УДК 004.934

НУТҚНИ АВТОМАТИК ТАНИБ ОЛИШДА QUARTZNET МОДЕЛИ

Маматов Н.С[1], Абдуллаев Ш.Ш[2], Йўлдошев Ю.Ш[3], Ҳомидов А.Ш[4].
[1]"ТИҚХММИ" Миллий тадқиқот университети, кафедра мудири,
m_narzullo@mail.ru
[2]Ориентал университети, доцент, abdulla.sherzod.87@gmail.com
[3]Олий аттестация комиссияси, инспектор, yusuf_yuldoshev@mail.ru

Ушбу мақолада нутқни таниб олишда NVIDIA компанияси томонидан таклиф этилган нейрон тармоғи моделини ўзбек тилидани нутқ маълумотлар тўплами билан ўқитиш орқали ўтказилган тадқиқотлар баён этилган. Ҳозирги кунда нейрон тармоқларнинг турли моделларидан нутқни автоматик таниб олиш тизимларини яратишда кенг фойдаланилмоқда. Бу эса нутқни таниб олиш сифатини кескин ошишига сабаб бўлмоқда. Нутқни таниб олишнинг мавжуд тизимлари турли нейрон тармоқ моделлари асосида ишлашга мўлжалланган. Мазкур мақола ана шундай моделлардан бирини нутқни автоматик таниб олишда қўлланилишига бағишланган бўлиб, унда quartznet модели ва ундан фойдаланиш бўйича маълумотлар келтирилган.

Калит сўзлар: нутқни автоматик таниб олиш, нейрон тармоқ модели, QuartzNet, спектограмма, хатолик функцияси, оптимизатор

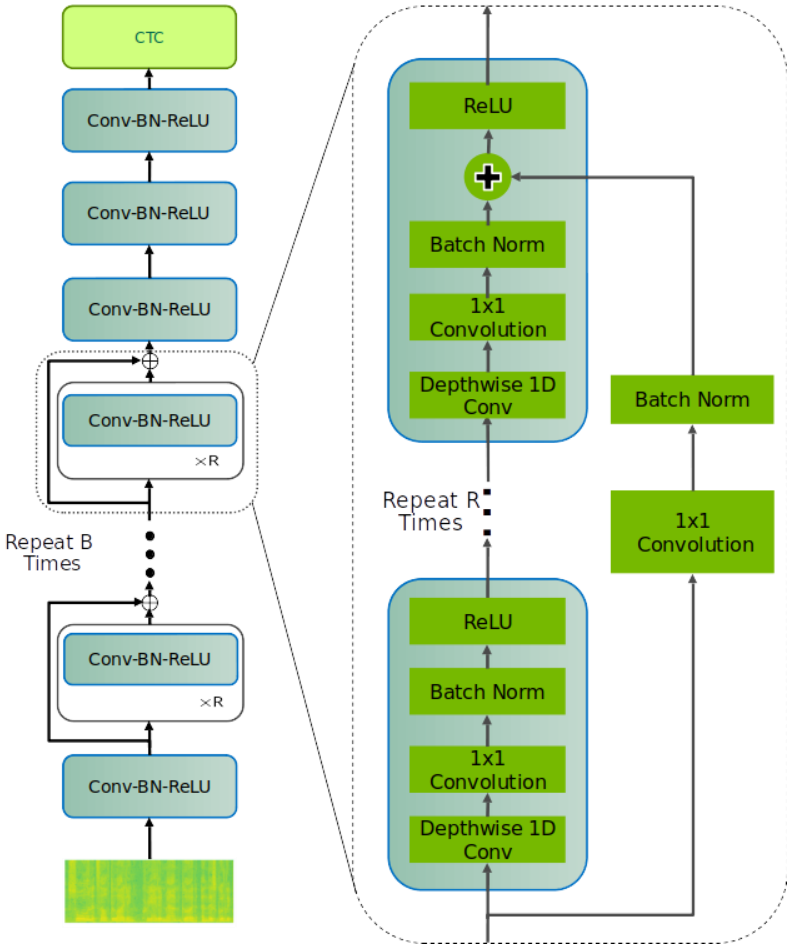
1.Кириш

Сўнгги йилларда кўплаб компаниялар нутқни автоматик таниб олишга мўлжалланган тизимларини таклиф этишмоқда. Жумладан, АҚШнинг NVIDIA компанияси тадқиқотчилари нутқни автоматик таниб олиш тизимида йўналтирилган тўлиқ нейрон тармоқдан ташкил топган акустик моделдан фойдаланишни таклиф этишди [1]. Таклиф этилган модел бир нечта қолдиқли блоклар [2] ва улар орасидаги боғланишлардан ташкил топган. Моделнинг ҳар бир блоки бир ёки бир нечта бир ўлчамли (1D) ўрамли қатламлардан, пакетли нормаллаштириш [3] ва ReLU (Rectified Linear Unit) [4] қатламлардан ташкил топган бўлиб, моделни ўқитиш CTC (Connectionist Temporal Classification) хатолик функцияси орқали амалга оширилади [5]. Ушбу моделда Jasper (Just Another Speech Recognizer) архитектурасидаги кодловчи ва декодловчи модулларидан фойдаланилган [6]. Ушбу нейрон тармоқ модели LibriSpeech ва Wall Street Journal нутқ корпусларида юқори аниқлик кўрсатган ва моделни оптималлаштириш натижасида унинг параметрлари ва гиперпараметрлари инглиз тилидаги нутқни автоматик таниб олиш учун созланган. Тадқиқотчилар модел гиперпараметрлари қийматларига боғлиқ ҳолда QUARTZNETнинг бир нечта вариантларини таклиф этишган. Бунда нейрон тармоқ модели параметрлари сони 6,7 - 18,9 миллион оралиғида ўзгариши мумкин. Бу нутқни автоматик таниб олиш учун йўналтирилган машҳур нейрон тармоқли моделлардан сезиларли даражада кичик бўлиб, у моделнинг нисбатан ихчамлигини таъминлайди. Масалан, wav2letter++ [7] да 208 милион, LAS (Listen, Attend and Spell) [8] 360 миллион, Deep Speech 2 моделида эса параметрлар сони 38 миллионтани ташкил этади [9]. Модел параметрларини

камлигига қарамай, бошқа моделлардан таниб олиш аниқлиги бўйича қолишмайди. Моделнинг ихчамлиги уни ўқитиш учун сарфланадиган вақт ва ҳисоблаш ресурсларини тежашга имкон беради. Бундан ташқари, ушбу моделни бошқа маълумотлар тўпламига ҳам самарали созлаш мумкинлиги тадқиқотчилар томонидан таъкидлаб ўтилган.

2.Асосий қисм

Олиб борилган тажрибаларда QuartzNet 15x5 WER (Word error rate) кўрсаткичи бўйича тил моделидан фойдаланмаган ҳолда халақитсиз нутқ маълумотлари LibriSpeech dev-cleanda ўртача 4.19%, test-other тестлаш тўпламида 10,98% хатоликни кўрсатди. Бу натижаларга эришиш учун моделни ўқитишдаги эпохалар сони 400га тенг, яъни ўқитиш маълумотлари тўплами моделга 400 марта қайта-қайта юкланган. Quartznet модели ўн беш марта такрорланган беш блокли архитектурадан иборат бўлиб, жами 79 та турли қатламлардан ташкил топган ва унинг архитектураси қуйидагича:



1-rasm. Quartznet модели архитектураси

QuartzNet моделини ўқитишда маълумотлар аугментацияси учун [10] ишда SpecCutout ва Speed perturbation алгоритмларидан фойдаланилган. SpecCutout алгоритми спектограммадан кичик ўлчамли тўртбурчак бўлакларни кесиб олиш, Speed perturbation эса нутқ тезлигини маълум

фоизларга ўзгартиришга хизмат қилувчи алгоритм бўлиб, моделни ўқитишда NovoGrad оптимизаторидан фойдаланиш [11] ишда тавсия этилган.

3. Тажрибавий тадқиқотлар

QuartzNet нейрон тармоқ моделини ўқитиш учун “Маълумотларга ишлов бериш тизимлари” лабораториясида яратилган ўзбек тилидаги нутқ корпусидан фойдаланилди. Ундаги нутқлар давомийлиги 432 соатдан иборат аудиокитоблар ва 72 соатли афоризм ҳамда мақолалар аудиоёзувидан иборат.Ушбу нейрон тармоғи моделини ўқитиш ва тажрибавий тадқиқотларни ўтказишда i9 процессорли, 64 ГБ оператив хотирага эга, 2 та GTX 3050 Ti видеокарталар кластерида ишловчи компьютердан фойдаланилган. Бунда нейрон тармоғи моделларини қуриш ва амалга оширишда Pytorch чуқур ўқитиш фреймворки асосида яратилган NVIDIA NeMo (**Ne**ural **Mo**dules) воситалар тўпламидан фойдаланилди. Нутқ сигналига дастлабки ишлов бериш ва белгилар фазоси ўлчамини камайтиришда [12-13] ишларда келтирилган алгоритмлардан фойдаланилди. Тажрибавий тадқиқотлар QuartzNet15x5 моделини ўзбек тилидаги нутқ корпуси билан ўқитиш ва ушбу моделнинг инглиз тилига ўқитилган шаклини қайта ўзбек тилига ўқитиш, яъни Transfer learning усулидан фойдаланган ҳолда ўтказилди.

1-жадвал.

Тажрибавий тадқиқот натижалари

Нейрон тармоқ архитектураси	**Ўқитиш вақти (соат)**	**Ўқитиш аниқлиги (%)**	**Тестлаш аниқлиги (%)**	**Тестлаш хатолиги (WER) (%)**
QuartzNet 15x5	67	83	72	28
QuartzNet 15x5 (transfer learning)	36	88	79	21

Олиб борилган тадқиқотлар QuartzNet 15x5 моделини ўқитишда Transfer learning усулидан фойдаланиш ўқитиш вақти ва аниқлик бўйича самаралироқ эканлигини кўрсатди. Ушбу усулда ўқитиш ушбу моделнинг фақат сўнгги қатламлари учун амалга оширилганлиги нисбатан кам вақт сарфини асослайди.

Хулоса

QuartzNet модели бошқа машҳур моделларига ихчам бўлганлиги учун у ҳисоблаш ресурсларни кам талаб қилади. Ўтказилган тажрибавий тадқиқотлар мазкур моделни нутқни автоматик таниб олишда ўзининг самарадорлигини кўрсатди ва уни ўзбек тилидаги нутқ сигналларини таниб олишнинг автоматик тизимини яратишда асос сифатида олиш мумкин. Бундан ташқари, ушбу модел кам ҳисоблаш ресурсларини талаб қилиганлиги учун ундан мобил ва портатив қурилмаларга мўлжалланган нутқни таниб олиш дастурий воситаларини яратишда фойдаланиш мақсадга мувофиқдир.

Фойдаланилган адабиётлар

[1] S. Kriman, S. Beliaev, B. Ginsburg J. Huang and etc. QuartzNet: deep automatic speech recognition with 1d time-channel separable convolutions. https://arxiv.org/pdf/1910.10261.pdf

[2] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian (2015-12-10). "Deep Residual Learning for Image Recognition". arXiv:1512.03385.

[3] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv:1502.03167, 2015.

[4] Bing Xu, Naiyan Wang, Tianqi Chen, Mu Li “Empirical Evaluation of Rectified Activations in Convolutional Network”, arxiv.org/abs/1505.00853

[5] Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, ´ “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in ICML, 2006.

[6] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J.M. Cohen, H. Nguyen, and R.T. Gadde, “Jasper: An end-to-end convolutional neural acoustic model,” arXiv:1904.03288, 2019.

[7] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, “End-to-end speech recognition from the raw waveform,” arXiv:1806.07098, 2018.

[8] Wiliam Chan et al. Listen, Attend and Spell arXiv: 1508.01211

[9] Dario Amodei et al. Deep Speech2: End-to-End Speech Recognition in English and Mandarin arXiv: 1512.02595

[10] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” Interspeech, 2015

[11] B. Ginsburg, P. Castonguay, O. Hrinchuk, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, H. Nguyen, and Cohen J. M., “Stochastic gradient methods with layerwise adaptive moments for training of deep networks,” arXiv:1905.11286, 2019.

[12] Mamatov, N., Niyozmatova, N., & Samijonov, A. (2021). Software for preprocessing voice signals. International Journal of Applied Science and Engineering, 18(1). https://doi.org/10.6703/IJASE.202103_18(1).006

[13] Fazilov, S., Mamatov, N., Samijonov, A., & Abdullaev, S. (2020). Reducing the dimensionality of feature space in pattern recognition tasks. Journal of Physics: Conference Series, 1441(1), 012139. https://doi.org/10.1088/1742-6596/1441/1/012139