УДК 004.855.5

МОДЕЛИРОВАНИЕ ДИАЛОГОВ НА ОСНОВЕ НОВОЙ АРХИТЕКТУРЫ ИЕРАРХИЧЕСКИХ НЕЙРОННЫХ СЕТЕЙ-ТРАНСФОРМЕРОВ

Варакин Я.С.
(Пятигорский государственный университет, г. Пятигорск)
Научный руководитель – к. пед. н., доцент Писаренко Е.А.
(Пятигорский государственный университет, г. Пятигорск)

В работе предложена новая архитектура диалогового агента на основе нейронных сетей Self-Attention. Новизна заключается в том, что при подаче в систему контекстного диалога поток разделяется на две ветви: одна ветвь обрабатывает текущий ввод данных (подача текущей реплики напрямую в основной энкодер), другая обеспечивает сохранение, обновление и использование истории диалога (составление тензора истории диалога с помощью энкодера контекста).

В задачах обработки естественного языка (распознавание речи, генерация текстов, машинный перевод) и задачах распознавания образов обычно используются рекуррентные нейронные сети (RNN). Однако наиболее распространенные рекуррентные сети на основе архитектуры HRED (Hierarchical Recurrent Encoder Decoder) имеют серьезные недостатки – трудности в оптимизации и обучении, высокую ресурсоемкость, их использование приводит к снижению эффективности работы аппаратных ресурсов. Поэтому на смену рекуррентным сетям пришли сети-трансформеры (Self- Attention Networks), которые позволяют строить модели с большим числом параметров благодаря более полному использованию параллелизма GPU. Трансформеры значительно легче оптимизируются, т.к. в них отсутствует проблема взрывающихся и угасающих градиентов. Такие сети характеризуются большей биологической правдоподобностью и способны давать State-of-the-art результаты.

Мы предлагаем новую архитектуру нейронной сети – HTED (Hierarchical Transformer Encoder Decoder). Ее особенность заключается в разделении обрабатываемого потока на две ветви, одна из которых хранит историю диалога с более низким уровнем гранулярности, чем другая. В первой ветви кодирование осуществляется из расчета один вектор на одно слово, а во второй – один вектор на одну реплику.

В предлагаемой архитектуре присутствует энкодер контекста, а декодер остается общим и получает данные одновременно и о контексте, и о текущей реплике. Кроме того, в энкодер контекста всегда поступает неизменная реплика, текст которой составляет т.н persona, или детали личности агента.

С точки зрения нейрофизиологии такая модель гораздо точнее имитирует работу мозга человека, чем обычные сети трансформеры, так как в некоторой степени воспроизводит функционирование рабочей памяти, в которой хранятся данные за некоторый период времени.

После обучения в течении 10 эпох система показывала Macro f1 в 0.41 и Perplexity score в 2,54. На тестовых данных система показывала Macro f1 в 19,33 и Perplexity score в 19,45. Сеть показывает способности к прямому zero-shot и few-shot обучению на основе словесных описаний задачи, не требуя дополнительных градиентных обновлений весов.

Предложенный механизм позволяет выстраивать и поддерживать личность диалогового агента, создавая имитацию естественного поведения, что дает человеку ощущение большей естественности поведения бота. В зависимости от содержания лингвистических корпусов, используемых для обучения системы, бот сможет выполнять функции консультанта в различных областях – экономике, банковском деле, туризме, образовании и т.д.

Варакин Я. С. (автор) Подпись

Писаренко Е. А. (научный руководитель) Подпись