Transformer Feed-Forward Layers Are Key-Value Memories: Рецензия

В статье изучают feed-forward слои языковых трансформеров. Авторы показывают, что эти слои ведут себя как key-value memories, где каждый ключ отвечает за определенный паттерн во входных данных, а значения за распределения над словарем. Эксперименты показывают, что усвоенные паттерны могут быть интерпретированы человеком.

Статья на 90 процентов состоит из экспериментов, однако идея на которой они основаны теоретически обоснована и не вызывает вопросов. Полнота эмпирического анализа также не вызывает вопросов: все эксперименты поставлены корректно и хорошо описаны. В рамках рассматриваемого авторами окружения сложно придумать не рассмотренный аспект.

Результаты статьи привносят новизну в область, поскольку никто до этого не занимался интерпретацией именно feed-forward слоев. Данная работа помогает лучше понять языковые трансформеры, но не совсем понятно, какую пользу это может принести на текущий момент. На мой взгляд, существование работы почти никак не влияет на состояние области в целом, поэтому нельзя назвать её значимой.

Вопросов по исследованию конкретной задачи поставленной в статье нет, однако возможно в следующих работах хотелось бы увидеть подобное же исследование для других задач, кроме языковых моделей.

Моя оценка статье – 8, поскольку это хорошее исследование и проделана качественная работа, заслуживающая внимания. Тем не менее, оценить работу выше нельзя из-за ее слабой практической применимости. Я слабо погружен в область NLP и возможно не осознал каких-то аспектов этой работы и ее влияния, поэтому моя уверенность в оценке 3.