

УДК 614.8.084

ПОРІВНЯННЯ ТРАНСФОРМЕРІВ І GAN У СИСТЕМАХ TTS

Федорук Н. С.¹, студ., Плахтій Є. Г.², ст. викл.

ННІ «Придніпровська державна академія будівництва та архітектури»

Українського державного університету науки і технологій

¹ nikitafedoruk784@gmail.com, ² plakhtii.yevhen@pdaba.edu.ua

Постановка проблеми. Синтез мовлення (TTS, text-to-speech) – технологія, що переводить текст у штучно згенерований голос, наближений до природного. Сучасні TTS використовуються в голосових асистентах (Siri, Alexa, Google Assistant) та доступних технологіях для користувачів із вадами зору. Завдяки розвитку глибокого навчання з'явилися два основні підходи: трансформери та генеративні змагальні мережі (GAN). Трансформери, спочатку успішні в NLP, зараз використовують для моделювання послідовностей символів або фонем з довготерміновими залежностями [1]. GAN, що здобули популярність у генерації зображень, тепер адаптовані для аудіо – генератор створює сигнал, а дискримінатор перевіряє його реалістичність [2]. Обидва підходи важливі для досягнення натурального і виразного синтезованого мовлення, хоча кожен має свої сильні та слабкі сторони.

Мета дослідження. Метою роботи є порівняльний аналіз трансформерних і GAN-підходів у системах синтезу мовлення. Зокрема, дослідження спрямоване на виявлення ключових характеристик цих моделей, їх відмінностей, переваг і недоліків при генерації мовлення, а також визначення шляхів подолання наявних проблем обох підходів.

Результати дослідження. Отже, у сучасних системах TTS спостерігаються наступні особливості:

1. Архітектура: Transformer-моделі для TTS використовують механізм багатоголової самоуваги, що дозволяє ефективно моделювати довготермінові залежності в послідовності тексту та здійснювати обчислення паралельно [1]. Замінивши рекурентні мережі Tacotron2 на самоувагу, перша трансформерна TTS-модель усунула дві головні проблеми: низьку швидкодію RNN під час навчання та генерування, а також труднощі моделювання далеких залежностей [1]. Натомість GAN-підхід базується на змагальному навчанні: модель-генератор генерує аудіосигнал (наприклад, мовленнєвий спектрограм або безпосередньо хвилю), а модель-дискримінатор намагається відрізнити згенероване мовлення від справжнього та перевіряє відповідність мовлення заданому тексту [2].

2. Якість синтезу: Обидва підходи здатні генерувати високо натуральний голос. Transformer-архітектури у поєднанні з потужними вокодерами демонструють якість, близьку до людської: так, модель Transformer TTS досягла середньої оцінки MOS = 4,39, що майже дорівнює якості живої мови (4,44) [1]. GAN-базовані системи також досягли аналогічного рівня природності. Зокрема, модель GAN-TTS (DeepMind) генерує мовлення з високою достовірністю та натуральністю, порівняною з найкращими авторегресивними підходами [2]. Сучасні GAN-вокодери (наприклад, HiFi-GAN) здатні відтворювати 22,05 кГц аудіо практично студійної якості – за суб'єктивними оцінками, якість майже не відрізняється від людської [3]. Таким чином, при достатньому навчанні обидва підходи забезпечують природне звучання вихідної мови.

3. Продуктивність: Трансформери дозволяють прискорити синтез. Завдяки паралельній обробці послідовності, неголосові (non-autoregressive) трансформерні моделі усувають покрокове генерування, притаманне RNN. Наприклад, модель FastSpeech запровадила прямий (feed-forward) Transformer з прогнозуванням тривалостей, що дало змогу генерувати мел-спектрограму в 38 разів швидше у порівнянні з авторегресивним аналогом, при цьому зберігаючи якість та усуваючи помилки пропусків/повторів слів [4]. GAN-вокодери також надзвичайно ефективні: генератор формує весь аудіосигнал паралельно, що

забезпечує реальний час або швидше. Модель HiFi-GAN, наприклад, синтезує мовлення більш ніж у 160 разів швидше за реальний час на GPU [3]. На відміну від авторегресивних WaveNet-вокодерів, що генерують один семпл за раз, GAN і паралельні трансформери ефективно масштабуються на сучасному обладнанні, що забезпечує низьку затримку для онлайн-озвучення та інтерактивних сервісів.

4. Стійкість навчання: Transformer-моделі потребують великих наборів даних та надійного вирівнювання тексту з мовленням. Ранні TTS, як Tacotron2, страждали через неправильне налаштування уваги, що призводило до пропусків або повторень слів. Запровадження моделі тривалості, як у FastSpeech із прогнозуванням тривалості фонем і регуляцією довжини, усунуло ці проблеми і покращило синтез мовлення [4]. Трансформери успішно навчаються за допомогою стандартних функцій втрат (L1/MSE для спектрограм) і забезпечують високу розбірливість мовлення. GAN потребують тонкого балансу між генератором і дискримінатором, і без спеціальних заходів можуть виникати артефакти. Сучасні дослідження вирішують це через використання кількох дискримінаторів і додаткових функцій втрат (feature matching, багатомасштабні спектральні втрати), що підвищує стабільність змагального навчання та якість звуку.

5. Роль у системах TTS: В сучасних TTS-системах трансформери та GAN нерідко працюють спільно, виконуючи різні завдання. Як правило, трансформер використовується на етапі акустичного моделювання – перетворює вхідний текст (або фонему) у проміжне представлення (наприклад, мел-спектрограму) [1].

На фінальному етапі застосовується нейронний вокодер, котрий часто побудований на GAN, для перетворення спектрограми в аудіосигнал високої якості. Такий поділ дозволяє поєднати сильні сторони обох підходів: трансформер відповідає за контекстну і просодичну коректність мовлення, а GAN-вокодер – за природний тембр і чіткість звуку. Існують також повністю end-to-end рішення на основі GAN, де весь тракт «текст–аудіо» навчається змагально. Хоча такі системи все ще поступаються модульним підходам у гнучкості налаштування, вони демонструють перспективність об'єднання трансформерної уваги та GAN-генерації в єдиній моделі.

Вирішення проблем. Основні проблеми трансформерних TTS – помилки вирівнювання та обмежена керованість просодією – можуть бути вирішені архітектурними та алгоритмічними методами: експліцитне вирівнювання між текстом і аудіо, зокрема з використанням окремої «вчительської» моделі для отримання вирівнювань і навчання предиктора тривалостей (як у FastSpeech [4]), усуває залежність від непрямої уваги та виключає пропуски і повтори, а додаткові входи чи керуючі параметри (напр., стилізовані токени, інформація про емоції, паузи) покращують інтонацію, компенсуючи усереднену просодію; оптимізовані механізми уваги (прозорлива чи площинна) дозволяють ефективно моделювати контекст при обробці довгих послідовностей, а для генеративно-змагальних моделей виклики стабільності і відповідності контенту долають багаторівневими дискримінаторами (аналіз сигналу як у часовій, так і у частотній області [2]), поєднанням з традиційними функціями помилки та розширенням тренувального корпусу або застосуванням методів попереднього навчання і переносу знань; альтернативним рішенням є використання дифузійних моделей вокодерів, які, хоча і поступаються GAN за швидкістю, забезпечують вищу якість звуку та стабільність навчання, що відкриває перспективи поєднання переваг обох підходів.

Висновки. Трансформери та GAN є ключовими для сучасних систем TTS, наближаючи якість синтезованого мовлення до людського рівня. Transformer забезпечує швидкість, ефективність та контроль за просодією, а GAN-вокодери – природний тембр і високу роздільну здатність звуку. Їх інтеграція у гібридних системах дозволяє об'єднати переваги кожного підходу, проте залишаються виклики щодо стабільності та універсальності, що є напрямком для майбутніх досліджень.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Naihan Li et al. Neural Speech Synthesis with Transformer Network. *Proc. AAAI Conference on Artificial Intelligence*. Vol. 33 (01). Pp. 6706–6713. 2019. URL: <https://arxiv.org/abs/1809.08895>
2. Mikołaj Binkowski et al. High Fidelity Speech Synthesis with Adversarial Networks. *ArXiv preprint arXiv*. Vol. 1909. P. 11646. 2019. URL: <https://arxiv.org/abs/1909.11646>
3. Jungil Kong, Jaehyeon Kim, Jaekyoung Bae. HiFi-GAN : Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Advances in Neural Information Processing Systems*. Vol. 33. 2020. URL: <https://arxiv.org/abs/2010.05646>
4. Yi Ren et al. FastSpeech : Fast, Robust and Controllable Text to Speech. *Advances in Neural Information Processing Systems*. Vol. 32. 2019. URL: <https://arxiv.org/abs/1905.09263>

УДК 004.738

РОЗРОБКА ІНТЕРНЕТ-МАГАЗИНУ З ПЕРСОНАЛІЗОВАНОЮ СИСТЕМОЮ РЕКОМЕНДАЦІЙ НА ОСНОВІ ІСТОРІЇ ПЕРЕГЛЯДІВ І ПОКУПОК

Чорний Г. С.¹, студ., Кривенкова Л. Ю.², ст. викл.

ННІ «Придніпровська державна академія будівництва та архітектури»

Українського державного університету науки і технологій

¹gordeyblack@gmail.com, ²kryvenkova.liudmyla@pdaba.edu.ua

Постановка проблеми. Зі стрімким розвитком електронної комерції та зростанням кількості онлайн-користувачів постає проблема релевантного підбору товарів для конкретного споживача. Стандартна фільтрація за категоріями є малоефективною для утримання уваги користувача. Розв'язанням проблеми є впровадження інтелектуальної системи персоналізованих рекомендацій, що базується на аналізі поведінки користувача – історії переглядів, клієнтські дії, доданих до кошика товарів та завершених покупок.

Мета дослідження. Основною метою є розробка функціонального веб-застосунку інтернет-магазину з вбудованою інтелектуальною системою персоналізованих рекомендацій, яка аналізує індивідуальні дії користувачів, зокрема історію переглядів, покупки, уподобання та час взаємодії з контентом. Запропонована система повинна адаптуватися до змін у поведінці користувача, формуючи актуальні пропозиції товарів, що відповідають інтересам кожного конкретного відвідувача. Особлива увага приділяється розробці алгоритмів гібридного фільтрування, які поєднують переваги колаборативного та контент-орієнтованого підходів для підвищення точності рекомендацій. Також досліджується вплив персоналізації на ключові бізнес-показники: коефіцієнт конверсії, середню тривалість сесії користувача та глибину перегляду [1; 2].

Метою є не лише технічна реалізація інтернет-магазину, а й дослідження ефективності впровадження рекомендаційної системи як інструменту цифрової трансформації торгівлі та підвищення конкурентоспроможності електронної комерції.

Результати дослідження. У процесі розробки інтернет-магазину з інтелектуальною системою персональних рекомендацій було визначено оптимальні технології та етапи створення функціонального і зручного для користувача програмного продукту. Побудова архітектури включає такі ключові компоненти: