Контролируемая генерация графов при помощи VAE

Бишук Антон Юрьевич

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель к.ф.-м.н. Зухба А.В.

Москва
2023 г

Задача генерации графов с фиксированными статистиками

Цель

Генерировать графы с заданными статистиками.

Задача

Предложить метод генерации графа, который позволит задавать ряд стандартных числовых характеристик графа на этапе генерации, а также теоретически обосновать работу такого метода.

Статистики графа

Простые признаки

Простые признаки (или же простые статистики) в предложенном методе - это числовые характеристики используемые в теории графов, которые могут быть вычислены не более, чем за квадратичное время.

Смешанные признаки

Смешанными признаками назовем любой способ численно описать граф.

Сложные признаки

Сложными статистиками назовем вектор $\vec{d}$, такой что каждая компонента вектора $\vec{d}$ независима от компонент вектора простых статистик $\vec{s}$ и при этом вектор смешанных статистик m выражается через $\vec{d}$ и $\vec{s}$ линейно.

Простые статистики

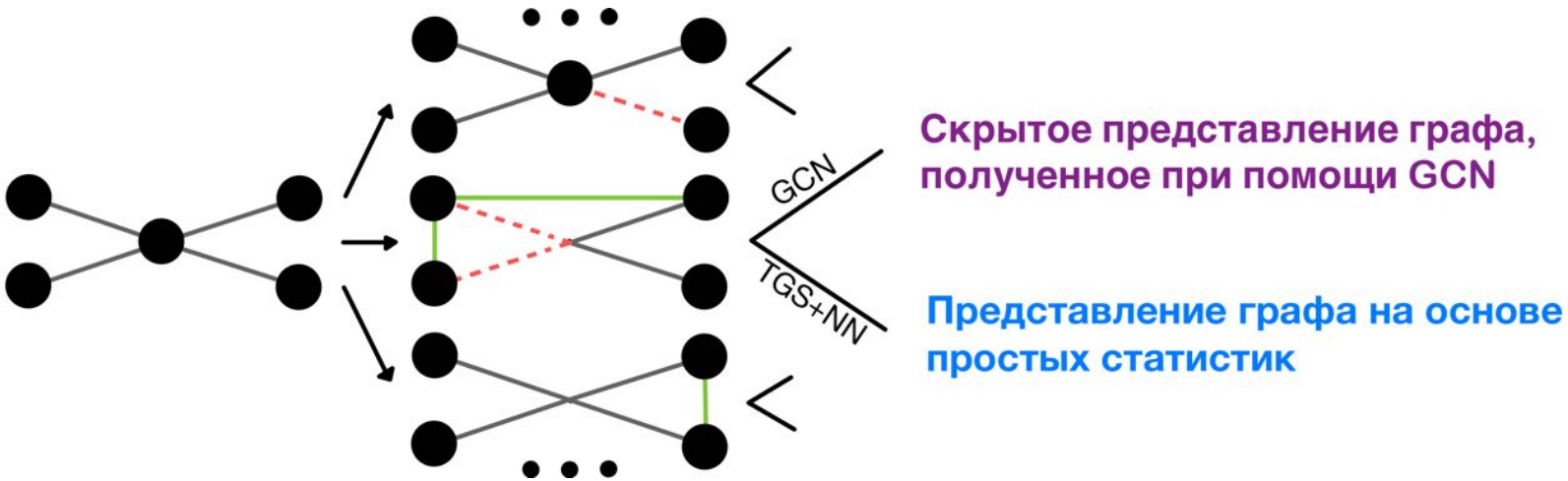
- Размерные показатели $[O(1)]$:
  - Число ребер
  - Число вершин
- Вершины специального вида $[O(V)]$:
  - Изолированные вершины – вершины без единого ребра
  - Висячие вершины – вершины с одним ребром
  - Промежуточные вершины – вершины с двумя ребрами
  - Вершины, связанные с каждой вершиной графа
- Статистики на степенях вершин $[O(V)]$:
  - Максимальная степень вершины
  - Средняя степень вершины
  - Медианная степень вершины
  - Модальная степень вершины
  - Стандартное отклонение степеней вершин в графе
- Гистограмма степеней вершин графа $[O(V)]$ (здесь μ – средняя степень вершин в графе, σ – среднеквадратичное отклонение степеней вершин в графе): Доля вершин со степенью на интервалах: $(\mu - \sigma, \mu)$, $(\mu, \mu + \sigma)$, $(\mu - 2\sigma, \mu - \sigma)$, $(\mu + \sigma, \mu + 2\sigma)$, $(\mu - 3\sigma, \mu - 2\sigma)$, $(\mu + 2\sigma, \mu + 3\sigma)$
- Коэффициент кластеризации $[O(V^2)]$ [1]

Построения распределения графа

Дан граф G и соответствующая ему матрица смежности A.

Необходимо получить распределение этого графа

Для этого n раз удалим либо добавим случайное число ребер в граф, получив тем самым набор из n графов, близких к исходному.



По итогу операции получения распределения графа

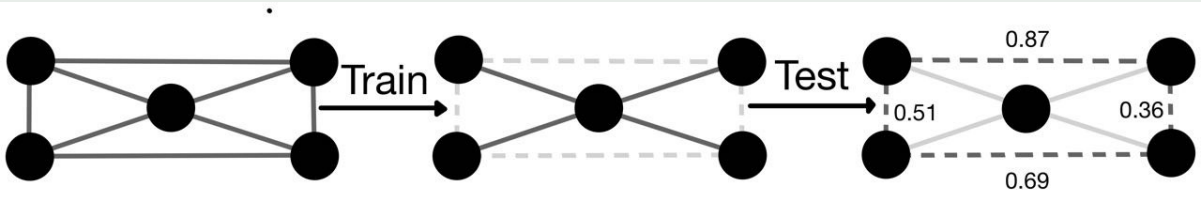
Было получено n графов, близких к исходному и для каждого из них:

- Вектор представления графа, полученный при помощи *графовых сверток*[2];
- Вектор представления графа, полученных на основе *простых статистик*.

Постановка задачи

Реконструкция

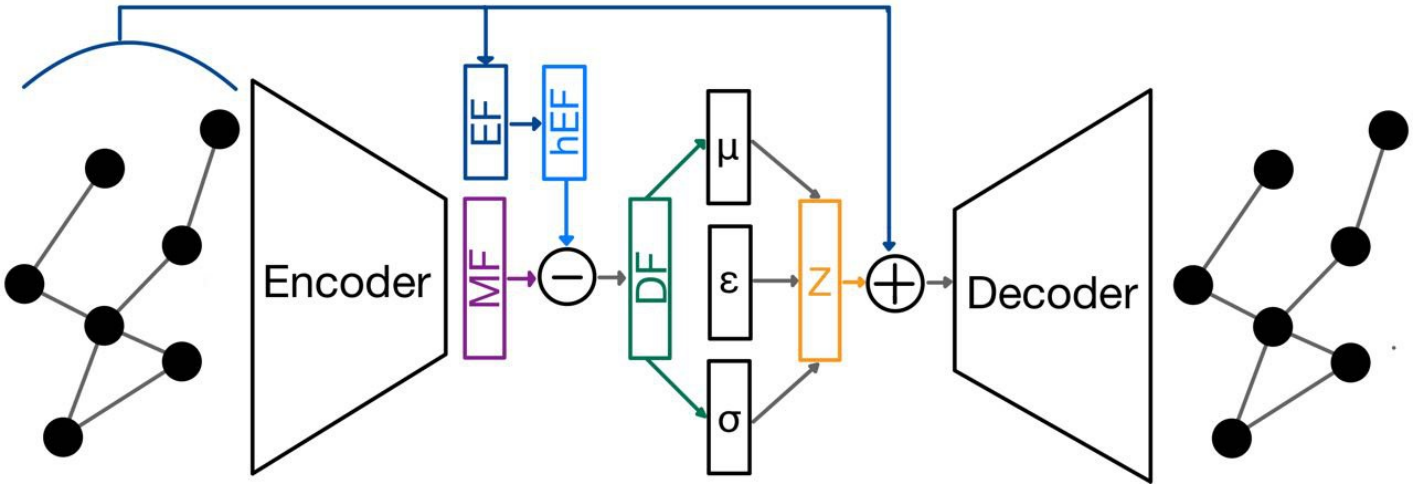
Дано:
Граф G с матрицей смежности $A \in \mathcal{R}^{n \times n}$, где $A_{ij} = 1$, если ребро (i, j) существует в графе, и 0 в противном случае. Матрица признаков вершин $V_f \in \mathcal{R}^{n \times k}$, а также набор скрытых ребер $E = \{(i, j)\}$.
Задача:
Построить модель, предсказывающую наличие ребра в графе на основе признаков вершин и существующих ребер.



Генерация графов

Дано:
Множество $\{A_i\}_{i=0}^{N} \in \mathcal{R}^{n \times n}$ матриц смежности графов G_i из неизвестного распределения $\pi(G)$, построенного на основе графа G.
Задача:
Получить распределение $\pi(G)$ для того, чтобы оценить $\pi(\hat{G})$ для нового графа $\hat{G}$, а также для генерации новых графов из распределения $\pi(G)$.

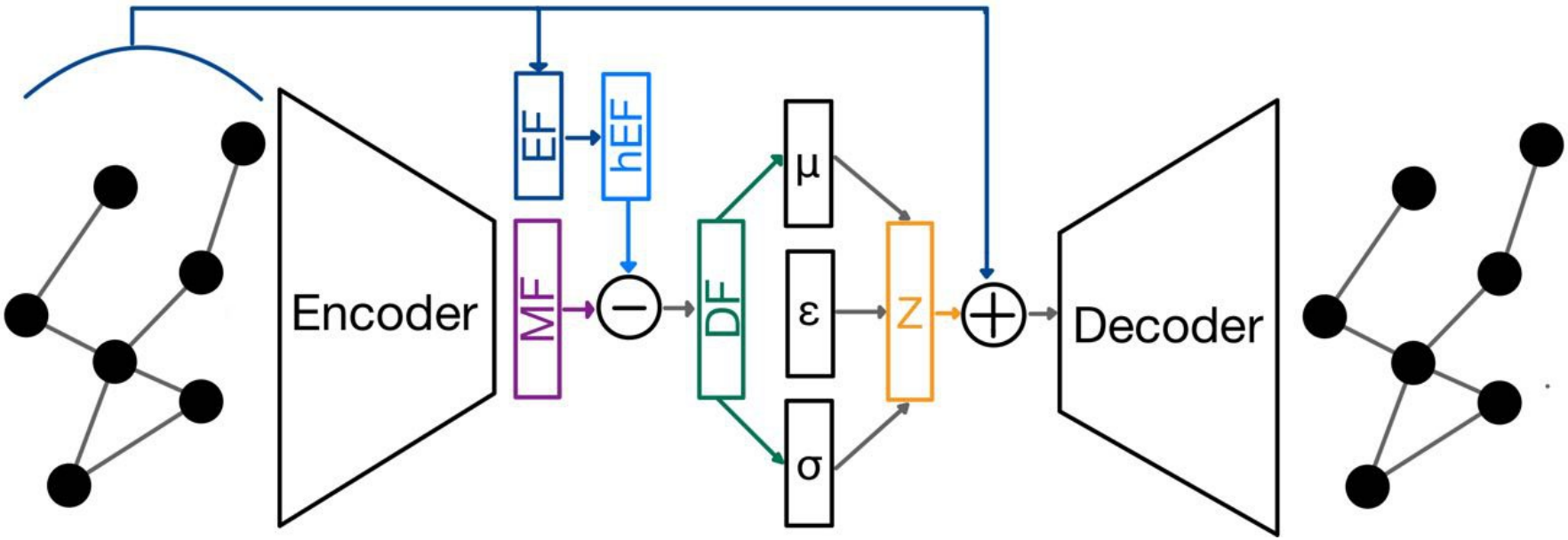
Предложенный метод: ControlVAE



Здесь:

- G и $\hat{G}$ – оригинальный и сгенерированный графы соответственно (A, $\hat{A}$ их матрицы смежности);
- MF – матрица смешанных статистик графа;
- EF – вектор простых статистик;
- hEF – скрытое представление вектора простых статистик;
- DF – матрица сложных статистик;
- ε – случайная величина $\in N(0, 1)$;
- Z – матрица из распределения $N(\mu, \sigma)$, полученная при помощи трюка репараметризации[3].

ControlVAE



Функция потерь метода ControlVAE

$$\text{LOSS}_{CtrlVAE} = \text{BCELoss}(A, \hat{A}) \text{ (реконструкция графа)}$$

$$+\text{KL-div}(DF, \varepsilon) \text{ (нормальность скрытого представления)}$$

$$+\text{MSELoss}(hEF, MF) \text{ (приближение смешанных простыми)}$$

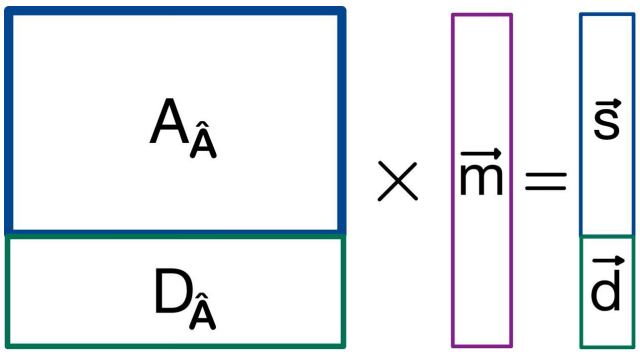
Существование разбиения на простые и сложные статистики

Гипотеза

Существует линейное отображение вектора смешанных статистик в вектор простых статистик.

Теорема (Бишук А.Ю. 2023): О разделении признаков графа

Пусть существует линейное отображение $A_{\hat{A}}$, переводящее вектор смешанных статистик $\vec{m}$ в вектор простых статистик $\vec{s}$. Тогда существует невырожденное преобразование $\hat{A}$, которое отображает вектор $\vec{m}$ в вектор $\hat{\vec{s}}$ таким образом, что подвектор из первых $|s|$ компонент совпадает с вектором $\vec{s}$, а оставшиеся $|m| - |s|$ компонент образуют вектор сложных статистик $\vec{d}$.



Построенное линейное преобразование смешанных статистик.

Идея доказательства теоремы о разделении признаков графа

Замечание

Независимость элементов вектора простых статистик мы можем гарантировать по построению.

Лемма 1

Пусть дан набор независимых, одинаково распределенных случайных величин $p_1, p_2, \ldots p_n$. *Случайная величина* $\xi = a_1 p_1 + a_2 p_2 + \ldots + a_n p_n$ *статистически зависима от каждой из случайных величин* p_i, *коэффициент перед которой* $a_i \neq 0$.

Лемма 2

Матрица $A_{\hat{A}}$ *имеет максимально возможный ранг.*

Построим ортогональное дополнение к $A_{\hat{A}}$. В силу Леммы 1 новая полученная часть $\vec{d}$ вектора $\vec{s}$ будет статистически независима от каждой компоненты $\vec{s}$. А силу теоремы о линейном преобразовании нормального вектора[4], она будет также нормальным вектором.

Таким образом, получившейся вектор $\vec{d}$ будет вектором сложных статистик.

Результаты вычислительного эксперимента

Наборы данных, используемые в экспериментах:

- Cora[5] (2708 статей и 10556 ссылки между ними);
- Citeseer[6] (3327 статьи и 9228 ссылки между ними).

Метрики:

	Dataset	ROC-AUC	AP	MAE (GT statistics)
VAE	Cora	75.18 %	**75.81%**	0.066
Our method		**76.68 %**	75.18 %	**0.046 (-30%)**
VAE	Citeseer	**82.09 %**	**79.94 %**	0.072
Our method		76.28 %	76.44 %	**0.060 (-17%)**

Метод ContolVAE генерирует графы с, более близкими к заданным, простыми статистиками (на 30% точнее для графа Cora и на 17% для Citeseer). При это слабо качество реконструкции падает незначительно – на 0.63% для графа Cora и на 3.5% для графа Citeseer)

Выносится на защиту

1. Предложен новый подход к генерации графов. ControlVAE позволяет генерировать графы с более точными статистиками, выбранными из теории графов;

2. Приведено теоретическое обоснование предложенного метода. Выдвинута гипотеза и в рамках нее сформулирована теорема о существовании линейного преобразования, разделяющего простых и сложные статистики;

3. Реализован алгоритм на основе GraphVAE[7]. Модифицирован классический метод генерации графов, позволяющий учитывать заданные статистики графов;

4. Проведены исследования на известных датасетах (Cora и Citeseer). Продемонстрировано преимущество метода в задаче генерации графов с заданными статистиками.

Список литературы

[1] Jari Saramäki и др. «Generalizations of the clustering coefficient to weighted complex networks». В: *Physical Review E* 75.2 (2007), с. 027105.

[2] Thomas N Kipf и Max Welling. «Semi-supervised classification with graph convolutional networks». В: *arXiv preprint arXiv:1609.02907* (2016).

[3] Diederik P Kingma и Max Welling. «Auto-encoding variational bayes». В: *arXiv preprint arXiv:1312.6114* (2013).

[4] Александр Алексеевич Боровков. *Теория вероятностей*. URSS, 2009.

[5] Prithviraj Sen и др. «Collective classification in network data». В: *AI magazine* 29.3 (2008), с. 93—93.

[6] Ryan Rossi и Nesreen Ahmed. «The Network Data Repository with Interactive Graph Analytics and Visualization». В: *AAAI Conference on Artificial Intelligence*. Т. 29. New York, NY, USA, 2015, с. 4292—4293.

[7] Thomas N Kipf и Max Welling. «Variational graph auto-encoders». В: *arXiv preprint arXiv:1611.07308* (2016).