Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

НА ТЕМУ:

«Метод построения поисковых индексов в реляционной базе данных на основе глубоких нейронных сетей»

Студент: ИУ7-83Б _____ М. Д. Маслова
(группа) (подпись, дата) (И. О. Фамилия)

Руководитель: _____ А. А. Оленев
(подпись, дата) (И. О. Фамилия)

Нормоконтролер: _____ Д. Ю. Мальцева
(подпись, дата) (И. О. Фамилия)

2023 г.

1 Исследовательская часть

1.1 Предмет исследования

Характеристиками, определяющими эффективность индекса являются:

— время выполнения основных операций:

  — построения;
  — поиска;
  — вставки (являющейся комбинацией первых двух операций);

— память, занимаемая индексом.

Так как метод основан на использовании глубокой нейронной сети, характеристикой также является средняя абсолютная ошибка предсказания позиции ключей, получаемая в ходе обучения.

Предполагается зависимость описанных характеристик от объема индексируемых данных, а также от распределения ключей, поэтому исследование проводится на различном количестве ключей для распределений, описанных в подразделе **??**: равномерного, нормального и распределения реальных данных OpenStreetMap. Также проводится сравнение времени поиска при моделях с различным числом скрытых слоев, описанных в подразделе **??**.

1.2 Исследование времени построения индекса

На рисунке 1.1 приведен график зависимости времени построения индекса от количества ключей в индексируемом наборе данных при различных распределениях с использованием модели с двумя скрытыми слоями.

На рисунке 1.2 приведен график зависимости времени построения индекса от количества ключей в индексируемом наборе реальных данных при различных количествах скрытых слоев в модели.

Из приведенных графиков можно сделать вывод, что распределение данных не оказывает влияние на время построения индекса в силу необходимости прохода по всему набору данных при обучении. При этом добавление дополнительного слоя в модель увеличивает время обучения, а следовательно и построения. По сравнению с индексом на основе модели с двумя скрытыми слоями индекс на основе модели с тремя увеличивает время построения на 10%.

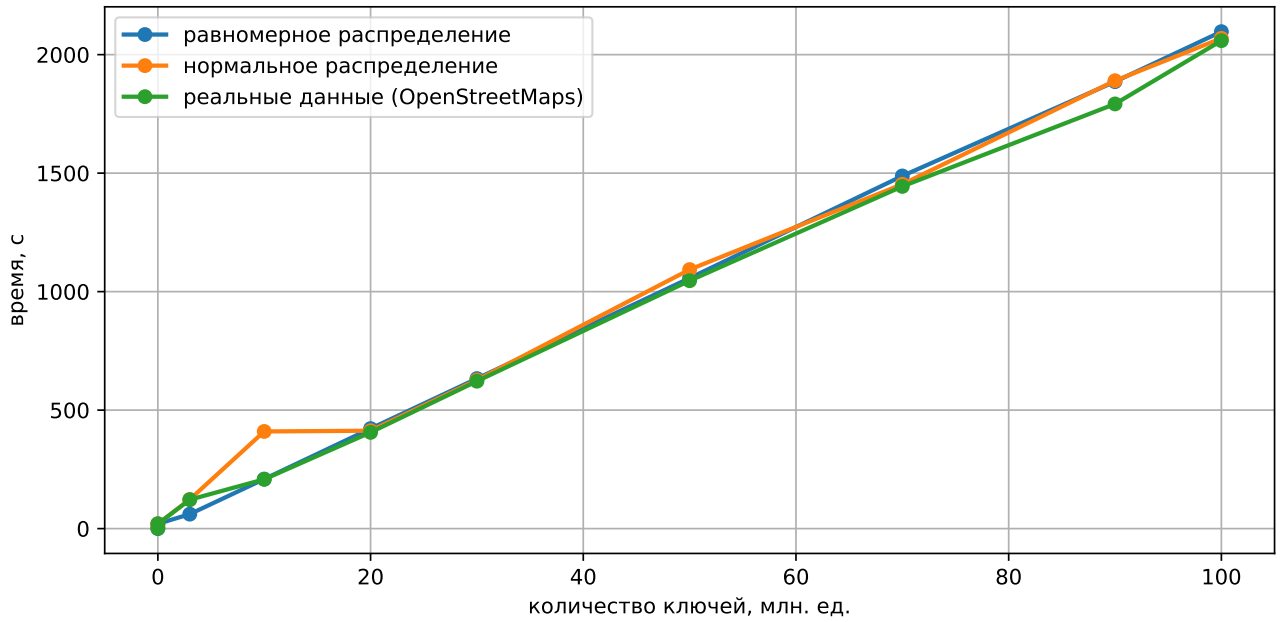

Рисунок 1.1 – График зависимости времени построения индекса от количества ключей (распределения, 2 скрытых слоя)

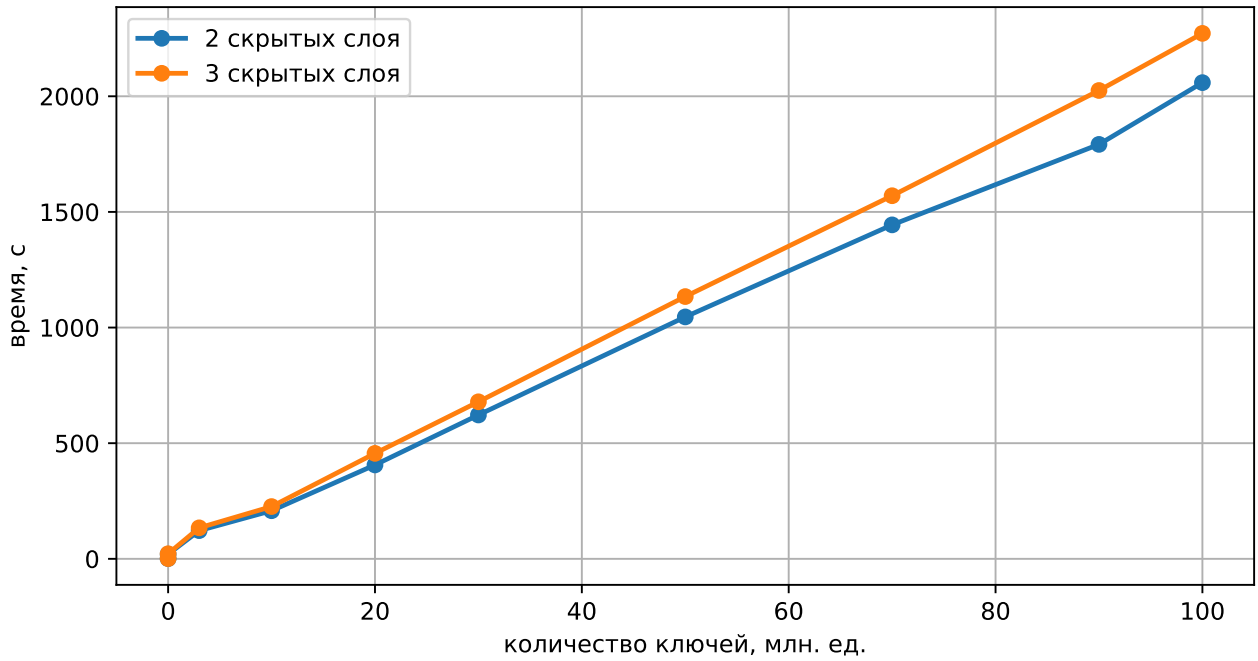

Рисунок 1.2 – График зависимости времени построения индекса от количества ключей (модели, реальные данные)

Основным выводом из приведенных графиков является наблюдаемая линейная зависимость времени построения индекса от количества ключей в наборе, что объясняется достижением необходимой точности модели за 1-2 эпохи обучения при каждом размере данных, за которые происходит проход по всем значениям ключей.

1.3 Исследование времени поиска

На время поиска с использованием индекса, построенного с помощью разработанного метода, должна оказывать влияние абсолютная ошибка предсказания позиции ключа моделью глубокой нейронной сети, так как она определяет диапазон, в котором осуществляется уточнение с помощью бинарного поиска.

График зависимости среденей абсолютной ошибки от количества ключей представлен на рисунке **??**. Нормированное распределение абсолютной ошибки представлено на рисунке **??**.

Подвывод об ошибке... Принимает некоторое постоянное значение в процентах к числу ключей. => диапазон бинарного поиска линейно растет.

На рисунке **??** представлен график зависимости времени поиска от числа ключей.

ДОБАВИТЬ ГРАФИК ВРЕМЕНИ БИНАРНОГО ПОИСКА???

Вывод по времени поиска...

1.4 Исследование времени вставки

ПОИСК + ПОСТРОЕНИЕ

На рисунке **??** представлен график зависимости времени вставки от числа ключей.

1.5 Исследование памяти, используемой индексом

На рисунке **??** представлен график зависимости размера индекса от количества ключей.

`модель + размер массива`