ГЕНЕРАТИВНЫЕ НЕЙРОСЕТИ: ОТ ИЗОБРАЖЕНИЯ ДО МУЗЫКИ

Малов Вадим Андреевич

студент, Поволжский государственный университет телекоммуникаций и информатики, РФ, г. Самара

Семионова Екатерина Сергеевна

студент, Поволжский государственный университет телекоммуникаций и информатики, РФ, г. Самара

Стефанова Ирина Алексеевна

научный руководитель, канд. техн. наук, доцент, Поволжский государственный университет телекоммуникаций и информатики, РФ, г. Самара

GENERATIVE NEURAL NETWORKS: FROM IMAGES TO MUSIC

Ekaterina Semionova

Student, Volga Region State University of Telecommunications and Informatics, Russia, Samara

Vadim Malov

Student, Volga Region State University of Telecommunications and Informatics, Russia, Samara

Irina Stefanova

Scientific Supervisor, Candidate of Technical Sciences, Associate Professor, Volga State University of Telecommunications and Informatics, Russia, Samara

Аннотация. Генеративные нейросети стали ключевым направлением в развитии искусственного интеллекта (ИИ), демонстрируя способность к созданию контента, сопоставимого с работами человека. В статье рассматриваются архитектуры и принципы работы основных классов генеративных моделей – вариационных автокодировщиков, генеративно-состязательных сетей, диффузионных моделей и трансформеров. Освещаются практические применения этих подходов в генерации изображений и музыкального контента. Анализируются современные достижения, ограничения технологий, а также этические вызовы, связанные с авторским правом, фейковым контентом и социальными последствиями распространения генеративного ИИ.

Abstract. Generative neural networks have become a key area in the development of artificial intelligence (AI.), demonstrating the ability to create content comparable to human work. The article discusses the architectures and principles of operation of the main classes of generative models – variational autoencoders, generative-adversarial networks, diffusion models and transformers. The practical applications of these approaches in image generation and music content are highlighted. The article analyzes modern achievements, limitations of technology, as

well as ethical challenges related to copyright, fake content and the social consequences of the spread of generative AI.

Ключевые слова: генеративные модели; GAN; VAE; трансформеры; диффузионные модели; генерация изображений; синтез музыки; искусственный интеллект.

Keywords: generative models; GAN; VAE; transformers; diffusion models; image generation; music synthesis; artificial intelligence.

Генеративный искусственный интеллект (Generative Artificial Intelligence, или GenAI) представляет собой совокупность моделей машинного обучения, способных создавать оригинальный контент – текст, изображения, музыку, аудио и видео – на основе ранее изученных данных. В отличие от традиционных алгоритмов, ограниченных задачами классификации или прогнозирования, генеративные модели ориентированы на продуктивную деятельность: они не только анализируют существующую информацию, но и производят новые, ранее не существовавшие элементы. Это позволяет говорить о GenAI как о важном шаге на пути к формированию машинного «творчества» [10].

Фундаментальная особенность генеративного искусственного интеллекта (ИИ) заключается в способности выявлять скрытые закономерности в больших объёмах данных. Нейросеть обучается на датасетах, содержащих миллионы образцов – будь то тексты, изображения или аудиофайлы – и формирует представление о структуре этих данных: стиле, композиции, логике и взаимосвязях. При взаимодействии с пользователем, например, при вводе текстового запроса в таких системах, как Midjourney или ChatGPT, модель использует накопленные представления для создания нового артефакта, визуально или семантически соответствующего исходному запросу. Таким образом, процесс генерации становится не просто механическим воспроизведением фрагментов обучающей выборки, а статистически обоснованным актом синтеза, имитирующим человеческую интуицию и креативность.

Сферы применения генеративного ИИ стремительно расширяются. Он используется в разработке цифрового дизайна, прототипировании, создании маркетинговых материалов [9, 10] (рис. 1), генерации музыки и звуков для кино- и игровой индустрии, а также в автоматизации рутинных творческих процессов. При этом, по мере интеграции таких технологий в повседневную практику, возникает всё больше вопросов, связанных с авторским правом, достоверностью создаваемого контента и изменением роли человека в творческой деятельности.

Рисунок 1. Графика, созданная ИИ

Примитивные генеративные модели уже несколько десятилетий используются в статистике для анализа числовых данных. Нейронные сети и глубокое обучение были предшественниками современного генеративного ИИ. Вариационные автокодировщики, разработанные в 2013 году, стали первыми глубокими генеративными моделями, способными генерировать реалистичные изображения и речь.

Эффективность генеративного ИИ во многом определяется архитектурными принципами, лежащими в основе моделей. Существует несколько ключевых подходов, каждый из которых обладает своими преимуществами и ограничениями, определяющими его область применения. Ниже кратко рассмотрены наиболее влиятельные архитектуры: генеративно-состязательные сети, вариационные автокодировщики, диффузионные модели и трансформеры.

Генеративно-состязательные сети (GAN), предложенные Иэном Гудфеллоу в 2014 году, состоят из двух взаимодействующих компонентов – генератора и дискриминатора. Генератор стремится создать реалистичные данные, в то время как дискриминатор пытается отличить сгенерированные образцы от настоящих. Это "состязание" приводит к прогрессивному улучшению обоих компонентов, позволяя создавать изображения высокого качества. GAN нашли широкое применение в синтезе фотореалистичных лиц (например, в модели StyleGAN), редактировании изображений и генерации deepfake-контента.

Вариационные автокодировщики (VAE) являются вероятностными моделями, которые обучаются сжимать данные в компактное представление (латентное пространство), а затем восстанавливать их. В отличие от обычных автокодировщиков, VAE используют статистические распределения, что делает возможной генерацию новых, оригинальных объектов. Эти модели применяются в задачах, требующих контролируемой генерации и интерполяции – например, в изменении характеристик изображений, аудиосигналах или генерации рукописного текста.

Диффузионные модели основываются на идее обратимого преобразования: они постепенно "зашумляют" данные в процессе обучения, а затем учатся восстанавливать их. На этапе генерации модель начинает с шума и шаг за шагом превращает его в осмысленное изображение. Такие модели, как Stable Diffusion и Imagen, продемонстрировали

впечатляющие результаты в генерации изображений на основе текстовых описаний. Они обеспечивают высокую степень детализации и контролируемости, хотя и требуют значительных вычислительных ресурсов.

Трансформеры стали архитектурным фундаментом большинства современных генеративных моделей. Изначально разработанные для задач обработки естественного языка, трансформеры быстро адаптировались к мультимодальному применению: от текста до изображений и музыки. Модели вроде GPT, DALL·E и MusicLM [1, 4] используют механизм самовнимания для построения контекста и генерации нового контента. Главное преимущество трансформеров – способность эффективно обрабатывать большие объёмы данных и учитывать сложные взаимосвязи между элементами входа.

Одним из наиболее активно развивающихся направлений генеративного искусственного интеллекта является синтез изображений. С появлением мощных моделей, таких как DALL·E, MidJourney, Stable Diffusion и Imagen, стало возможным создавать фотореалистичные или стилизованные визуальные сцены по простому текстовому описанию. Эти системы широко применяются в дизайне, цифровом искусстве, рекламе, архитектуре и даже в научной визуализации.

В основе большинства современных моделей лежат диффузионные или трансформерные архитектуры. Так, модель DALL·E 2, разработанная OpenAI, основана на трансформере и сочетает в себе возможности CLIP (Contrastive Language–Image Pretraining) – модели, обученной соотносить изображения и текстовые описания. Такой подход позволяет DALL·E интерпретировать текстовый запрос не просто как метку, а как сложную семантическую структуру, содержащую информацию о цвете, стиле, ракурсе, контексте и эмоциях изображения. Stable Diffusion, напротив, использует диффузионную модель – она добавляет к изображению шум, а затем учится его поэтапно устранять, приближаясь к целевой визуальной структуре. Одним из её достоинств является способность к локальной генерации изображений: модель может работать даже на персональных компьютерах с видеокартой, в отличие от облачных решений, таких как DALL·E или MidJourney. Stable Diffusion стала популярной в сообществе художников, так как позволяет детально управлять параметрами изображения – от выбора начального шума до точной генерации по шаблону.

Генерация изображений с помощью нейросетей выходит далеко за рамки развлечений. Эти технологии активно применяются в архитектурном проектировании (создание концептуальных планов зданий), медицине (синтез медицинских изображений для обучения моделей распознавания заболеваний), а также в образовании, где визуализация абстрактных понятий может значительно повысить понимание материала. Например, в химии или биологии нейросети способны создавать изображения молекул, клеток или анатомических структур [9] по описанию. Одновременно с ростом качества изображений возрастают и вызовы. Главный из них – проблема достоверности и манипуляции визуальным контентом. Генерация реалистичных, но фальшивых фотографий [9, 10] (deepfake), может быть использована в дезинформационных компаниях. Это требует разработки систем маркировки, распознавания искусственно созданных изображений и внедрения этических норм при использовании подобных инструментов. Генерация видео с помощью нейросетей представляет собой одну из самых технологически сложных и перспективных задач в области генеративного искусственного интеллекта. Если синтез изображений требует согласованности в пространстве, а генерация музыки – во времени, то видеоконтент требует соблюдения и пространственной, и временной непрерывности одновременно. Каждое видео состоит из последовательности изображений (кадров), которые должны быть не только реалистичны по отдельности, но и логично выстроены в динамике. Это требует от моделей способности понимать движение, взаимодействие объектов, контекст сцены и даже кинематографические приёмы, такие как смена фокуса или движение камеры.

Первые подходы к генерации видео были основаны на трёхмерных свёрточных нейросетях (3D-ConvNets), способных обрабатывать несколько кадров одновременно как единый объём данных. Однако этот метод оказался ограниченным в длине видео и детализации движения. В дальнейшем в генерацию видео стали внедряться архитектуры трансформеров, хорошо зарекомендовавшие себя в текстовых и аудиозадачах [2, 8], адаптированные к работе с временными зависимостями. Одним из современных решений стали диффузионные модели,

обучающиеся пошагово восстанавливать видео из шума. Такие системы, как Imagen Video или Sora от OpenAI, позволяют по текстовому описанию создавать короткие, но детализированные видеоролики с реалистичным движением и глубиной сцены. Несмотря на серьёзные достижения, область генеративного видео по-прежнему сталкивается с рядом технических ограничений. Большинство моделей способны генерировать видео продолжительностью всего несколько секунд, и даже при этом возникает риск артефактов: объекты могут искажаться, исчезать или вести себя физически неправдоподобно. Кроме того, процесс генерации остаётся крайне ресурсоёмким и требует значительных вычислительных мощностей, что делает технологию менее доступной для широкого круга пользователей по сравнению, например, с генерацией изображений. Тем не менее потенциал этих систем огромен. В киноиндустрии генеративные нейросети могут применяться для создания раскадровок, анимаций или даже целых сцен, не требующих участия актёров. В игровой индустрии они позволяют разрабатывать динамический визуальный контент или живые кат-сцены. В образовательных и научных целях возможна генерация визуальных симуляций процессов, которые трудно или невозможно наблюдать в реальности. При этом необходимо учитывать и потенциальные риски – в частности, распространение поддельного видео, нарушающего этические и правовые нормы. Именно поэтому развитие генеративных моделей должно сопровождаться разработкой надёжных методов распознавания искусственно созданного контента и системой регулирования его использования.

Ранее создание мелодий, гармоний, аранжировок и вокала было исключительно областью человеческой интуиции, слуха и культурного опыта, но теперь нейросети способны не просто имитировать отдельные элементы музыкального произведения, но и создавать их с нуля – с учётом структуры, стиля, ритма и даже эмоций. Музыка, как и видео, представляет собой временную форму искусства, где важны не только отдельные звуки, но и их последовательность, динамика, развитие тем и мотивов. Именно поэтому модели, работающие с аудио, должны учитывать временные зависимости и целостность звуковой композиции. Технически генерация музыки реализуется через глубокие нейросетевые архитектуры, чаще всего основанные на трансформерах или вариационных автокодировщиках. MusicLM от Google* способна генерировать музыку по текстовому описанию с учётом жанра, инструментов и настроения. Модель обучена на масштабных аудиодатасетах и может создавать уникальные композиции. Jukebox от OpenAI дополняет этот подход генерацией вокала и стилистическим подражанием реальным исполнителям, что расширяет границы применения ИИ в музыкальной сфере. Аудиогенерация выходит далеко за рамки музыки. Современные модели способны синтезировать человеческую речь [5] с учётом тембра, ритма, акцента и даже эмоций. Эти технологии применяются в голосовых ассистентах, навигационных системах, озвучивании фильмов и создании аудиокниг. Особое внимание уделяется моделям, способным персонализировать голос – например, воспроизвести голос конкретного человека по короткой голосовой выборке. Такие возможности открывают потенциал для инклюзивных технологий (например, для людей с нарушениями речи), но одновременно создают серьёзные этические риски, включая угрозу появления достоверных, но поддельных аудиозаписей.

Генеративный ИИ становится не просто технологией, а новой формой творческого и культурного производства. Нейросети уже применяются для создания изображений, видео, музыки, текста и речи, всё больше влияя на сферы дизайна, науки, образования и медиа. При этом возникает всё больше вопросов, связанных с авторским правом, достоверностью и этикой. Поскольку границы между машинным и человеческим творчеством постепенно размываются, важно вырабатывать механизмы регулирования и критического осмысления новых инструментов. Генеративные нейросети вряд ли заменят человека, но уже сегодня становятся его активным партнёром в расширении границ возможного.

Список литературы:

1. Agostinelli A., Copet A., Pascual R. и др. MusicLM: Generating Music From Text / Agostinelli A. и др. – [электронный ресурс]. – arXiv:2301.11325, 2023. – Режим доступа. – URL: https://arxiv.org/abs/2301.11325 (дата обращения: 06.06.2025).

2. Benaim S., Goel S., Singh A. и др. From Text to Video: Zero-Shot Generation using Diffusion Models / S. Benaim и др. – [электронный ресурс]. – arXiv:2209.03150, 2022. – Режим доступа. – URL: https://arxiv.org/abs/2209.03150 (дата обращения: 06.06.2025).

3. Dhariwal P., Nichol A. Diffusion Models Beat GANs on Image Synthesis / Prafulla Dhariwal, Alex Nichol – [электронный ресурс]. — arXiv:2105.05233, 2021. – Режим доступа. – URL: https://arxiv.org/abs/2105.05233 (дата обращения: 06.06.2025).

4. Introducing Sora: a text-to-video model – [электронный ресурс]. – OpenAI, 2024. – Режим доступа. – URL: https://openai.com/sora (дата обращения: 06.06.2025).

5. Jukebox: A Neural Net That Generates Music, Including Singing – [электронный ресурс]. – OpenAI, 2020. – Режим доступа. – URL: https://openai.com/research/jukebox (дата обращения: 06.06.2025).

6. Karras T., Laine S., Aila T. Analyzing and Improving the Image Quality of StyleGAN / T. Karras, S. Laine, T. Aila. – [электронный ресурс]. – In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2020. – Режим доступа. – URL: https://arxiv.org/abs/1912.04958 (дата обращения: 06.06.2025).

7. MusicGen: Simple and Controllable Music Generation – [электронный ресурс]. – Meta AI, 2023. – Режим доступа. – URL: https://github.com/facebookresearch/audiocraft (дата обращения: 05.06.2025).

8. Singer Y., Polyak A., Hayes T. и др. Make-A-Video: Text-to-Video Generation without Text-Video Data / Y. Singer и др. – [электронный ресурс]. – arXiv:2209.14792, 2022. – Режим доступа. – URL: https://arxiv.org/abs/2209.14792 (дата обращения: 05.06.2025).

9. Поспелова Е. А., Отоцкий П. Л., Горлачева Е. Н., Файзуллин Р. В. Генеративный искусственный интеллект в образовании: текущие тенденции и перспективы – [электронный ресурс] // Профессиональное образование и рынок труда. – 2024. – Т. 12, № 3. – С. 6–21. – Режим доступа. – URL: https://www.po-rt.ru/articles/2173 (дата обращения: 05.06.2025).

10. Что такое генеративный искусственный интеллект? – [электронный ресурс]. – Amazon Web Services. – Режим доступа. – URL: https://aws.amazon.com/ru/what-is/generative-ai/ (дата обращения: 04.06.2025).

*По требованию Роскомнадзора информируем, что иностранное лицо, владеющее информационными ресурсами Google является нарушителем законодательства Российской Федерации – прим. ред.)