**Грачев Артем Михайлович Методы сжатия рекуррентных нейронных сетей для задач обработки естественного языка**

ОГЛАВЛЕНИЕ ДИССЕРТАЦИИ

кандидат наук Грачев Артем Михайлович

Введение

Глава 1. Рекуррентные нейронные сети для задачи

моделирования языка

1.1 Задача моделирования языка

1.2 Рекуррентные нейронные сети для задачи моделирования языка

1.3 Оценка качества языковых моделей

1.3.1 Перплексия

1.3.2 Точность

1.3.3 Дивергенция Кульбака-Лейблера

1.4 Методы оптимизации

1.5 Алгоритм обратного распространения ошибки через время

1.6 Разновидности рекуррентных нейронных сетей

1.7 Практические приёмы для обучения нейронных сетей

1.8 Описание датасетов

Глава 2. Введение в методы сжатия. Особенности сжатия

рекуррентных нейронных сетей для моделирования

языка

2.1 Анализ числа параметров в нейронной сети

2.1.1 Проблема входного и выходного слоя

2.1.2 Использование одной матрицы для входного и выходного слоя

2.2 Основные подходы к сжатию нейронных сетей

2.3 Прунинг и квантизация как базовые методы сжатия

Глава 3. Матричные и тензорные разложения для сжатия

нейронных сетей

3.1 Введение

3.2 Обзор методов матричных разложений

3.2.1 Адаптивное преобразование Fastfood

3.2.2 Унитарная ЯШ

3.3 Методы матричного разложения для сжатия рекуррентных нейронных сетей

3.4 Алгоритм обратного распространения ошибки в случае матриц низкого ранга

3.5 Сжатие нейронной сети с помощью ТТ разложения

3.5.1 Описание ТТ-разложения

3.5.2 Применение ТТ для сжатия нейронной сети

3.6 Эксперименты по сжатию рекуррентных нейронных сетей с использованием низкорангового и ТТ-разложения

3.6.1 Описание экспериментов и детали реализации

3.6.2 Результаты прунинга и квантизации

3.6.3 Результаты экспериментов

3.6.4 Результаты разложения софтмакс слоя

3.7 Общая схема сжатия рекуррентных нейронных сетей

3.8 Заключение

Глава 4. Байесовские методы для сжатия нейронных сетей

4.1 Введение

4.2 Байесовский подход. Вариационная нижняя граница (ЕЬБО)

4.3 Вариационный дропаут для сжатия нейронных сетей

4.4 Автоматическое определение значимости

4.5 Детали экспериментов и результаты

4.5.1 Обучение моделей и оценка качества

4.5.2 Результаты экспериментов

4.6 Выводы

Заключение

Список литературы

Благодарности

Список рисунков

Список таблиц