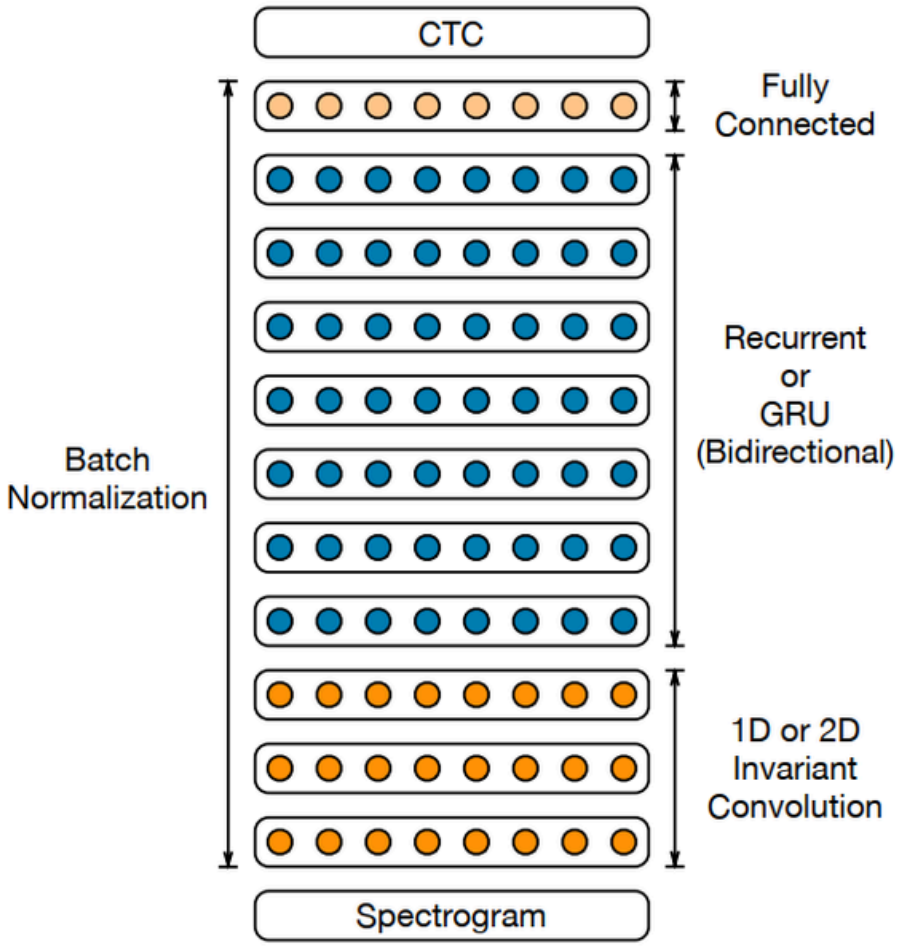
Выполнил: Стамбеков Алмас, БПМИ213

Проделанная работа, кратко

В данной домашней работе была реализована архитектура deepspeech2, которая изначально состояла из слоев конволюций и rnn. В финальной версии модели осталась только двунаправленная рекуррентная сеть, обученная посредством BPE токенайзера (ByteBPE) и обыкновенного английского алфавита. Был реализован beam_search, но в целях экономии ресурсов и времени удалось прологгировать его в wandb только с использованием beam_size=3 на "алфавитной" моделе, который дал хоть и небольшой, но все же прирост (оно и очевидно, так как beam_size всего 3). Был реализован инференс, в котором доступно использование beam_search произвольной длины (стоит учитывать, что для bpe размер словаря = 1024, поэтому это может занять огромное количество времени).
Итоговый результат на test-clen: CER: 12.5%, WER: 35,4%

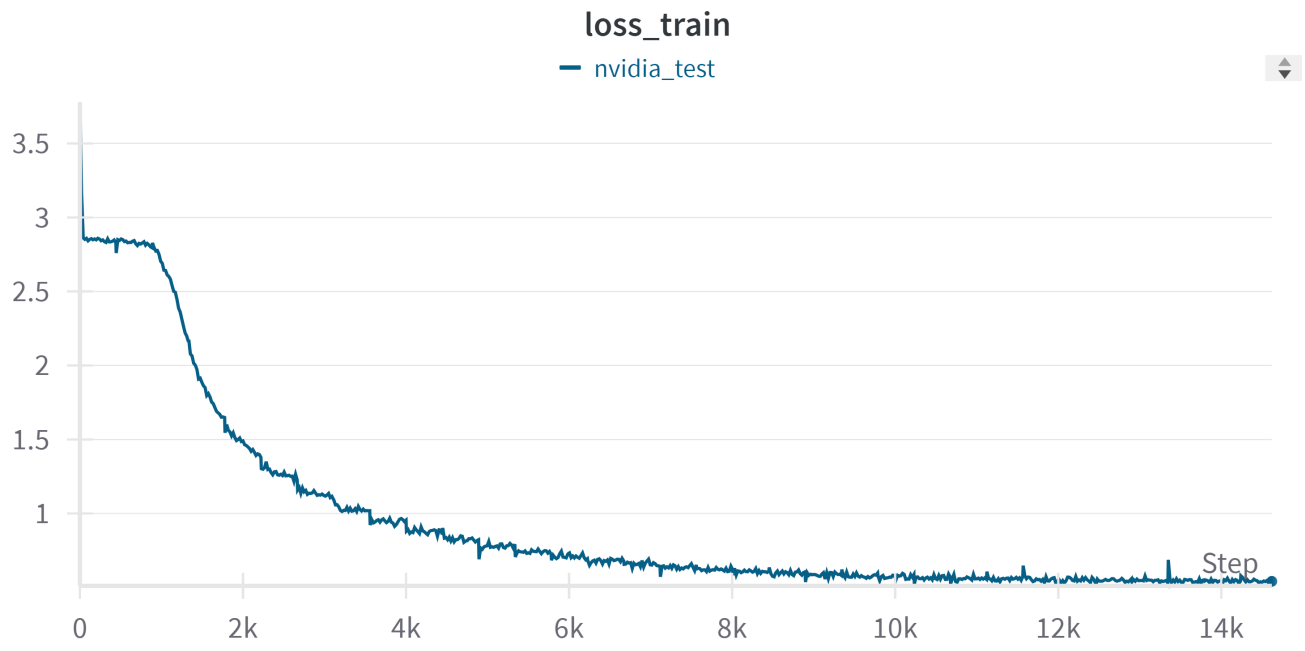
Архитектура

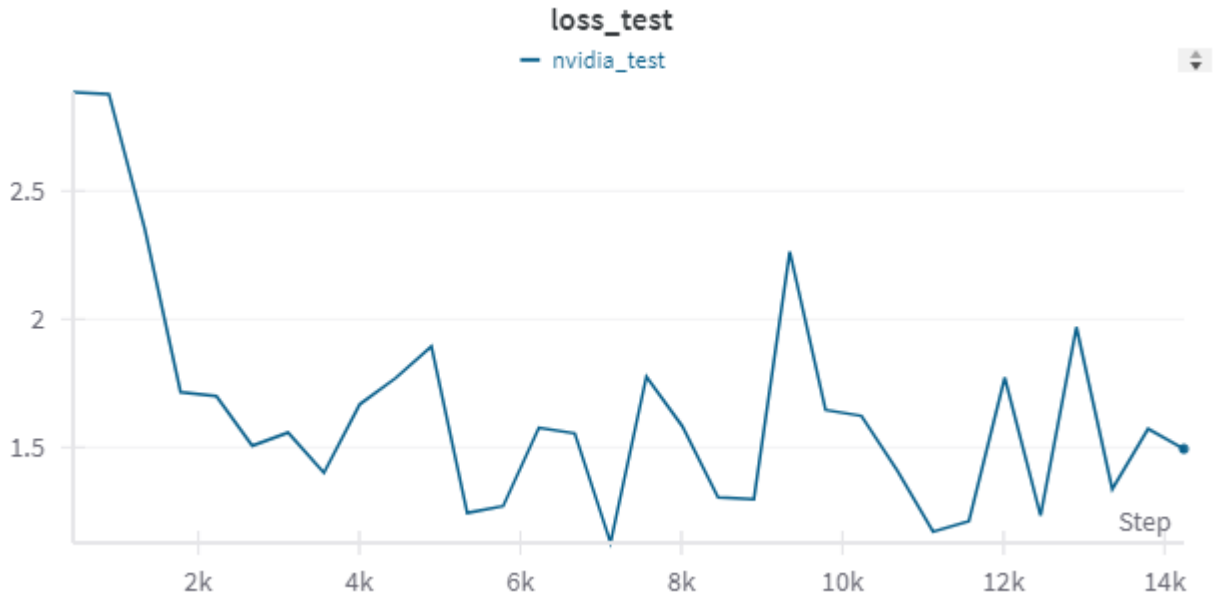


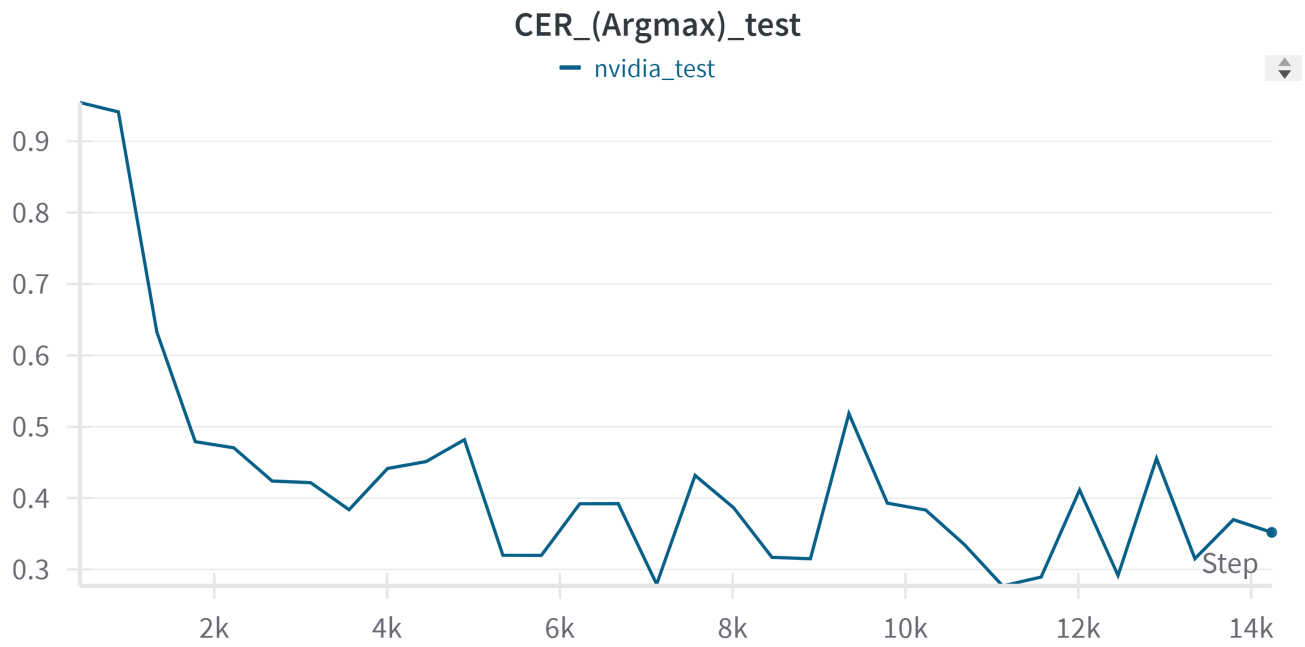
В данной работе для архитектуры модели применялась схема выше. В качестве конволюционных слоев выступали две 2D конволюции (которые не применялись в "лучшей модели", хоть функционал и присутствует), в качестве рекуррентной "сердцевины" - 5 bidirectional gru, а далее линейный слой с выходной размерностью равной размеру словаря. Изначально, большое количество параметров (такие как оптимайзеры, размерность ядра, количество слоев в rnn и тд) были "подсмотрены" у описания архитектуры от nvidia, но в дальнейшем большинство параметров были подобраны эмпирически.
В конеченой модели осталось все, кроме конволюционных слоев, которые вносили нестабильность и несходимость по сравнению с моделью без их использования.

Обучение

Сначала обучалась модель "бейзлайн" с параметрами, рекомендованными nvidia (в wandb называется nvidia_test):


loss_train
nvidia_test
3.5
3
2.5
2
1.5
1
Step
0
2k
4k
6k
8k
10k
12k
14k


loss_test
nvidia_test
2.5
2
1.5
Step
2k
4k
6k
8k
10k
12k
14k



Видно, насколько нестабильным получилось обучение исходной архитектуры.

Далее применялись методы аугментаций, BPE, изменялись параметры модели (модель расширялась, ужималась и тд), но такая нестабильная тенденция все равно сохранилась. После несольких ("облегчающих" модель) экспериментов было замечено, что такая нестабильность убирается путем избавления от конволюционных слоев, что и стало финальной архитектурой модели. Так получилось стабилизировать процесс и минимизировать метрики качества wer-cer:

(остальные графики и параметры в ране bpe_rnn_only_models_big - финальная версия).

Такая модель обучалась 22 эпохи с конфигурациями из deepspeech2.yaml

В итоге модель стала выдавать какие-то осмысленные предложения (соответствующие ее метрикам), но больше из данной архитектуры выжать не получилась (оно и ясно, ведь использовался deepspeech2 без конволюций, а по статьям одних rnn'ок недостаточно). Модель обучалась с небольшой вероятность аугментации шума (ColorNoise) и ByteLevelBPETokenizer'ом размера 1024.

Что не получилось.

1. Не получилось номрально завести стандартные вариации deepspeech'a, которые должны работать в разы лучше.
2. Не удалось реализовать полноценное внедрение LM. Отмечу. что идеи были, даже способ использования, но под конец стало понятно, что выбранный пайплайн не сработал с ByteLevelBPETokenizer (анлак, что поздно это понял).
3. Не удалось все по-честному посчитать для больших размеров beam_size (потому что бесконечный компьют для большого словаря). Напомню, что для алфавита маленьки прирост имеется.

PS отмечу, то очень душным было все это умещать в kaggle, поэтому были небольшие трудности со временем вычисления.

Веса:
https://drive.google.com/drive/folders/1WuoNFy2Q-xq3Bjvr8vUMVUipzWrU9TgW?usp=sharing

wandb логи (все веса имеются там же):
https://wandb.ai/dungeon_as_fate/pytorch_template_asr_example