь человеческого представления с каким-то числовым значением, а другие направлены на выявление проблем, которые могут возникнуть во время обучения (overfitting, mode collapse). Мы рекомендуем при работе с генеративными сетями использовать несколько метрик для оценки качества модели, чтобы получать большую информативность о поведении модели. Отдельно хочется выделить Kernel MMD и NDB. Эти метрики имеют приемлемую вычислительную и временную сложность и вместе могут хорошо отобрать качество модели.

В качестве будущей работы можно попробовать доказать или опровергнуть гипотезы, выдвинутые в 3 пункте.

**Список литературы**

[1] T. Salimans, W. Goodfellow, Ian gotand Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, in: Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.

[2] W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, I. Goodfellow, Many paths to equilibrium: Gans do not need to decrease adivergence at every step, arXiv preprint arXiv:1710.08446.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818– 2826.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 248–255.

[5] J. Yang, A. Kannan, D. Batra, D. Parikh, Lr-gan: Layered recursive generative adversarial networks for image generation, arXiv preprint arXiv:1703.01560.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in Neural Information Processing Systems, 2017, arXiv preprint arXiv:1706.08500.

[7] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, 2012, arXiv preprint arXiv:0805.2368

[8] G. Huang, Y. Yuan, Q. Xu, C. Guo, Y. Sun, F. Wu, K. Weinberger, An empirical study on evaluation metrics of generative adversarial networks, International Conference on Learning RepresentationsRejected. URL <https://openreview.net/forum?id=Sy1f0e-R->

[9] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, arXiv preprint arXiv:1701.07875.

[10] Rachev S.T. Duality theorems for Kantorovich-Rubinstein and Wasserstein functionals, 1990

[11] S. Arora, Y. Zhang, Do gans actually learn the distribution? an empirical study, arXiv preprint arXiv:1706.08224.

[12] E. Richardson, Y. Weiss, On gans and gmms, arXiv preprint arXiv:1805.12462.

[13] E. Zacharov, D. Volhonskiy, 8.3 Генеративно-состязательные сети (GAN), Яндекс Образование, URL <https://education.yandex.ru/handbook/ml/article/generativno-sostyazatelnye-seti-(gan)>

[14] Ali Borji, Pros and Cons of GAN Evaluation Measures, 2018, aeXiv preprint arXiv:1802.03446v5