УДК 004.855.5+004.93'12+519.676

Світлана Ткаченко,
студентка факультету математики, фізики
і комп'ютерних наук
Вінницького державного педагогічного університету
імені Михайла Коцюбинського

ОГЛЯД МЕТРИК ОЦІНЮВАННЯ ГЕНЕРАТИВНО-ЗМАГАЛЬНИХ МЕРЕЖ (GANs)

Анотація. *У статті охарактеризовано сучасні та найбільш популярні метрики оцінювання якості генеративно-змагальних мереж та описано переваги і недоліки кожного з них.*

Ключові слова: *генеративно-змагальні мережі, метрики оцінювання, генератор, дискримінатор.*

Abstract. *The article describes evaluation metrics of the modern generative adversarial networks (GANs) and defines their main advantages and disadvantages.*

Keywords: *GANs, evaluation metrics, generator, discriminator.*

Вступ. Генеративно-змагальні мережі (*Generative Adversarial Networks*, *GANs*) відносяться до технологій генеративного моделювання з використанням методів глибинного навчання (глибинних нейронних мереж). Вперше генеративно-змагальні мережі були представлені Яном Ґудфелоу в 2014 році [5]. Після того з'явилось чимало нових ідей та моделей, які в основі своїй використовують генеративно-змагальні мережі або повністю на них базуються. Таким чином, відносна популярність *GANs* призвела не лише до розширення самої технології, а й до урізноманітнення сфер її застосування, навіть часом неочікуваним шляхом.

Однією з цікавих, але як не дивно ще не до кінця розв'язаних універсально задач, яка виникає при роботі з архітектурою генеративно-змагальної мережі, є визначення її якості. Загалом з цією метою (для оцінки якості) зазвичай використовують спеціальні метрики, які допомагають якісно або кількісно визначити ефективність створених моделей. Для генеративно-змагальних мереж було запропоновано уже чимало різних метрик, які володіють тим чи іншим

рівнем експертизи. Однак кожна з них має свої особливості та обмеження, тому дуже важливо з'ясувати ці властивості та визначити рівень корисності тієї чи іншої метрики.

Отже, **метою статті** є огляд сучасних (та найбільш частіше використовуваних) метрик оцінювання генеративно-змагальних мереж, з'ясування основних переваг та недоліків кожної з них, а також визначення рівня універсальності, який вони надають.

Виклад основного матеріалу. Для того, щоб метрика, яка використовується для оцінки генеративно-змагальної мережі була якісною та корисною на неї накладається цілий перелік вимог, а саме [1]:

- вміння розрізняти згенеровані та справжні зразки та відповідним чином оцінювати моделі, які здатні генерувати зразки високої якості;
- вміння виділяти та оцінювати моделі, які здатні генерувати різні, не схожі між собою зразки (а тому ці моделі можуть бути схильними до проблем з перенавчанням чи колапсу мод);
- вміння виділяти та оцінювати здатністю моделей до контрольованого семплювання;
- наявність чітко визначених меж;
- інваріантність до спотворення чи трансформації зразків, які не змінюють їхню семантику;
- відповідність людському судженню щодо якості згенерованих зразків;
- відносно низька обчислювальна складність.

Загалом усі існуючі метрики оцінювання генеративно-змагальних мереж можна поділити на дві групи: кількісні та якісні. Кількісні метрики на відміну від якісних є менш суб'єктивними. Однак якісні метрики дозволяють зробити оцінку в тому числі і з перспективи сприйняття згенерованих зразків людиною, що кількісні метрики не завжди гарантують.

Розглянемо спочатку кількісні метрики. Основна увага буде направлена на деякі сучасні метрики та найбільш часто використовувані на практиці.

1. Inception Score (*IS*). Метрика *IS* була запропонована Salimans та ін. [11] і є однією з найбільш популярних метрик для оцінювання генеративно-змагальних мереж. В її основі лежить використання попередньо натренованої для класифікації зображень на датасеті *ImageNet* нейронної мережі *Inception Net* (*Inception v3*). Для знаходження величини *IS* згенеровані зображення подають на вхід класифікатору і для кожного з зображень знаходиться ймовірність належності до того чи іншого класу. Потім ці передбачення перетворюють у метрику *IS*, щоб оцінити якість поданих зображень (чи були зображення класифіковані як певні об'єкти) та загальне різноманіття (наскільки широкий спектр зображень було згенеровано) [7].

Найменше можливе значення *IS* дорівнює 1, а найбільше визначається кількістю класів, яку підтримує модель класифікації, що використовується. Оскільки модель *Inception v3* підтримує 1000 класів, то найвище можливе значення *IS* дорівнює 1000. Загалом метрика *IS* демонструє зв'язок своїх значень з якістю та різноманітністю згенерованих зображень, але їй властива і певна кількість недоліків [1]:

- за допомогою *IS* не можна визначити перенавчання;
- за допомогою *IS* не можна визначити чи відбувся колапс мод;
- *IS* є асиметричною метрикою;
- IS є чутливою до роздільної здатності зображення.

2. Frechet Inception Distance (*FID*). Метрика *FID* була представлена Heusel та ін. [9]. *FID* базується на порівнянні розподілу згенерованих зображень з розподілом справжніх зображень, які використовувались для тренування генератора. Вибірка згенерованих зразків подається в простір ознак, заданий певним шаром мережі *Inception Net* (або будь-якої іншої згорткової нейронної мережі). Цей шар (*embedding layer*) розглядається як багатовимірний неперервний розподіл Гаусса. Далі між двома розподілами Гаусса (розподілами справжніх та згенерованих зображень) рахується відстань Фреше, як показник якості згенерованих зразків:

$$FID(r,g) = \left\| \mu_r - \mu_g \right\|_2^2 + Tr\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{\frac{1}{2}} \right),$$

де $\left(\mu_r, \Sigma_r\right)$ та $\left(\mu_q, \Sigma_q\right)$ – середнє та коваріація розподілів справжніх та згенерованих даних відповідно.

Чим нижчим є значення *FID*, тим менша відстань між розподілами згенерованих та справжніх зображень. Загалом *FID* є хорошою метрикою в плані надійності та обчислювальної ефективності [1]. Було з'ясовано, що метрика *FID* є більш стійкою до шуму, ніж *IS* [9]. Однак ефективність *FID* може знижуватись, якщо надати зображенням різного роду спотворень (рис. 1).

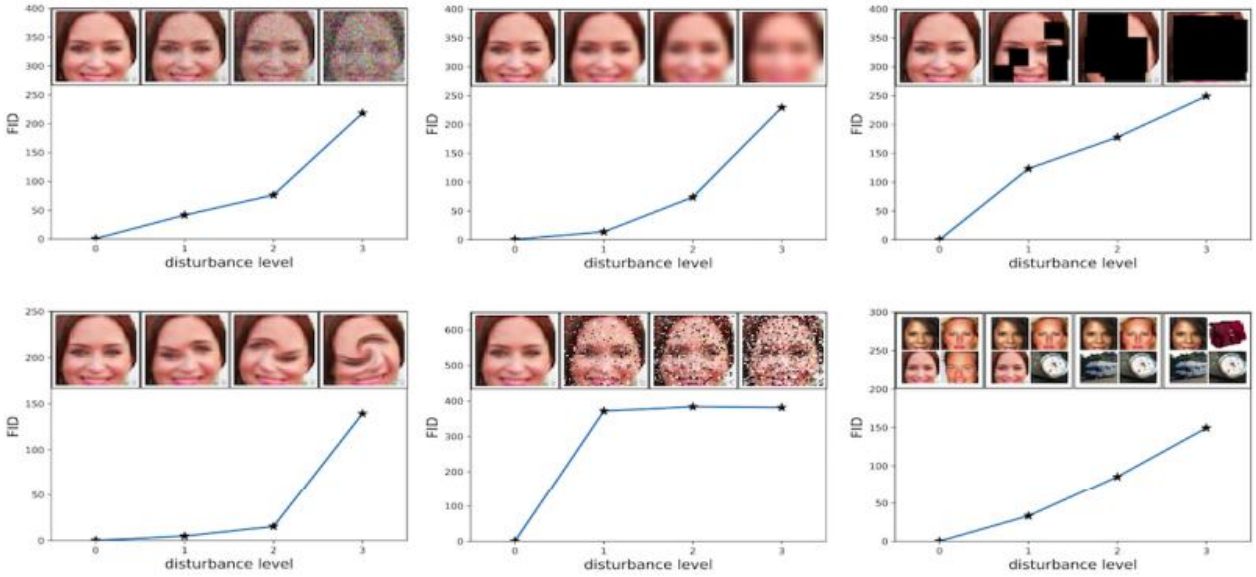

Рис. 1. Демонстрація вразливості *FID* до різного роду спотворень зображень [9]

3. Perceptual Path Length (*PPL*). Метрика *PPL* була вперше запропонована авторами генеративно-змагальної мережі *StyleGAN3* (Karras та ін., 2019 р.) [10]. *PPL* вимірює те, наскільки *заплутаним* є латентний простір генератора (чи він є гладким і фактори варіації правильно розподілені) [2]. Інтуїтивно зрозуміло, що менш вигнутий латентний простір призводить до більш плавного переходу, ніж сильно вигнутий латентний простір. Більш формально, *PPL* – це емпіричне середнє перцептивної різниці між відповідними зображеннями латентного простору $\mathcal{Z}$:

$$l_{\mathcal{Z}} = \mathbb{E}\left[\frac{1}{\epsilon^2} d\left(G\left(\text{slerp}\left(\mathbf{z}_1, \mathbf{z}_2; t\right)\right), G\left(\text{slerp}\left(\mathbf{z}_1, \mathbf{z}_2; t+\epsilon\right)\right)\right)\right],$$

де $\mathbf{z}_1, \mathbf{z}_2 \sim P(\mathbf{z})$, $t \sim U(0,1)$, G – генератор, $d(\cdot,\cdot)$ – перцептивна відстань між отриманими зображеннями, як *slerp* позначено сферичну інтерполяцію (Shoemake, 1985), а $\epsilon = 10^{-4}$ – розмір кроку. Як можна побачити на рис. 2, метрика *PPL* добре розпізнає семантику та якість зображення. Крім того, багато експериментів демонструють перевагу метрики *PPL* над *FID* [10].

(a) Low PPL scores (b) High PPL scores

Рис. 2. Зв'язок між *PPL* та якістю згенерованого зображення [10]

Тепер перейдемо до розгляду деяких якісних метрик оцінювання генеративно-змагальних мереж. Суть якісних метрик полягає у візуальному огляді згенерованих зразків людиною, що на перший погляд здається інтуїтивно зрозумілим методом. Однак в порівнянні з кількісними методами йому притаманна велика кількість недоліків, що пов'язані з суб'єктивністю та упередженістю оцінювання, що здійснюється людиною, а також з тим, що проведення подібних оцінювань може бути доволі затратним.

1. Nearest Neighbors (Метод найближчих сусідів). Цей метод полягає в тому, що згенеровані зображення для оцінювання моделі на перенавчання розглядаються поруч зі своїми найближчими сусідами з тренувального датасету (з вибірки реальних зображень) (рис. 3) [1].

Серед основних недоліків методу найближчих сусідів є його вразливість до навіть незначних перцептивних збурень. Так як зазвичай найближчі сусіди шукаються на основі евклідової відстані, то дуже легко потрапити у ситуацію, коли візуально згенерований та еталонний зразки схожі між собою, але при цьому евклідова відстань між ними достатньо велика.

Рис. 3. Справжні зображення, яким у відповідність поставлено згенеровані зображення (найближчі сусіди) [1]

2. Rapid Scene Categorization (Розбиття на категорії зі швидким реагуванням). Метрика, яка пов'язана з розбиттям на категорії зі швидким реагуванням базується на дослідженнях, які доводять, що людина здатна виявляти необхідні візуальні характеристики швидким поглядом [3, 12].

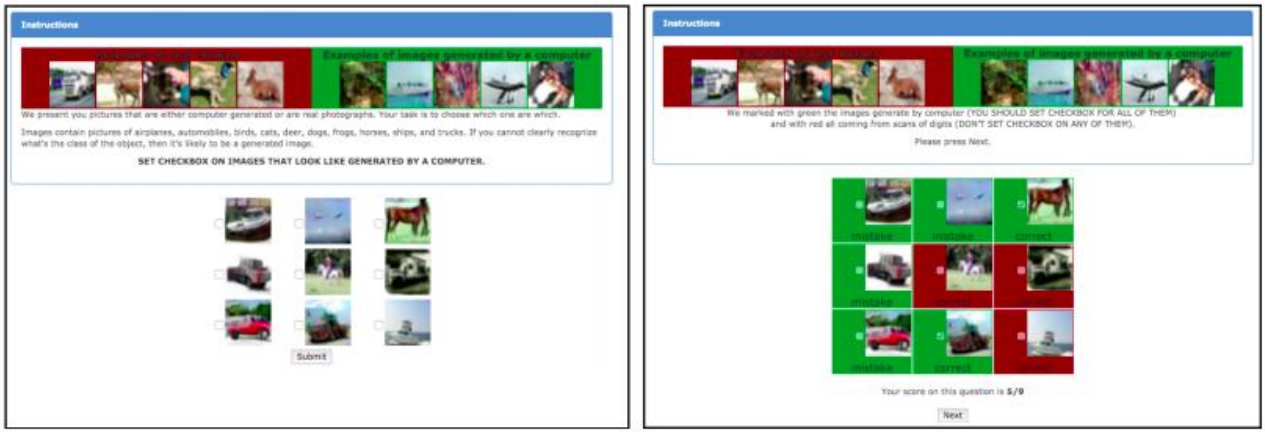

Рис. 4. Приклад запитання на виявлення згенерованих зображень, що пропонували Denton та ін. респондентам [4]

Наприклад, Denton та ін. [4] для того, щоб якісно оцінити свої зображення, попросили респондентів відрізнити згенеровані зображення від справжніх (рис. 4). Учасникам оцінювання було запропоновано обрати зображення, які на їхній погляд є згенерованими. Час показу запитання варіювався від 50 мс до 2000 мс. На основі отриманих відповідей автори зробили висновок, що модель, яку вони запропонували була краща, ніж та, з якою велось порівняння, оскільки респонденти частіше відносили згенеровані нею зображення до справжніх.

Подібні експериментальні оцінювання чимось нагадують тест Тюрінга. Їхня користь полягає в тому, що таким чином ми можемо побачити, а чи справді генеративні моделі здатні створити зразки, які навіть в людському сприйнятті можуть здаватись справжніми. Однак подібні оцінювання є досить ненадійними та важко контрольованими.

3. Rating and Preference Judgement (Ранжування та визначення переваг). В типах експериментів, які використовують ранжування, суб'єктів просять проранжувати зразки згенеровані різними моделями. Наприклад, Snell та ін. [8] попросили респондентів обрати, яким зображенням серед поданих вони віддають перевагу. Вони підготували трійки зображень, де по центру розмістили оригінальне зображення, а по бокам зображення створені двома різними нейронними мережами. Серед цих двох потрібно було обрати одне, якому повинна бути надана перевага.

Висновки. Метрик оцінювання якості генеративно-змагальних мереж існує чимало. Найбільш часто на практиці використовують кількісні метрики такі як *IS*, *FID*, *PPL* через відсутність суб'єктивності в оцінюванні, однак якісні метрики теж заслуговують на окрему увагу через врахування людського фактору сприйняття згенерованих зразків.

Література:

1. A. Borji. Pros and Cons of GAN Evaluation Measures, arXiv preprint arXiv: 1802.03446v5.
2. A. Borji. Pros and Cons of GAN Evaluation Measures: New Developments, arXiv preprint arXiv: 2103.09396v3.
3. A. Oliva. Gist of the scene, in: Neurobiology of attention, Elsevier, 2005, pp. 251-256.
4. E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks, in: Advances in neural information processing systems, 2015, pp. 1486-1494.

5. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014.
6. I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press. 2016.
7. J. Brownlee, Generative Adversarial Networks with Python, Deep Learning Generative Models for Image Synthesis and Image Translation. 2019.
8. J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, R. S. Zemel. Learning to generate images with perceptual similarity metrics, arXiv preprint arXiv: 1511.06409.
9. 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, pp. 6629-6640.
10. T. Karras, S. Laine, T. Aila. A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 4401-4410.
11. 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.
12. T. Serre, A. Oliva, T. Poggio. A feedforward architecture accounts for rapid categorization, Proceedings of the national academy of sciences 104 (15) (2007) 6424-6429.
13. Tkachenko S., Bak S. Implementation of Cycle-GAN model for image transformation into image with Anime style. Abstracts of III International Scientific and Practical Internet Conference «Mathematics and Informatics in Higher Education: Challenges of Modernity», dedicated to the memory of Professors Pankov O. A. and Trokhymenko V. S. (May 20-21, 2021, Vinnytsia, Ukraine) [Electronic network scientific publication]: book of abstracts. Vinnytsia, 2021. P. 107-110.
14. Ткаченко С. В. Використання генеративно-змагальних мереж (GANs) для генерації зразків рукописних цифр. *Науково-популярний альманах «Математика та інформатика навколо нас»* / Вінницький державний педагогічний університет імені Михайла Коцюбинського; [редкол.: М. М. Ковтонюк (голова) та ін.]. Вінниця: ФОП Рогальська І. О., 2020. Вип. 4. С. 154-162.
15. Ткаченко С. В. Реалізація процесу перетворення звичайних зображень на зображення з ефектом авторського стилю картин Ван Гога за допомогою генеративно-змагальних мереж. *Концептуальні шляхи розвитку наукових знань: матеріали III Міжнародної науково-практичної конференції* (Київ, 6-7 лютого 2021 р.). Київ: МЦНіД, 2021. С. 49-51.
16. Ткаченко С. В. Реалізація процесу перетворення зображень на зображення з ефектом аніме за допомогою генеративно-змагальних мереж. Актуальні проблеми математики, фізики і комп'ютерних наук: зб. наук. пр. Вінниця, 2021.

Науковий керівник: докт. фіз.-мат. наук, професор Бак Сергій Миколайович.