Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
Высшего образования
**«Северо-Осетинский государственный университет
имени Коста Левановича Хетагурова»**

Дипломная работа

Seq2seq подход для задач Машинного Перевода

Выполнил:
Студент 4 курса направления:
«Прикладная математика и информатика»
Гамосов Станислав Станиславович _____

Научный руководитель:
Кандидат физико-математических наук:
Басаева Елена Казбековна _____

Консультант
Старший преподаватель:
Макаренко Мария Дмитриевна _____

Владикавказ 2022

Содержание

1 Введение

Seq2seq - это семейство подходов машинного обучения, используемых для обработки естественного языка. Основные задачи для которого используется данные методы: нейронный перевод, субтитры к изображениям, разговорные модели и обобщение текста.

Первоначальный алгоритм, который в процессе породил целое семейство методов, был разработан *Google* для использования в машинном переводе. Как уже можно заметить за последнюю пару лет коммерческие системы стали удивительно хороши в переводе - посмотрите, например, *Google Translate*, *Яндекс*-переводчик, переводчик *DeepL*, переводчик *Bing Microsoft*.

Так же **seq2seq** технология несет в себе огромный потанцевал, помимо привычного машинного перевода между естественными языками, вполне реализуем перевод между языками программирования (*Facebook AI "Глубокое обучение переводу между языками программирования"*). Поэтому возможности применений такого рода подходов довольно велики. В связи с этим под машинным переводом будет подразумеваться любая задача **seq2seq**, если точнее, то перевод между последовательностями любой природы.

2 Формализация задачи машинного перевода

Формально в задаче машинного перевода у нас есть входная последовательность $x_1, x_2, ...x_m$ и последовательность вывода $y_1, y_2, ...y_n$, само собой длинна данных последовательностей может отличатся. Саму процедуру *перевода* можно рассматривать как нахождение искомой последовательности, которая является наиболее вероятной с учетом входных данных. Формально искомая последовательность, которая максимизирует условную вероятность $p(y|x) : y^{'} = argmax[p(y|x)]$.

Когда человеку известны уже два языка с которыми он работает, то уже при переводе можно сказать насколько хорошо справилась модель, является ли перевод естественным и насколько он приятен на слух. Однако такой вид анализа неприемлем для машины, поэтому нам стоит проанализировать уже имеющуюся функцию $p(y|x, \theta)$ с неким параметром θ, а затем найти его $argmax$ для $y^{'} = argmax_y[p(y|x, \theta)]$.

Прежде чем перейти к самой задачи перевода, нужно ответить на 3 вопроса:

- **Моделирование**: Как работает модель для $p(y|x, \theta)$?
- **Обучение**: Как найти параметр θ?
- **Вывод**: Как понять, что текущий y лучший?

3 Задача машинного перевода

$source = (x_1, x_2, ..., x_n)$ - The cat sits on the floor

$target = (y_1, y_2, ..., y_m)$ - Кошка сидит на полу

Основная задача машинного перевода - найти наиболее вероятную последовательность для $target$ языка при условии, того что входная последовательность будет на $source$ языке.

$target^{'} = argmax_{target} P(target|source, \theta)$

$P(target|source) = P(y_1, y_2, ..., y_3|source) =$

$= P(y_1|source)\dot{P}(y_2|y_1, source)...P(y_m|y_1, ..., y_{m-1}, source)$

4 Структура Encoder-Decoder

Наиболее распространенная модель **Sequence-to-sequence (seq2seq)** являются модель **Encoder-Decoder**, в которой обычно используют **рекуррентную нейронную сеть (RNN)** для кодирования исходной последовательности в один вектор.

На самом деле полученный вектор можно представить как набор образов сущностей с образами взаимоотношений между ними. Этот вектор затем декодируется вторым **RNN**, который учится выводить выходное предложение, генерируя его по одному слову за раз.