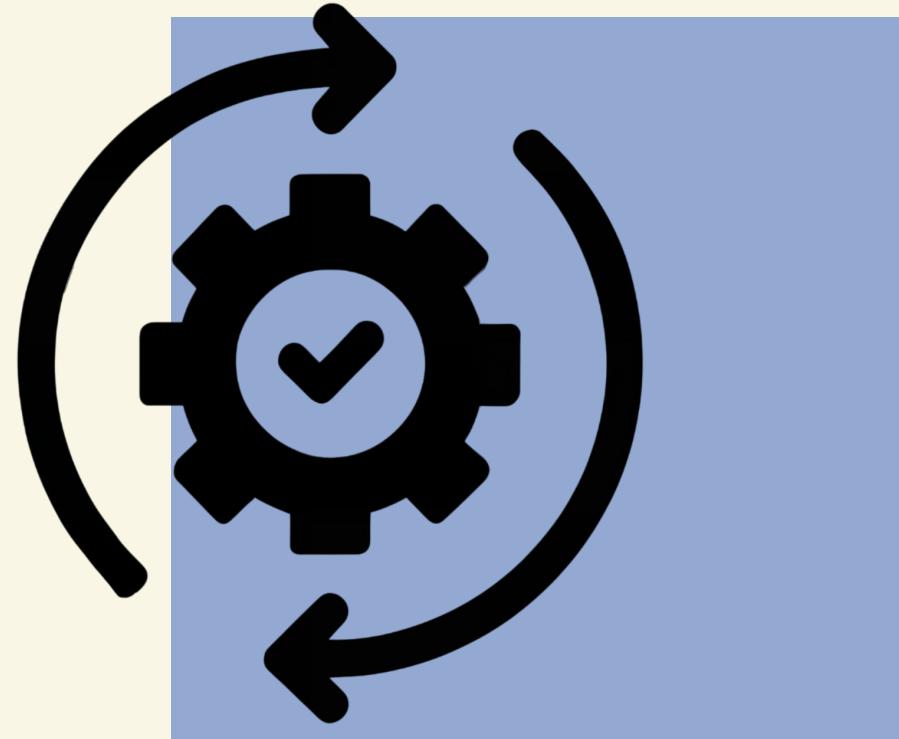


Эффективные системы машинного обучения



Лекция 7

“Dealing with graphs”

Преподаватель

Феоктистов Дмитрий

8 ноября 2024

ВМК МГУ

Графы



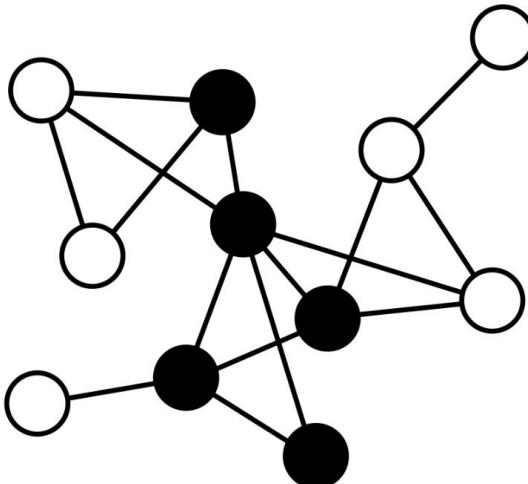
- 1 Задачи на графах
- 2 Графовые эмбеды
- 3 GNNs
- 4 О наболевшем
- 5 Эффективность в графах
- 6 Примеры успехов

На основе материалов Людмилы Прохоренковой,
Yandex Research

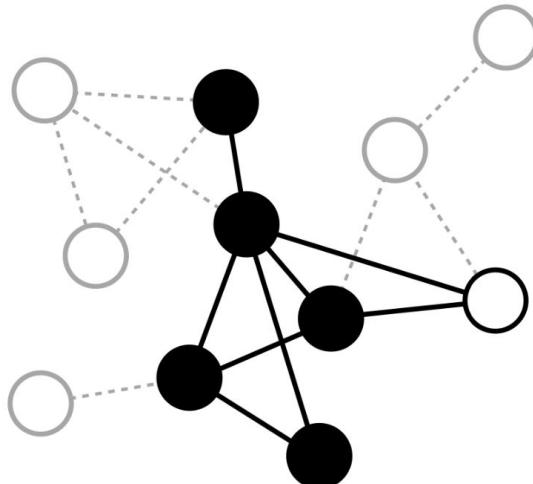
Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

Классификация вершин



Transductive

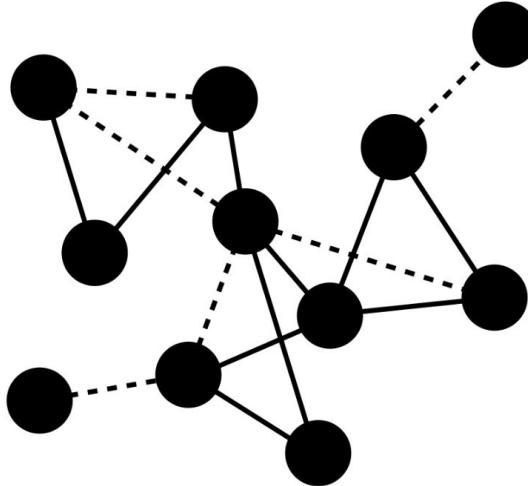


Inductive

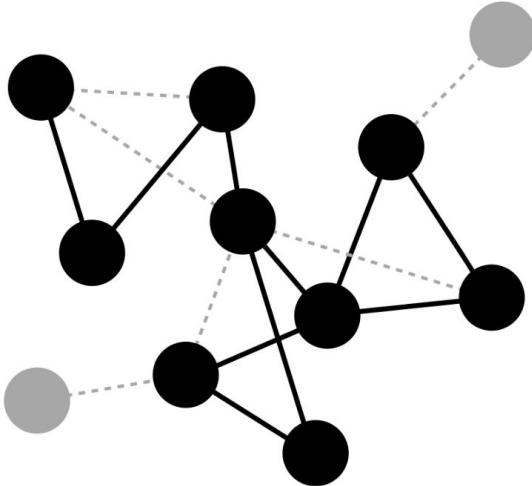
Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

Классификация ребер



Transductive

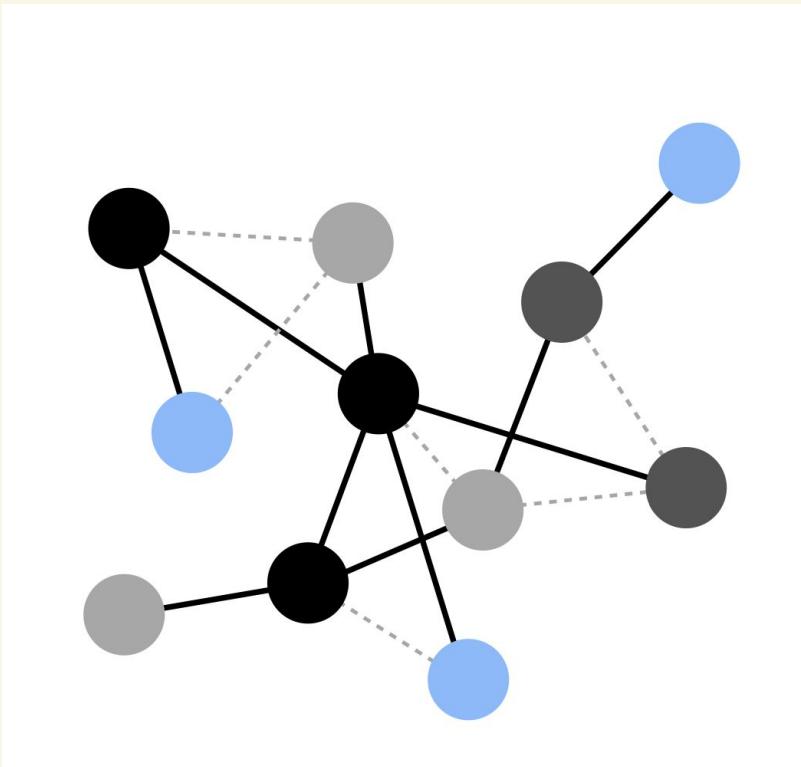


Inductive

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

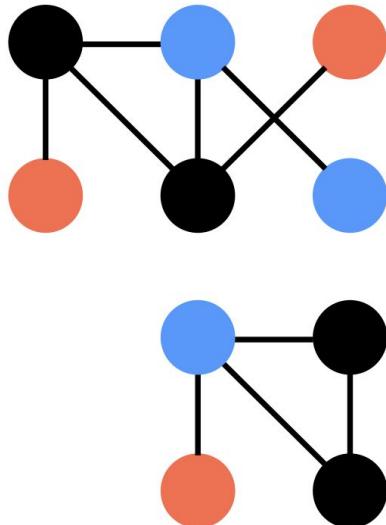
Предсказание ребер



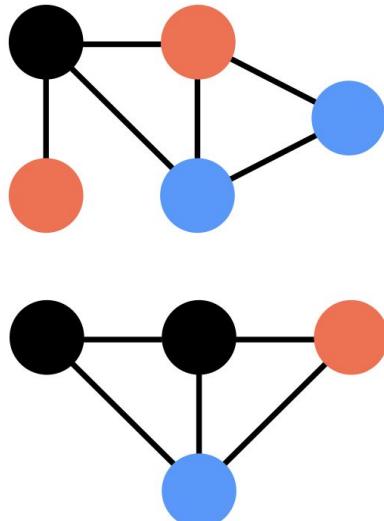
Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

Классификация графов



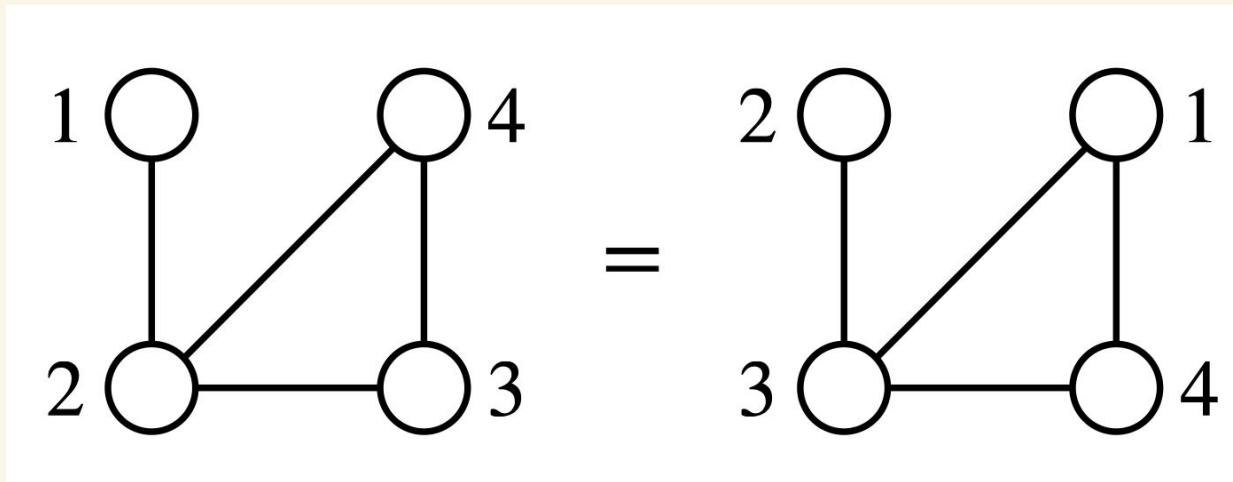
Train graphs



Test graphs

В чем вообще сложность?

- У вершин нет порядка и это нужно учитывать
- Фичи могут быть у вершин, ребер и у самих графов
- У графов разный масштаб и структура
- И они могут даже меняться со временем



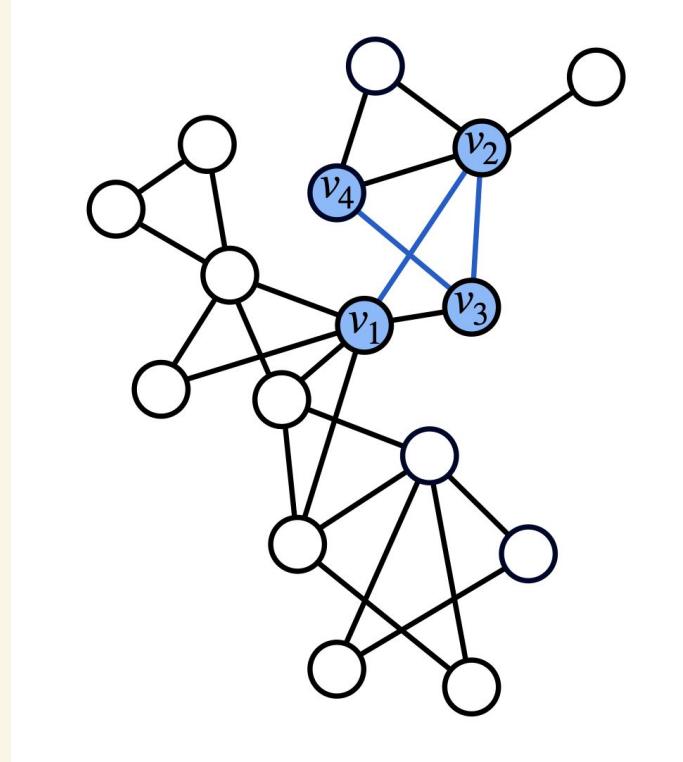
Графовые фичи

- Степени вершин
- Среднее расстояние до других вершин
- Фичи соседей
-

$$c_v = \sum_{s \neq v \neq e} \frac{|\{P \in \text{SP}(s, e) : v \in P\}|}{|\text{SP}(s, e)|}$$

DeepWalk

Рассмотрим случайное блуждание, стартующее из вершины v_1



DeepWalk

Рассмотрим случайное блуждание, стартующее из вершины v_1 . И давайте максимизировать следующее правдоподобие, где $f(v)$ – эмбеддинг

$$\log P(\{v_{i-w}, \dots, v_{i-1}, v_{i+1}, \dots, v_{i+w}\} | f(v_i))$$

DeepWalk

Рассмотрим случайное блуждание, стартующее из вершины v_1 . И давайте максимизировать следующее правдоподобие, где $f(v)$ – эмбеддинг

$$\log P(\{v_{i-w}, \dots, v_{i-1}, v_{i+1}, \dots, v_{i+w}\} | f(v_i))$$

И предположим независимость v_j при условии $f(v_i)$

DeepWalk

Тогда получим следующий алгоритм

- Generate random walks of fixed length starting from each node $u \in V(G)$
- For each $v_i \in RW(u)$ let
$$N_{RW}(v_i) = \{v_{i-w}, \dots, v_{i-1}, v_{i+1}, \dots, v_{i+w}\}$$
- Objective:

$$\mathcal{L} = \sum_{RW} \sum_{v_i \in RW} \sum_{v_j \in N_{RW}(v_i)} -\log P(v_j | f(v_i))$$

$$\text{where } P(v_j | f(v_i)) = \frac{\exp(f(v_i)^T f(v_j))}{\sum_{u \in V(G)} \exp(f(v_i)^T f(u))}$$

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

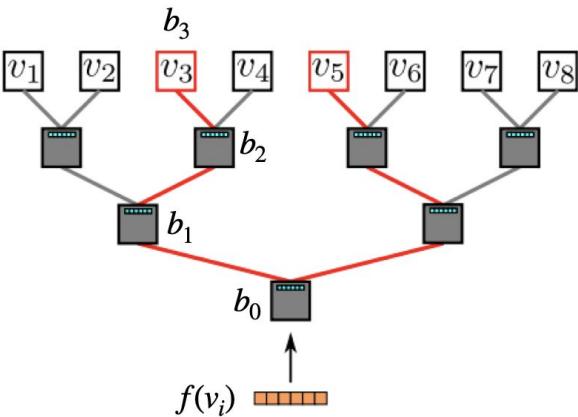
DeepWalk

$$P(v_j | f(v_i)) = \prod_{l=1}^{\lceil \log(n) \rceil} P(b_l | f(v_i))$$

$$P(b_l | f(v_i)) = \frac{1}{1 + \exp(-f(v_i) \cdot g(b_l))}$$

$g(b_l)$ — representation of b_l

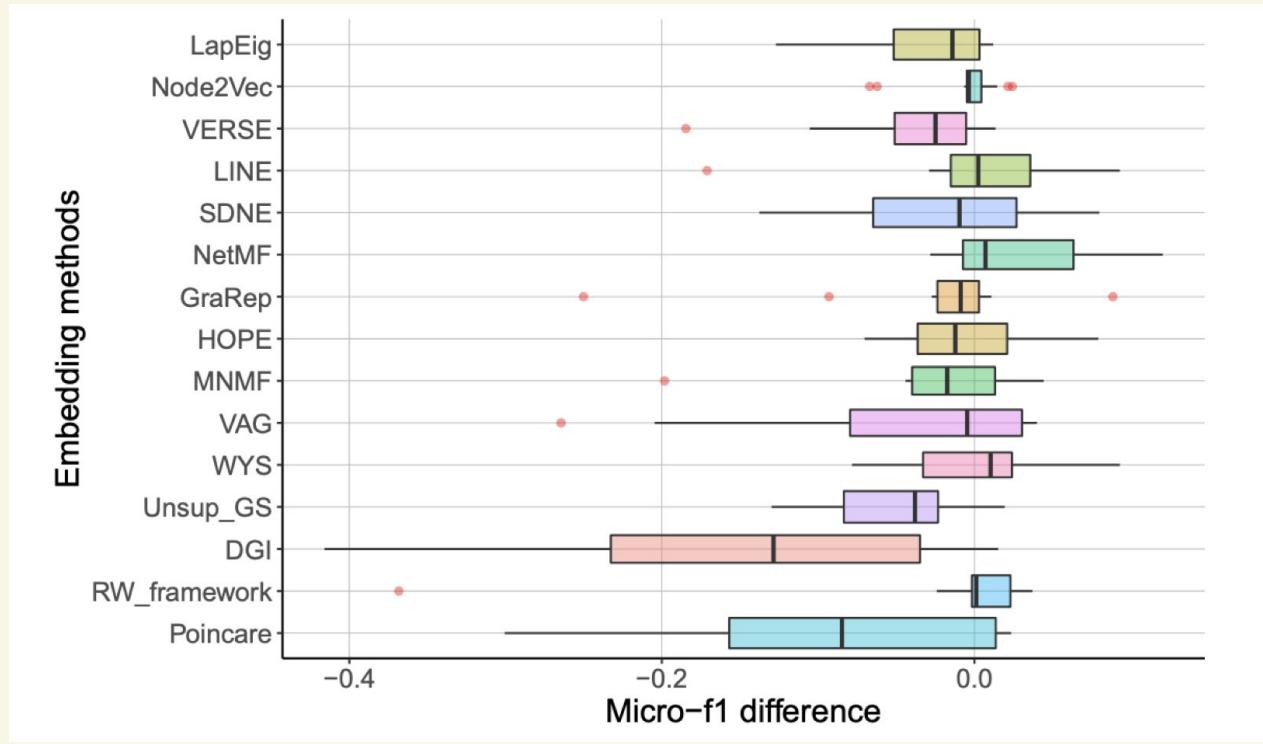
Complexity becomes $O(\log(n))$ instead of $O(n)$



Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

DeepWalk



Промежуточные итоги

- Взять фичи вершин (уже неплохо, если честно)
- Добавить к ним графовые фичи
- Добавить к ним графовые эмбеды
- И уже будет хорошо

	tolokers-tab	questions-tab	city-reviews	browser-games	hm-categories	web-fraud
ResNet	45.17 ± 0.61	84.01 ± 0.26	64.33 ± 0.32	78.82 ± 0.32	70.45 ± 0.24	14.21 ± 0.24
XGBoost	48.79 ± 0.25	85.03 ± 5.78	65.55 ± 0.18	79.73 ± 0.17	71.08 ± 0.70	16.95 ± 0.18
LightGBM	48.49 ± 0.27	87.24 ± 0.14	66.17 ± 0.17	79.28 ± 0.15	71.09 ± 0.10	17.08 ± 0.11
CatBoost	48.61 ± 0.25	87.59 ± 0.04	66.05 ± 0.16	80.46 ± 0.22	71.14 ± 0.12	TLE
MLP-PLR	47.72 ± 0.45	87.34 ± 0.42	66.36 ± 0.11	80.69 ± 0.24	71.02 ± 0.08	16.24 ± 0.12
TabR-PLR	48.50 ± 0.69	85.56 ± 0.52	66.50 ± 0.26	80.29 ± 0.26	71.38 ± 0.22	MLE
LightGBM-NFA	57.99 ± 0.43	87.79 ± 0.19	71.66 ± 0.11	83.09 ± 0.26	81.72 ± 0.12	23.72 ± 0.16
MLP-PLR-NFA	57.70 ± 0.20	87.43 ± 0.07	71.93 ± 0.12	83.36 ± 0.26	81.35 ± 0.21	22.33 ± 0.29

Промежуточные итоги

- Эмбеды нельзя применить к новый вершинам (т.е. только трансдуктивный подход)
- Нет шаринга параметров между вершинами => дорого
- Эмбеды не знают ничего о фичах вершин, т.е. графовая часть несколько независима от не графовой

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

GNNs

Input graph



Node features

- Node structural characteristics
- Node embeddings

$$G = (V, E)$$

ML algorithm

- SVM
- GBDT
- Neural Network

Prediction

- Node classification
- Node regression
- Link prediction

Input graph

$$G = (V, X, E)$$

GNN model

- GCN
- GraphSAGE
- GAT
- GIN

Prediction

- Node-level
- Graph-level
- Link-level

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

GNNs

Пусть у вершины есть признаки x , а также вектор соседей a . Построим функцию предиктор, как $y = F(x, a)$. Какие проблемы могут быть у такой функции?

a_1	0	1	1	0	0
a_2	1	0	1	1	0
a_3	1	1	0	0	1
a_4	0	1	0	0	1
a_5	0	0	1	1	0

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

GNNs

Пусть у вершины есть признаки x , а также вектор соседей a . Построим функцию предиктор, как $y = F(x, a)$. Какие проблемы могут быть у такой функции?

a_1	0	1	1	0	0
a_2	1	0	1	1	0
a_3	1	1	0	0	1
a_4	0	1	0	0	1
a_5	0	0	1	1	0

GNNs

Пусть у вершины есть признаки x , а также вектор соседей a . Построим функцию предиктор, как $y = F(x, a)$. Какие проблемы могут быть у такой функции?

- Ответ может зависеть от порядка соседей
- Не факт, что выйдет использовать в индуктивном сэтапе

a_1	0	1	1	0	0
a_2	1	0	1	1	0
a_3	1	1	0	0	1
a_4	0	1	0	0	1
a_5	0	0	1	1	0

GNNs

Пусть P матрица перестановки, тогда функция F называется инвариантной к перестановке, если $F(PX) = F(X)$; Эквивариантной, если $F(PX) = P(FX)$. Как эти свойства обобщаются на функцию GNN?

a_1	0	1	1	0	0
a_2	1	0	1	1	0
a_3	1	1	0	0	1
a_4	0	1	0	0	1
a_5	0	0	1	1	0

GNNs

Пусть P матрица перестановки, тогда функция F называется инвариантной к перестановке, если $F(PX) = F(X)$; Эквивариантной, если $F(PX) = PF(X)$. Как эти свойства обобщаются на функцию GNN?

- $F(X, A) = F(PX, PAP.T)$
- $PF(X, A) = F(PX, PAP.T)$

В каких задачах какое из свойств нужно?

a_1	0	1	1	0	0
a_2	1	0	1	1	0
a_3	1	1	0	0	1
a_4	0	1	0	0	1
a_5	0	0	1	1	0

GNNs

Можно показать, что любую инвариантную к перестановке функцию можно задать в виде

$$F(X) = G\{\stackrel{+}{\oplus} H(X_i)\})$$

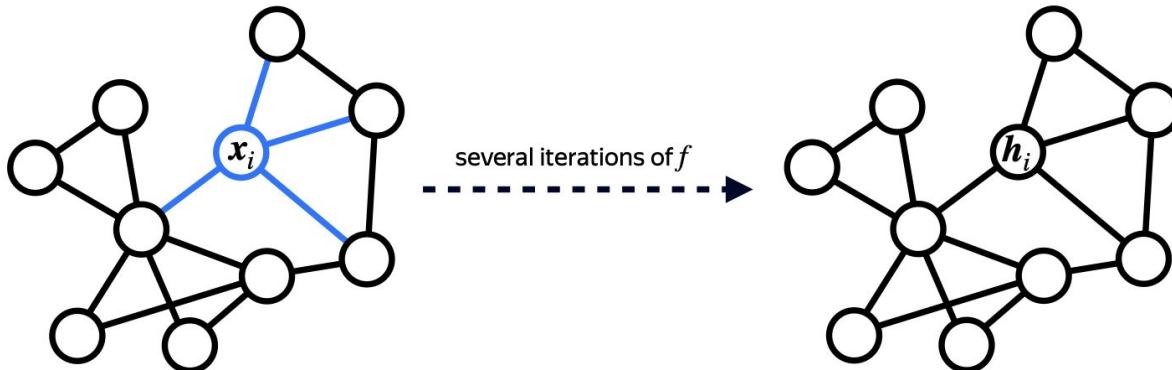
Где $\stackrel{+}{\oplus}$ – операция взятия суммы; Но по факту нас устраивает усреднение или max/min

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

MPNNs

$$\mathbf{h}_i = f(\mathbf{X}, \mathbf{A})_i = \phi\left(\mathbf{x}_i, \text{AGG}(\{\mathbf{X}_{N_i}\})\right)$$

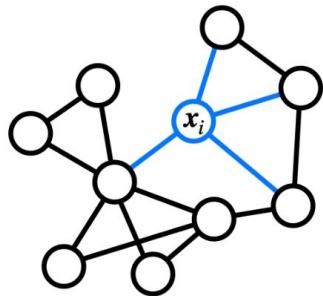


Эффективные системы машинного обучения, ВМК МГУ

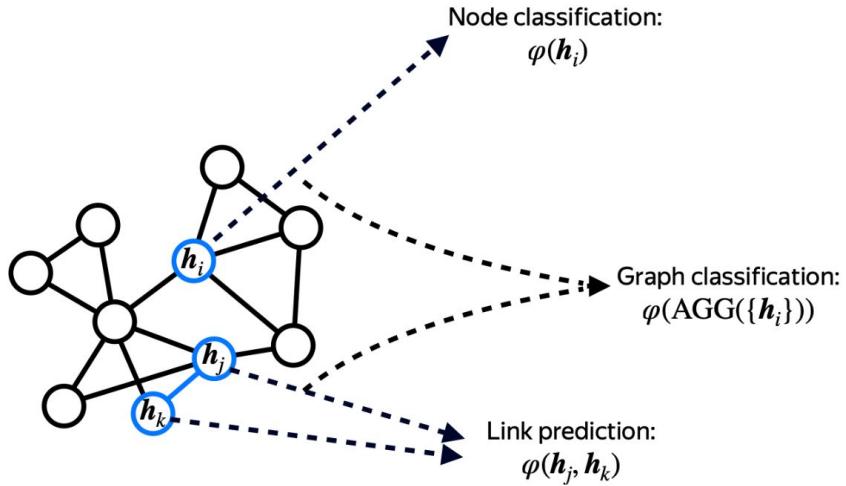
Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

GNNs

$$\mathbf{h}_i = f(X, A)_i = \phi\left(\mathbf{x}_i, \text{AGG}(\{X_{N_i}\})\right)$$



several iterations of f



GCN

$$\mathbf{h}_i = \phi \left(\mathbf{x}_i, \bigoplus_{j \in N(i)} c_{ij} \psi(\mathbf{x}_j) \right)$$

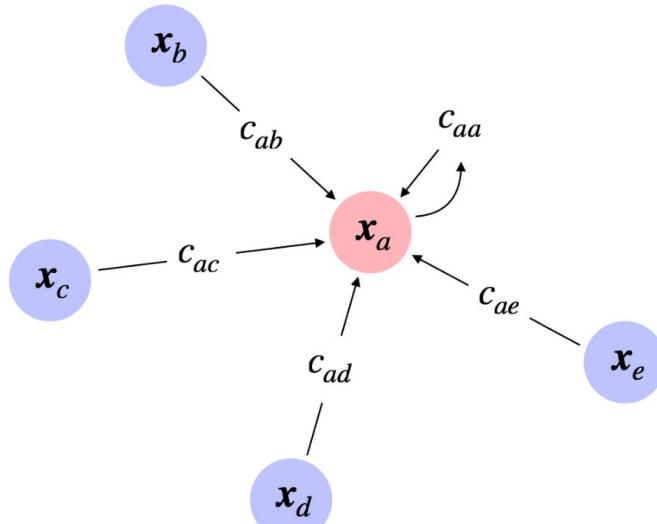
$$\mathbf{h}_i = \sigma \left(\sum_{j \in N_i} \frac{W\mathbf{x}_j}{\sqrt{d_i d_j}} \right)$$

σ — activation function

W — learnable parameters

Matrix form:

$$H = \sigma(D^{-1/2} A D^{-1/2} W X)$$



GAT

$$\mathbf{h}_i = \phi \left(\mathbf{x}_i, \bigoplus_{j \in N(i)} a(\mathbf{x}_i, \mathbf{x}_j) \psi(\mathbf{x}_j) \right)$$

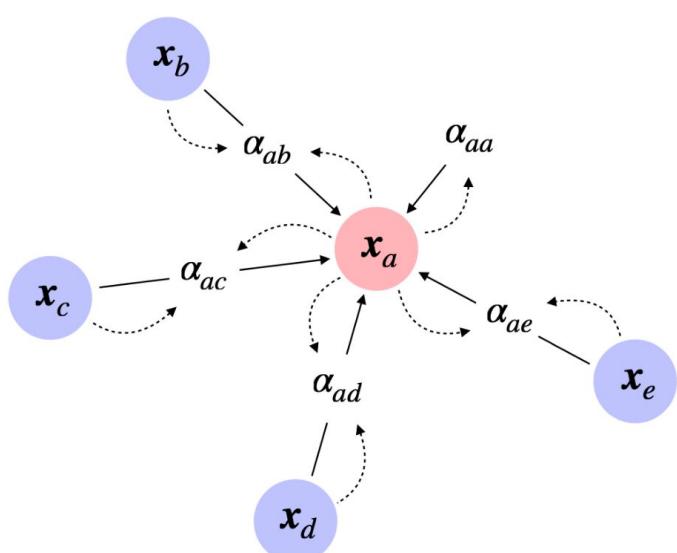
$$\mathbf{h}_i = \sigma \left(\sum_{j \in N_i} \alpha_{ij} W \mathbf{x}_j \right)$$

$$\alpha_{ij} = \text{softmax}(m_{ij})$$

$$m_{ij} = \text{LeakyReLU} \left(\mathbf{a}^T \text{Concat} \left(W \mathbf{x}_i, W \mathbf{x}_j \right) \right)$$

σ — activation function

W and \mathbf{a} — learnable parameters



Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

GNN: результаты

(a) Results for classification datasets. Accuracy is reported for multiclass classification datasets (**browser-games** and **hm-categories**), Average Precision (AP) is reported for binary classification datasets (all other datasets).

	tolokers-tab	questions-tab	city-reviews	browser-games	hm-categories	web-fraud
ResNet	45.17 ± 0.61	84.01 ± 0.26	64.33 ± 0.32	78.82 ± 0.32	70.45 ± 0.24	14.21 ± 0.24
XGBoost	48.79 ± 0.25	85.03 ± 5.78	65.55 ± 0.18	79.73 ± 0.17	71.08 ± 0.70	16.95 ± 0.18
LightGBM	48.49 ± 0.27	87.24 ± 0.14	66.17 ± 0.17	79.28 ± 0.15	71.09 ± 0.10	17.08 ± 0.11
CatBoost	48.61 ± 0.25	87.59 ± 0.04	66.05 ± 0.16	80.46 ± 0.22	71.14 ± 0.12	TLE
MLP-PLR	47.72 ± 0.45	87.34 ± 0.42	66.36 ± 0.11	80.69 ± 0.24	71.02 ± 0.08	16.24 ± 0.12
TabR-PLR	48.50 ± 0.69	85.56 ± 0.52	66.50 ± 0.26	80.29 ± 0.26	71.38 ± 0.22	MLE
LightGBM-NFA	57.99 ± 0.43	87.79 ± 0.19	71.66 ± 0.11	83.09 ± 0.26	81.72 ± 0.12	23.72 ± 0.16
MLP-PLR-NFA	57.70 ± 0.20	87.43 ± 0.07	71.93 ± 0.12	83.36 ± 0.26	81.35 ± 0.21	22.33 ± 0.29
GCN	61.09 ± 0.38	84.92 ± 0.95	71.08 ± 0.32	79.17 ± 0.41	86.42 ± 0.31	14.65 ± 0.24
GraphSAGE	57.08 ± 0.24	85.70 ± 0.30	71.15 ± 0.27	82.56 ± 0.11	86.35 ± 0.18	20.28 ± 0.48
GAT	58.77 ± 1.00	84.44 ± 0.68	71.38 ± 0.53	82.60 ± 0.26	87.84 ± 0.23	19.95 ± 0.51
GT	58.92 ± 0.57	83.59 ± 1.17	71.72 ± 0.23	83.29 ± 0.33	89.00 ± 0.23	20.19 ± 0.44
GCN-PLR	60.81 ± 0.56	88.80 ± 0.25	70.40 ± 0.58	80.50 ± 0.58	83.85 ± 0.28	MLE
GraphSAGE-PLR	60.28 ± 0.97	88.55 ± 0.48	72.26 ± 0.40	83.19 ± 0.34	86.77 ± 0.12	MLE
GAT-PLR	60.99 ± 0.82	88.69 ± 0.63	71.66 ± 0.66	83.59 ± 0.33	87.94 ± 0.20	MLE
GT-PLR	61.95 ± 0.73	82.41 ± 1.60	71.80 ± 0.21	83.26 ± 0.36	89.01 ± 0.15	MLE
BGNN	47.45 ± 1.29	47.20 ± 5.24	51.59 ± 3.42	75.92 ± 0.34	84.60 ± 0.50	3.21 ± 0.15
EBBS	43.86 ± 1.63	79.03 ± 3.57	57.40 ± 1.99	64.56 ± 0.16	41.77 ± 1.89	6.00 ± 0.68

Expressive power

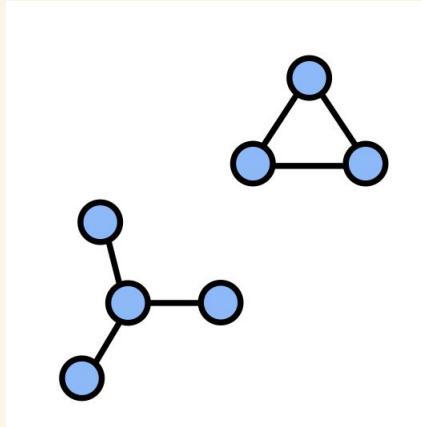
Существует WL-тест (Weisfeiler–Lehman). Его суть – итеративная процедура, которая сопоставляет каждому графу гистограмму. Известно, что у изоморфных графов эта гистограмма одинакова всегда, а у неизоморфных она может быть одинакова.

Можно показать, что MPNN эквивалентны по экспрессивности WL тесту

Expressive power

Можно показать следующие два факта:

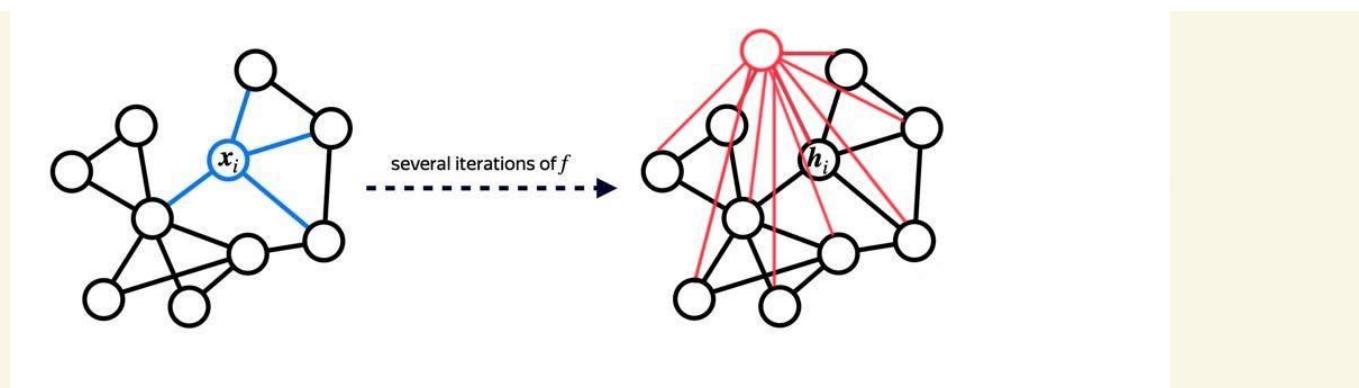
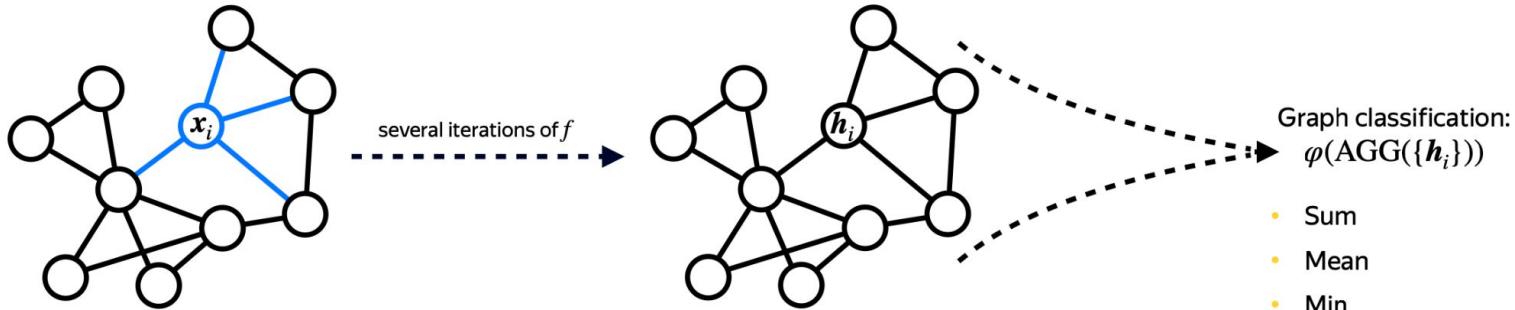
- 1) WL и MPNN не могут считать подструктуры, содержащие 3 и более вершин
- 2) WL и MPNN не могут считать подструктуры в форме звезды



Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

GNNs for graph classification



Link prediction

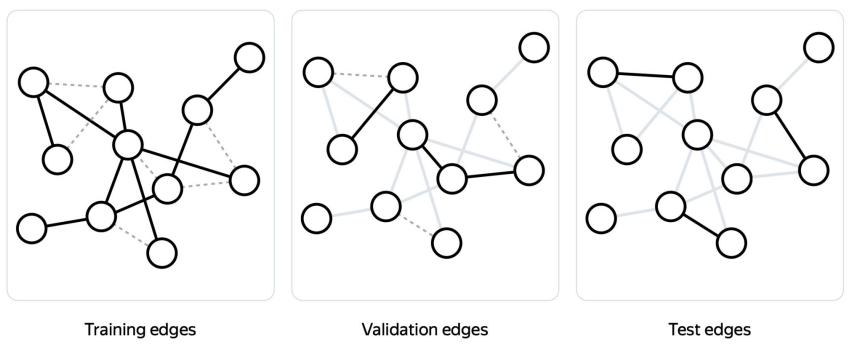
Очень похоже на рекурсис (и часто применяется в нем)

- 1) Для метрик нужны позитивные и негативные ребра.
Негативные равномерно сэмплируются
- 2) Метрики: F1, ROC, MRR, Hits@k

We are given M positive edges and N negative edges

For each positive edge e , compute $r(e)$ — rank of e among the negative edges

$$\text{Hits}@k = \frac{1}{M} \sum_e \mathbf{1}\{r(e) \leq k\}$$



Link prediction: similarity

Очень похоже на рекурсис:

- 1) Количество общих соседей
- 2) Мера Жаккара для соседей
- 3) Адамик-Адар

$$s(u, v) = \sum_{w \in |N(u) \cap N(v)|} \frac{1}{\log d_w}$$

Link prediction: similarity

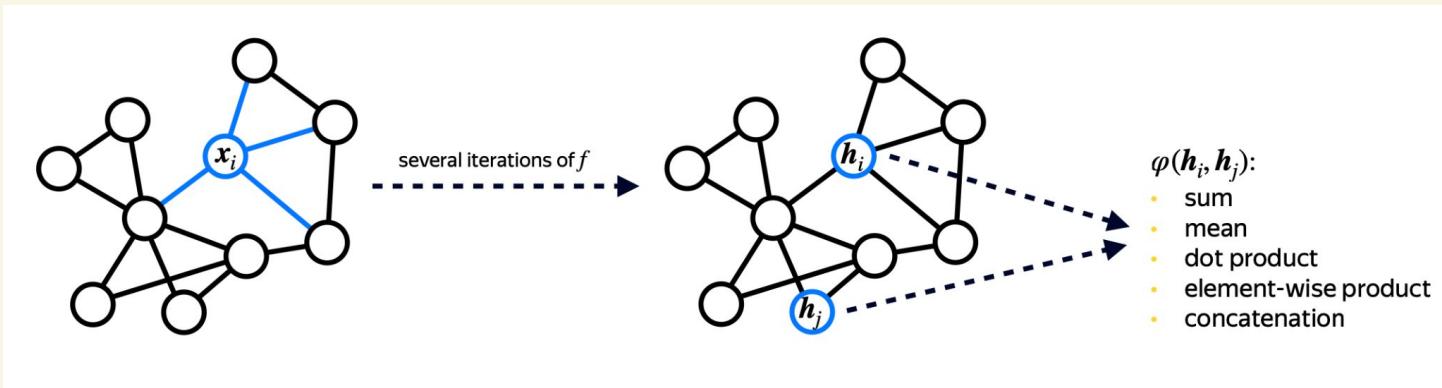
Очень похоже на рекурсис:

- 1) Скалярное произведение эмбедов
- 2) Также можно учить классификатор на поэлементном
произведении/сумме/etc эмбедов
- 3) Матричные разложения :)

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

Link prediction: GNN



Link prediction: GNN

В этом типе задач становится сильно более
существенно, что MPNN не могут различать подструктуры
и ничего не знают о позиции, а это может многое говорить
о вершинах, для которых мы хотим провести ребро =>
давайте как-то добавим эту информацию

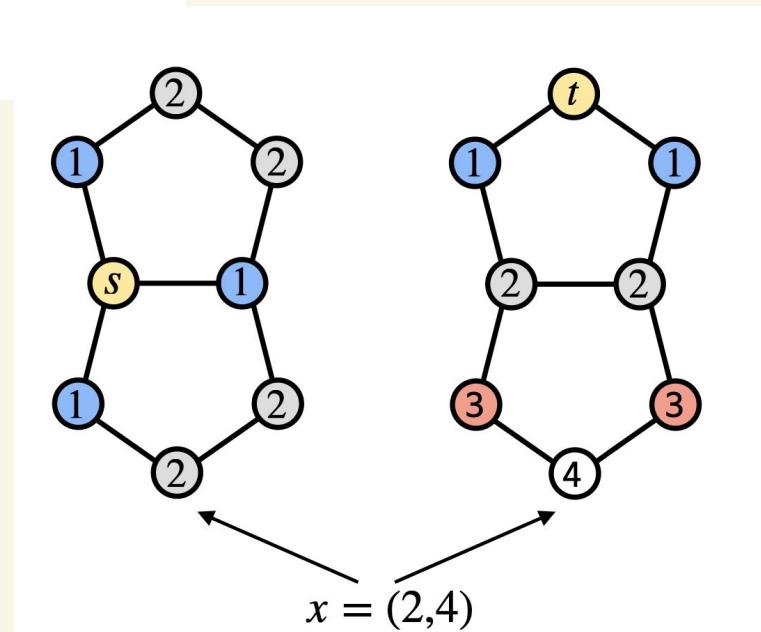
Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

Link prediction: GNN

Embedding $f(v)$ is *position-aware* if there exists a function $g(\cdot, \cdot)$ such that

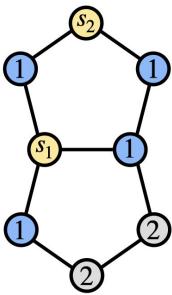
$$d_{\text{SP}}(u, v) = g(f(u), f(v))$$



Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

Link prediction: GNN



$$f(v) = \left(\frac{d(v, S_{1,1})}{k}, \frac{d(v, S_{1,2})}{k}, \dots, \frac{d(v, S_{\log n, c \log n})}{k} \right) \in \mathbb{R}^{c \log^2 n}$$

$S_{i,j} \in V$ includes each node in V independently with probability $\frac{1}{2^i}$

Link prediction: GNN

Как использовать такие эмбеды?

- 1) Добавим к вершинам => не будет инвариантности, но зато хорошо работает на практике
- 2) Но можно модифицировать архитектуры так, что инвариантность вернется
- 3) Как использовать для индуктивного сетапа? Во время обучения ресэмплировать якорные вершины, во время теста сэмплировать новые

GNN: выдуманные проблемы?

Besides, the repeated aggregation of local messages can either make node representations indistinguishable—a process known as ‘oversmoothing’ [40, 41]—or limit the models’ ability to capture long-range dependencies—a phenomenon called ‘oversquashing’ [1].

Relating the rich literature on RNN behavior and graph topology, we theoretically show and experimentally verify that RUM attenuates the aforementioned symptoms and is more expressive than the Weisfeiler–Lehman (WL) isomorphism test.

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

GNN: лучшая статья 2024

Table 2: Node classification results over homophilous graphs (%). * indicates our implementation, while other results are taken from [12, 76]. The top 1st, 2nd and 3rd results are highlighted.

	Cora	CiteSeer	PubMed	Computer	Photo	CS	Physics	WikiCS								
# nodes	2,708	3,327	19,717	13,752	7,650	18,333	34,493	11,701								
# edges	5,278	4,732	44,324	245,861	119,081	81,894	247,962	216,123								
Metric	Accuracy↑															
GraphGPS	82.84 ± 1.03	72.73 ± 1.23	79.94 ± 0.26	91.19 ± 0.54	95.06 ± 0.13	93.93 ± 0.12	97.12 ± 0.19	78.66 ± 0.49								
GraphGPS*	83.87 ± 0.96	72.73 ± 1.23	79.94 ± 0.26	91.79 ± 0.63	94.89 ± 0.14	94.04 ± 0.21	96.71 ± 0.15	78.66 ± 0.49								
NAGphormer	82.12 ± 1.18	71.47 ± 1.30	79.73 ± 0.28	91.22 ± 0.14	95.49 ± 0.11	95.75 ± 0.09	97.34 ± 0.03	77.16 ± 0.72								
NAGphormer*	80.92 ± 1.17	70.59 ± 0.89	80.14 ± 1.06	91.69 ± 0.30	96.14 ± 0.16	95.85 ± 0.16	97.35 ± 0.12	77.92 ± 0.93								
Exphormer	82.77 ± 1.38	71.63 ± 1.19	79.46 ± 0.35	91.47 ± 0.17	95.35 ± 0.22	94.93 ± 0.01	96.89 ± 0.09	78.54 ± 0.49								
Exphormer*	83.29 ± 1.36	71.85 ± 1.11	79.67 ± 0.73	91.80 ± 0.35	95.69 ± 0.39	95.92 ± 0.25	97.06 ± 0.13	79.38 ± 0.62								
GOAT	83.18 ± 1.27	71.99 ± 1.26	79.13 ± 0.38	90.96 ± 0.90	92.96 ± 1.48	94.21 ± 0.38	96.24 ± 0.24	77.00 ± 0.77								
GOAT*	83.26 ± 1.24	72.21 ± 1.29	80.06 ± 0.67	92.29 ± 0.37	94.33 ± 0.21	93.81 ± 0.19	96.47 ± 0.16	77.96 ± 0.63								
NodeFormer	82.20 ± 0.90	72.50 ± 1.10	79.90 ± 1.00	86.98 ± 0.62	93.46 ± 0.35	95.64 ± 0.22	96.45 ± 0.28	74.73 ± 0.94								
NodeFormer*	82.73 ± 0.75	72.37 ± 1.20	79.59 ± 0.92	87.29 ± 0.58	93.43 ± 0.56	95.69 ± 0.27	96.48 ± 0.34	75.13 ± 0.93								
SGFormer	84.50 ± 0.80	72.60 ± 0.20	80.30 ± 0.60	91.99 ± 0.76	95.10 ± 0.47	94.78 ± 0.20	96.60 ± 0.18	73.46 ± 0.56								
SGFormer*	84.82 ± 0.85	72.72 ± 1.15	80.60 ± 0.49	92.42 ± 0.66	95.58 ± 0.36	95.71 ± 0.24	96.75 ± 0.26	80.05 ± 0.46								
Polynormer	83.25 ± 0.93	72.31 ± 0.78	79.24 ± 0.43	93.68 ± 0.21	96.46 ± 0.26	95.53 ± 0.16	97.27 ± 0.08	80.10 ± 0.67								
Polynormer*	83.43 ± 0.89	72.19 ± 0.83	79.35 ± 0.73	93.78 ± 0.10	96.57 ± 0.23	95.42 ± 0.19	97.18 ± 0.11	80.26 ± 0.92								
GCN	81.60 ± 0.40	71.60 ± 0.40	78.80 ± 0.60	89.65 ± 0.52	92.70 ± 0.20	92.92 ± 0.12	96.18 ± 0.07	77.47 ± 0.85								
GCN*	85.10 ± 0.67	3.50↑	73.14 ± 0.67	1.54↑	81.12 ± 0.52	2.32↑	93.99 ± 0.12	4.34↑	96.10 ± 0.46	3.40↑	96.17 ± 0.06	3.25↑	97.46 ± 0.10	1.28↑	80.30 ± 0.62	2.83↑
GraphSAGE	82.68 ± 0.47	71.93 ± 0.85	79.41 ± 0.53	91.20 ± 0.29	94.59 ± 0.14	93.91 ± 0.13	96.49 ± 0.06	74.77 ± 0.95								
GraphSAGE*	83.88 ± 0.65	1.20↑	72.26 ± 0.55	0.33↑	79.72 ± 0.50	0.31↑	93.25 ± 0.14	2.05↑	96.78 ± 0.23	2.19↑	96.38 ± 0.11	2.47↑	97.19 ± 0.05	0.70↑	80.69 ± 0.31	5.92↑
GAT	83.00 ± 0.70	72.10 ± 1.10	79.00 ± 0.40	90.78 ± 0.13	93.87 ± 0.11	93.61 ± 0.14	96.17 ± 0.08	76.91 ± 0.82								
GAT*	84.46 ± 0.55	1.46↑	72.22 ± 0.84	0.12↑	80.28 ± 0.64	1.28↑	94.09 ± 0.37	3.31↑	96.60 ± 0.33	2.73↑	96.21 ± 0.14	2.60↑	97.25 ± 0.06	1.08↑	81.07 ± 0.54	4.16↑

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024



Figure 1: Graph Processing Pipeline.

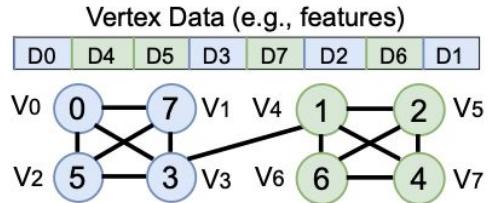
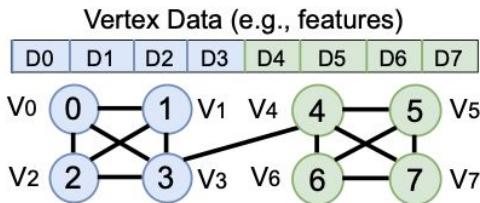
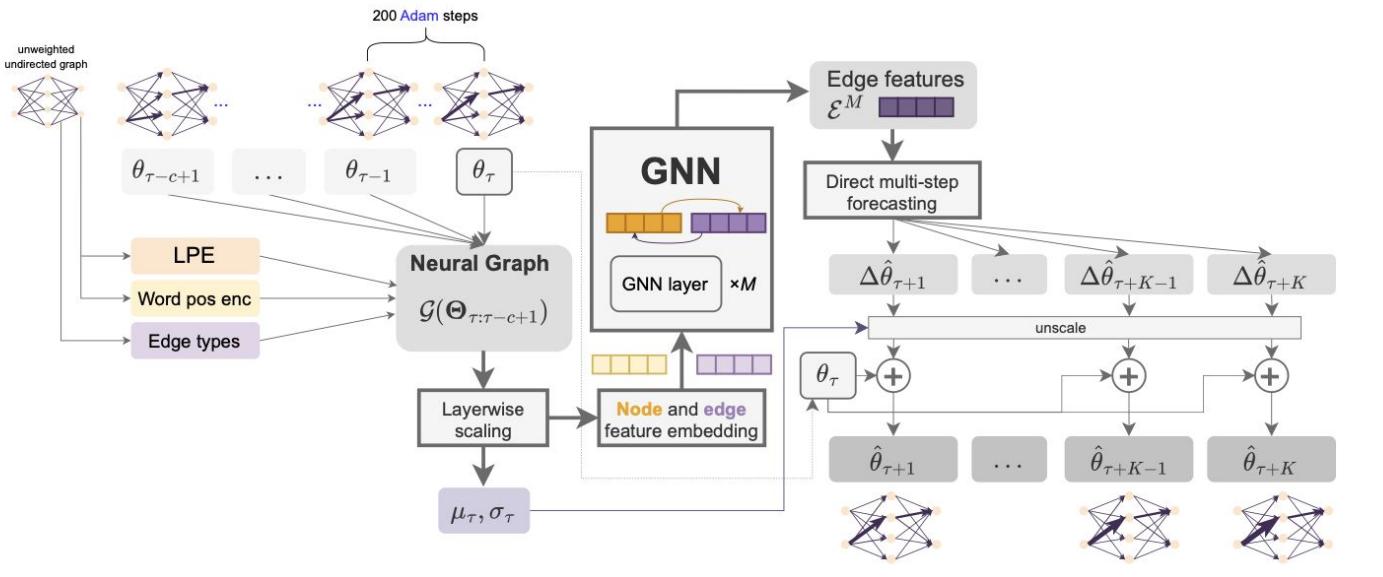


Figure 2: Two different orderings of the same graph.

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

NN parameters forecasting



Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

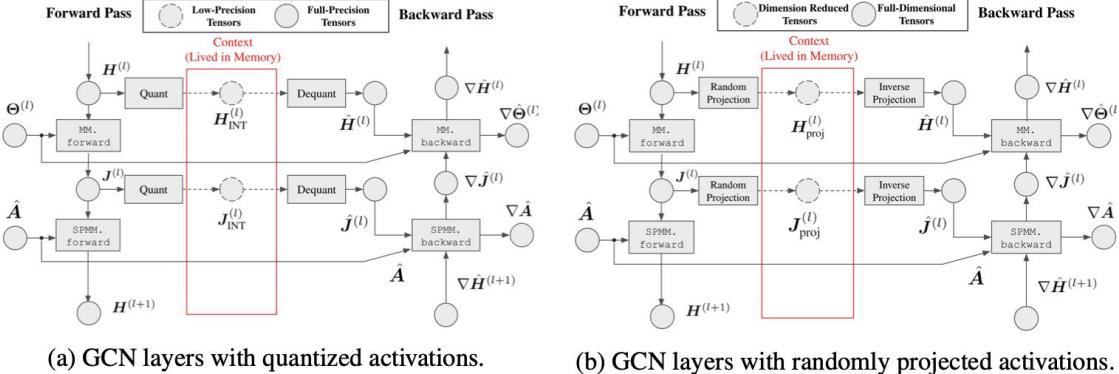


Figure 1: The training procedure of GCN layers, where only compressed activations are stored in memory. For illustration convenience, ReLU is ignored. During the forward pass, each layer's *accurate activations* ($H^{(l)}$) is used to compute those for the subsequent layer. Then the accurate activations will be compressed into the *compressed activations* ($H_{\text{INT}}^{(l)}$ and $H_{\text{proj}}^{(l)}$), which overwrites the the accurate activations and is retained in the memory. During the backward pass, we recover the compressed activations back to *decompressed activation* ($\hat{H}^{(l)}$) for computing the gradient.

Activations compression

Specifically, each node embedding vector $\mathbf{h}_v^{(l)}$ will be quantized and stored using b -bit integers. Let $B = 2^b - 1$ be the number of quantization bins. The integer quantization can be expressed as

$$\mathbf{h}_{v_{\text{INT}}}^{(l)} = \text{Quant}(\mathbf{h}_v^{(l)}) = \lfloor \frac{\mathbf{h}_v^{(l)} - Z_v^{(l)}}{r_v^{(l)}} B \rfloor, \quad (3)$$

where $Z_v^{(l)} = \min\{\mathbf{h}_v^{(l)}\}$ is the zero-point, $r_v^{(l)} = \max\{\mathbf{h}_v^{(l)}\} - \min\{\mathbf{h}_v^{(l)}\}$ is the range for $\mathbf{h}_v^{(l)}$, and $\lfloor \cdot \rfloor$ is the stochastic rounding operation (Courbariaux et al., 2015). $\mathbf{H}_{\text{INT}}^{(l)}$ in Figure 1a is the matrix containing all quantized $\mathbf{h}_{v_{\text{INT}}}^{(l)}$. During the backward pass, each $\mathbf{h}_{v_{\text{INT}}}^{(l)}$ will be dequantized as

$$\hat{\mathbf{h}}_v^{(l)} = \text{Dequant}(\mathbf{h}_{v_{\text{INT}}}^{(l)}) = r_v^{(l)} \mathbf{h}_{v_{\text{INT}}}^{(l)} / B + Z_v^{(l)}. \quad (4)$$

Table 1: The test accuracy of GCN, GraphSAGE, and GAT trained on the *ogbn-arxiv* dataset with compressed activations storing in different precision. All results are averaged over ten random trials.

Model	FP32 (Baseline)	INT8	INT4	INT2	INT1
GCN	72.07±0.16	72.06±0.29	71.96±0.26	71.93±0.20	71.68±0.17
GraphSAGE	71.85±0.24	71.83±0.15	71.85±0.27	71.58±0.22	71.34±0.24
GAT	72.35±0.12	72.39±0.14	72.34±0.12	72.35±0.12	72.17±0.12

Activations compression

$$\mathbf{h}_{v_{\text{proj}}}^{(l)} = \text{RP}(\mathbf{h}_v^{(l)}) = \mathbf{h}_v^{(l)} \mathbf{R}, \quad (5)$$

where $\mathbf{R} \in \mathbb{R}^{D \times R}$ is a random matrix ($R < D$) which satisfies $\mathbb{E}[\mathbf{R}\mathbf{R}^\top] = \mathbf{I}$. $\mathbf{H}_{\text{proj}}^{(l)} \in \mathbb{R}^{|\mathcal{V}| \times R}$ is the matrix containing all projected node embeddings $\mathbf{h}_{v_{\text{proj}}}^{(l)}$. In this paper, \mathbf{R} is set as the **normalized Rademacher random matrix** (Achlioptas, 2001) due to its low sampling cost (detailed introduction in Appendix D). After projection, we store only $\mathbf{H}_{\text{proj}}^{(l)}$ in memory, and hence the compression ratio is $\frac{D}{R}$. During the backward pass, the projected node embeddings are inversely transformed by

$$\hat{\mathbf{h}}_v^{(l)} = \text{IRP}(\mathbf{h}_{v_{\text{proj}}}^{(l)}) = \mathbf{h}_{v_{\text{proj}}}^{(l)} \mathbf{R}^\top, \quad (6)$$

where $\text{IRP}(\cdot)$ is the inverse projection operation. $\hat{\mathbf{H}}^{(l)} = \mathbf{H}_{\text{proj}}^{(l)} \mathbf{R}\mathbf{R}^\top \in \mathbb{R}^{|\mathcal{V}| \times D}$ is the recovered

Model	Full-Dimension (Baseline)	$\frac{D}{R} = 2$	$\frac{D}{R} = 4$	$\frac{D}{R} = 8$	$\frac{D}{R} = 16$
GCN	72.07±0.16	71.87±0.15	71.72±0.18	71.71±0.22	71.55±0.13
GraphSAGE	71.85±0.24	71.58±0.28	71.46±0.28	71.29±0.25	71.13±0.28
GAT	72.35±0.12	72.18±0.14	72.15±0.10	72.02±0.12	71.89±0.12

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

Activations compression

Model	Methods	# nodes	230K		89K		717K		169K		2.4M	
		# edges	11.6M		450K		7.9M		1.2M		61.9M	
		Reddit		Flickr		Yelp		<i>ogbn-arxiv</i>		<i>ogbn-products</i>		
		Acc.	Act Mem.	Acc.	Act Mem.	F1-micro	Act Mem.	Acc.	Act Mem.	Acc.	Act Mem.	Act Mem.
Cluster-GCN	Baseline	95.62 \pm 0.10	14.5 (1 \times)	49.61 \pm 0.47	16.5 (1 \times)	63.98 \pm 0.14	29.3 (1 \times)	—	—	78.62 \pm 0.26	35.2 (1 \times)	
	EXACT(INT2)	95.58 \pm 0.09	2 (7.3 \times)	49.69 \pm 0.20	1.5 (11 \times)	63.90 \pm 0.15	4 (7.3 \times)	—	—	78.47 \pm 0.40	2.5 (14 \times)	
	EXACT(RP+INT2)	95.32 \pm 0.07	1.4 (10.4 \times)	49.31 \pm 0.21	0.9 (18.3 \times)	63.61 \pm 0.21	3 (9.8 \times)	—	—	77.86 \pm 0.28	2.2 (16 \times)	
Graph-SAINT	Baseline	96.02 \pm 0.08	44.3 (1 \times)	51.11 \pm 0.28	88.7 (1 \times)	63.78 \pm 0.12	33.5 (1 \times)	71.49 \pm 0.20	270 (1 \times)	79.03 \pm 0.23	516 (1 \times)	
	EXACT(INT2)	95.96 \pm 0.05	6.6 (6.7 \times)	50.86 \pm 0.32	7.8 (11.4 \times)	63.77 \pm 0.14	4.3 (7.8 \times)	71.76 \pm 0.15	20 (13.5 \times)	78.94 \pm 0.28	40.5 (12.7 \times)	
	EXACT(RP+INT2)	95.69 \pm 0.06	3.6 (12.3 \times)	50.65 \pm 0.17	3.4 (26 \times)	63.41 \pm 0.19	3.3 (10.2 \times)	71.44 \pm 0.16	10.8 (25 \times)	78.50 \pm 0.41	29.5 (17.5 \times)	
GCN	Baseline	95.39 \pm 0.04	1029 (1 \times)	53.08 \pm 0.14	378.8 (1 \times)	40.22 \pm 0.47	6429 (1 \times)	72.07 \pm 0.16	729.4 (1 \times)	—	—	
	EXACT(INT2)	95.36 \pm 0.03	122.8 (8.4 \times)	52.92 \pm 0.20	37 (10.2 \times)	40.20 \pm 0.38	640 (10 \times)	72.04 \pm 0.21	54.5 (13.4 \times)	—	—	
	EXACT(RP+INT2)	95.30 \pm 0.03	67 (15.4 \times)	52.90 \pm 0.22	17.8 (21.3 \times)	39.89 \pm 0.56	427 (15 \times)	71.67 \pm 0.16	30.2 (24.1 \times)	—	—	
Graph-SAGE	Baseline	96.44 \pm 0.04	1527.4 (1 \times)	51.74 \pm 0.13	546.9 (1 \times)	62.05 \pm 0.14	6976 (1 \times)	71.85 \pm 0.24	786.2(1 \times)	78.78 \pm 0.19	OOM	
	EXACT(INT2)	96.40 \pm 0.05	155.5 (9.8 \times)	51.97 \pm 0.22	49.3 (11.1 \times)	61.95 \pm 0.12	680 (10.3 \times)	71.71 \pm 0.38	60.8 (12.9 \times)	78.79 \pm 0.12	1144	
	EXACT(RP+INT2)	96.34 \pm 0.03	71.7 (21.3 \times)	51.83 \pm 0.21	20.4 (26.8 \times)	61.59 \pm 0.12	466.5 (15 \times)	71.32 \pm 0.26	30.5 (25.8 \times)	78.76 \pm 0.13	572	
GCNII	Baseline	96.71 \pm 0.07	5850 (1 \times)	54.23 \pm 0.77	4067 (1 \times)	OOM	OOM	72.85 \pm 0.27	14409 (1 \times)	—	—	
	EXACT(INT2)	96.65 \pm 0.06	388 (15 \times)	54.55 \pm 0.47	256.4 (15.9 \times)	64.79 \pm 0.09	2236	72.85 \pm 0.43	899 (16 \times)	—	—	
	EXACT(RP+INT2)	96.51 \pm 0.09	197.8 (29.6 \times)	53.96 \pm 0.58	127.6 (31.9 \times)	64.01 \pm 0.17	1649	72.67 \pm 0.45	451.2 (32 \times)	—	—	

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

Traffic forecasting

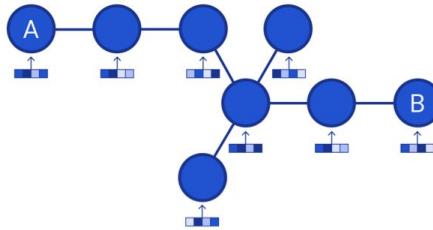
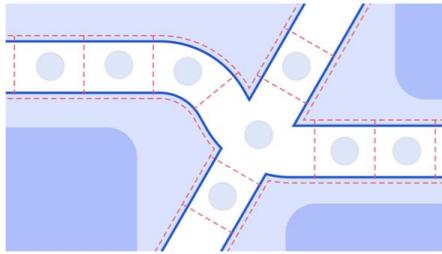
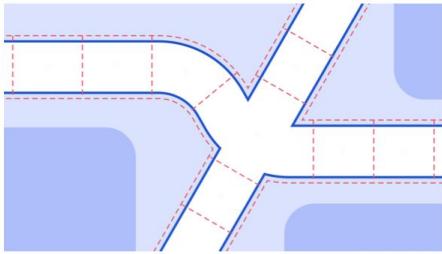
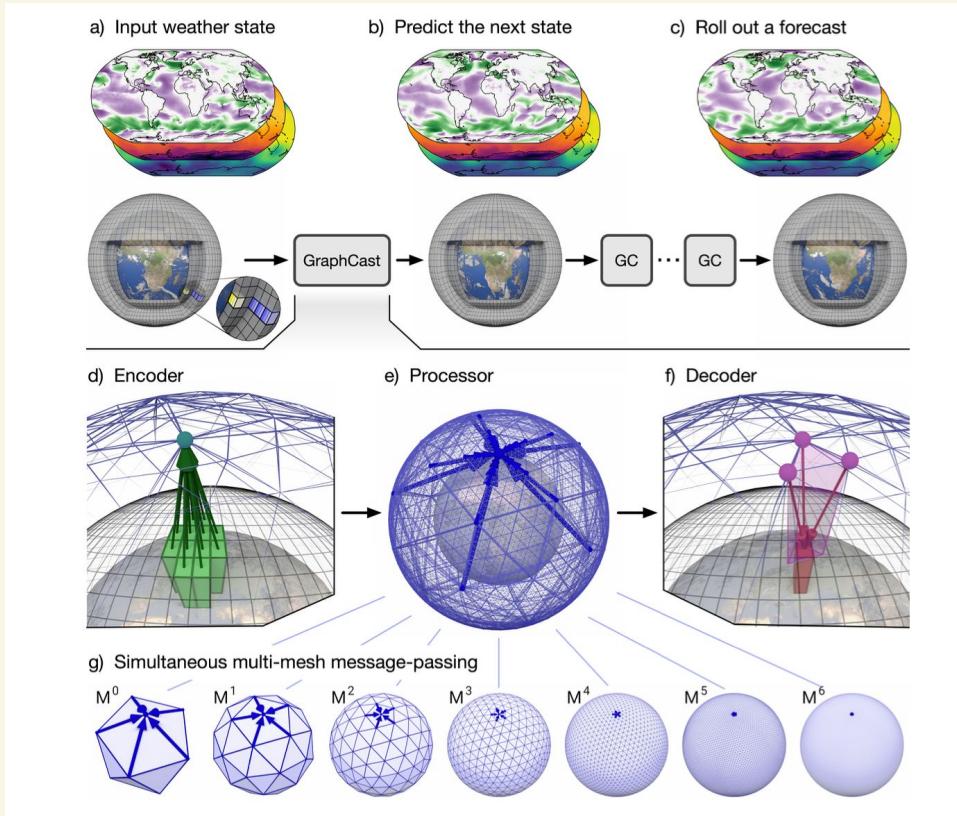


Figure 2: An example road network with shared traffic volume, which is partitioned into segments of interest (left). Each segment is treated as a node (middle), with adjacent segments connected by edges, thus forming a *supersegment* (right). Note that for *extended supersegments* (discussed in Section 4.1.4), extra off-route nodes may be connected to the graph.

Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

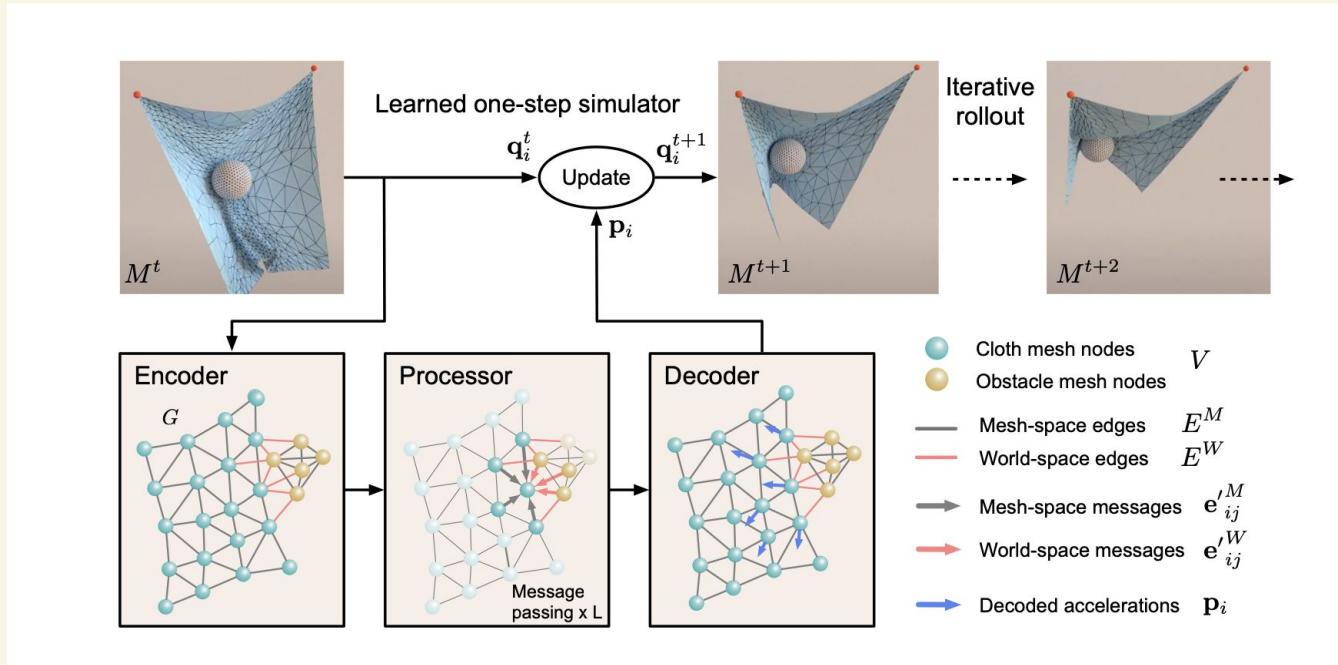
Weather forecasting



Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

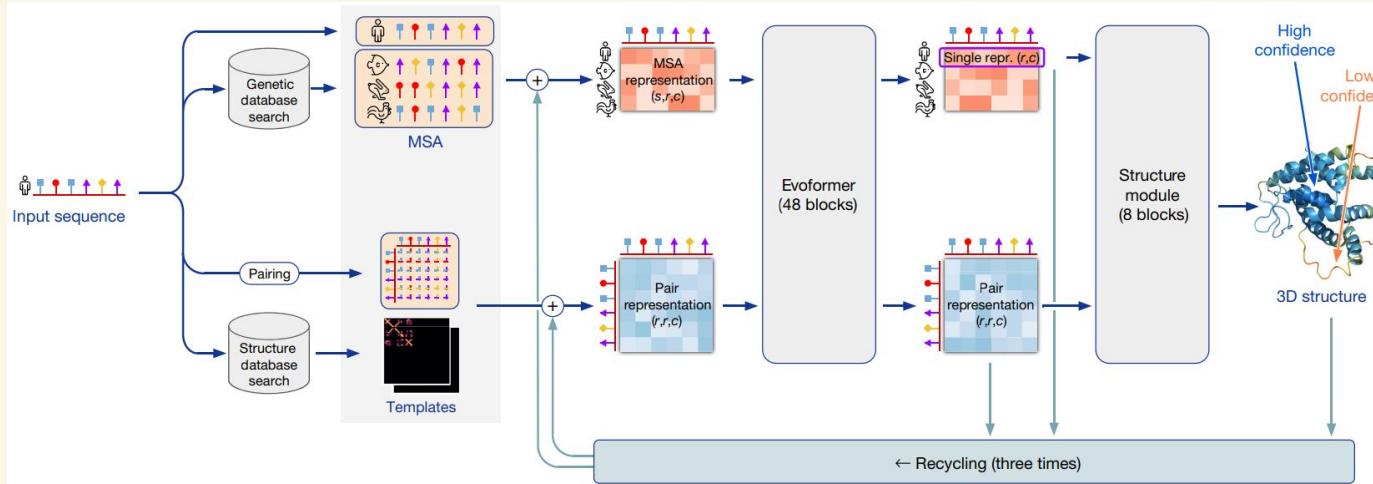
3D simulation



Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

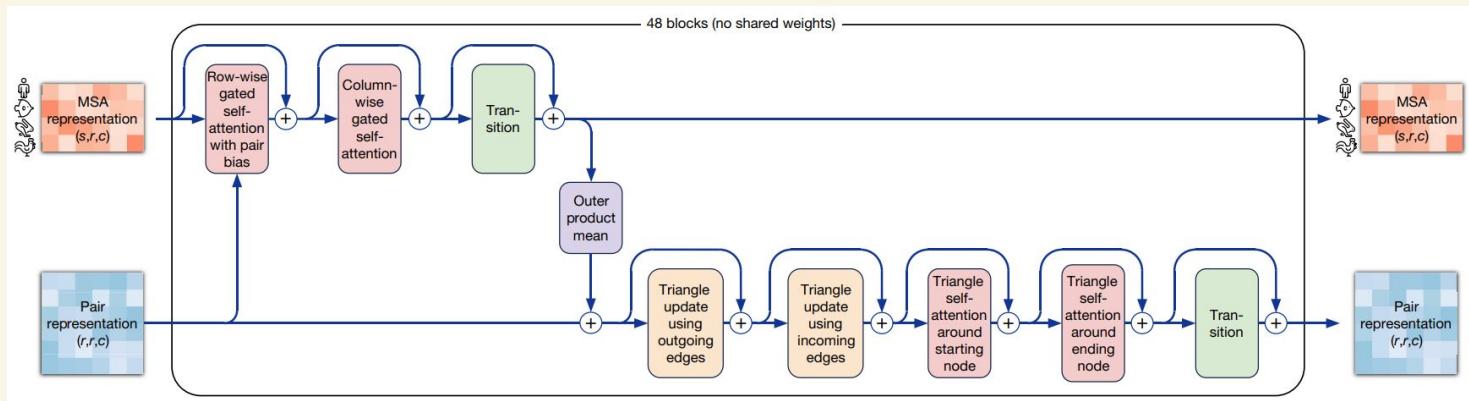
AlphaFold2



Эффективные системы машинного обучения, ВМК МГУ

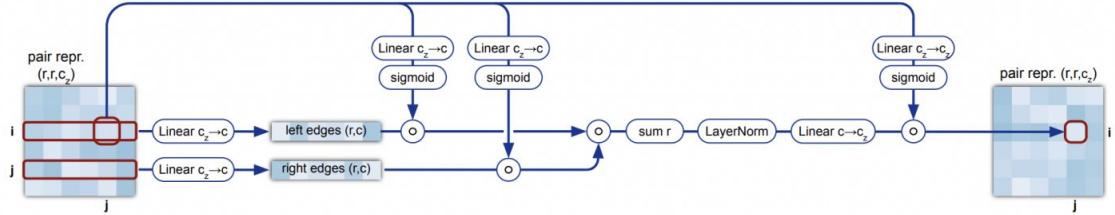
Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

AlphaFold2



Эффективные системы машинного обучения, ВМК МГУ

Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

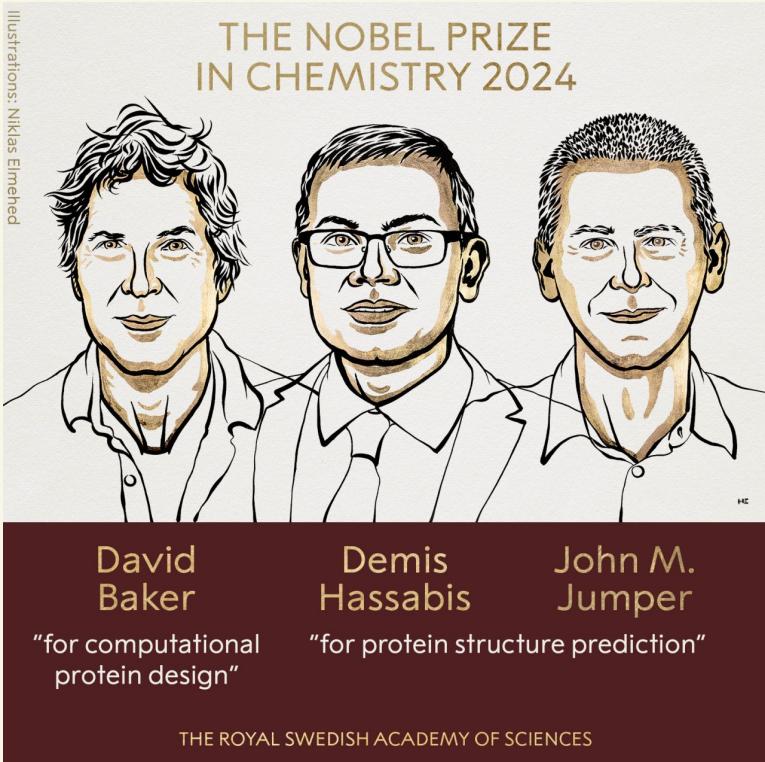


Supplementary Figure 6 | Triangular multiplicative update using “outgoing” edges. Dimensions: r : residues, c : channels.

Algorithm 11 Triangular multiplicative update using “outgoing” edges

```
def TriangleMultiplicationOutgoing({ $\mathbf{z}_{ij}$ },  $c = 128$ ) :
    1:  $\mathbf{z}_{ij} \leftarrow \text{LayerNorm}(\mathbf{z}_{ij})$ 
    2:  $\mathbf{a}_{ij}, \mathbf{b}_{ij} = \text{sigmoid}(\text{Linear}(\mathbf{z}_{ij})) \odot \text{Linear}(\mathbf{z}_{ij})$   $\mathbf{a}_{ij}, \mathbf{b}_{ij} \in \mathbb{R}^c$ 
    3:  $\mathbf{g}_{ij} = \text{sigmoid}(\text{Linear}(\mathbf{z}_{ij}))$   $\mathbf{g}_{ij} \in \mathbb{R}^{c_z}$ 
    4:  $\tilde{\mathbf{z}}_{ij} = \mathbf{g}_{ij} \odot \text{Linear}(\text{LayerNorm}(\sum_k \mathbf{a}_{ik} \odot \mathbf{b}_{jk}))$   $\tilde{\mathbf{z}}_{ij} \in \mathbb{R}^{c_z}$ 
    5: return { $\tilde{\mathbf{z}}_{ij}$ }
```

AlphaFold2



Эффективные
системы машинного
обучения, ВМК МГУ
Занятие 7. Dealing with graphs
Феоктистов Дмитрий
8 ноября 2024

Спасибо за внимание!

Если остались вопросы, то задавайте их в чат или пишите преподавателю в tg:
[@tradelik](https://t.me/trandelik)