

**АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ ОБУЧЕНИЯ ПРЕДСТАВЛЕНИЙ ГРАФОВ****Ильясова О.С.** (Университет ИТМО)**Научный руководитель – к.т.н., доцент Муравьев С. Б.**

(Университет ИТМО)

В работе систематизируются знания о подходах в обучении представлений графов (Graph Representation Learning) от методов, основанных на случайных блужданиях, до алгоритмов получения графовых эмбедингов при помощи графовых нейронных сетей (Graph Neural Networks).

**Введение.** Графы являются одним из самых выразительных способов представления данных, поскольку позволяют естественным образом представлять любые взаимосвязанные данные, будь то банковские транзакции или молекулярные структуры. Графовые данные могут использоваться как для решения типовых графовых задач, таких как предсказание связей (link prediction) или нахождение сообществ (community detection), так и передаваться в модели машинного обучения. Традиционно, для извлечения структурной информации из графов использовались различные топологические характеристики, например меры центральности, однако, во-первых, конструирование специальных эвристик, описывающих вершины, под каждую конкретную задачу является неэффективным, во-вторых, эвристики зачастую агрегируют только локальную информацию, и наконец в случае больших графов расчет эвристик более высоких порядков становится практически невозможен.

**Основная часть.** В последние годы появились методы, позволяющие автоматически преобразовывать граф в векторные низкоразмерные представления с сохранением информации о семантических отношениях. Семейство таких алгоритмов получило название graph representation learning или обучение представления графов, поскольку в процессе оптимизации некоторой параметрической функции (энкодера), они "выучивают" новое представление вершин графа – эмбединги. Оптимизационная задача в данном случае заключается в том, чтобы выбрать параметры энкодера таким образом, чтобы близкие вектора в пространстве эмбедингов соответствовали соседним вершинам в исходном графе.

Первой группой алгоритмов получения графовых эмбедингов стали модели, основанные на случайных блужданиях. В 2014 году был представлен алгоритм DeepWalk, адаптирующий хорошо зарекомендовавшую себя для решения задач моделирования языка архитектуру word2vec под графовые данные. Рассматривая посещенные вершины графа как слова в предложениях, а само случайное блуждание как предложение, для каждой вершины методом Skip-gram определялся ее эмбединг. Обобщение метода, параметризующее процесс случайных блужданий, было представлено в алгоритме node2vec. Оба метода максимизируют вероятность нахождения двух вершин на одном пути; в отличие от них, предложенный позднее алгоритм VERSE оптимизирует распределение вероятностей появления вершин на пути случайного блуждания и поддерживает различные меры схожести, что дает state-of-the-art результаты для задачи link-prediction. Недостатками моделей на случайных блужданиях является, их вычислительная сложность, занимающая  $O(n \log n)$  на каждом шаге оптимизации, что на практике ограничивает размер графа примерно тремя миллионами вершин, а также необходимость выбора меры схожести для конкретной задачи. Существуют и другие алгоритмы, использующие, например матричные разложения, агрегации эмбедингов вершин и «анонимные» блуждания, но все эти методы, основанные на неглубоком обучении, объединяют трансдуктивность – то есть невозможность генерировать представления для вершин, которые не были доступны в процессе обучения, а также неиспользование дополнительной информации, содержащейся в признаках вершин.

Перечисленные проблемы отсутствуют в моделях на основе глубоких нейронных сетей – так называемые графовые нейронные сети (Graph Neural Networks, GNNs), которые итеративно вычисляют эмбединги вершин. Для каждой вершины запускается случайное блуждание, после чего свойства смежных вершин комбинируются с помощью нелинейных агрегаторов (нейронных сетей). На сегодняшний день предложен целый ряд всевозможных способов агрегации, помимо простого суммирования состояний соседних вершин. GNNs в свою очередь также не лишены недостатков, основным из которых является чрезмерное сглаживание – чрезмерная схожесть эмбедингов соседних вершин из-за схожих окрестностей. Для предотвращения этой проблемы используют приёмы, применяемые в глубоких нейронных сетях – “прокидывание” слоев (residual connections), использование ячеек памяти, аугментации и регуляризации.

**Выводы.** Существует множество алгоритмов получения графовых эмбедингов, каждый из которых имеет свои ограничения и преимущества. Выбор наилучшего метода получения графовых представлений является непростой задачей и в настоящее время подбирается под конкретную задачу экспертами вручную. Исходя из изложенного выше, создание автоматических методов получения графовых эмбедингов, без привязки к конкретным приложениям, является актуальной задачей, которая будет решена в ходе дальнейшего исследования в рамках диссертационной работы.

Ильясова О. С. (автор)

Подпись

Муравьёв С. Б. (научный руководитель)

Подпись