

1 Введение

С ростом вычислительных мощностей, задачи связанные с данными представленными в виде графов, все больше набирают популярность. Развиваются как старые подходы (детерминированные алгоритмы), так и новые (подходы, основанные на моделях и принципах искусственного интеллекта), в том числе подходы, основанные на нейронных сетях. Важной задачей в области графов, является задача генерации графов – она может применяться как для создания выборки для обучения других моделей, так и, например, для предсказания состояния графа на следующем временном шаге. Уже существует ряд работ, посвященных генерации графов, основанной на идее вариационного автокодировщика, но все они имеют ряд недостатков: первый – графовые нейронные сети не могут иметь большую глубину, что не позволит выучить сложные зависимости графов, второй – генерация, происходит из скрытого вектора, который включает, как интересую для нас информацию (содержащую сложные зависимости в графах), так и довольно простую, легко рассчитываемую статистику – например число вершин, ребер и так далее. Мы презентуем нашу идею контролируемой генерации графа на основе сепаративного скрытого представления, а также архитектуру, которая реализует эту идею.

2 Обзор литературы

Задача генерации графов стала особенно острой в последние годы в связи с ростом популярности нейросетевых моделей, которые требуют большое количество данных для обучения.

Для генерации графов было опробовано множество подходов, популярных в задачах генерации других типов данных: авторегрессионные модели[], модели вариационного автокодировщика[], модель нормализующих потоков[], GAN[] и другие.

Однако предложенные методы, не учитывают специфику данных. В отличии алгоритмов, работающих с изображениями, алгоритмы, принимающие на вход граф, могут использовать в качестве дополнительной информации структурные характеристики графа для того, чтобы точнее улавливать скрытые структурные признаки.

Среди всех подходов, нами была выбрана модель вариационного автокодировщика, поскольку она допускает несложную модификацию, позволяющую учитывать дополнительные признаки при генерации.

3 Предложенный метод

На этапе подготовки данных мы предлагаем вычислить характеристики, называемые нами простыми признаками: статистики графа, которые можно рассчитать за линейное время от входа. В работе рассматривается гипотеза о том, что если учесть их отдельно, возможно выделить из скрытого вектора составляющие, характеризующие сложные зависимости.

3.1 Conditioning ELBO

При использовании моделей генерации, мы часто прибегаем к оценке распределения объектов(ELBO). В нашем случае объектами являются графы или матрицы смежности графов:

$$\begin{aligned}\log(p_\theta(x)) &\geq ELBO = E_{q_\phi(z|x)} \log \left[\frac{p_\theta(z, x)}{q_\phi(z|x)} \right] = \\ &= E_{q_\phi(z|x)} [\log p_\theta(x|z)] - KL(q_\phi(z|x) || p_\theta(z))\end{aligned}$$

Использование дополнительных признаков y предполагает, что нас интересует условное распределение $p_\theta(x|y)$ (Conditioning ELBO):

$$\log(p_\theta(x|y)) \geq E_{q_\phi(z|x,y)} [\log p_\theta(x|y, z)] - KL(q_\phi(z|x, y) || p_\theta(z|y))$$

В нашем же случае, дополнительные признаки могут быть получены из входа детерминированным способом, поэтому CELBO переписывается следующим образом:

$$\log(p_{\theta}(x|y)) \geq E_{q_{\phi}(z|x)}[\log p_{\theta}(x|y, z)] - KL(q_{\phi}(z|x)||p_{\theta}(z|y))$$

3.2 Теоретическое обоснование раздельной генерации

Теорема 1: Класс графов, полученных при фиксировании простых параметров меньше, чем класс графов, полученных при фиксировании сложных параметров. (Надо очень аккуратно расставить ограничения)

Теорема 2: (О неоптимальности нахождения положения скрытых векторов)

Теорема 3: (Об оценке разницы положения вектора, найденного по двум переменным одновременно и по каждой в отдельности)

3.3 Архитектура модели

В работе было предложено несколько вариантов архитектур на основе GraphVAE и DSSM. GraphVAE используется для создание скрытого представления и генерации, а DSSM для разделения информации внутри скрытого представления. Матрица смежности и признаки вершин используются в качестве входа в GraphVAE. Энкодер GraphVAE преобразовывает исходный граф в скрытый вектор. Декодер трансформирует вектор в новую матрицу смежности. После применения энкодера, мы приводим скрытый вектор и вектор простых признаков в одно пространство при помощи структуры на основе DSSM. Мы строим приближение скрытого вектора на основе вектора простых признаков и вычитаем одно из другого. Полученную разность назовем сложными признаками. Они отвечают за микроструктуру графа.

Далее будет представлен поэтапный пайплайн модели генерации графов при помощи нетривиальных распутанных представлений. Полный пайплайн можно видеть на Рис. 1-2.

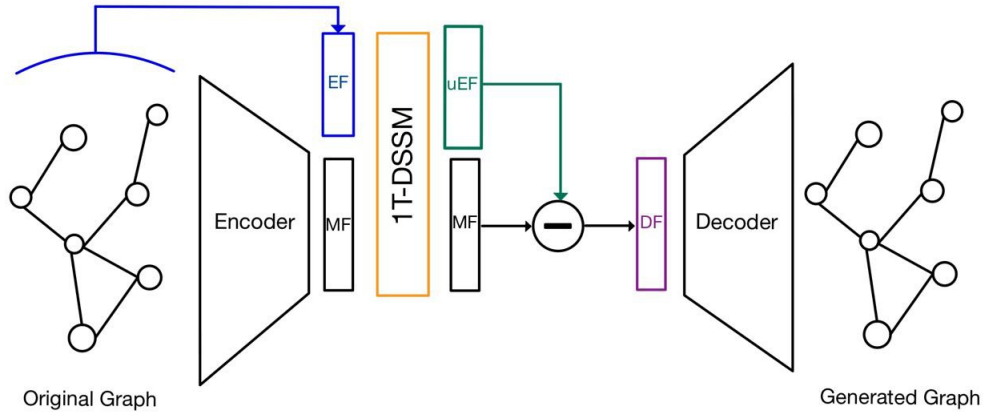


Рис. 1: Схема модели с генерацией только на основе сложных признаков.

Здесь EF – это представление простых признаков, которые можно достаточно быстро посчитать точными алгоритмами (число вершин, число ребер, максимальная степень и т.п.). DF – это представления сложных признаков графа (например максимальная смежная центральность). MF – это смешанное представление сложных и простых признаков графа. uEF – представление, которое построено на основе EF, и максимально приближено к MF.

Первым этапом является обучение представления на основе смешанных признаков графа. Это будет происходить при помощи модели VAE (Рис. 3).

Следующим этапом идет обучение модели 1T-DSSM (Рис. 4). Модель будет использовать обученные представления из прошлого этапа и будет стремиться используя только простые признаки, получить представление смешанных. Часть пайплайна обучения предсказана на Рис. 5.

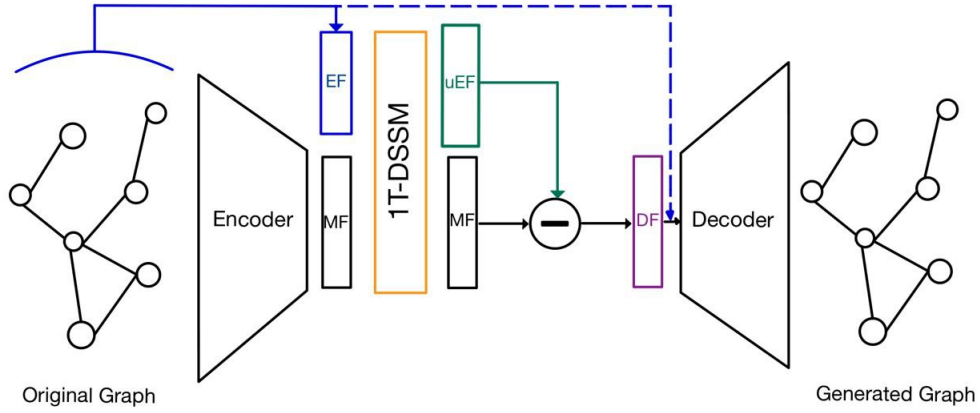


Рис. 2: Схема модели с генерацией на основе сложных и простых признаков.

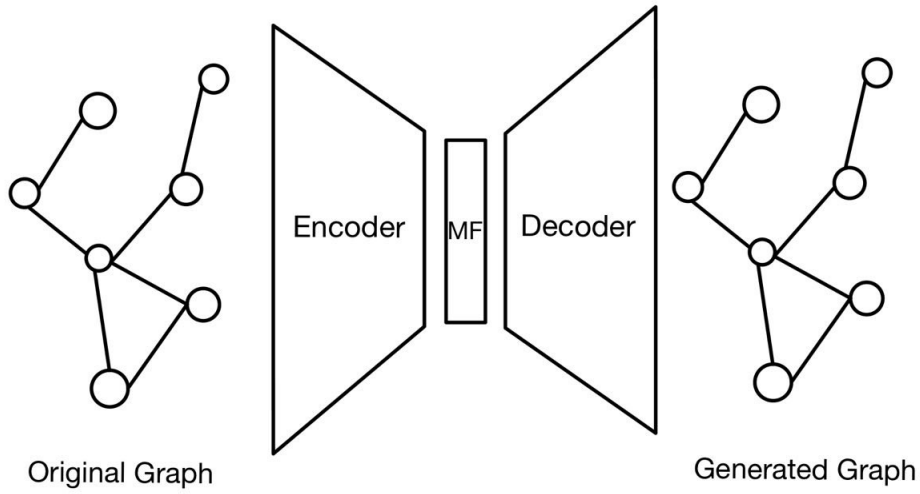


Рис. 3: GraphVAE.

При обучении 1T-DSSM мы будем оптимизировать следующий функционал:

$$\text{sim}(C, D) = \frac{C^T \cdot \bar{D}}{\|C\| \cdot \|\bar{D}\|} = \frac{MF^T \cdot uEF}{\|MF\| \cdot \|uEF\|}$$

Благодаря использованию 1T-DSSM мы можем исключить из MF большую часть влияния простых признаков просто отняв от MF uEF и получить представление только сложных признаков графа DF.

После подготовительных этапов, весь пайплайн будет обучаться вместе в два этапа:

1. Несколько эпох обучения VAE с фиксированными весами 1T-DSSM
2. Несколько эпох обучения 1T-DSSM при фиксированных весах VAE

Итогом этого пайплайна будет возможность генерировать графы с фиксированными простыми признаками и случайными сложными, что позволит генерировать похожие графы с одной макроструктурой, но сильно отличающейся микроструктурой.

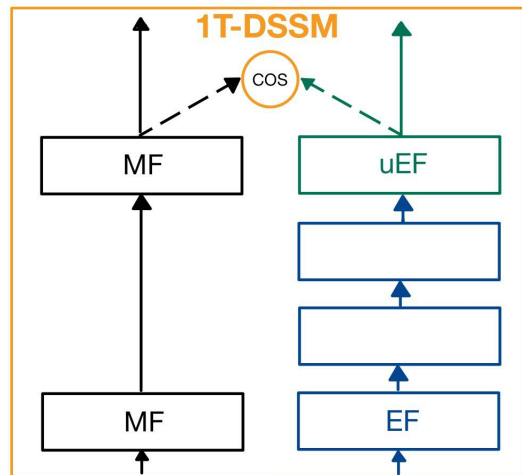


Рис. 4: Модель однобашенной DSSM.

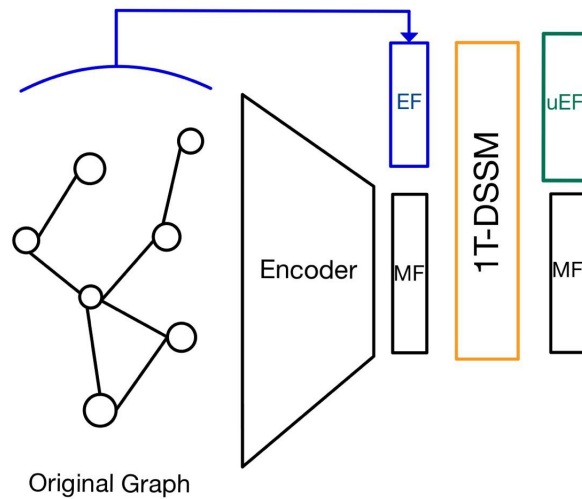


Рис. 5: Часть модели на которой будет обучаться модель 1T-DSSM.

4 Вычислительный эксперимент

TODO

5 Заключение

TODO