УДК 004.8

ИССЛЕДОВАНИЕ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ РЕШЕНИЯ ЗАДАЧИ КОНТЕКСТУАЛЬНОГО ПОИСКА ПО НОВОСТНЫМ СТАТЬЯМ (RECORD)

Олюнин В.В. (Университет ИТМО), **Быков И.А.** (Университет ИТМО), **Милантьев С.А.** (Университет ИТМО)

Научный руководитель – к. т. н., доцент Платонов А.В.
(Университет ИТМО)

Описывается исследование нейросетевых алгоритмов решения задачи контекстуального поиска по новостным статьям (*ReCoRD*). Представлены алгоритмы нейронных сетей, показывающие наилучшие результаты, и описаны особенности их работы. Выдвинута гипотеза по оптимизации алгоритмов.

Введение. Машинное понимание текста (*MRC*) является центральной задачей в понимании естественного языка, и в последнее время использует методы, основанные на большом количестве крупномасштабных наборов данных, обычно формализованная как задача ответа на вопросы по отрывку. Исследования показали, что на большую часть вопросов в этих наборах данных можно ответить, просто сопоставив шаблоны между вопросом и предложением в отрывке. Один из основных типов вопросов, которых не хватает в этих наборах данных, — это те, которые требуют рассуждений на основе здравого смысла или понимания нескольких предложений в отрывке

Для преодоления данного ограничения создан крупномасштабный набор данных по контекстуальному поиску типа вопрос-ответ (*Reading Comprehension with Commonsense Reasoning Dataset, ReCoRD*), состоящий из более чем 120 000 примеров, требующих глубокого контекстного анализа. В отличие от большинства существующих наборов данных *MRC*, все запросы и переходы в автоматически извлекаются из новостных статей, что максимально снижает предвзятость со стороны человека. Модель способна имитировать человеческий «здравый смысл», если она позволяет вывести вероятные следствия всего, что ей рассказывают и что уже проанализировал алгоритм. После контекстного анализа модель может выбрать наиболее вероятное следствие.

Структура для набора данных *ReCoRD* состоит из отрывков с отмеченными отрезками текста, закрытых запросов и справочных ответов. Сбор данных происходит в четыре этапа:
1) курирование новостных статей *CNN* / *Daily Mail*;
2) создание троек (контекст, запрос, ответ) на основе новостных статей;
3) фильтрация запросов, на которые можно легко ответить современные модели MRC;
4) фильтрация запросов, неоднозначных для читателей, удаление шумов.

Основная часть. Для оценки результатов работы алгоритмов с набором данных *ReCoRD* используются две основные метрики оценки: *EM* и *F1*. Оба игнорируют знаки препинания и статьи. Точное соответствие (*Exact match, EM*) измеряет процент прогнозов, которые точно соответствуют любому из справочных ответов: *F1*-мера измеряет среднее перекрытие между предсказанием и эталонными ответами, объединяя метрики *precision* и *recall* в агрегированный критерий качества.

По данным платформы *SuperGLUE*, а также авторов исходной задачи, лидерующие позиции по решению данной задачи занимают модификации алгоритма *BERT: LUKE, RoBERTa, SKG-BERT*, а также алгоритм *XLNet*. *Bidirectional Encoder Representations from Transformers (BERT)* от *Google* — это усовершенствованная сеть *GPT* от *OpenAI* (двунаправленная вместо однонаправленной и т. д.), основана на архитектуре Transformer. BERT избавляется от рекуррентности, трансформеры для каждого слова строят признаки, используя для этого механизм внимания.

Для того выдвинуть возможный путь оптимизации нейросетевого алгоритма в решении задачи *ReCoRD* были проанализированы недостатки представленных алгоритмов. Особое

внимание уделялось алгоритмам на основе модели *BERT*, так как именно их модификации занимают лидирующие позиции.

Предложена следующая гипотеза по оптимизации работы нейросетевого алгоритма в решении задачи *ReCoRD*. Основная идея заключается в том, что при расчетах в программном обеспечении большая часть расчетов избыточна. Чтобы оптимизировать это, можно использовать отрицательную выборку. Суть этого подхода заключается в том, что мы максимизируем вероятность встречи для желаемого слова в типичном контексте (тот, который часто встречается в нашем корпусе) и в то же время сводим к минимуму вероятность встречи в нетипичном контексте (который маловероятен или не происходит). Соответственно, в процессе дополнительного обучения модели замаскированное слово будет явно «отрицательным». Похожая идея использования отрицательной выборки на уровне предложения в предварительно обученной модели *BERT* описана в многоязычных внедрениях предложений с помощью *BERT*. Двунаправленные двойные кодеры были обучены с аддитивным запасом *softmax loss* с отрицательной выборкой.

Таким образом, основное предложение по оптимизации работы алгоритмов на датасете *ReCoRD* - это дополнительное подключение методов отрицательных выборок к работе предварительно обученного алгоритма. Особое внимание будет уделено оценке работы модификации *BERTa + Negative sampling*. Для дальнейшего анализа будет проведена серия экспериментов.

Ввиду наличия небольших вычислительных ресурсов и сокращения времени обучения модели, в качестве нейросетевого алгоритма может быть использована модификация *BERT – DistilBERT*. Алгоритм *DistilBERT* находится в открытом доступе в библиотеке *Transformers v2.* от *HuggingFace*.

В ходе выполнения эксперимента необходимо пройти основные этапы:

1. Постановка задачи и выбор архитектуры нейросетевого алгоритма
2. Определение количественного и качественного составов входов и выходов
3. Формирование исходной выборки данных
4. Предварительная обработка и нормализация исходной выборки
5. Разделение исходной выборки на обучающую и тестовую составляющие
6. Определение структуры нейронной сети
7. Настройка параметров нейросетевого алгоритма для обучения
8. Обучение нейронной сети
9. Оценка работы нейросетевого алгоритма, получение метрик оценки

Выводы. По результатам проведения эксперимента оценки работы модели имеют следующие значения: *exact match* = 76,71, *F1*-мера = 85,1. Есть небольшое улучшение прогнозов до постобработки, где *exact match* = 75,32, *F1*-мера = 83,67. Следующим шагом в продолжении исследования является осуществление дополнительной оптимизации работы полученного нейросетевого алгоритма с использованием методов глубокого обучения, применение техники *Wide and Deep* с использованием самописных признаков (*handcrafted features*), прикладываемых к нейросетевой модели.

Олюнин В.В. (автор) Подпись

Платонов А.В. (научный руководитель) Подпись