Машинный перевод

Задача: реализовать модель Transformer (базово: из Attention is All you Need) для задачи машинного перевода.
Чужой код использовать можно, если вы понимаете, что там происходит и готовы ответить по нему (не забудьте указать ссылку)

Можно взять:

1. пары языков (например https://lionbridge.ai/datasets/25-best-parallel-text-datasets-for-machine-translation-training/)
2. нормализация текстов для предобработки перед подачей в языковую модель распознавания речи (“сегодня 23 сентября и я заработал 33 руб” -> “сегодня двадцать третье сентября и я заработал тридцать три рубля”
3. (*) расстановка пунктуации (или смайликов) (энкодер из трансформера + линейный слой на классификацию). Тут без прунинга, но пошатать архитектуру надо будет
4. Для ленивых: обучить transformer из HF (но это максимум **20%** баллов, то есть минимальная тройка)

Напишите в чат в slack, какую тему вы выбираете

Часть 1 (20% баллов)

Результат: готовы bpe-токенизация (опционально), эмбеддинги и датасет для обучения

1. В зависимости от задачи выбрать данные
2. Обучить bpe-токенизатор (sentencepiece/youtokentome/hugginface) (опционально)
3. Обучить word2vec вашим любимым способом

Часть 2 (50% баллов)

1. Написать пайплайн для обучения. В интернетах есть масса мест куда подглядеть, поэтому я предлагаю подглядывать сюда если очень захочется (это https://arxiv.org/abs/1904.10509) https://drive.google.com/drive/folders/1qXkPrDyjH3fp0ufkvUspqPQllCJRt7NM?usp=sharing (открою доступ на почту по запросу)
2. Запустить обучение. При необходимости видеокарты можно арендовать на Google Cloud, однако скорее всего вам хватит колаба. Будут вопросы про аренду сервера -- пишите
3. Реализовать рабочее демо идеально через бота (PyTelegramBotAPI) либо через colab

После выполнения первых двух частей вы получаете допуск к третьей части

Часть 3 (30% баллов и шоколадка)

1. Рассказать мне в тг @Nestyme, как работает прунинг своими словами. Это дает допуск к выполнению третьей части
2. Написать/раз прунинг для голов трансформера. Посмотреть на результаты http://jalammar.github.io/illustrated-transformer/

https://lena-voita.github.io/posts/acl19_heads.html

3. Реализовать рабочее демо идеально через бота (PyTelegramBotAPI) либо через colab

Финал

Если вы сделали ленивый подход, то вы просто присылаете мне проект в любом читаемом виде в телеграм @Nestyme и получаете максимально 3/10 баллов

Если вы сделали 1+2 либо 1+2+3 части, то проект нужно защитить. Вы присылаете мне в телеграм @Nestyme демо и код, по итогам которого я могу письменно прислать вопросы. После ответа на них ставится оценка по десятибалльной шкале (если вы сделали 1+2, то максимально получаете 7/10, если сделали 1+2+3, то максимально получаете 10/10). Если вы хотите повысить балл, то мы созваниваемся и я задаю рандомные вопросы по курсу

Стараюсь смотреть slack, но быстрее отвечу в телеграме. Пишите по любым вопросам :)