Одеський національний університет імені І. І. Мечникова
Факультет математики, фізики та інформаційних технологій
Кафедра оптимального керування та економічної кібернетики

Кваліфікаційна робота

на здобуття ступеня вищої освіти «магістр»

Генеративно-змагальна нейронна мережа для перетворення тексту в мову

Generative adversarial network for text to speach

Виконав: здобувач денної форми навчання
спеціальності 113 Прикладна математика
Освітня програма «Прикладна математика»
Григорян Костянтин Ашотович

Керівник: канд. техн. наук, доц. Мазурок І. Є.
Рецензент: канд. техн. наук, доц. Мороз В. В.

Рекомендовано до захисту:

Протокол засідання кафедри

№ ___ від _____ 2022 р.

Завідувач кафедри

_____ _____

Захищено на засіданні ЕК № _____

Протокол № ___ від _____ 2022 р.

Оцінка _____ / _____ / _____

Голова ЕК

_____ _____

Одеса — 2022 р.

ЗМІСТ

ВСТУП

Актуальність

Останнім часом інтелектуальні інформаційни системи все більше і більше проникають в життя людини та стають невід'ємною частиною її повсякденного життя. Окремою галузю таких систем є голосові помічники, які служать для пошуку необхідної інформації в Інтернеті, коригування маршруту тощо. Незважаючи на розповсюдженність, більшість голосових помічників має суттевий недолік: їхні голоси занадто роботизовані, що робить їх використання не дуже зручним.

Проблема роботизованності машиного голосу актуальна не тільки для голосових помічників. Саме це заважає реалізовувати автоматичну озвучку фільмів та мультфільмів: голоси не схожі на реальних персонажів, не передають необхідну інтонацію тощо.

Не менш актуальним є застосування не роботизованого машиного голосу для вивчення іноземних мов: такий підхід дозволяє імітувати спілкування з різними носіями мови.

Рішення, наявні на ринку, на даний час в більшості своєї або не мають необхідного функціоналу, надаючи, наприклад невелекий вибір заздалегідь готових голосів, що значно звужує можливі сфери використання, або платні та в закритому доступі.

Не менш важливою є проблема шквидкості донавчання існуючих рішень. Воно вимагає підготовку начального датасету дуже великого обсягу, що значно звужує можливі сфери використання.

Актуальність проблематика TTS та використання GAN для вирішення цієї задачі підтверджується великою кількістю досліджень на цю тематику. Наприклад, стаття Glow-TTS[1] вивчає можливість застосування генеративно-змагальної мережі для перетворення текста в мову. Також застосування ГЗМ для цієї задачі можна зустріти у High fidelity speech synthesis with adversarial networks[9]. Застосування досить нових мереж трансформерів досліджені у FastSpeech[2] і FastSpeech 2[3]. Duration informed attention network for multimodal synthesis[4] надає опис стратегії багатоголосної пара-

лельної генерації поверх моделі WaveRNN.

Мета

Розробка системи перекладу тексту в мову певним голосом для озвучки медіафайлів.

Для досягненя цієї мети, потрібно вирішити такі задачі: розробка системи перекладу тексту в мову певним голосом, аналіз існуючих рішень: виявленяя їх сильних сторін та недоліків, порівняння метрик якості роботи, швидкості навчання та процесингу, аналіз попереднього дослідження.

Об'єкт дослідження

Перетворення тексту до мови за допомогою генеративно-змагальних мереж. Застосування технології Transfer Learning (донавчання моделі).

Предмет дослідження

Нейроні мережі. Генеративно-змагальнихі мережи.

ВИСНОВКИ

Ця робота присвячена рішеню проблеми відтворення тексту в мову певним голосом. Це може допомогти сліпим людям в читанні книг або в користуванні комп'ютером, де голосовий помічник може говорити голосом, обраним цією людиною.

В першому розділі розглянути більшість популярних рішень, деякі з яких використовують технолгию GAN. Проведено аналіз та виявлено сильні сторони та недоліки кожного з рішень.

Другий розділ присвячений аналізу попереднього нашого рішення та підкреслення недоліків, які заважали зручному користуванню моделлю.

Третій розділ пояснює нове рішення, описується процес треннінгу, а саме гіперпараметри для генератора та дискримінатора, loss-функції. Проведені експерименти з різними відеокартами і різною кількістю оперативної пам'яті. Також розглянуто деякі підхіди, які не дали бажанного результату. За допомогою експериментів проведено аналіз якості звуку створеного з написаного тексту. Введено спеціальні метрики, такі як схожість голоса та якість вимови. Показано універсальність рішення, яке вирішує не одну, а дві задачі.

В ході дослідження були розглянуті та вирішені задачі для виконання поставленної цілі.

Проведено аналіз таких рішень: WaveNet, WaveGlow, Parallel WaveNet (PWN), Parallel WaveGAN, GAN-TTS, виявлено їх сильні сторони та недолікі, порівняно метрикі якості роботи, швидкості навчання та процесингу.

Проведено аналіз попереднього дослідження, в основі якого були Tacotron 2 та WaveGlow. Виявлення мінусів цього підходу, основними з яких були якість вимови тексту, незручність тренінгу моделі та відсутність робастності.

Розроблено систему перекладу тексту в мову певним голосом за допомогою технології GAN. Описано архітектури моделей та процес тренінгу. Описано спроби об'єднання GAN та Tacatron 2. Була натренована модель, а саме генератор та дискримінатор. Проведено оцінка якості згенерованого

аудіосемплу.

Таким чином ми маємо рішення перетворення тексту в мову певним голосом, яке не вимагає багатьох годин трейн аудіосеплів і дає якісну аудіозапис, та також може використовуватися у реальному часі, а не в офлайн режимі.

СПИСОК ЛІТЕРАТУРИ

1. Kim J., Glow-TTS: A Generative Flow for Text-to-Speech viaMonotonic Alignment Search / J. Kim, S. Kim, J. Kong, S. Yoon – 2020. – Resource access mode: https://arxiv.org/pdf/2005.11129.pdf
2. Ren Y. FastSpeech: Fast, Robust and ControllableText to Speech / Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, Tie-Yan Liu – 2019. – Resource access mode: https://arxiv.org/pdf/1905.09263.pdf.
3. Ren Y. FastSpeech 2: Fast and hight-quality end-to-end text to speech / Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, Tie-Yan Liu – 2021. – Resource access mode: https://arxiv.org/pdf/2006.04558.pdf.
4. Yu C. Durian: duration informed attention network for multimodal synthesis / C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, D. Su, D. Yu – 2019. – Resource access mode: https://arxiv.org/pdf/1909.01700.pdf.
5. Aaron van den Oord: WaveNet: A generative model for raw audio / A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu - 2016 - Resource access mode: https://arxiv.org/pdf/1609.03499.pdf.
6. Prenger R. WaveGlow: a flow-based generative network for speech synthesis / R. Prenger, R. Valle, B. Catanzaro. – 2018. – Resource access mode: https://arxiv.org/pdf/1811.00002.pdf.
7. Aaron van den Oord: Parallel WaveNet: Fast High-Fidelity Speech Synthesis / A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu - 2017 - Resource access mode: https://arxiv.org/pdf/1711.10433.pdf.
8. Ryuichi Yamamoto: Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram / R. Yamamoto, E. Song, J. Kim - 2020 - Resource access mode: https://arxiv.org/pdf/1910.11480.pdf.
9. Donahue J. High fidelity speech synthesis with adversarial networks / J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, K. Simonyan – 2019. – Resource access mode:

https://arxiv.org/pdf/1909.11646.pdf.

10. Ian J. Goodfellow: Generative Adversarial Nets / I. Goodfellow, J. Pouget-Abadie, M. Mirza, B, Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio - 2014 - Resource access mode: https://arxiv.org/pdf/1406.2661v1.pdf.
11. Shen J. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions / [J. Shen, R. Pang, W. Ron]. – 2018. – Resource access mode: https://arxiv.org/pdf/1712.05884.pdf.
12. Abadi M. TensorFlow: A system for large-scale machine learning / M. Abadi, P. Barham, J. Chen – (2016), pp. 265-283 – Resource access mode: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.
13. Paszke A. PyTorch: An Imperative Style, High-Performance Deep Learning Library / A. Paszke, S. Gross, F., A. Lerer, J. Bradbury, Massa – 2018. – Resource access mode: https://arxiv.org/abs/1912.01703.
14. Chandeep S. Transfer Learning and its application in Computer Vision / S. Chandeep, P. Sayonee // Electrical and Computer Engineering University of Waterloo Waterloo – 2022
15. Hryhorian K., Volkov K., Mazurok I. Tacotron 2 and WaveGlow for text-to-speech for PC game characters // Інформатика, інформаційні системи та технології: тези доповідей вісімнадцятої всеукраїнської конференції студентів і молодих науковців. – Одеса. – 2021. – 62-63 с.
16. Григорян К.А., Мазурок І. Є., Волков К.С., Масальський Р.О. Tacotron 2 І WaveGlow для перетворення тексту до речи для персонажів комп'ютерних ігор // Стан, досягнення та перспективи інформаційних систем і технологій / Матеріали XXI Всеукраїнської науково-технічної конференції молодих вчених, аспірантів та студентів. - Одеса, Видавництво ОНАХТ. – 2021. – 207-208 с.
17. Григорян К., Майдан А., Масальський Р., Мазурок І. Моделювальня системи для навчання нейронних мереж // Інформатика, інформаційні системи та технології: тези доповідей дев'ятнадцятої всеукраїнської конференції студентів і молодих науковців. – Одеса. – 2022. – 65-67 с.
18. Hryhorian K., Maidan A., Masalskyi R., Mazurok I. Simulating systems for training neural networks // Стан, досягнення та перспективи інформаційних систем і технологій / Матеріали XXII Всеукраїнської