

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:
Басеева Елена Казбековна _____

Консультант
Старший преподаватель:
Макаренко Мария Дмитриевна _____

Владикавказ 2022

Содержание

1	Введение	2
2	Формализация задачи машинного перевода	3
3	Структура Encoder-Decoder	4

1 Введение

Seq2seq - это семейство подходов машинного обучения, используемых для обработки естественного языка. Основные задачи для которого используются 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 потенциал, 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, \dots, x_m и последовательность вывода y_1, y_2, \dots, y_n , само собой длина данных последовательностей может отличаться. Саму процедуру *перевода* можно рассматривать как нахождение искомой последовательности, которая является наиболее вероятной с учетом входных данных. Формально искомая последовательность, которая максимизирует условную вероятность $p(y|x) : y' = \operatorname{argmax}[p(y|x)]$.

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 $y' = \operatorname{argmax}_y[p(y|x, \theta)]$.

Прежде чем перейти к самой задаче 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 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.