Обзор метрик с целью оценки качества работы генеративных моделей для создания изображений

А.В. Катаев, Ю.М. Власова, Д.А. Гусынин, В.А. Ким

Волгоградский государственный технический университет

Аннотация: Статья представляет обзор метрик, используемых для оценки качества изображений, полученных с помощью генеративных моделей. Для них требуются специализированные метрики, позволяющие объективно оценивать качество изображений. Сравнительный анализ показал, что для комплексной оценки качества генерации необходимо сочетание различных метрик. Перцептивные метрики эффективны для оценки качества изображений с точки зрения машинных систем, в то время как метрики, оценивающие структуру и детали, полезны для анализа восприятия человеком. Метрики, основанные на текстовом описании, позволяют оценить соответствие изображений заданным текстам, однако они не могут заменить метрики, направленные на визуальную или структурную оценку. Результаты исследования будут полезны специалистам в области машинного обучения и компьютерного зрения, а также способствуют улучшению алгоритмов генерации и расширению областей применения диффузионных моделей.
Ключевые слова: глубокое обучение, метрика, генеративная модель, качество изображений, изображение.

Современные диффузионные модели демонстрируют прогресс в области генерации изображений, который позволяет использовать новые возможности при генерации изображений. Однако объективная оценка качества этих изображений остается актуальной задачей. В отличие от традиционных задач компьютерного зрения, где качество может быть измерено с помощью прямых метрик, например, точности классификации, оценка генеративных моделей требует более сложных подходов.

Для оценки генеративных моделей ранее активно использовались метрики, разработанные для оценки генеративных состязательных сетей, такие как индекс Inception и расстояние Фреше между распределениями признаков (Frechet Inception Distance - FID) [1]. Однако диффузионные модели обладают иной архитектурой и динамикой генерации, что делает некоторые из этих метрик менее применимыми. В отличие от генеративных состязательных сетей, диффузионные модели позволяют контролировать

процесс генерации через условия, например, текстовые описания, маски или исходные изображения. Это требует разработки и адаптации метрик, способных учитывать не только реалистичность сгенерированных изображений, но и их соответствие заданным условиям [2-4].

Данная статья представляет обзор существующих метрик, применяемых для оценки качества изображений, сгенерированных диффузионными моделями. В работе рассматриваются принципы работы этих метрик, их применимость к различным сценариям генерации.

Особенности генерации диффузионных моделей

Обучение и генерация изображений с помощью диффузионных моделей отличаются от традиционных генеративных подходов. В основе этих различий лежат особенности архитектур моделей и их процессы генерации [2, 3]. Диффузионные модели используют итеративный процесс, в котором изображения постепенно разрушаются путем добавления гауссовского шума, а затем восстанавливаются с помощью обученного нейросетевого преобразования, предсказывающего и удаляющего шум на каждом шаге [2]. В отличие от генеративных состязательных сетей, где генерация происходит за один шаг через трансформацию случайного латентного вектора с помощью сверточной нейросети, диффузионные модели требуют выполнения многократных итераций восстановления изображения [3].

Ключевым преимуществом диффузионных моделей является их устойчивость к коллапсу моды. В генеративных состязательных сетях обучение происходит в форме соперничества между генератором и дискриминатором, что часто приводит к ситуации, когда генератор начинает воспроизводить лишь ограниченное подмножество образцов, игнорируя разнообразие данных. В диффузионных моделях генерация основана на моделировании вероятностного перехода между распределениями, что

обеспечивает более равномерное покрытие множества возможных решений и улучшает воспроизведение редких или сложных деталей [3, 5].

Диффузионные модели обладают высокой гибкостью в управлении процессом генерации. В отличие от генеративных состязательных сетей, где управление контентом происходит через модификацию латентного пространства, диффузионные модели позволяют добавлять условия на каждом шаге генерации. Следовательно, при оценке качества работы диффузионных моделей необходимо учитывать не только реалистичность изображений, но и степень их соответствия заданным условиям [4].

Традиционные метрики, такие как FID, недостаточны при анализе диффузионных моделей. В связи с этим в исследованиях используются альтернативные метрики, включая оценку CLIP для семантического соответствия и метрику обученного перцептивного сходства фрагментов изображений (Learned Perceptual Image Patch Similarity - LPIPS) [6, 7].

Метрики для оценки качества генеративных моделей можно разделить на две категории: основанные на изображении и основанные на тексте.

Метрики, основанные на изображении

Метрики, основанные на изображении, оценивают качество изображений с точки зрения визуальных характеристик, не учитывая текстовые условия генерации. Эти метрики делятся на перцептивные, оценивающие визуальное качество изображения, сравнивая с реальными образцами, и метрики оценки структуры и деталей, которые измеряют сохранение локальных и глобальных структурных характеристик [8].

Перцептивные метрики оценивают, насколько сгенерированные изображения визуально близки к реальным, основываясь на сравнении распределений признаков, извлеченных нейросетевыми моделями.

FID вычисляет расстояние между распределениями признаков, извлеченных из реальных и сгенерированных изображений с помощью сети Inception. Чем ниже значение FID, тем выше качество генерации. Метрика вычисляется как расстояние Фреше между двумя многомерными гауссовскими распределениями (1) [1]:

$$FID = \left\| \mu_r - \mu_g \right\|^2 + Tr\left(\Sigma_r + \Sigma_g - 2 \cdot \sqrt{\Sigma_r + \Sigma_g}\right), \quad (1)$$

где μ_r, $\sum_r$ - среднее и ковариационная матрица для реальных изображений, μ_g, $\sum_g$ - среднее и ковариационная матрица для сгенерированных изображений.

Ядровое расстояние Inception (Kernel Inception Distance - KID) является альтернативой FID и использует методы максимального среднего расхождения с ядерными функциями для сравнения распределений (2). Она отличается устойчивостью к небольшим размерам выборки [9, 10]:

$$KID(P_X, P_G) := \sup_{f \in H_\kappa : \|f\|_{H_\kappa} \le 1} \left(E_{X \sim P_X}[f(X)] - E_{X' \sim P_G}[f(X')]\right)^2, \quad (2)$$

где P_X - распределение признаков, извлечённых из реальных изображений, P_G - распределение признаков, извлечённых из сгенерированных изображений; H_k – гильбертово пространство с воспроизводящим ядром, соответствующему ядру k.

Метрики оценки структуры и деталей оценивают, насколько хорошо сгенерированные изображения сохраняют текстуры, границы объектов и другие структурные особенности, сравнивая их с реальными изображениями.

Индекс структурного сходства (Structural Similarity Index - SSIM) - измеряет структурное сходство между двумя изображениями, учитывая яркость, контраст и структуру (3) [11].

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (3)$$

где μ_x и μ_y - средние значения яркости для изображений x и y, σ_x^2 и σ_y^2 – дисперсии значений яркости для x и y, σ_{xy} – ковариация между изображениями, C_1 и C_2 – малые константы для стабилизации.

LPIPS оценивает сходство изображений, сравнивая активации нейросетевых слоев на локальном уровне (4) [7]. Она позволяет оценивать восприятие человеком:

$$d(x,x_0)=\sum_l \frac{1}{H_l W_l}\sum_{h,w}\left\| w_l \circ \left(\hat{y}_{hw}^{l} - \hat{y}_{hw}^{l'} \right) \right\|_2^2, \quad (4)$$

где l – индекс слоя нейросети, $\hat{y}^l$ и $\hat{y}^{l'}$ – нормализованные по каналам карты признаков, полученные из изображений x и x_0 на слое l, w_l – обучаемые веса слоя, H_l и W_l - размеры карты признаков.

Метрики, основанные на тексте

Метрики, основанные на тексте, оценивают соответствие сгенерированных изображений входному текстовому описанию.

Оценка CLIP - метрика, измеряющая семантическое соответствие между текстом и изображением с использованием модели CLIP, которая обучаются на больших наборах данных текст-изображение (5) [6]:

$$CLIP-S(c,v)=w\cdot\max(\cos(c,v),0), \quad (5)$$

где v - эмбеддинг для изображения, c - эмбеддинг текстового описания, w - весовой коэффициент, обычно установленный на 2.5.

Сравнительный анализ и ограничения метрик

Для оценки качества сгенерированных изображений важно учитывать, как особенности каждой метрики, так и их ограничения при анализе восприятия человеком и машиной. Сравнительный анализ метрик FID, KID, SSIM и LPIPS был проведен на основе набора данных "Общие объекты в контексте" [12].

Для сравнения работы метрик были созданы искаженные версии изображений с добавлением гауссовского шума и эффекта размытия. Параметры были подобраны для минимизации расхождений с исходными изображениями в восприятии человека. Также на основе исходных изображений были сгенерированы новые с использованием модели Stable Diffusion в режиме изображения в изображение с применением ControlNet для контроля процесса генерации [13, 14]. На рисунке 1 представлен полученный набор изображений, в котором для каждого выбранного уникального изображения из набора данных показаны его искаженный вариант и сгенерированное изображение, расположенные по столбцам.

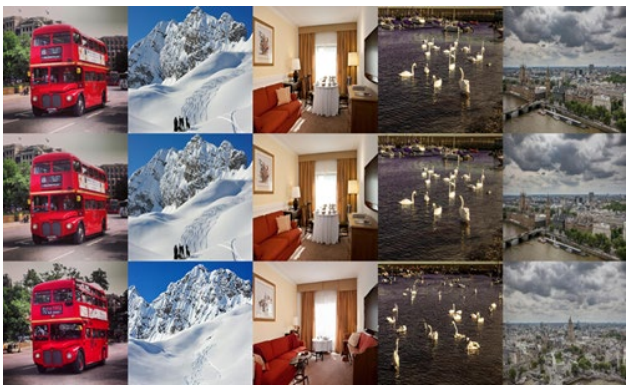

Рис. 1. - Оригинальные, измененные и сгенерированные изображения

Для последующего анализа были рассчитаны попарно метрики для оригинальных и зашумленных изображений (таблица №1).

Таблица №1

Сравнение попарно оригинальных изображений с модифицированными шумом, замыливанием, контрастом

	Автобус	Гора	Комната	Лебеди	Город
FID	1.6469	9.9046	14.9005	25.7672	16.6934
KID	0.0031	0.0175	0.0311	0.0547	0.0328
SSIM	0.9900	0.9640	0.9800	0.9233	0.9699
LPIPS	0.0120	0.0617	0.0271	0.0697	0.0558

Согласно таблице №1, значения FID и KID оставались на низком уровне, что указывает на сохранение распределений признаков, извлеченных с помощью нейросетевых моделей. Особенно низкие значения KID подтверждают устойчивость метрики к выбросам и коррелируют с результатами FID. Значения SSIM указывают на высокую степень структурного сходства, а низкие значения LPIPS указывают на сохранение перцептивной близости к оригиналу.

Аналогичным образом были рассчитаны метрики оригинальных и сгенерированных изображений (таблица №2).

Таблица №2

Сравнение попарно оригинальных изображений со сгенерированными с помощью Stable Diffusion и ControlNet

	Автобус	Гора	Комната	Лебеди	Город
FID	29.9462	36.2245	96.8077	260.7752	165.7741
KID	0.0546	0.0632	0.1973	0.6287	0.3330
SSIM	0.3789	0.4034	0.5179	0.1570	0.3374
LPIPS	0.3058	0.4428	0.3150	0.4272	0.3936

Однако значения ухудшились. Значения FID увеличились в десятки раз, что указывает на значительное расхождение распределений признаков. Существенное увеличение KID подтверждает это, свидетельствуя о значительном отклонении от эталонного распределения. Значения SSIM оказались значительно ниже, что указывает на потерю локальной структурной информации. Перцептивная метрика LPIPS также показала значительное увеличение, отражая снижение визуального сходства с оригиналом.

Для проведения сравнительного анализа оценки CLIP были созданы для каждого оригинального изображения по три описания так, чтобы они не

совпадали, частично совпадали и полностью совпадали с изображением. В таблице № 3 приведен перевод текстовых описаний с английского языка.

Таблица №3

Сравнение оригинальных изображений с текстовым описанием

Изображение	Переведенное описание	Оценка CLIP
Автобус	Вечерний закат на море	11.9997
	Автомобиль на городской улице	21.0862
	Красный двухэтажный автобус на фоне городской улицы	31.5303
Гора	Автомобиль на городской улице	12.9942
	Горная местность	25.2028
	Снежная гора с людьми у подножия	27.4848
Комната	Кот спит на подоконнике	13.7362
	Светлая комната	22.5477
	Комната с белым потолком и бежевой стеной. В комнате красный диван, стол со стульями и телевизор на стене	28.9883
Лебеди	Город с высоты птичьего полета	19.5406
	Белые лебеди в природе	29.8923
	Белые лебеди на фоне пришвартованных лодок в реке	31.1981
Город	Снежная гора с людьми у подножия	17.0383
	Город с высоты птичьего полета	25.5117
	Городской пейзаж в пасмурную погоду с часовой башней, рекой и несколькими лодками, пришвартованными вдоль набережной.	27.0417

Полученные значения показали, что степень семантической релевантности варьируется в зависимости от сложности сцены и формулировки текстового описания. Градация значений отражает степень сходства между описанием и изображением, что позволяет более точно оценивать соответствие визуального контента и текстового описания.

Следовательно, метрики FID, KID и LPIPS наиболее чувствительны к глобальным и перцептивным искажениям, а SSIM эффективно отображает изменения в локальной структуре. Оценка CLIP дополняет анализ, но не заменяет метрики, направленные на визуальную или структурную оценку.

Заключение

Для комплексной оценки качества сгенерированных изображений важно применять несколько метрик, ориентированных на разные аспекты восприятия. FID и KID обеспечивают количественную оценку качества изображений, LPIPS и SSIM акцентируют внимание на восприятии человеком, в частности в задачах, где важны структура и детали. Оценка CLIP дополняет анализ, устанавливая связь текстового описания с изображением, но не может заменить метрики для визуальной или структурной оценки. Выбор метрики должен быть обусловлен задачей и целью генерации, а применение комбинированного подхода способствует более объективной и точной оценки качества генерации.

Литература (References)

1. Eric J. N., Pejman K., Shadrokh S. Compound Frechet Inception Distance for Quality Assessment of GAN Created Images // URL: doi.org/10.48550/arXiv.2106.08575

2. Martin H., Hubert R., Thomas U., Bernhard N., Sepp H. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium // Conference: 31st Conference on Neural Information Processing Systems (NIPS

2017), Long Beach, CA, USA. 2018. pp. 1-38. URL: doi.org/10.48550/arXiv.1706.08500.

3. Jonathan H., Ajay J., Pieter A. Denoising Diffusion Probabilistic Models // Conference: 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. pp. 1-12. URL: proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

4. Shanchuan L., Xiao Y. Diffusion Model with Perceptual Loss // URL: doi.org/10.48550/arXiv.2401.00110.

5. Mario L., Karol K., Marcin M., Sylvain G., Olivier B. Are GANs Created Equal? A Large-Scale Study // Conference: 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada. 2018. pp. 1-21. URL: doi.org/10.48550/arXiv.1711.10337.

6. Jack H., Ari H., Maxwell F., Ronan L. B., Yejin C. CLIPScore: A Reference-free Evaluation Metric for Image Captioning // URL: doi.org/10.48550/arXiv.2104.08718.

7. Richard Z., Phillip I., Alexei A. E., Eli S., Oliver W. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric // URL: doi.org/10.48550/arXiv.1801.03924.

8. Sebastian H., Dominik E., Leon S., Hannah K., Tristan P., Poonam P., Michael G., Alex B., Timo R. A Survey on Quality Metrics for Text-to-Image Generation // URL: doi.org/10.48550/arXiv.2403.11821.

9. Mikołaj B., Danica J. S., Michael A., Arthur G. Demystifying MMD GANs // Conference: ICLR 2018. 2021. pp. 1-36. URL: doi.org/10.48550/arXiv.1801.01401.

10. Zixiao W., Farzan F., Zhenghao L., Yunheng S., Bei Y. On the Evaluation of Generative Models in Distributed Learning Tasks // URL: doi.org/10.48550/arXiv.2310.11714.

11. Zhou W., Alan C. B., Hamid R. S., Eero P. S. Image quality assessment: from error visibility to structural similarity // IEEE Transactions on Image Processing. 2004. Vol. 13. No. 4. pp. 600-612. URL: ieeexplore.ieee.org/document/1284395.

12. COCO - Common Objects in Context // URL: cocodataset.org/#home.

13. stable-diffusion-v1-5 // URL: huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5.

14. Lvmin Z., Anyi R., Maneesh A. Adding Conditional Control to Text-to-Image Diffusion Models // URL: doi.org/10.48550/arXiv.2302.05543.

Дата поступления: 13.04.25

Дата публикации: 25.05.25