Тема Работы

Исследование и разработка алгоритма распознавания устной речи на базе метода переноса обучения в глубокой нейронной сети типа wav2vec

Состав коллектива

Студент группы 18133 ММФ НГУ, Болдинов Артём Константинович. Старший преподаватель кафедры ФиПЛ ГИ НГУ Бондаренко Иван

Аннотация работы

Цель нашей работы состоит в исследовании и разработке многозадачного метода обучения глубокой сети *wav2vec2* при дообучении на русскоязычном наборе данных *Golos* с учётом иерархичности речи после предварительного обучения на многоязычном наборе данных.

Объектом исследования в работе являются методы распознавания речи на основе глубоких нейронных сетей.

Предметом исследования является иерархическое многозадачное обучение нейронной сети типа *wav2vec* распознаванию русской речи.

Методы исследования. Для решения поставленных задач использованы методы анализа данных и машинного обучения, методы математической статистики, и методы программной инженерии.

Научная новизна полученных результатов состоит в следующем. Предложена архитектура многозадачной модели, позволяющая повысить качество распознавания символов и слов по сравнению с базовой моделью.

Обоснованность и достоверность научных положений и выводов подтверждается корректным использованием математических методов, применением способов разработки корректных программ, результатами экспериментальной проверки.

Постановка задачи

Исследование и разработка многозадачного алгоритма на языке Python с использованием библиотеки Pytorch на базе алгоритма wav2vec2 при дообучении на русскоязычном наборе данных Golos с учётом иерархичности речи после предварительного обучения на многоязычном наборе данных.

Современное состояние проблемы

Современное состояние проблемы: С развитием технологий появилась большая потребность в переводе речевого сигнала в текст. Голосовые помощники, расшифровка голосовых сообщений, транскрибирование интервью, автоматический перевод видео на другой язык и т. д. Сама проблема распознавания речи возникла ещё в 1952 году. Но даже в наши дни эта проблема ещё не решена. Классический подход распознавания речи основан на соединении нескольких компонентов. В начале звуковой сигнал подается на вход акустико-фонетическому блоку для распознавания фонем, после этого лингвистический блок переводит цепочку фонем в цепочку слов, используя языковые модели. Однако данный подход имеет несколько недостатков. В их числе проблема поиска наиболее вероятной последовательности слов, которых в русском языке более ста пятидесяти тысяч, и каждое слово может иметь несколько десятков словоформ. Поэтому поиск по такому набору элементов весьма трудозатратный. На смену данному подходу постепенно приходит сквозной подход, т.е. использующий одну большую нейронную сеть, которая преобразовывает звуки напрямую в буквы. Глубокие нейронные сети во время процесса обучения также обладают некоторой иерархией: нижние слои отвечают за более низкоуровневые признаки, а верхние за более высокоуровневые. Так что можно сказать, что такие модели, решая задачу распознавания речи, неявно моделируют акустическую и языковую модели.

Представленная авторами статьи [1] нейронная сеть $wav2vec2$ показала себя на практике как одно из лучших решений для задачи распознавания речи на английском языке, но подход к многозадачному обучению с учётом иерархичности речи не исследовался.

Описание работы

Основной задачей данной работы является проверка гипотезы о повышении качества распознавания русской речи, при подходе с иерархическим обучением, т.е. более нижние слои нейронной сети обучаются на более низкоуровневых признаках, например, распознавание фонем. На средних слоях обучается задача распознавания частей речи. И с последнего слоя идёт распознавание самой речи. В качестве предобученной модели была выбрана модель *wav2vec2-large-xlsr-53* компании *facebook*(ныне *Meta*), загруженная с *Hugging Face*, состоящая из 24 блоков трансформера и содержащая около 300 млн обучаемых параметров. Данная модель была предобучена на 53 языках.

На последний слой модели мы добавили линейный классификатор с выходом на количество элементов в словаре символов. Тренировка на разме-

ченных данных проходит с помощью специальной нейросетевой функции потерь для структурной классификации временной последовательности [2]. Далее для краткости будем обозначать её как и в источнике английской аббревиатурой *CTC*.

Коэффициент скорости обучения (lr) изменяется пошагово на 20, 100, 150, 220 эпохах с шагом 0.3162. Данный шаг был выбран, исходя из того, что $0.3162^2 \approx 0.1$. В качестве оптимизатора был выбран $AdamW$ [5] с начальным $lr = 1e{-}3$.

Одной из задач многозадачного обучения стало распознавание фонем. В качестве алгоритма преобразования графем (букв) в фонемы был выбран проект *russian_g2p* [6]. Т.к. фонемы являются более низкоуровневыми признаками, линейный слой, отвечающий за классификацию фонем, мы добавили на второй скрытый слой. В качестве целевой функции для задачи распознавания фонем нами также была выбрана *CTC*. В качестве третьей задачи была выбрана задача распознавания частей речи. Разметка осуществлялась с помощью библиотеки *spacy* [3] моделью $ru\_core\_news\_lg$, точность разметки частей речи которой составляет 99% на наборе данных серебряного стандарта *Nerus*. Соответствующий классификатор получал выход с десятого слоя трансформера. Схема итоговой модели представлена на рисунке 1.

Набор данных

В качестве набора данных был взят *Golos* [4], представленный компанией *Sber* в 2021 году и на данный момент являющийся самым большим набором данных размеченной вручную речи на русском языке. Он содержит около 1240 часов размеченной речи с частотой дискретизации 16кГц. В тренировочной и тестовой частях может встречаться речь одного и того же пользователя. В нашей работе для уменьшения количества вычислений была выбрана опция с обучением на 10 часах и тестированием на всей выборке. В дополнение к данным из набора нами была проведена дополнительная разметка частей речи и фонем. В таблице 1 приведены примеры распознавания речи и меры соответствующих ошибок.

Во время тренировочного процесса значение функции потерь у многозадачной модели уменьшалось быстрее, это видно на рисунке 2(a). Действительно, ведь дополнительные линейные блоки, присоединенные к нижнему и среднему слоям, способствуют предотвращению проблемы затухающего градиента.

График зависимости значения функции потерь от эпохи в процессе тестирования показан на рисунке 2(b).

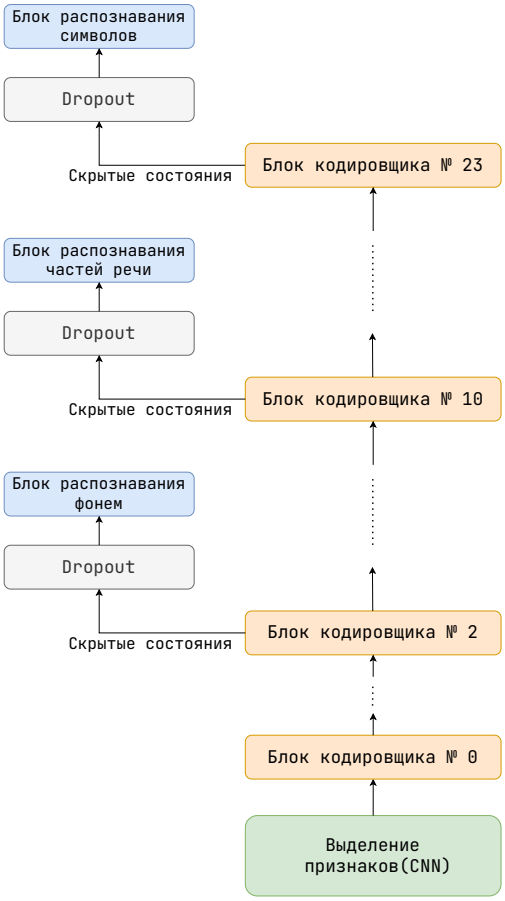
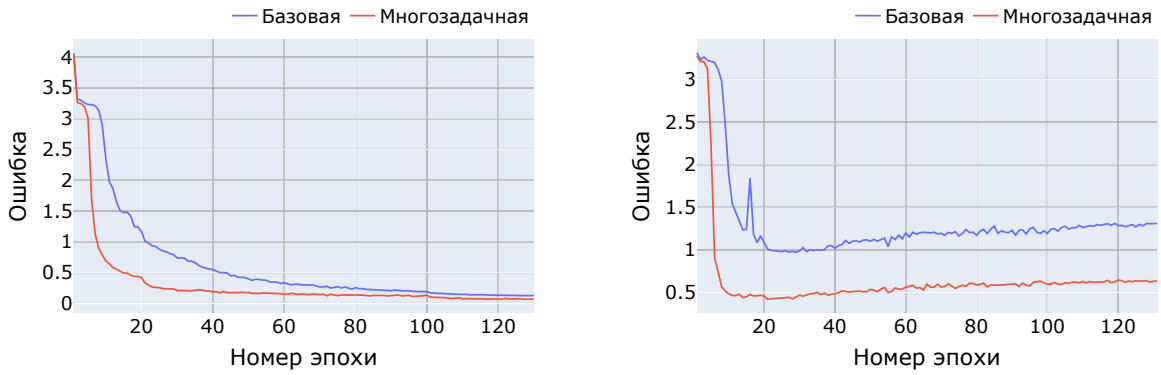


Рис. 1: Диаграмма многозадачной модели

По нему видно, что приблизительно после 20 эпохи значение функции потерь начинает расти, однако мера ошибки распознавания символов и слов продолжает падать, как это видно на рисунках 3(a) и 3(b). Мы предполагаем, такое несоответствие значений говорит об отсутствии сильной корреляции между данными величинами. Например, зависимость может иметь сложный нелинейных характер, который корреляция не выявляет. Ещё одной причиной может быть то, что значения функции потерь вычисляются между "сырыми" предсказаниями (вещественные значения) и известными метками (целочисленные значения), а значения мер ошибок считается между округлёнными данными (целочисленные значения) и известными

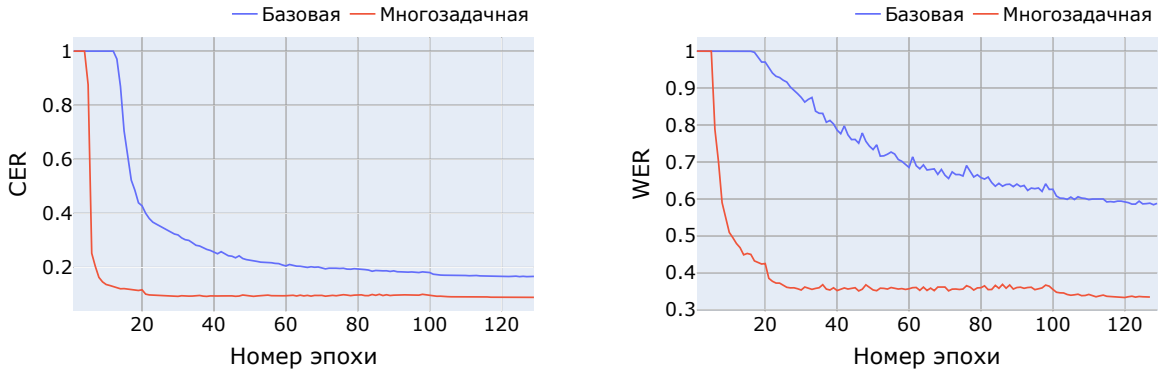
Оригинал и предсказание	CER	WER
хочу посмотреть фильм касл сезон четыре серия тринадцать хочу посмотреть фильм касл сезон четыре серия тринадцать	0.0	0.0
три триста восемьдесят пять семьсот четыре шестьдесят один девять пять три триста **восемьдеся** пять семьсот четыре шестьдесят один девять пять	0.014	0.1
у тебя найдется одиннадцатая серия мастера меча онлайн у тебя найдется одиннадцатая серия **мастеровича отлайн**	0.09	0.375
список кинофильмов александра котта список кинофильмов александра **кота**	0.03	0.25

Таблица 1: Примеры распознавания речи и меры соответствующих ошибок



(a) (b)

Рис. 2: Значения функции ошибок

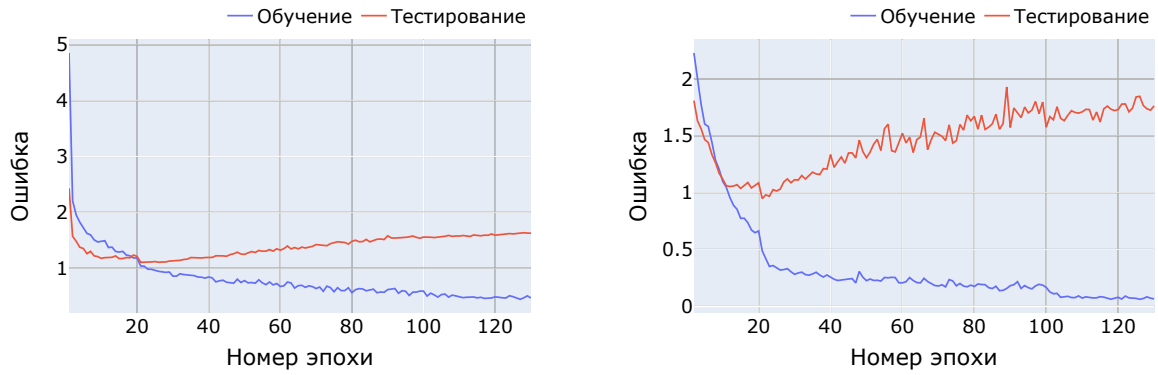


(a) (b)

Рис. 3: Значения мер ошибок

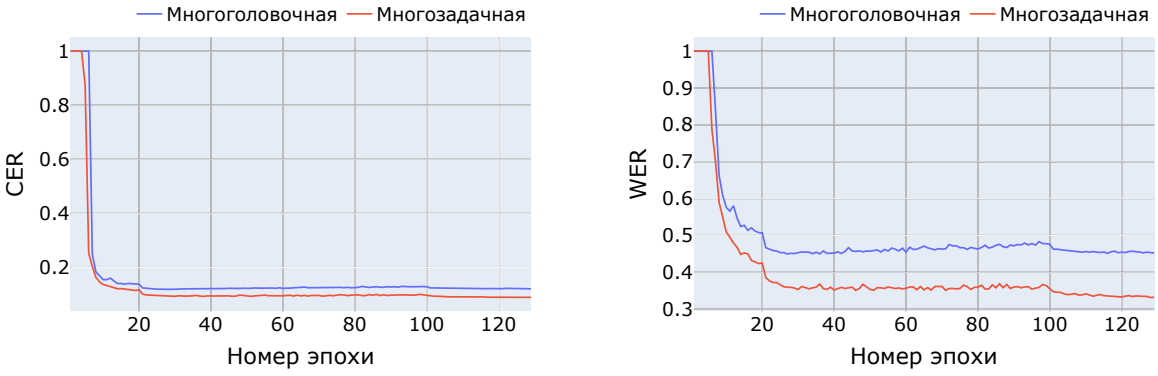
метками (целочисленные значения). Поэтому меры ошибок в отличие от функции потерь более "устойчивы". Исходя из этих причин, мы решили продолжать обучение до тех пор, пока меры ошибок продолжают падать. Также, данная проблема возникла при обучении многозадачной модели для задачи распознавания фонем и частей речи. Графики приведены на рисунках 4(a) и 4(b).

Чтобы убедиться, что на результат работы многозадачной модели повлиял иерархический подход к обучению, нами был проведен ещё один эксперимент, в котором дополнительные линейные классификаторы с нижнего и среднего слоев вместо задач распознавания фонем и частей речи также непосредственно занимались задачей распознавания текста. Графики зависимости приведены на рисунках 5(a) и 5(b)



(a) (b)

Рис. 4: Переобучение на задачах распознавания фонем и частей речи



(a) (b)

Рис. 5: Переобучение на задачах распознавания фонем и частей речи

Результаты

Итоговые результаты, представлены в таблице 2. Также, после проведения серии экспериментов с различными начальными приближениями нулевая гипотеза, говорящая о том, что многозадачная модель работает не лучше чем базовая, отвергается в пользу альтернативной с реально допустимым уровнем значимости равным 0.03125, что меньше порога 0.05. Важным моментом является то, что мы не сравниваем нашу многозадачную модель с другими моделями распознавания речи, поскольку для ускорения экспериментов обучение производилось только на 10 часах вместо полной обучающей выборки. Данное исследование было направлено на проверку эффективности иерархического многозадачного обучения в сравнении с однозадачным подходом дообучения глубокой нейронной сети, при равных остальных условиях, включая размер обучающей выборки.

Модель	CER	WER
Базовая	15.6	56.2
Многозадачная	**8.6**	**32.1**
Многоголовочная	12.2	45.8

Таблица 2: Меры ошибок

Список литературы

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020.

[2] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks. volume 2006, pages 369–376, 01 2006.

[3] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[4] Nikolay Karpov, Alexander Denisenko, and Fedor Minkin. Golos: Russian dataset for speech research, 2021.

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[6] Olga Yakovenko, Ivan Bondarenko, Mariya Borovikova, and Daniil Vodolazsky. Algorithms for automatic accentuation and transcription of russian texts in speech recognition systems. In *International Conference on Speech and Computer*, pages 768–777. Springer, 2018.