Проект caption generation.

Цель проекта – обучить модель, способную генерировать адекватные с точки зрения человека описания для картинок, не включенных в тестовую выборку.

Основным подходом к решению этой задачи является использование предобученного энкодера (модели, обученной на задаче классификации картинок) и рекуррентной нейронной сети. Такой метод решения описан в статье Show and Tell: A Neural Image Caption Generator Oriol Vinyals (Alexander Toshev Google, Samy Bengio, Dumitru Erhan, 2015). Авторы статьи отмечают, что для данной задачи предобученные эмбеддинги не способствуют повышению качества работы модели.

**Формат данных**

Я использовала датасет cococo 2014 annotations (тренировочный и валидационный), представленный в виде json. Фактически использовались два поля, одно из которых являлось массивом с данными о картинках (id, название файла, ссылки на изображение), а другое – массив с данными об аннотациях (текст аннотации, id соответствующей картинки). Для более быстрого соотнесения я использовала словарь, созданный на основании поля с аннотациями, ключами которого являлись индексы картинок, а значениями - тексты аннотаций.

**Успешные попытки**

Я обучила две модели одинаковой архитектуры: в качестве энкодера использовала модель Alexnet, из выходных данных предпоследнего слоя которой с помощью обучаемого линейного преобразования формировала memory модели и обучаемых параметров lstm. В первой модели я обучала эмбеддинги одновременно с моделью (ограничила словарь до 1500 слов), во втором случае взяла предобученные английские эмбеддинги (500000).

Предсказания для оценки генерировались с помощью beam search.

# Оценивание

# Оценивалась адекватность с точки зрения человека (меня), bleu (метрика, использованная в вышеупомянутой статье), bert score (метрика, которая не штрафует чрезмерно использование близких по значению слов, в статье «BERTScore: Evaluating Text Generation with BERT» ([Felix Wu](https://www.researchgate.net/profile/Felix_Wu6), [Kilian Weinberge](https://www.researchgate.net/profile/Kilian_Weinberger), 2019) при анализе коэффициента согласия используется датасет CoCoCo, следовательно, метрика применима к данной задаче).

# Обе модели при генерации часто создают повторяющиеся описания.

# С точки зрения человека результаты с предобученными эмбеддингами чаще соответствуют тематике картинки, но сгенерированный текст часто зацикливается и становится грамматически неправильным. Случаи практически идеального описания с точки зрения человека замечены только у модели с ограниченным словарем, но их крайне мало и это могло быть случайностью.

# Эти наблюдение подтверждается результатами bleu: 25-й, 50-й и 75-й перцентили выше для модели с предобученными эмбеддингами, но наилучший результат существенно выше у модели с ограниченным словарем.

# Значения Bert score сопоставимы, но несколько выше у модели с ограниченным словарем.

# Неуспешные попытки

# Я пыталась обучать веса, преобразующие выходной слой энкодера в past key values gpt2. Результат генерации не отличался от генерации без заданного значения past key values.

# Также я пыталась воспроизвести метод, описанный в статье «From Captions to Visual Concepts and Back» (Forrest Iandola, Rupesh K. Srivastava, Li Deng Piotr Dollar, Jianfeng Gao Xiaodong, He Margaret Mitchell John C. Platt, C. Lawrence Zitnick Geoffrey Zweig), где авторы сначала по картинке предсказывали ключевые слова, а затем из них составляли наиболее вероятное предложение. В статье использовалась предварительная сегментация картинок. Так как знакомство с инструментами сегментации значительно увеличило бы трудоемкость проекта, я опустила этот шаг и не получила удовлетворительного результата. Авторы после фильтрации отсекали самые частые слова, которые модель пытается предсказывать для всех картинок, но в моем решении в число самых частых слов входили слова man и people, при этом других частотных номинаций людей не было, а большую часть датасета составляли картинки с людьми.

# Я остановилась на этом этапе, полагая, что не смогу получить подходящий список ключевых слов для выполнения следующего шага.

# Бейзлайн

# Если бы я смогла реализовать метод, основанный на определении ключевых слов, бейзлайном можно было бы считать оценку на этом наборе слов или части этого набора(от генерации связного текста ожидалось бы, что она повышает качество).

# Кажется, что для генерации описаний с помощью энкодера и декодера бейзлайна нет

# Что было сложно

# Я впервые использовала нетекстовые данные и впервые работала с данными такого большого объема (папка с тестовыми картинками весит более 8Гб). Пришлось потратить много времени на попытки считать картинки в момент инициализации датасета, прежде чем я поняла, что стоит делать это в момент получения данных. После этого я слишком поздно поняла, как можно было обрабатывать исключения при получении данных при составлении батчей, и все обучение провела на батчах из одного примера (с аккумуляцией градиента).

# Кроме того, было сложно понять, как контролировать происходящее и в какой момент уже не рано оценивать промежуточные результаты. Поэтому все мои модели обучались одну эпоху (это занимало около суток) и кажется, что этого было недостаточно.