

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

Отчет по БДЗ-2 по курсу Deep Learning на тему:
Машинный перевод

Выполнил студент:

группы #БПМИ223, 3 курса

Тань Сипэн

Содержание

1	Введение	3
2	LSTM with Attention	4
3	LSTM with Attention update	4
4	Transformer	5
5	Заключение	6

1 Введение

Данную работу я поделю на 3 основные части:

- LSTM with Attention
- LSTM with Attention update
- Transformer

В итоге Bleu достиг 28.27. Данная работа я потратил больше всего времени на перебор параметров, а не написание модели.

2 LSTM with Attention

Первая модель представляет собой базовую реализацию sequence-to-sequence (seq2seq) архитектуры с механизмом внимания. Рассмотрим ее компоненты более подробно:

- **Encoder:** Двухнаправленный LSTM с 2 слоями
- **Decoder:** Однонаправленный LSTM с 2 слоями
- **Attention:** Аддитивное внимание

Основные параметры модели:

- **Размерность эмбеддингов:** 256
- **Размерность скрытого состояния:** 512
- **Регуляризация:** 0.6

Эта модель представляет собой классический подход к нейронному машинному переводу, популярный до появления трансформеров. Несмотря на свои ограничения, она демонстрирует базовую способность к переводу с немецкого на английский с BLEU 16.7, что является приемлемым результатом для начальной точки.

3 LSTM with Attention update

Вторая модель сохраняет основную парадигму seq2seq с вниманием, но вносит несколько критических улучшений.

Значительно увеличенная емкость модели:

- **Размерность эмбеддингов:** 512 (увеличение в 2 раза)
- **Размерность скрытого состояния:** 1024 (увеличение в 2 раза)
- Сохранение двухнаправленного LSTM в энкодере и однонаправленного в декодере

Улучшенный механизм внимания:

- Сохранение аддитивного внимания, но с улучшенной параметризацией
- Более точное маскирование для обработки паддинга
- Более эффективное объединение контекстного вектора с выходами декодера

Оптимизированная структура декодера:

- Улучшенное объединение информации из различных источников
- Более глубокая интеграция контекстного вектора в процесс генерации

И еще dropout 0.6, но вообще наверно надо было сделать dropout поменьше (просто мне в прошлой бдз помог большой dropout). Вторая модель демонстрирует, что даже в рамках традиционной архитектуры LSTM с вниманием можно достичь значительных улучшений путем увеличения емкости модели и продолжительности обучения. Это важное наблюдение, показывающее, что не всегда требуется радикальное изменение архитектуры для достижения лучших результатов. В итоге получил BLEU 20.

4 Transformer

После многочисленных попыток выбить BLEU больше 20 я понял что пора делать transformer. Третья модель представляет собой полный переход от рекуррентных нейронных сетей к архитектуре трансформера, что является парадигматическим сдвигом в подходе к машинному переводу.

Фундаментальные компоненты:

- **Энкодер:** 4 идентичных слоя, каждый с многоголовым вниманием и полносвязной сетью
- **Декодер:** 4 идентичных слоя с дополнительным механизмом маскированного внимания
- **Многоголовое внимание:** 4 головы внимания, каждая работает с проекцией размерности $512/4 = 128$
- **Позиционное кодирование:** Синусоидальные функции для кодирования позиций токенов

На данном этапе уже происходит поиск лучшего гиперпараметра. У меня модель достиг BLEU 25, осталось немного поменять параметры:

- **NUM_HEADS:** $8 \Rightarrow 4$
- **NUM_LAYERS:** $6 \Rightarrow 4$

- **FF_DIM: 2048 \implies 1024**

В итоге результат действительно стало лучше, получил Bleu 28.27. Это может быть из за того что модель была сложной и на маленьких эпохах не смог успеть обучиться.

5 Заключение

- **Базовая модель LSTM с вниманием (BLEU 16.7):** Представляет собой фундаментальный подход, обеспечивающий приемлемые результаты с ограниченными вычислительными ресурсами.
- **Улучшенная модель LSTM с вниманием (BLEU 20.0):** Демонстрирует, что увеличение емкости модели и продолжительности обучения может значительно повысить качество перевода даже при сохранении базовой архитектуры.
- **Трансформер (BLEU 28.0):** Представляет революционный подход, полностью отказывающийся от рекуррентных связей в пользу механизма внимания, что приводит к значительному улучшению качества перевода.