Image captioning Bot

Semen Semenov
semenov.ss@phystech.edu

Daniil Merkulov
daniil.merkulov@skoltech.ru

Project Proposal

В рамках данного проекта мы реализуем модель, основанную на глубокой рекуррентной архитектуре, которая сочетает в себе последние достижения в области компьютерного зрения и машинного перевода и которая может быть использована для генерации естественных предложений, описывающих изображение.

1 Problem

Задача автоматического описания изображений является фундаментальной в области машинного обучения, объединяет в себе классы задач распознавания изображений и обработки естественного языка, имеет неоценимую важность в современных прикладных вопросах и заслуживает большого внимания.

Данный проект является первым шагом на пути к решению этой основополагающей проблемы и является естественной ступенью в изучении методов искусственного интеллекта.

1.1 Идея

В процессе анализа нейронных сетей было показано, что глубокие сверточные нейронные сети способны воспроизводить входное изображение с помощью вектора фиксированной длины. Такие нейронные сети носят название автоэнкодеров и состоят из двух принципиальных блоков: энкодера и декодера. Энкодер кодирует изображение некоторым вектором, а декодер, в свою очередь, декодирует его, то есть преобразовывает его обратно в исходное изображение.

В основе проекта лежит именно эта концепция. Здесь, энкодером будет являться глубокая сверточная нейронная сеть, а декодером - глубокая рекуррентная нейронная сеть. Таким образом, подавая на вход изображение, на выходе мы получим его естественное текстовое описание.

2 Outcomes

Чтобы повысить доступность, упросить способ взаимодействия пользователя с нейронной сетью, обладающей описанной выше архитектурой, и сделать результат самодостаточным, проект будет упакован в Telegram - бота. Реализация бота будет выполнена с помощью библиотеки python - telegram - bot.

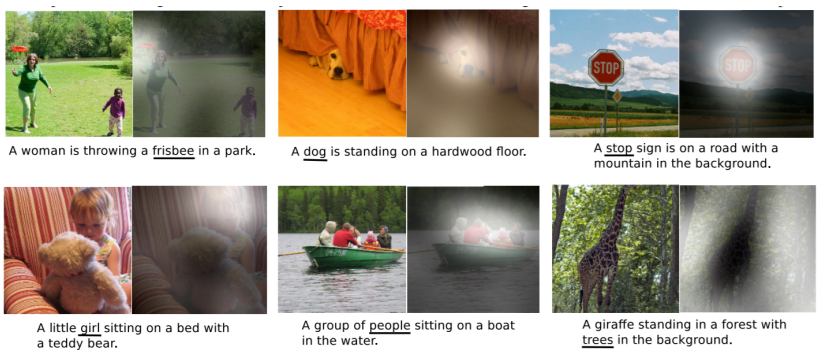

Figure 1: Итоговая цель проекта.

3 Литературный обзор

Обратимся к оригинальной статье [1], в которой изложены основные идеи решения задачи image captioning. Кроме того, приведем классические статьи, в которых подробно изучены важнейшие реализации сверточной архитектуры, используемой здесь в качестве энкодера: Inception-v3 [2], VGG16 [3] и рекуррентной архитектуры в качестве декодера: LSTM [4], GRU [5]. В рамках проекта используем принципиально важный подход - Attention mechanism [6]. Данный метод позволяет "обратить внимание" нейронной сети на ключевые детали изображения. Расположив "слой внимания" между энкодером и декодером, можно существенно улучшить точность модели.

4 Метрики качества

В качестве показателя точности описания будем использовать BLEU Score [7]. BLEU - алгоритм оценки качества текста, который был автоматически переведен с одного языка на другой. Баллы рассчитываются для отдельных переведенных предложений путем сравнения их с набором качественных справочных переводов. Затем эти оценки усредняются по всей выборке, чтобы получить оценку общего качества перевода. Разборчивость или грамматическая правильность не принимаются во внимание.

5 План

- Предобработка текста из датасета MSCOCO.
- Реализация архитектуры CNN и RNN.
- Обучение нейронной сети.
- Создание Telegram - бота.
- Хостинг бота. Это программа минимум.
- Attention mechanism - если успею программу минимум.

References

[1] Vinyals, Oriol and Toshev, Alexander and Bengio, Samy and Erhan, Dumitru. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence.

[2] Xia, Xiaoling and Xu, Cui and Nan, Bing. Inception-v3 for flower classification. 2017 2nd International Conference on Image, Vision and Computing (ICIVC).

[3] Qassim, Hussam and Verma, Abhishek and Feinzimer, David. Compressed residual-VGG16 CNN model for big data places image recognition. 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC).

[4] Hasim Sak, Andrew Senior, Francoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition.

[5] Fu, Rui and Zhang, Zuo and Li, Li. Using LSTM and GRU neural network methods for traffic flow prediction. 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC).

[6] Chorowski, Jan and Bahdanau, Dzmitry and Serdyuk, Dmitriy and Cho, Kyunghyun and Bengio, Yoshua. Attention-based models for speech recognition.

[7] Zhang, Ying and Vogel, Stephan and Waibel, Alex. Interpreting bleu/nist scores: How much improvement do we need to have a better system?