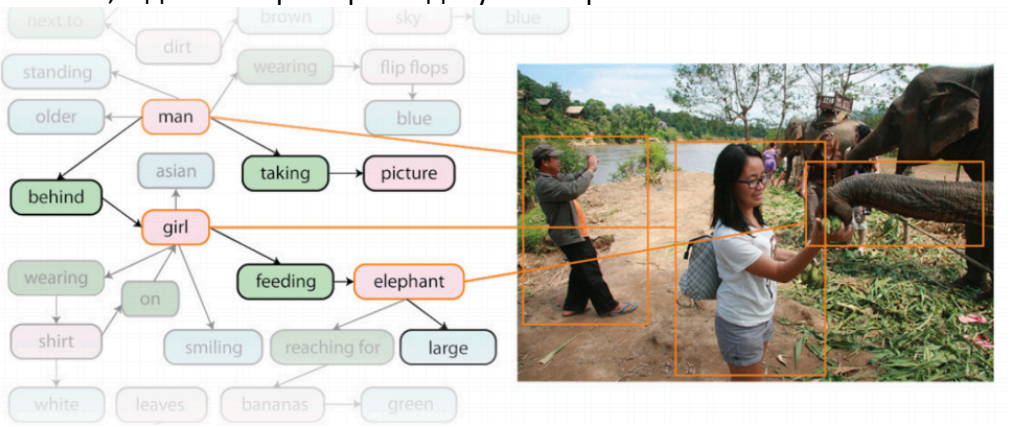
Разбор статьи «Attentive Gated Graph Neural Network for Image Scene Graph Generation» опубликованной 2 апреля 2020г. https://www.mdpi.com/2073-8994/12/4/511/pdf

Статья Attentive Gated Graph Neural Network for Image Scene Graph Generation
Описывает построение графа сцены изображения, на котором определяется взаимоотношение между выявленными объектами. К функции обнаружения объектов путём сегментации добавляется описание того, как они взаимодействуют или относятся друг к другу путем анализа соприкасающихся краёв. Объекты находятся в узлах графа, а рёбра отображают взаимоотношения.
Граф сцены представлен на изображении. Граф состоит из триплетов «субъект-предикат-объект», в данном примере – «девушка-кормление-слон».



В качестве предикатов выявляются виды взаимодействия между объектами сцены, в частности:

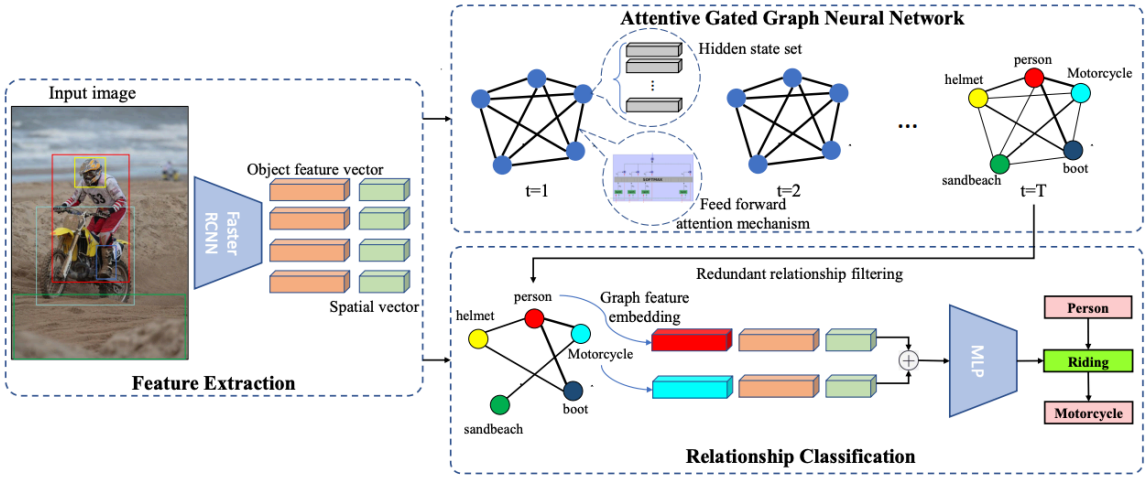
- глаголы (например, взятие, кормление);
- сравнительные (например, больше, меньше);
- относительные позиции (например, выше, позади)

Особенность подхода, описанного в статье в том, что рассматривается весь контекст изображения, т.е. весь граф с помощью сверточной сети, которая может распространять информацию в обоих направлениях. Глобальный граф сцены кластеризуется, затем для отношений устанавливаются различные веса, а избыточные связи убираются.
Предлагается использовать нейронную сеть Attentive Gated Graph Neural Network (AGGNN) для одновременного распознавания объектов и фильтрации избыточной информации. На узлах закрытого графа рекуррентными последовательными архитектурами Graph Long Short-Term Memory Network (GLSTM) сегментируется набор объектов в прямоугольниках. Для генерации набора ограничивающих рамок прямоугольников используется Faster R-CNN.

Рисунок иллюстрирует конвейер из трёх модулей:

- модуль извлечения признаков
- модуль нейронной сети для построения полного графа взаимосвязей
- модуль классификации взаимосвязей

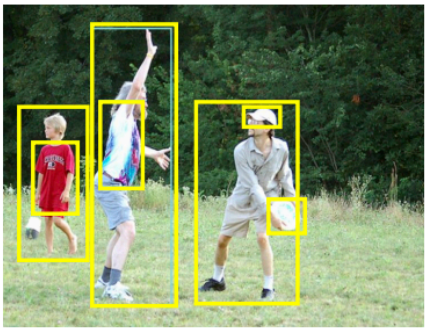
В последнем модуле граф разрежается с помощью Faster R-CNN классифицируя и отбирая важные отношения между объектами. Процесс повторяется для всех пар объектов графа сцены.

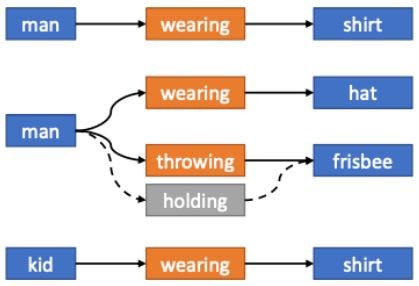


Обучали сеть и проводили оценку на датасете Visual Genome Dataset (VG). Авторы дополнительно улучшили датасет (доработанные версии CVG и DCVG) унифицируя слова, чтобы разные слова могли выразить одно и то же значение (использовали герундий вместо глагола), приводили к одной формы единственного и множественного числа, абстрагировали категории зависимостей («имеет», «включено» и «из»). Использовали 150 категорий объектов и 50 предикатов для оценки.При обучении использовалась разметка статистической вероятности совпадения объектов из разных категорий и их отношений.

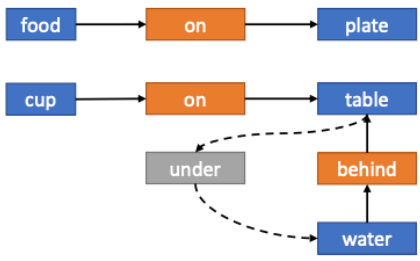
Оба сета CVG и DCVG делили на 70% трейн и 30% тест. Затем на 5000 изображениях из тренировочного набора проводили настройку гиперпараметров.
Модель реализована на TensorFlow, считалась на. GPU NVIDIA 2080 Ti.
Использовалась сеть R-CNN предобученная на наборе данных ImageNet.
Для оценки качества использовали Top-K Recall.

Цель модели в том, чтобы предсказать категории взаимодействия всех объектов и отношений на изображении, т.е. правильно определить триплеты «Субъект-предикат-объект».





(a)



В нашей модели выявлено два распространенных недостатка.

1) На рисунке «человек держит фрисби» ошибочно идентифицируется как «человек бросает фрисби». Действительно, действие на рисунке можно определить и как «удержание» и как «бросание».
2) В сложных сценах модель ошибается в логике взаимодействия объектов. Связь между «столом» и «водой» классифицируется ошибочно.

Тем не менее Модель способна генерировать графы сцен с высоким качеством в большинстве сценариев.

В документе предложена новая сквозная нейронная сеть, которая может автоматически генерировать граф сцены изображения. Метод состоит из трех модулей. Избыточные взаимодействия объектов отсекаются. Метод имеет сопоставимые результаты метрик со state-of-the-arts.

/Дьяченко Даниил https://github.com/daniilstv/neural_intro