DOI 10.37882/2223-2966.2024.11.07

# СИНТЕЗИРОВАНИЕ ВИДЕО С ПОМОЩЬЮ ВАРИАЦИОННОГО АВТОЭНКОДЕРА И ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНОЙ МОДЕЛИ

## GENERATION OF SYNTHETIC VIDEOS ACCORDING TO THE NECESSARY PARAMETERS

***R. Vysotsky***
***V. Demichev***

*Summary*. The basic principles of operation of generative neural networks are considered using the example of developing a project for frame-by-frame video generation based on the first frame and a video sample. Using a variational autoencoder, the image is transformed into a feature space, and then a decoder is used to assemble a new image based on the sample. Additional use of a generative adversarial network allows for a significant improvement in the quality of the images created.

*Keywords*: neural networks, generation, video, autoencoder, discriminator.

***Высоцкий Роман Николаевич***
*Еврейский университет, г. Москва*
*roman0810.r@gmail.com*
***Демичев Василий Анатольевич***
*Доцент, Псковский государственный университет;*
*Еврейский университет, г. Москва*
*vademichev@gmail.com*

*Аннотация*. Рассматриваются основные принципы работы генеративных нейронных сетей на примере разработки проекта по покадровой генерации видео на основе первого кадра и видео-образца. С помощью применения вариационного автоэнкодера осуществляется преобразование изображения в пространство признаков, а затем с помощью декодера производится сборка нового изображения по образцу. Дополнительное применение генеративно-состязательной сети позволяет значительно повысить качество создаваемых изображений.

*Ключевые слова*: нейронные сети, генерация, видео, автоэнкодер, дискриминатор.

## Введение

Быстрое развитие сферы искусственного интеллекта и машинного обучения повлекло за собой появления множества сервисов по генерации различного рода контента, в том числе — синтетических видео, в которых, например, возможно заменить актера в фильме на другого человека, при этом он будет выполнять те же действия.

Некоторые такие сервисы, например Creative Reality Studio [1] имеют довольно хорошее качество работы. Большой интерес представляет методика генерации таких видео, но как правило подробности их работы в открытых источниках не раскрываются.

Настоящая работа представляет собой практическое исследование принципов создания генеративных сетей для синтеза видео, тонких моментов их настройки и обучения.

Мы рассмотрим преимущества и недостатки вариационного автоэнкодера (VAE), его устройство, а также объединим эту модель с генеративно-состязательной (GAN) для сведения к минимуму этих недостатков.

Демонстрируются полученные практические результаты.

## Вариационный автоэнкодер

Вариационный автоэнкодер применим для работы со многими типами данных, и не ограничивается изображениями, однако так как мы собираемся применять его только для обработки последних, все следующие примеры и решения будут касаться только изображений.

Принцип работы вариационного автоэнкодера заключается в том, что каждое изображение может быть сжато в определенную точку на пространстве признаков изображений и восстановлено из такой точки при необходимости. Причем размерность данного пространства должна быть намного меньше, чем количество признаков, которыми описано исходное изображение в виде описания цветов каждого отдельного пикселя. Таким образом, при сильном сжатии данных, описывающих картинку, остаются только ее важнейшие признаки, которые намного проще обрабатывать. Нечто подобное происходит, когда человек вспоминает увиденную картинку: в памяти откладывается не точная копия увиденного, а набор замеченных паттернов. Так, например, увидев красивую машину на дороге, потом легко вспомнить какого она была цвета, но невозможно вспомнить какой у нее был номер. Информация про номер машины была просто несущественной и не была отложена в памяти как признак увиденного.

Для реализации такой модели необходимо обучить две нейронные сети с зеркальной архитектурой — энкодер (encoder) и декодер (decoder). Энкодер будет отвечать за то, чтобы найти точку на пространстве признаков соответствующую входному изображению, а декодер, наоборот, должен будет найти изображение, соответствую-

щее такой точке. Обучение этих нейронных сетей в паре дает большое преимущество с точки зрения маркировки данных — одно и то же изображение будет являться как входными данными, так и правильным ответом для модели. Однако, такой подход не обеспечивает никакого контроля за тем, каким образом будут распределяться признаки на латентном пространстве (так же известном как пространство признаков или эмбеддингов) [2].

В конечном итоге целью разработки VAE является генерация новых изображений по заданным признакам, а это является неоправданно сложной задачей если заранее не будет известно каким образом они распределены. Для того, чтобы обойти данную трудность существует подход известный как репараметризация (reparametrization trick) [3]. Его суть заключается в том, чтобы энкодер вместо точки на латентном пространстве находил параметры распределения точек, соответствующих входному изображению. При этом декодер будет получать на вход одну из таких точек полученную случайным выбором на основе описанных энкодером параметров распределения (см. схему на рис. 1).

Таким образом, если в общую функцию потерь добавить значение метрики удаленности полученного энкодером распределения от стандартного нормального (дивергенции Кульбака-Лейблера [4]), то модель будет вынуждена распределять признаки изображений более кучно так как это будет снижать значение потерь. Это даст возможность более плавно переходить от одной картинки к другой при применении линейных операций к их представлениям

Данную модель можно применять как основу для генерации видео следующим образом: пусть есть видео-образец, на котором человек совершает какие-то движения, каждый кадр такого видео будет иметь свое представление, которое его описывает, а значит для любой пары соседних кадров разность будет выражена только тем, как сдвинулся человек. Так, если взять представление фотографии другого человека и последовательно добавлять к нему разности представлений между парами соседних кадров видео-образца, то на каждом шаге мы будем иметь представление человека с фотографии, двигающегося подобно человеку с видео. Восстановив полученные представления при помощи декодера, мы получим множество упорядоченных изображений, которые могут быть смонтированы в видео с человеком из начальной фотографии.

Для изучения работы данного метода мы создали вариационный автоэнкодер, обрабатывающий трехканальные изображения 64x64 пикселя. Энкодер и декодер данной модели работающие с 200-мерным признаковым пространством и состоящие из 7 сверточных и 3 линейных слоев были обучены на 250 тысячах изображений людей на 30 эпохах (см. схему энкодера и декодера на рис. 2). Реализация модели выполнялась на языке программирования Python при помощи библиотеки PyTorch с задействуем модуля DataLoader для осуществления потоковой подачи данных из постоянной памяти устройства в модель. Оптимизация реализованной модели выполнялась при помощи алгоритма Adam с функцией потерь при реконструкции BCELoss.

Как можно видеть на рис. 3, нейронная сеть обучилась воссоздавать изображения людей лишь в общих чертах — видны элементы одежды и основные черты лица, однако все остальное остается очень размытым и не детальным. Данную проблему удалось бы минимизировать, увеличив число обучаемых параметров в модели, но данное решение является как очень затратным с точки зрения вычислительных ресурсов, так и не устраняющим корень проблемы. С данным несовершенством полученного результата связан основной недостаток модели VAE — нечеткость генерируемых изображений. Его причина может крыться в слишком малой размерности признакового пространства и в несовершенстве части функции потерь, рассчитывающей качество получившегося изображения. И если размерность признако-

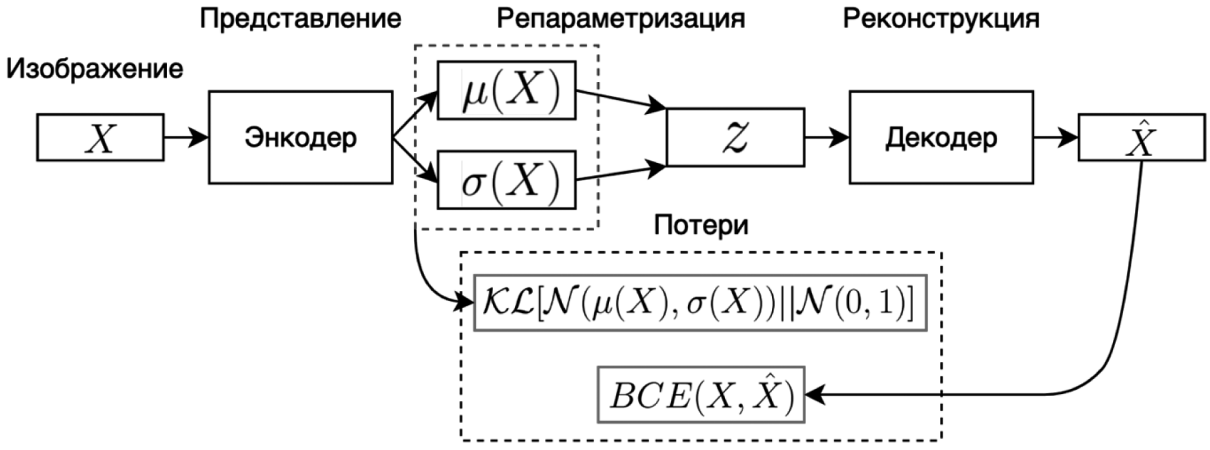


Рис. 1. Схема устройства вариационного автоэнкодера

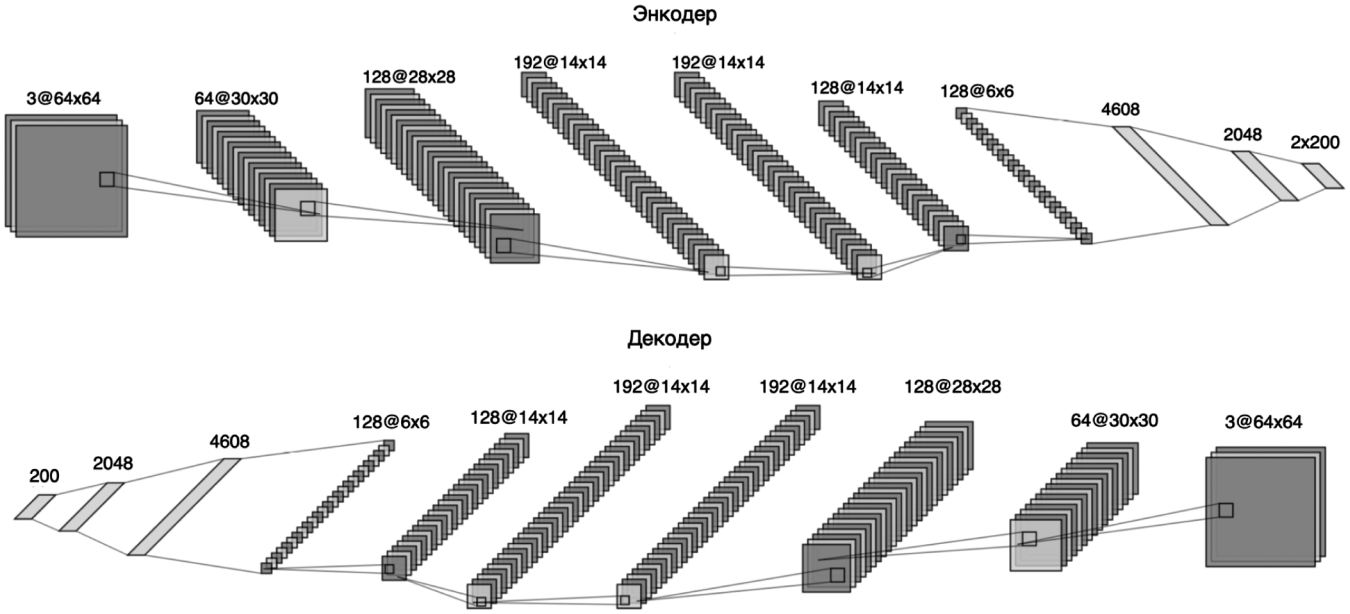


Рис. 2. Схема структуры энкодера и декодера

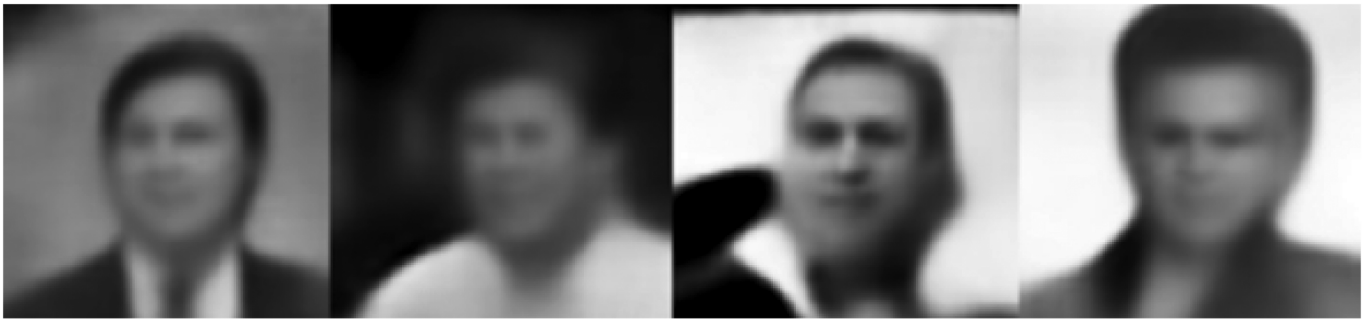

Рис. 3. Визуализация реконструкции изображений при помощи VAE

го пространства при необходимости можно просто увеличить, то с оценкой качества изображений возникают определенные трудности.

Обычно для такой оценки применяют метод суммы квадратов разностей или перекрестной энтропии между значениями цветов пикселей начального и полученного изображения, однако все это является не совсем подходящими критериями того, что мы сами считаем качественным изображением. Данные критерии можно по истине назвать чрезвычайно сложными, однако нам ничто не запрещает вновь прибегнуть к помощи нейронных сетей для того, чтобы определить их.

По своей сути эта идея не является новой, именно она лежит в основе генеративно-состязательных моделей, где две нейронные сети — генератор и дискриминатор обучаются в паре. Генератор создает изображения из случайно созданного шума, а дискриминатор дает оценку того, на сколько получившееся изображение похоже на настоящее. Так, если генератор создаст изображение, которое дискриминатор сочтет реальным, то значение его потерь будет минимальным и наоборот. За счет этого принципа и происходит обучение. Самое главное, что применение данной концепции не противоречит VAE, поэтому объединив их можно добиться лучших свойств от обоих этих моделей.

## Объединение VAE И GAN.

Прежде чем приступать к реализации улучшенной модели необходимо отметить, что дискриминатор сам по себе не сможет заменить функцию потерь вариационного автоэнкодера. Согласно принципам устройства состязательных моделей в этом случае вместо желаемых реконструкций получатся какие-то случайные изображения людей. Для предотвращения подобного исхода потери модели должны состоять из трех частей: потерь при репараметризации, стандартных потерь при реконструкции и потерь дискриминатора при целевом значении соответствующему не сгенерированному изображению. Каждая часть функции потерь в таком случае может иметь разные порядки, поэтому для них необходимо подобрать правильные гиперпараметры в соответствии с их значимостью для модели.

Таким образом, реализовав дискриминатор, состоящий из 4 сверточных и 3 линейных слоев была усовер-

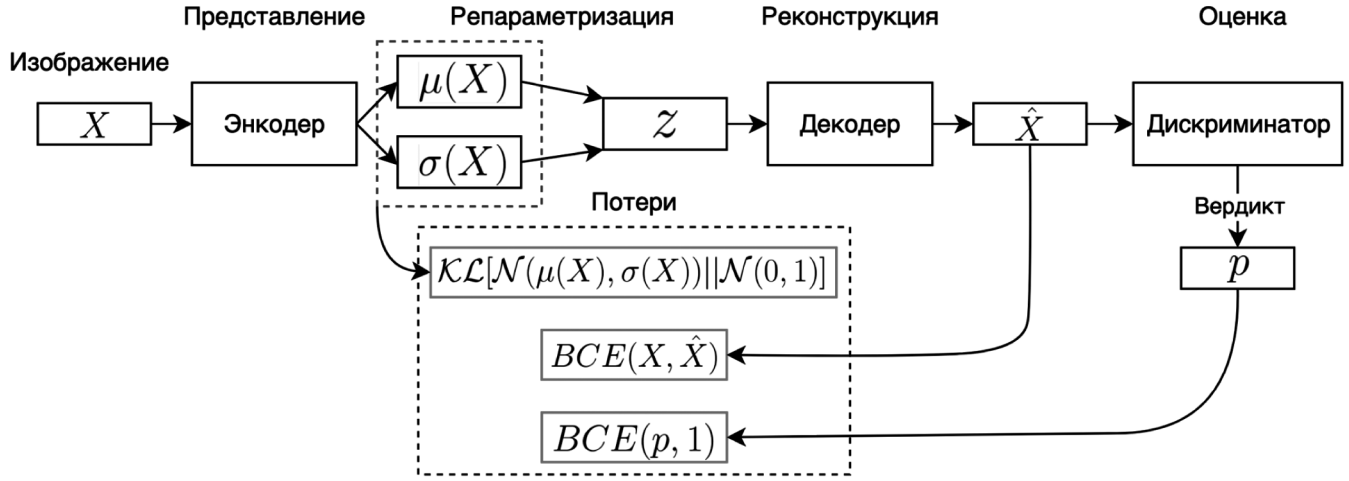


Рис. 4. Схема устройства VAE с дискриминатором

Рис. 5. Визуализация реконструкции изображений при помощи VAE+GAN

шенствована описанная в предыдущей главе модель (см. схему на рис. 4). Итерации обучения при этом были разбиты на сегменты обучения дискриминатора по реальным и созданным VAE фотографиям и, затем, обучения VAE при помощи вердикта, данного дискриминатором. Рассмотрим результаты такого обучения при все той же обучающей выборке на 30 эпохах обучения:

Как можно видеть на рисунке 5, качество реконструкций заметно улучшилось по сравнению с реконструкциями VAE изображенными на рисунке 3. Теперь, когда полученные реконструкции получились достаточно детальными, чтобы иметь возможность различать выражения лиц, можно генерировать кадры видео, согласно принципу изложенному в начале статьи.

Рис. 6. Исходное изображение человека

Рис. 7. Раскадровка видео-образца (сверху) и сгенерированного видео моделью VAE+GAN (снизу)

Несмотря на некоторые погрешности, на рисунке 7 видно, как человек с изображенный на рисунке 6 двигается в такт человеку на видео-образце. На сгенерированном видео прослеживаются как изменения в мими-

ке, так и движения головы. Тем не менее стоит обратить внимание на то, что человек на сгенерированном видео слишком много улыбается — данный эффект, предположительно, связан с особенностью используемой обучающей выборки The IMDB-WIKI dataset [5]. Поэтому, для лучшего качества генерации видео стоит использовать обучающие выборки, собранные из кадров других видео. Такие данные дадут возможность модели лучше выучить мимические движения.

Для лучшего понимания преимуществ модели VAE+GAN по сравнению с VAE, также приведем раскадровку видео сгенерированную VAE по аналогичным входным данным:

Рис. 8. Раскадровка сгенерированного видео моделью VAE

Описанная технология генерации видео может быть применена в индустрии видео-хостинга, киноиндустрии и в сфере развлечений.

Полученное решение можно усовершенствовать, повысив качество генерации видео. Для этого необходимо увеличить размер обрабатываемых изображений до 1920x1080 и увеличить число обучаемых параметров, задействовав более совершенные вычислительные мощности, а также повысить качество обучающих данных, дополнив их изображениями людей с разнообразными выражениями лиц, что является целью дальнейшей работы.

## Заключение

Рассмотрена задача синтезирования видео с использованием нейронных сетей.

С этой целью создана модель в виде вариационного автоэнкодера, позволяющего представить кадры видео в пространстве признаков и затем собрать обратно по требуемому паттерну.

Для повышения качества генерируемых кадров добавлена генеративно-состязательная модель.

В результате, при совместном использовании указанных моделей возможно осуществлять покадровую генерацию видео с заменой действующих персонажей в автоматическом режиме. При этом персонаж будет выполнять те же действия и мимику, что и его прототип.

Программная часть реализована с использованием библиотек PyTorch, PIL и numpy.

Проект имеет хороший потенциал для развития, ограниченный на данном этапе лишь доступной вычислительной мощностью и трудоемкостью сбора обучающих данных.

ЛИТЕРАТУРА


1. Creative Reality Studio URL: https://www.d-id.com/creative-reality-studio/ (дата обращения: 11.04.2024).
2. Эмбединги для начинающих // HABR.COM: 2024 18 янв. URL: https://habr.com/ru/companies/otus/articles/787116/ (дата обращения 19.05.2024).
3. Репараметризация в вариационных автоэнкодерах // BAELDUNG.COM: 2023 11 июня URL: https://www.baeldung.com/cs/vae-reparameterization (дата обращения: 06.01.2024).
4. Энтропия и дивергенция Кульбака-Лейблера // YANDEX.RU: учебник по машинному обучению 2020. URL: https://education.yandex.ru/handbook/ml/article/eksponencialnyj-klass-raspredelenij-i-princip-maksimalnoj-entropii (дата обращения: 21.02.2024).
5. The IMDB-WIKI dataset // ETHZ.CH: URL: https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/ (дата обращения: 23.11.2023).
6. Мозговой И.В. Распознавание неопределенных объектов с использованием нейросетей / И.В. Мозговой // Материалы Ежегодной межвузовской студенческой научной конференции ОЧУ ВО «Еврейский университет» : сборник тезисов, Москва, 05 апреля 2020 года / ОЧУ ВО «Еврейский университет». — Москва: ООО «МАКС Пресс», 2020. — С. 317–323. — EDN DURWLI.
7. Соколов В.О. Использовние современных технологий для создания образовательных программ / В.О. Соколов, Э.Н. Замега, В.А. Демичев // Антропологическая дидактика и воспитание. — 2022. — Т. 5, № 2. — С. 232–247. — EDN SKMZTY.



Журнал «Современная наука: актуальные проблемы теории и практики»