Оценка качества вариационных авто-кодировщиков

М.Ю. Леонтьев[1,2], А.В. Михеев[3], К.В. Святов[3], С.В. Сухов[1]

[1]Ульяновский филиал Института радиотехники и электроники им. В.А. Котельникова РАН, Спасская 14, Ульяновск, Россия, 432011
[2]Научно-исследовательский технологический институт им. С.П. Капицы Ульяновского государственного университета, Университетская Набережная 1, Ульяновск, Россия, 432017
[3]Ульяновский государственный технический университет, Северный Венец 32, Ульяновск, Россия, 432027

Аннотация. Вариационные авто-кодировщики (ВАК) являются одними из популярных методов генерации реалистичных изображений. Обычно качество генерируемых изображений оценивается лишь визуально, поскольку методы количественной оценки их качества в полной мере не разработаны. В данной работе предложены и протестированы методы, которые предоставляют возможность объективно оценивать реалистичность и разнообразие генерируемых изображений. Соответствующие метрики качества найдены для нескольких разновидностей ВАК. Эксперименты проведены на общедоступных базах данных.

1. Введение

В последние годы были разработаны модели искусственных нейронных сетей (ИНС), которые могут воспроизводить данные, сходные по типу и по вероятностному распределению с тренировочными данными (генеративные ИНС). Подобные сети могут генерировать изображения трудно отличимые от реальных [1], и это свойство генеративных ИНС может быть использовано в различных областях компьютерного зрения [2, 3]. Генеративные свойства ИНС широко используются и в задачах непрерывного обучения и предотвращения катастрофического забывания (КЗ) [4]. Точность сохранения предыдущих знаний при непрерывном обучении зависит от точности воспроизведения генеративными сетями вероятностных свойств исходных тренировочных данных. При этом важны как реалистичность генерируемых изображений, так и их разнообразие. Таким образом, необходима объективная оценка качества генеративных сетей.

Одним из видов генеративных сетей являются вариационные авто-кодировщики (ВАК) [5], которые представляют собой две последовательно расположенные нейронные сети (кодировщик и декодер). ВАК обучается путем оптимизации целевых функций сравнения сигналов на входе кодировщика и на выходе декодера. Дополнительная целевая функция следит за тем, чтобы активация нейронов во внутреннем латентном слое была близка к нормальному распределению. После обучения, сеть способна воспроизводить данные из того же распределения, что и изначальные тренировочные данные, при подаче гауссова шума на латентный слой [5].

В научной литературе регулярно предлагаются множество новых модификаций вариационных автокодировщиков. Возникает критический вопрос - как эти модификации можно оценивать и сравнивать друг с другом? Нужна объективная оценка того, как модификация ВАК отражается на качестве генерируемых данных. Оценка и сравнение различных ВАК состоит в оценке и сравнении исходных и сгенерированных изображений, что является довольно сложной задачей. Ввиду высокой сложности подобной оценки, часто многими исследователями использовалась простая субъективная визуальная оценка изображений, синтезированных с помощью ВАК.

Недавно появились и количественные методы оценки качества изображений, синтезированных генеративными сетями. Известным недостатком ВАК является размытость генерируемых изображений. Для оценки размытости можно использовать метрику базирующуюся на дисперсии распределения Лапласа (LAP) [6]. LAP увеличивается при увеличении резкости изображения и уменьшается при увеличении размытости. В среднем, реальные изображения имеют большую метрику LAP, чем сгенерированные. Оценка размытости служит одной из простейших метрик качества сгенерированных изображений.

Другими использующимися в настоящее время метриками являются Inception score (IS) [7] и Frechet Inception distance (FID) [8]. Расчет IS не требует первоначальных изображений [7]. Таким образом, IS может оценить лишь разнообразие сгенерированных изображений – большее значение IS соответствует большему разнообразию. FID сравнивает распределение активностей нейронов внутри нейронной сети Inception для реальных и сгенерированных изображений. Меньшее значение FID говорит о более полном сходстве синтетических изображений с реальными. Однако распределения активностей аппроксимируются с помощью нормального распределения, что является довольно грубым приближением. Кроме того, и IS, и FID используют для расчета метрик нейронную сеть Inception, которая может оказаться непригодной для определенных типов данных. Таким образом, предложенные выше метрики не могут в общем случае адекватно предсказывать реалистичность и разнообразие генерируемых изображений.

2. Метод оценки качества ВАК с помощью дополнительных классификаторов

В работе [9] были предложены методы оценки качества генеративно-состязательных сетей (ГСС) GAN-train и GAN-test. Для вычисления GAN-train дополнительный классификатор обучался на изображениях, генерируемых ГСС. Затем эффективность обученного классификатора оценивалась на тестовом наборе реальных изображений, что и является метрикой GAN-train. Метрика GAN-train оценивает отличие между синтетическим и исходным распределениями данных. GAN-test представляет собой точность классификатора, обученного на реальных изображениях и протестированного на сгенерированных. Эта метрика оценивает реалистичность сгенерированных изображений.

Генерация искусственных данных и обучение на них классификатора используется в генеративных методах предотвращения катастрофического забывания [10]. Таким образом, метрики GAN-train и GAN-test напрямую определяют эффективность ГСС для задач предотвращения КЗ.

В настоящей работе мы распространили методы, предложенные в работе [9], для оценки моделей вариационных авто-кодировщиков. По аналогии с метриками GAN-train и GAN-test мы предлагаем метрики ВАК-трен, ВАК-тест. ВАК-трен предполагает обучение классификационной ИНС (КИНС) на генерированных тренировочных данных (ГТД), а затем оценку ее производительности на исходном проверочном наборе (ИПН), состоящем из реальных изображений (рисунок 1). ВАК-тест предполагает обучение КИНС с помощью реальных изображений, а затем оценку ее производительности на сгенерированных изображениях (рисунок 1).

3. Экспериментальные данные

В нашем исследовании метод оценки качества ВАК, основанный на проверке генерируемых изображений в КИНС, был апробирован на наборе Fashion-MNIST [11]. Fashion-MNIST состоит

из 600000 тренировочных и 10000 тестовых изображений предметов одежды, разбитых на 10 категорий (классов). Каждое изображение имеет размер 28×28 пикселей. Архитектуры сверточных сетей ВАК и КИНС были заимствованы из работы [12]. Процесс оценки различных модификаций ВАК состоял из нескольких этапов:

1) Одна из двух идентичных КИНС (КИНС-1) тренируется на ИТН до приемлемой точности в течение 20 эпох;
2) На ИТН до приемлемой точности в течение 20 эпох тренируется ВАК;
3) ВАК генерирует изображения в количестве 1000 штук на класс. КИНС-1, натренированная ранее на ИТН, присваивает им метки (лейблы), формируя тем самым ГТД, причем метки могут быть «жёсткие» (one-hot) либо «мягкие» (каждому выходу соответствует число в диапазоне (0,1), отражающее вероятность того, что изображение принадлежит этому классу);
4) На ГТД, которые сгенерировал ВАК, обучается классификационная сеть КИНС-2;
5) Оценивается производительность КИНС-1 на ГТД (ВАК-тест), а КИНС-2 на ИПН (ВАК-трен).

Итого, для получения оценки необходимо три процедуры обучения сетей (КИНС-1, ВАК и КИНС-2), одна процедура генерации изображений (ВАК) и две оценки производительности классификаторов (КИНС-1 и КИНС-2), а также процедура распознавания генерируемых изображений для восстановления меток, что в совокупности весьма вычислительно ощутимая задача. Общая схема всего процесса представлена на рисунке 1.

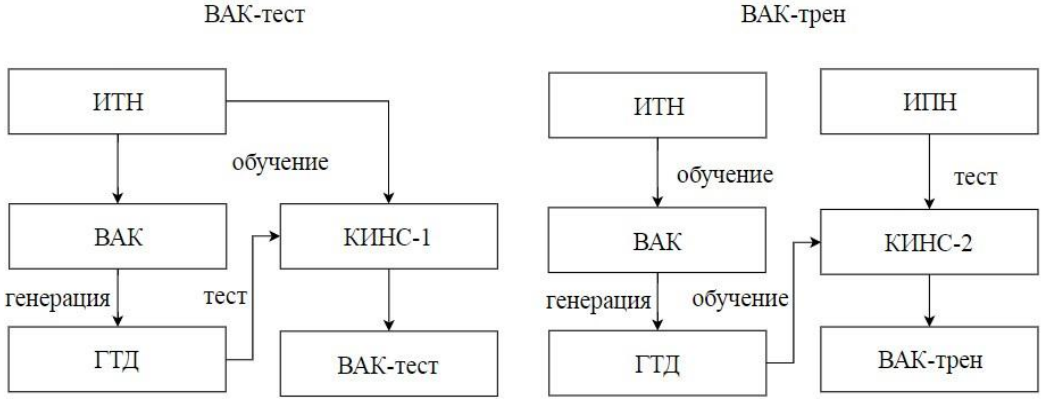

Рисунок 1. Общая схема процесса получения оценки ВАК-тест и ВАК-трен.

По полученным оценкам ВАК-тест и ВАК-трен можно уже достаточно объективно судить о качестве модели ВАК и корректировать параметры модели и её обучения в сторону улучшения исследуемых характеристик. Наибольшую ценность оценок ВАК-тест и ВАК-трен представляет то, что с помощью них можно напрямую оценивать эффективность ВАК при его использовании для предотвращения катастрофического забывания.

Помимо вычисления оценок ВАК-тест и ВАК-трен дополнительно оценивалась степень размытости реальных и генерируемых изображений путем вычисления метрики LAP для каждого изображения. Также были произведены вычисления оценок IS и FID. Оценки LAP и IS для реальных изображений набора Fashion-MNIST показаны в таблице 1. Дополнительно были исследованы следующие модели и подходы:

Модель 1. ВАК и стандартные («жёсткие») метки для генерируемых изображений;

Модель 2. ВАК и «мягкие» метки для генерируемых изображений;

Модель 3. Условный ВАК (УВАК). В УВАК к случайной латентной переменной добавляется вектор, кодирующий метку (лейбл) изображения, что позволяет в дальнейшем генерировать изображения заданного класса [13];

Модель 4. ВАК с «мягкими» метками и функцией ошибки, оценивающей «восприятие» нейронной сети (perception loss) [14]. В этом подходе вместо попиксельного сравнения оригинального и сгенерированного изображений, сравниваются активности на скрытых слоях КИНС-1, создаваемые исходным и сгенерированным изображениями;

Модель 5. ВАК с «мягкими» метками и с отбором лучших (точность от 0.9 на выходе при распознавании) из сгенерированных изображений с помощью КИНС-1.

Рисунок 2. Примеры реальных изображений и изображений, сгенерированных различными моделями ВАК.

Примеры сгенерированных изображений для каждого из пяти вариантов ВАК показаны на рисунке 2. Визуально все сгенерированные изображения выглядят почти одинаково, и невозможно отдать предпочтение тому или иному варианту ВАК. Объективные метрики оценки качества различных ВАК приведены в сводной таблице 1.

Таблица 1. Сводная таблица результатов исследования.

	LAP	IS	FID	ВАК-трен	ВАК-тест
Реальные изображения	0.184 ± 0.079	4.98 ± 0.07	-	-	-
Модель 1	0.042 ± 0.031	3.51 ± 0.05	105.69 ± 0.61	0.78 ± 0.02	-
Модель 2	0.042 ± 0.031	3.47 ± 0.07	107.0 ± 1.1	0.82 ± 0.01	-
Модель 3	0.037 ± 0.026	3.60 ± 0.03	113.63 ± 0.81	0.66 ± 0.02	0.65 ± 0.03
Модель 4	0.080 ± 0.037	4.17 ± 0.07	95.31 ± 0.94	0.84± 0.01	-
Модель 5	0.049 ± 0.033	3.53 ± 0.05	95.8 ± 1.5	0.786 ± 0.007	-

ВАК-тест может быть оценен только в случае условного ВАК (Модель 3). Во всех других случаях метки в ГТД расставлял тот же самый классификатор, который потом эти изображения и распознавал, поэтому метрика ВАК-тест теряла смысл. Эксперименты показали низкие значения ВАК-трен и ВАК-тест для УВАК (см. таблицу 1), поэтому большинство экспериментов не использовало эту архитектуру.

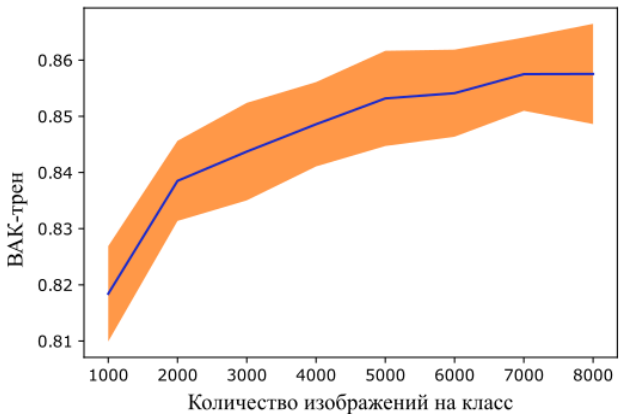

Рисунок 3. Зависимость ВАК-трен от количества генерируемых изображений. Линия показывает среднее значение, а заполненная цветом область означает стандартное отклонение.

Использование «мягких» меток (Модель 2) приводит к большему разнообразию сгенерированных изображение (большее ВАК-трен). Дополнительное использование внутренних активаций нейронной сети (Модель 4) увеличивает разнообразие еще больше. Для ВАК-трен можно заметить отсутствие прямой корреляции с остальными метриками.

С помощью метрики ВАК-трен можно определить примерное количество независимых изображений, которое может создать та или иная модель ВАК. Модель с малым разнообразием генерирует повторяющиеся образцы, и увеличение их количества не приведет к возрастанию значения ВАК-трен. И наоборот, генерация все большего и большего количества изображений из модели с большим разнообразием обеспечит постоянное увеличение значения ВАК-трен. Выберем Модель 2 и начнем менять количество генерируемых изображений в сторону увеличения (рисунок 3). Насыщение ВАК-трен достигается при 7000 сгенерированных изображений на класс, что близко к количеству изображений на класс в первоначальных тренировочных данных. Таким образом, Модель 2 способна сгенерировать около 7000 независимых изображений на класс.

4. Заключение

В данной работе изложены методы по решению проблемы оценки и сравнения изображений, сгенерированных вариационными автокодировщиками. Нами были использованы не только традиционные метрики (LAP, IS и FID), но и принципиально новые способы, базирующиеся на применении классификационных ИНС (метрики ВАК-трен и ВАК-тест).

Основные конкретные количественные показатели результатов экспериментов, иллюстрирующих эффективность предложенных методов изложены в таблице 1. Особенно в ней хорошо заметна обратная корреляция между метриками FID и LAP. В среднем изображения из исходного проверочного набора имеют наибольшую метрику LAP, что означает их наименьшее размытие или максимальный фокус. Напротив, сгенерированные ВАК изображения являются более размытыми и обладают, следовательно, меньшей метрикой LAP, а так как более размытые изображения менее похожи на настоящие, то имеют большую метрику FID – чем и объясняется их обратная корреляция с метрикой LAP. Эксперименты показали, что наилучшие по качеству изображения генерируются ВАК, обученным с «мягкими» метками и функцией «восприятия» (Модель 4). Обученный на изображениях этой модели классификатор показал точность классификации ИПН 0.84± 0.01.

Экспериментальный анализ показал, что ВАК-трен и ВАК-тест не только подчеркивают разницу в эффективности различных методов обучения ВАК, но и взаимодополняют прежние метрики. Метрики ВАК-трен и GAN-train являются оптимальными для оценки генеративных ИНС в задачах предотвращения катастрофического забывания, где сохранность знаний сетей-классификаторов поддерживается их непрерывным обучением на синтетических данных [4].

Метрики ВАК-трен и ВАК-тест открывают путь к разработке итеративных алгоритмов самомодификации ВАК для максимально эффективной подстройки её параметров под тот или иной набор данных как это реализовано например в EfficientNet [15].

5. Благодарности

Исследование выполнено при финансовой поддержке РФФИ и Правительства Ульяновской области в рамках научного проекта №18-47-732006.

6. Литература

[1] Karras, T. Progressive growing of GANs for improved quality, stability, and variation / T. Karras, T. Aila, S. Laine, J. Lehtinen [Electronic resource]. – Access mode: https:// arxiv.org/abs/1710.10196 (26.02.2018).

[2] Isola, P. Image-to-image translation with conditional adversarial networks / P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros // Computer Vision and Pattern Recognition. – 2017. – P. 1125-113.

[3] Ledig, C. Photo-realistic single image superresolution using a generative adversarial network / C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang // Computer Vision and Pattern Recognition. – 2017. – P. 4681-4690.

[4] Van de Ven, G.M. Generative replay with feedback connections as a general strategy for continual learning [Electronic resource]. – Access mode: https://arxiv.org/abs/1809.10635 (17.04.2019).

[5] Kingma, D.P. Auto-Encoding Variational Bayes [Electronic resource]. – Access mode: https:// arxiv.org/abs/1312.6114 (01.05.2014).
[6] Pertuz, S. Analysis of focus measure operators for shape-from-focus / S. Pertuz, D. Puig, M.A. Garcia // Journal Pattern Recognition. – 2013. – Vol. 46(5). – P. 1415-1432.
[7] Salimans, T. Improved techniques for training GANs. / T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen // Neural Information Processing Systems. – 2016. – P. 2234-2242.
[8] Heusel, M. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. / M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter // Neural Information Processing Systems. – 2017. – P. 6626-6637.
[9] Shmelkov, K. How good is my GAN? / K. Shmelkov, C. Schmid, K. Alahari // Proceedings of the European Conference on Computer Vision. – 2018. – P. 213-229.
[10] Shin, H. Continual learning with deep generative replay / H. Shin, Jung K. Lee, J. Kim, J. Kim // Advances in Neural Information Processing Syst ems. – 2017. – P. 2990-2999.
[11] Xiao, H. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms [Electronic resource]. – Access mode: https://arxiv.org/abs/1708.07747 (15.09.2017).
[12] Krasser, M. Deep feature consistent variational auto-encoder [Electronic resource]. – Access mode: http://krasserm.github.io/2018/07/27/dfc-vae/ (27.07.2018).
[13] Sohn, K. Learning structured output representation using deep conditional generative models / K. Sohn, H. Lee, X. Yan // Neural Information Processing Systems. – 2015. – P. 3483-3491.
[14] Hou, X. Deep Feature Consistent Variational Autoencoder / X. Hou, L. Shen, K. Sun, G. Qiu // IEEE Winter Conference on Applications of Computer Vision. – 2017. – P. 1133-1141.
[15] Tan, M. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks / M. Tan, Q. Le // International Conference on Machine Learning. – 2019. – P. 6105-6114.

Quality metrics of variational autoencoders

M.I. Leontev[1,2], A.V. Mikheev[3], K.V. Sviatov[3], S.V. Sukhov[1,3]

[1]Kotel'nikov Institute of Radio Engineering and Electronics of Russian Academy of Sciences (Ulyanovsk branch), Spasskaya Str.14, Ulyanovsk, Russia, 432011
[2]S.P. Kapitsa Research Institute of Technology (Technological Research Institute) of Ulyanovsk State University, Universitetskaya Naberejnaya, Ulyanovsk, Russia, 432000
[3]Ulyanovsk State Technical University, Severny Venets 32, Ulyanovsk, Russia, 432027

Abstract. Variational autoencoders (VAEs) are popular models for the generation of realistic images. Usually, the quality of generated images is estimated only by visual inspection, as the methods for the quantitative estimation are not fully developed. In this work, we present and test methods that allow the possibility to evaluate the quality and the diversity of generated images objectively. The corresponding quality metrics are calculated for several implementations of VAE. The experiments are performed on publicly available datasets.