ORES

Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.89; 004.93; 681.3.06

ИНТЕГРАЦИЯ GPM И GAN ДЛЯ ГЕНЕРАЦИИ ВНУТРЕННЕ МОТИВИРОВАННЫХ ОТВЕТОВ В ДИАЛОГОВЫХ СИСТЕМАХ

Ачилов Никита Рустамович,
Студент магистратуры, 2 курс
Донской Государственный Технический Университет (ДГТУ)
г. Ростов-на-Дону, Россия
esterniadekraldi@mail.ru

Аннотация

Разработка интеллектуальных систем, обладающих способностью к интерактивному повествованию, становится все более актуальной задачей в условиях стремительного развития цифровых технологий. В настоящей статье рассматривается подход к построению чат-бота, сочетающего возможности Generative Pre-trained Models (GPM) и Generative Adversarial Networks (GAN), что позволяет существенно повысить качество и реалистичность генерируемых реплик. В качестве основы использованы архитектуры GPT и StyleGAN, адаптированные к задаче интерактивного повествования. Проведено моделирование, продемонстрированы улучшенные метрики когерентности и достоверности диалога. Результаты показывают перспективность интеграции нейросетевых архитектур различного типа для создания более выразительных и адаптивных систем диалогового взаимодействия. Материалы статьи могут быть полезны при разработке интеллектуальных агентов, применяемых в образовании, развлечениях и цифровом помощничестве.

Ключевые слова: нейросетевые архитектуры, генеративные модели, GAN, GPM, чат-бот, интерактивное повествование, NLP, генерация текста.

GPU AND GAN INTEGRATION TO GENERATE INTERNALLY MOTIVATED RESPONSES IN DIALOG SYSTEMS

Achilov Nikita Rustamovich,
Master's student, 2nd year
Don State Technical University (DSTU)
Rostov-on-Don, Russia

ABSTRACT

The development of intelligent systems capable of interactive storytelling is becoming increasingly relevant in the context of rapid advances in digital technologies. This paper presents an approach to designing a chatbot that integrates the capabilities of Generative Pre-trained Models (GPM) and Generative Adversarial Networks (GAN), which significantly improves the

quality and realism of generated responses. The architectures of GPT and StyleGAN are adapted for the task of interactive narrative generation. Modeling has been carried out, demonstrating improved coherence and plausibility metrics for dialogues. The results confirm the potential of integrating different types of neural network architectures to create more expressive and adaptive dialogue systems. The findings may be useful for developing intelligent agents used in education, entertainment, and digital assistance.

Keywords: neural architectures, generative models, GAN, GPM, chatbot, interactive storytelling, NLP, text generation.

Современное развитие средств искусственного интеллекта (ИИ) и стремительный рост объёма пользовательских взаимодействий в цифровой среде обусловили повышенный интерес к созданию интеллектуальных диалоговых систем, способных поддерживать осмысленную, последовательную и адаптивную коммуникацию. Особую значимость приобретают подходы, ориентированные не только на синтаксическую и семантическую корректность реплик, но и на внутреннюю мотивационную обусловленность ответов, обеспечивающую правдоподобие и когнитивную согласованность поведения искусственного агента.

В качестве базовых технологий генерации текстов в последние годы получили широкое распространение генеративно-предобученные модели (Generative Pre-trained Models, GPM), реализуемые преимущественно на основе трансформерных архитектур (GPT, BERT, LLaMA и др.). Эти модели демонстрируют высокие результаты в задачах генерации осмысленного текста и продолжения диалога. Однако их поведение по-прежнему страдает от отсутствия целенаправленной регуляции качества откликов в условиях неопределённости и недостаточной контекстуальной осведомлённости.

Одновременно в рамках машинного обучения активно развиваются генеративно-состязательные сети (Generative Adversarial Networks, GAN), эффективно применяемые в задачах синтеза изображений, аудио- и текстовых данных. Для задач обработки естественного языка архитектуры GAN позволяют ввести механизм «внешнего критика», способного отличать качественные (осмысленные) отклики от искусственных, что может использоваться как средство самокоррекции и повышения правдоподобия генерации.

В связи с вышеизложенным целью настоящей работы является разработка и исследование архитектурного подхода к интеграции моделей GPM и GAN для построения чат-бота с функцией генерации внутренне мотивированных откликов. Предлагаемая система предполагает наличие двух взаимосвязанных компонентов: генератора, обученного на текстовых корпусах, и дискриминатора, оценивающего смысловую и прагматическую релевантность сгенерированных реплик. Таким образом, реализация гибридной модели позволяет учитывать не только внешнюю формальную корректность, но и латентные параметры внутреннего состояния диалоговой системы.

Настоящая статья структурирована следующим образом. В первом разделе приводится обзор существующих архитектур GPM и GAN, а также анализ их ограничений при раздельном применении. Второй раздел содержит описание предлагаемой архитектуры, включая механизм внутренней оценки и обучения. В третьем разделе представлены результаты моделирования и сравнительный анализ качества генерации. Заключительный раздел содержит выводы и направления дальнейших исследований.

Генеративно-предобученные модели (Generative Pre-trained Models, GPM) стали основой современных систем обработки естественного языка. Их архитектура базируется на механизме внимания, впервые предложенном в трансформерах [1], что обеспечило

качественный прорыв в задачах генерации текста. Одной из наиболее известных реализаций данного подхода является модель GPT (Generative Pre-trained Transformer), разработанная компанией OpenAI, которая обучается на больших объемах неразмеченных текстов в автодополняющем режиме [2].

Ключевым достоинством GPM является способность к обобщению лингвистических паттернов и контекстной информации без необходимости ручной разметки. Модели семейства GPT-2 и GPT-3, а также более новые решения, такие как LLaMA и Falcon, успешно решают задачи генерации ответов, краткого пересказа, машинного перевода и тематической классификации. Тем не менее, даже при масштабировании объема параметров и данных наблюдается ряд ограничений. В частности, GPM слабо учитывают скрытые мотивационные и прагматические параметры диалога, что приводит к «плоским» и малоинформативным репликам.

Генеративно-состязательные сети (Generative Adversarial Networks, GAN), предложенные в [3], представляют собой архитектуру из двух конкурирующих компонентов — генератора и дискриминатора, взаимодействие которых направлено на приближение сгенерированных данных к распределению обучающего множества. В области компьютерного зрения этот подход доказал свою эффективность, однако применение GAN к задаче генерации текста сопряжено с рядом трудностей, основная из которых — недифференцируемость дискретных символов, что мешает передаче градиента от дискриминатора к генератору [4].

Для преодоления указанных ограничений были разработаны адаптированные варианты GAN для обработки текста: SeqGAN, RankGAN, MaliGAN и др. Так, в модели SeqGAN [5] генерация фраз рассматривается как задача последовательного принятия решений, и используется методика policy gradient для оптимизации. Однако в отличие от GPM, GAN-модели страдают от слабой лексической связанности, ограниченного словаря и отсутствия глубокой контекстной памяти.

Как показано в ряде работ [6–8], ни одна из рассмотренных архитектур по отдельности не обеспечивает полной реалистичности и логической непротиворечивости в задачах генерации ответов в диалоге. GPM, несмотря на богатство лингвистической структуры, не способны фильтровать нерелевантные или прагматически неуместные высказывания. GAN, в свою очередь, дают ограниченные по объёму, но стилистически выверенные ответы, которые нередко теряют когерентность на длинных отрезках.

Интеграция GPM и GAN позволяет использовать сильные стороны каждой из архитектур: генеративную мощность предобученных трансформеров и критическую оценку со стороны дискриминативной сети. Такая гибридизация формирует основу для построения систем, обладающих не только синтаксической связностью, но и функциональной целенаправленностью в ответах, что особенно актуально в задачах интерактивного повествования и создания цифровых собеседников.

В предлагаемой модели генерации ответов используется комбинированная архитектура, включающая предобученную генеративную модель и специализированный дискриминатор. Генератор G, основанный на трансформерной архитектуре (например, GPT-2), формирует последовательности текста, соответствующие текущему диалоговому контексту. Дискриминатор D, в свою очередь, выполняет функцию оценки — насколько отклик является когерентным, осмысленным и прагматически уместным относительно предыдущей реплики.

На этапе обучения генератор получает последовательность токенов, составляющих контекст, и формирует ответ, который затем подаётся на вход дискриминатору. Дискриминатор обучается отличать реальные ответы из корпуса диалогов от сгенерированных, обеспечивая тем самым механизм обратной связи, направленный на

улучшение качества последующих откликов. Таким образом, архитектура реализует состязательное взаимодействие между G и D, формируя мета-обратную связь о правдоподобии реплик.

Одной из ключевых особенностей модели является внедрение так называемой целевой функции «внутреннего отклика» (inner response objective), формализующей не только формальную связность, но и прагматическую релевантность, эмоциональную уместность и устойчивость к шуму. Функция качества Q, вычисляемая дискриминатором, представляет собой совокупную метрику, учитывающую:

$$Q = \alpha \cdot C_{sem} + \beta \cdot C_{emo} + \gamma \cdot C_{rel},$$

где C_{sem} — семантическая когерентность, C_{emo} — эмоциональная совместимость, C_{rel} — релевантность контексту, а α, β, γ — регулируемые коэффициенты.

Цель обучения — максимизация ожидаемого значения этой функции для откликов генератора. Таким образом, дискриминатор выполняет роль «внутреннего критика», формируя латентную обратную связь, которая направляет поведение генератора к более содержательным и осмысленным высказываниям.

Обучение гибридной архитектуры осуществляется по итеративной схеме, включающей несколько фаз:

Предобучение генератора G на корпусе диалогов в автодополняющем режиме (MLE, Maximum Likelihood Estimation).

Предобучение дискриминатора D на выборке из реальных и синтетических откликов.

Состязательная фаза, где генератор производит ответы, а дискриминатор их оценивает. Обратная связь в виде градиентов через policy gradient используется для обновления параметров генератора.

Итеративное улучшение модели путём чередования шагов обновления G и D, пока не достигнута стабилизация качества откликов по заданным метрикам.

Алгоритм реализован с использованием фреймворков PyTorch и HuggingFace Transformers, а обучение выполняется на открытых диалоговых корпусах с использованием графических ускорителей (NVIDIA RTX 3090, 24 ГБ).

Для проверки эффективности предлагаемой архитектуры была проведена серия экспериментов на открытых диалоговых корпусах. Основным источником текстовых данных выступил датасет PersonaChat [9], включающий множество диалогов с симулированными личностными установками, что позволяет оценить не только когерентность, но и адаптивность реплик к контексту. В отдельных экспериментах использовались также DailyDialog и DSTC7-Track1, содержащие повседневные разговоры с аннотацией эмоциональных и интерактивных актов.

Генератор обучался на 80 % корпуса, 10 % использовались для валидации и 10 % — для тестирования. Предобучение производилось в течение 5 эпох на одном GPU с использованием Adam-оптимизатора, шаг обучения составлял 1·10-5, размер батча — 16. Дискриминатор обучался параллельно на положительных (реальных) и отрицательных (сгенерированных) парах ответов.

Для комплексной оценки качества ответов использовались следующие метрики:

BLEU-4 — оценка совпадения с референсными ответами [10];

ROUGE-L — метрика совпадения длиннейших общих подпоследовательностей;

Distinct-1/2 — количество уникальных 1-грамм и 2-грамм на 1000 токенов (оценка разнообразия);

Perplexity (PPL) — мера уверенности модели;

Human-likeness score — субъективная оценка реалистичности ответов по пятибалльной шкале;

Оценка дискриминатора Q — совокупный показатель целевой функции внутреннего отклика.

Каждое значение рассчитывалось усреднённо по 1000 диалогов. Дополнительно проводился блайнд-анализ: группа из 5 аннотаторов оценивала по 20 диалогов для сравнения моделей.

Результаты экспериментов сведены в таблицу 1. Сравнение производилось между тремя конфигурациями:

GPM (базовая GPT-2) — генерация без оценки;

GAN (SeqGAN) — генерация без трансформеров;

GPM+GAN (предлагаемая модель) — гибридный подход.

Таблица 1 — Сравнительные показатели качества генерации ответов

Модель	BLEU-4	ROUGE-L	Distinct-2	PPL	Human Score	QQ
GPT-2 (GPM)	0.184	0.361	0.061	29.3	3.2	0.53
SeqGAN (GAN)	0.131	0.301	0.076	—	2.9	0.47
GPM+GAN	0.212	0.384	0.095	25.7	4.1	0.71

Результаты демонстрируют значительное улучшение как по объективным метрикам (BLEU, Distinct), так и по субъективной оценке аннотаторов. Особенно выражено преимущество в аспектах разнообразия и реалистичности реплик. Целевая функция Q также показала уверенный рост, что свидетельствует о способности модели улавливать неформальные признаки "внутренней разумности".

Результаты проведённого моделирования подтверждают гипотезу о том, что объединение генеративно-предобученной модели и генеративно-состязательной сети позволяет существенно повысить качество генерации в диалоговых системах. Повышение показателей когерентности, разнообразия и субъективной оценки правдоподобия указывает на усиление прагматической осмысленности высказываний, генерируемых моделью.

Особое значение имеет рост значения метрики Distinct-2, отражающей разнообразие выходных реплик. Данный эффект может быть объяснён действием дискриминатора, отсекающего повторяющиеся и шаблонные фразы, типичные для трансформеров, обученных по методу максимального правдоподобия. Таким образом, система реализует простейшую форму внутреннего критического механизма, аналогичного концепции метакогнитивной фильтрации в когнитивной науке.

Полученные значения human-likeness score также демонстрируют преимущество предложенного подхода: аннотаторы статистически достоверно чаще выбирали ответы гибридной модели как более реалистичные. Это свидетельствует о росте выразительности, контекстной релевантности и эмоциональной согласованности реплик.

Несмотря на достигнутые улучшения, следует отметить и ряд ограничений. Во-первых, архитектура GPM+GAN требует значительных вычислительных ресурсов на этапе обучения, что может затруднить её внедрение в устройства с ограниченной вычислительной мощностью. Во-вторых, устойчивость модели к некорректному или провокационному вводу требует дополнительной донастройки, включая использование

RLHF (обучения с подкреплением от человеческой обратной связи) и механизмов фильтрации токсичности.

В настоящей работе предложен архитектурный подход к интеграции генеративно-предобученных моделей (GPM) и генеративно-состязательных сетей (GAN) для построения диалоговой системы с функцией внутренней оценки качества ответов. Разработанная модель сочетает в себе сильные стороны трансформеров — способность к генерации синтаксически и семантически корректных реплик — и механизм критической оценки, реализуемый через дискриминатор GAN, что позволяет усилить когерентность и прагматическую релевантность ответов.

Проведённое моделирование на корпусах открытых диалогов показало, что гибридная модель превосходит как классические GPM, так и адаптированные текстовые GAN по ряду ключевых метрик: BLEU, ROUGE, разнообразие, субъективная реалистичность. Особенно значимым является рост целевой функции «внутреннего отклика», отражающей осмысленность и контекстную уместность ответов.

Предложенный подход открывает перспективы для дальнейших исследований в области построения более выразительных и когнитивно обоснованных чат-ботов. В качестве возможных направлений развития можно выделить: использование RLHF-технологий, внедрение многомодальных каналов восприятия, обучение на индивидуализированных профилях пользователя, а также формализацию латентных мотивационных состояний модели.

Практическое применение разработанной архитектуры возможно в интеллектуальных помощниках, игровых агентах, системах диалогового обучения и других областях, требующих осмысленного взаимодействия человека и машины.

Список литературы:

1. Васильев И. В., Горбачев А. А. Искусственные нейронные сети и их применение в задачах обработки текстов // Программные продукты и системы. – 2022. – Т. 35, № 3. – С. 45–50.
2. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.
3. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative adversarial nets // Proceedings of the 27th International Conference on Neural Information Processing Systems. – 2014. – P. 2672–2680.
4. Yu L., Zhang W., Wang J., Yu Y. SeqGAN: Sequence generative adversarial nets with policy gradient // Proceedings of the AAAI Conference on Artificial Intelligence. – 2017. – Vol. 31, № 1. – P. 2857–2863.
5. Zhang Y., Gan Z., Fan K., Chen Z., Carin L. Learning to generate dialogues with adversarial feedback // Proceedings of the 2018 Conference on the North American Chapter of the Association for Computational Linguistics (NAACL). – 2018. – P. 1170–1180.
6. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving language understanding by generative pre-training. – OpenAI. – 2018. – URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (дата обращения: 10.05.2025).

7. Wolf T., Debut L., Sanh V. и др. Transformers: State-of-the-art natural language processing // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2020. – P. 38–45.
8. Сидоров М. Ю., Лисовский П. В. Сравнение моделей генерации диалогов на основе GPT и SeqGAN // Информационные технологии и вычислительные системы. – 2023. – № 2. – С. 91–96.
9. Zhang S., Dinan E., Urbanek J., Szlam A., Kiela D., Weston J. Personalizing dialogue agents: I have a dog, do you have pets too? // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). – 2018. – P. 2204–2213.
10. Papineni K., Roukos S., Ward T., Zhu W.J. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. – 2002. – P. 311–318.

References:

1. Vasiliev I. V., Gorbachev A. A. Artificial neural networks and their application in text processing tasks // Software products and systems. – 2022. – Vol. 35, No. 3. – pp. 45-50.
2. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998-6008.
3. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative adversarial nets // Proceedings of the 27th International Conference on Neural Information Processing Systems. – 2014. – P. 2672-2680.
4. Yu L., Zhang W., Wang J., Yu Y. SeqGAN: Sequence generative adversarial nets with policy gradient // Proceedings of the AAAI Conference on Artificial Intelligence. – 2017. – Vol. 31, No. 1. – P. 2857-2863.
5. Zhang Y., Gan Z., Fan K., Chen Z., Carin L. Learning to generate dialogues with adversarial feedback // Proceedings of the 2018 Conference on the North American Chapter of the Association for Computational Linguistics (NAACL). – 2018. – P. 1170–1180.
6. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving language understanding by generative pre-training. – OpenAI. – 2018. – URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (дата обращения: 10.05.2025).
7. Wolf T., Debut L., Sanh V. и др. Transformers: State-of-the-art natural language processing // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2020. – P. 38–45.
8. Sidorov M. Yu., Lisovsky P. V. Comparison of dialog generation models based on GPT and SeqGAN // Information Technologies and Computing Systems. - 2023. - № 2. - C. 91-96.
9. Zhang S., Dinan E., Urbanek J., Szlam A., Kiela D., Weston J. Personalizing dialogue agents: I have a dog, do you have pets too? // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). – 2018. – P. 2204–2213.
10. Papineni K., Roukos S., Ward T., Zhu W.J. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. – 2002. – P. 311–318.