Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

图匹配网络

沈宇辉

2023年06月18日

内容提要

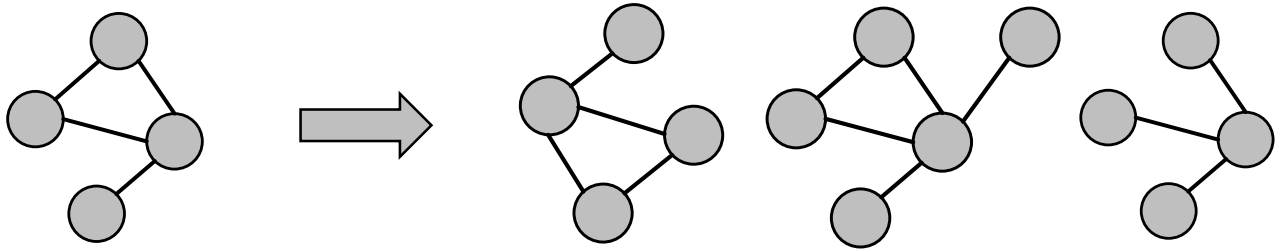
- 背景简介
- 基础概念
  - 图相似度问题
  - 图匹配网络
- 算法原理
  - GMN
  - MGMN
- 总结
- 参考文献

背景简介

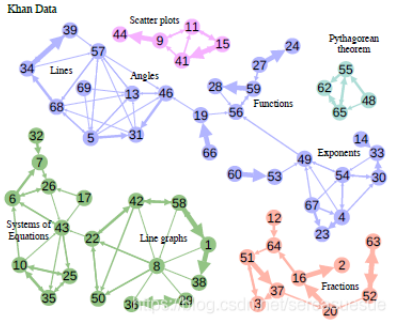
- 预期收获
  - 理解图匹配网络的定义
  - 理解图匹配网络的算法原理
  - 了解图匹配网络的应用
  - 了解图匹配网络的前沿发展

背景简介

- 图（graph）
  - $G=(V,E)$，由节点和边组成
  - 表示事物间联系的通用数据结构
- 图神经网络（Graph Neural Network）
  - 学习图结构数据的深度学习网络
  - 提取和发掘图结构数据中的节点特征、边特征
- 图相似度学习（graph similarity learning）
  - 计算两个图之间的相似性得分
  - 图结构对象的检索与匹配

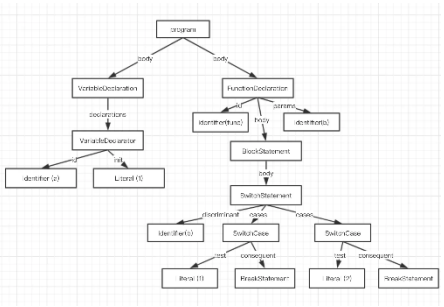


待匹配图　　候选图



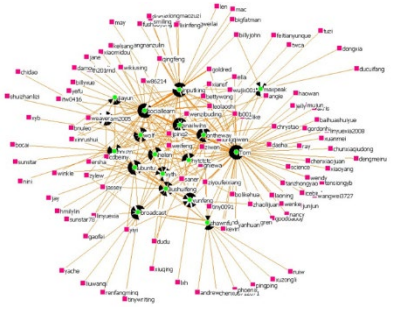
知识追踪

知识结构图



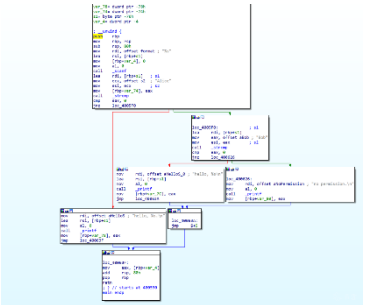
源代码漏洞挖掘

抽象语法树



推荐算法

协作知识图

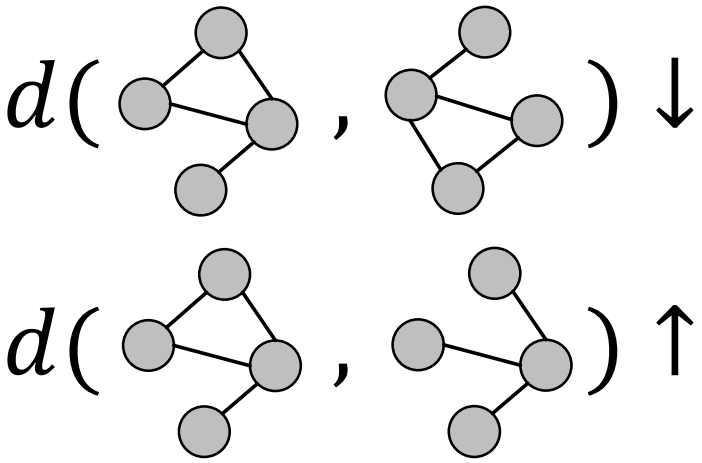


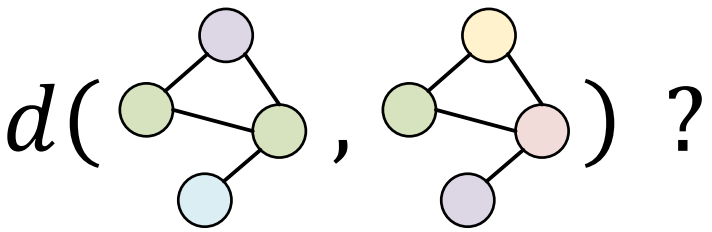
二进制漏洞挖掘

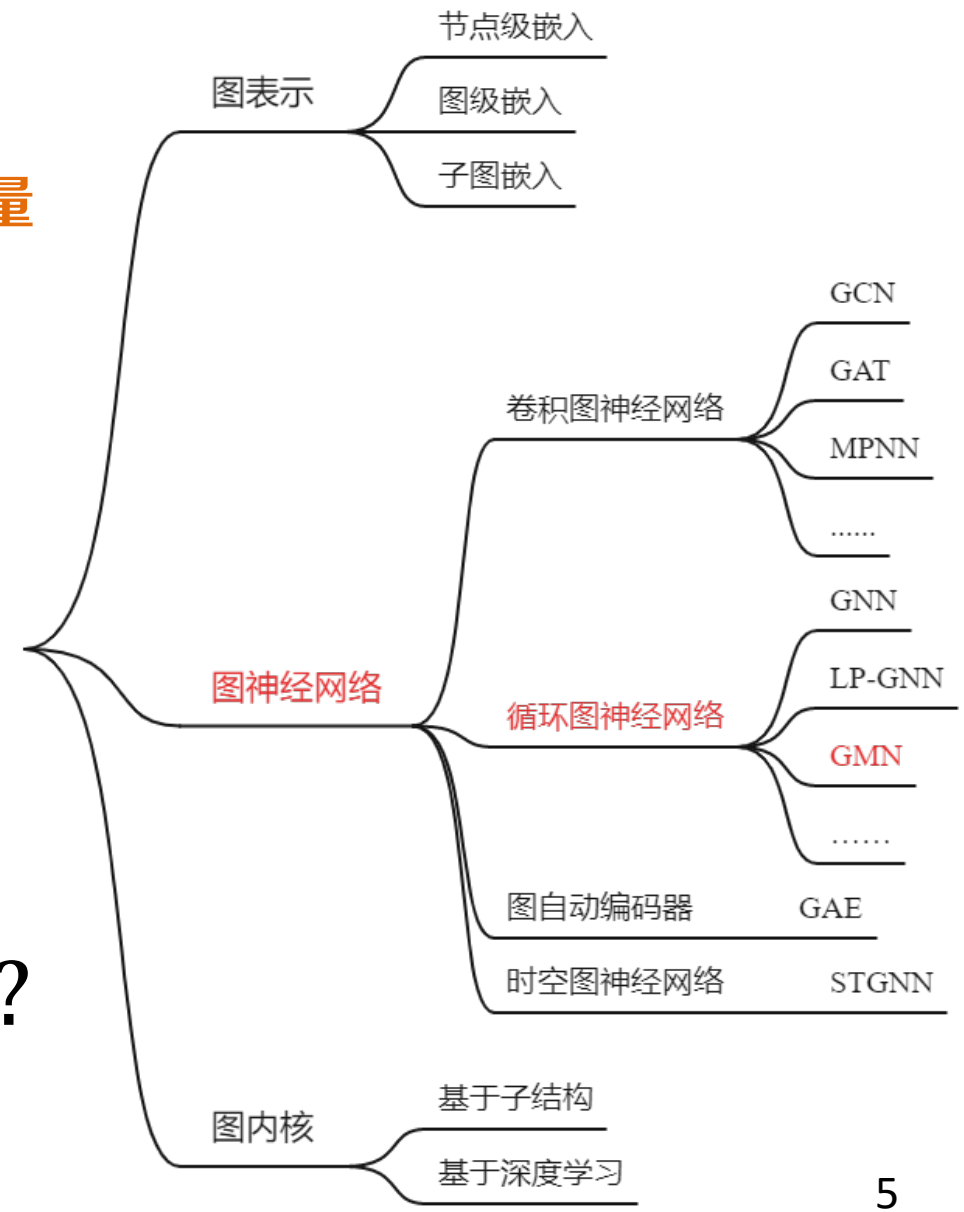
控制流图

基础概念

- 图相似度学习基本思路
  - 在保留图特征的前提下将图经过非线性变换成向量
  - 图相似度->数值向量之间的距离
- 图相似度学习难点
  - 图的节点和边都可带有属性特征
  - 图嵌入需综合考虑图结构和语义信息
  - 不同任务中“相似”的定义不同



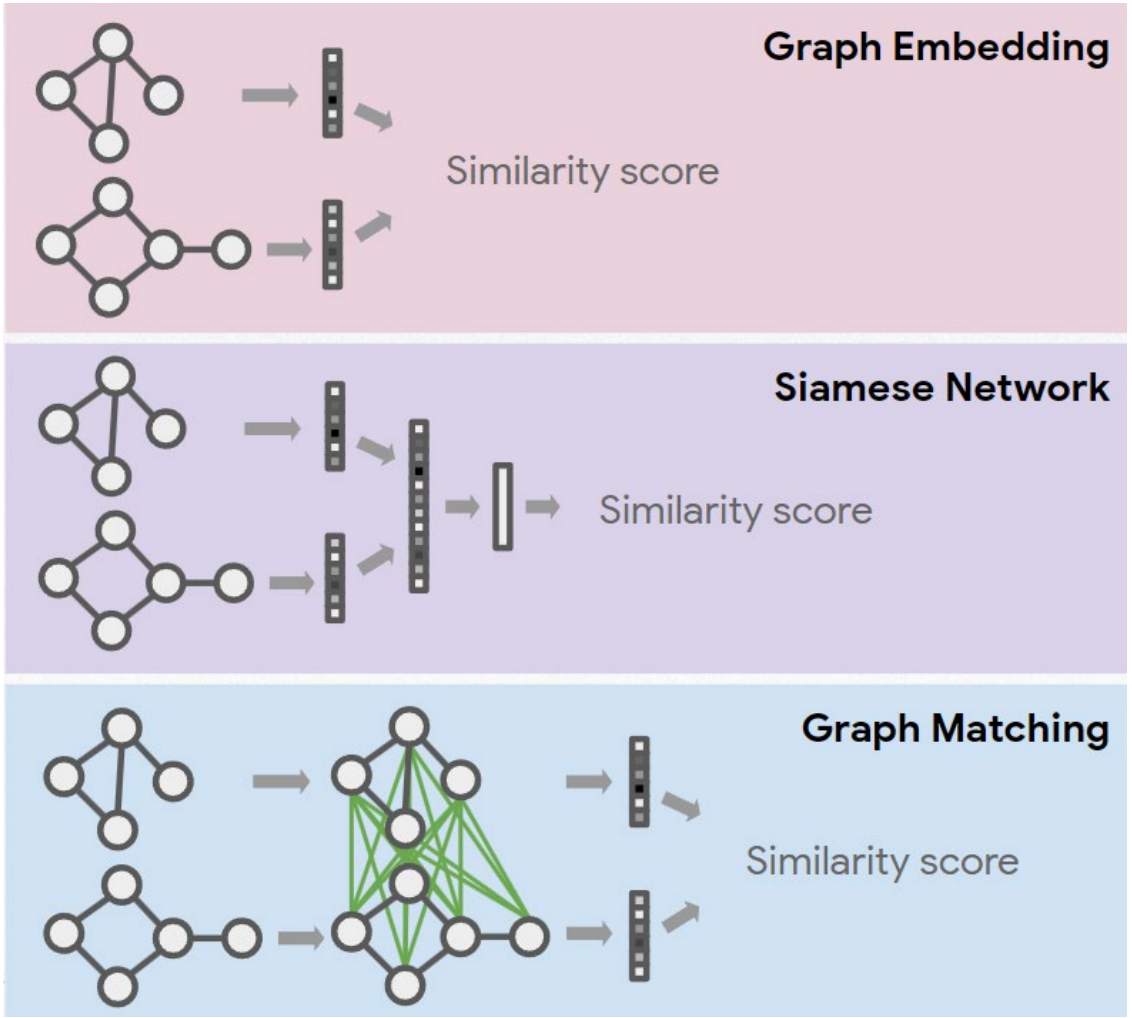




基础概念

- 常规相似度分析手段
  - 独立计算每个图的嵌入向量
  - 向量化过程中存在潜在的信息损失
- 图匹配
  - 在两个或多个图结构之间，建立节点与节点间对应关系
- 图匹配网络
  - 关联一对图之间的节点并识别差异
  - 通过联合推理计算相似度值

跨图节点关联和差异识别



T	目标	计算图之间的相似性
I	输入	待比较相似度的一对图
P	处理	1. 计算节点邻居节点的关联信息 2. 计算跨图节点关联信息 3. 依据邻居节点关联信息和跨图节点关联信息更新节点向量 4. 聚合各节点向量，生成图嵌入向量 5. 根据嵌入向量计算相似性分数
O	输出	图之间的相似性分数

P	问题	现有方法独立计算输入两个图的嵌入向量，未考虑图间关系
C	条件	有监督训练任务，训练时需提供图间相似/不相似的标签
D	难点	如何实现跨图节点关联与差异识别
L	水平	ICML2019（CCF-A）

GMN 邻居节点关联信息

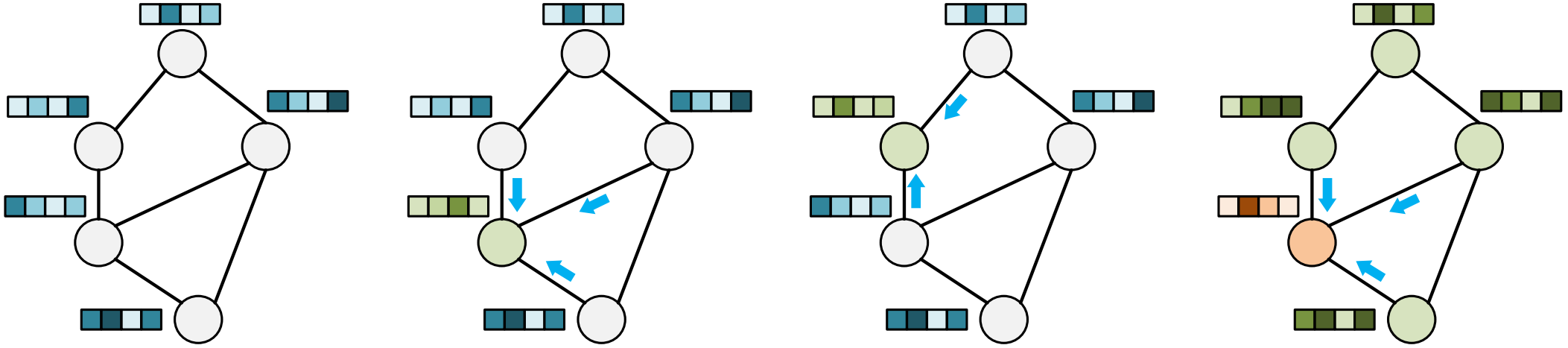
- GMN模型本质上是GNN的扩展
- 邻居节点关联信息
  - 利用周围节点信息来推导节点自身特征
  - h_i^t：节点i经过t轮迭代后的特征向量
  - 随着迭代次数增加，逐步涵盖k阶邻居节点信息

通常使用多层感知机（MLP）

$$m_{j\to i} = f_{message}(h_i^t, h_j^t, e_{ij})$$

$$h_i^{t+1} = f_{node}\left(h_i^t, \sum_{j\in E} m_{j\to i}\right)$$

可使用MLP/RNN/GRU/LSTM



GMN 跨图节点关联信息

- 跨图节点关联信息
  - 计算一个图中的节点与另一个图中各节点的匹配程度
  - 基于注意力的跨图匹配

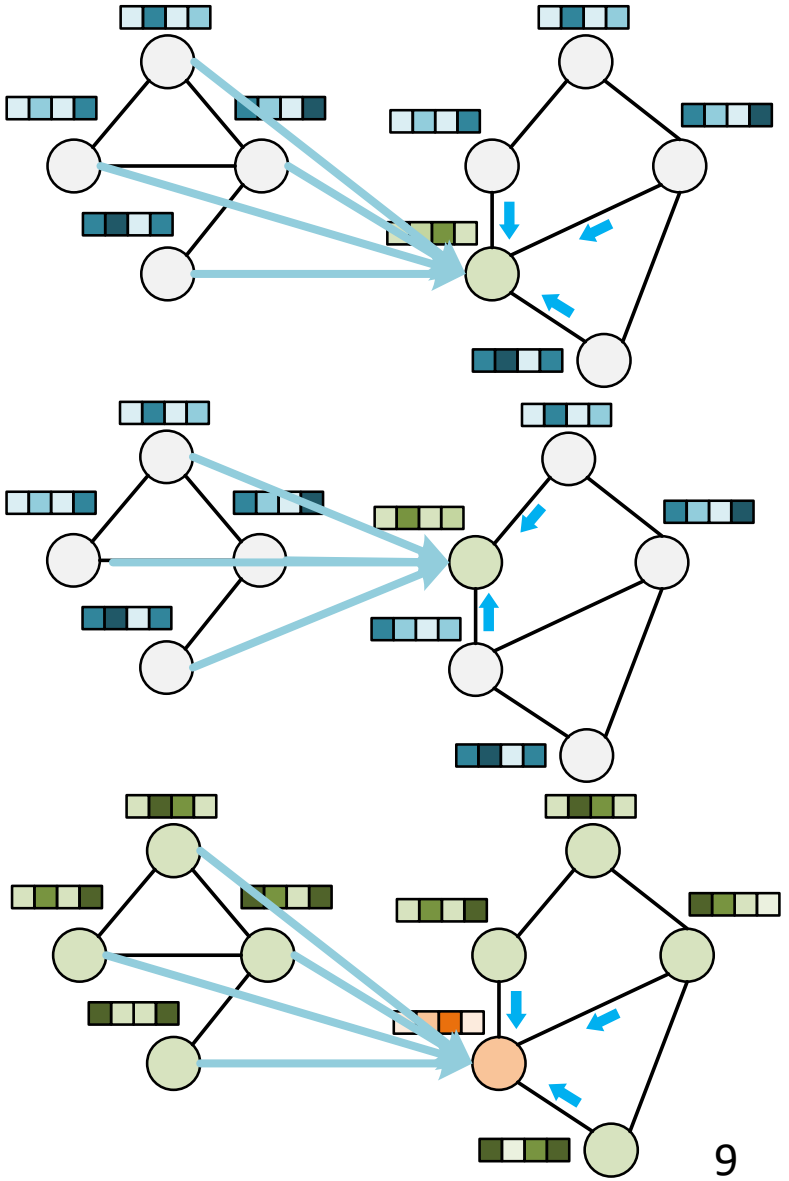
$$a_{j'\to i} = \mathrm{Softmax}_{j'}\left(s\left(\boldsymbol{h}_i^t, \boldsymbol{h}_j^t\right)\right)$$

$$\boldsymbol{\mu}_{j'\to i} = a_{j'\to i}\left(\mathbf{h}_\mathrm{i}^\mathrm{t} - \mathbf{h}_\mathrm{j}^\mathrm{t}\right)$$

  - 更新节点特征向量时，同时考虑两类关联信息

$$\boldsymbol{h}_i^{t+1} = f_{node}\left(\boldsymbol{h}_i^t, \sum_j \boldsymbol{m}_{j\to i}, \sum_{j'} \boldsymbol{\mu}_{j'\to i}\right)$$

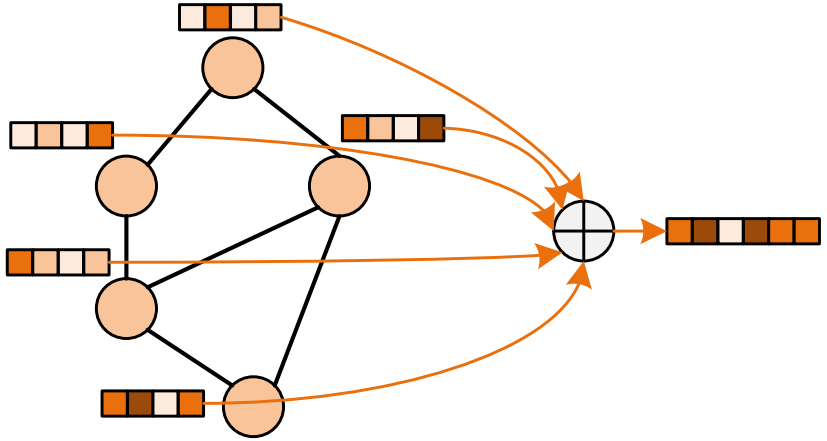
  - 图之间的差异将被捕获在交叉图匹配向量中，通过传播过程放大，使匹配模型对这些差异更加敏感



GMN 特征向量聚合

- 特征向量聚合
  - 聚合各节点特征向量，生成图嵌入向量

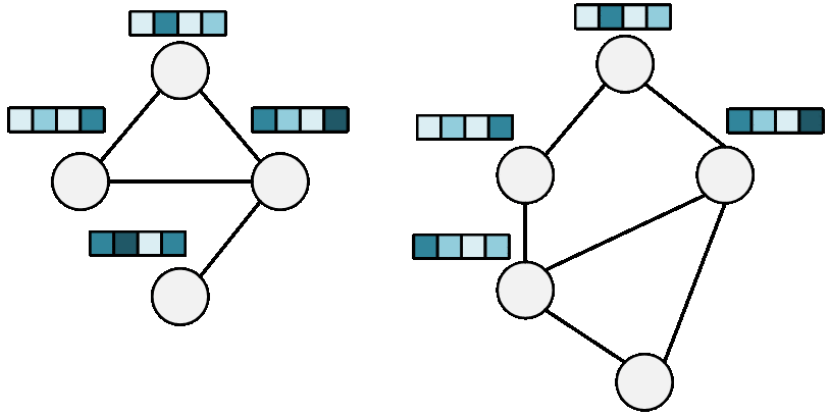
$$\boldsymbol{h}_G = \mathrm{MLP}\big(\mathrm{POOL}(\{\boldsymbol{h}_v\}_{v\in V})\big)$$



- 相似度分数计算
  - 使用欧氏距离或余弦相似度评价相似性

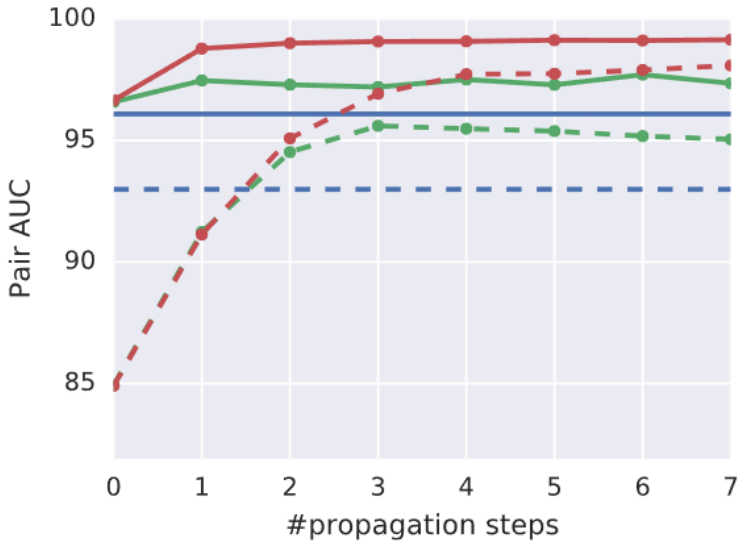
$$s = f_s\big(\boldsymbol{h}_{G_1}, \boldsymbol{h}_{G_2}\big)$$

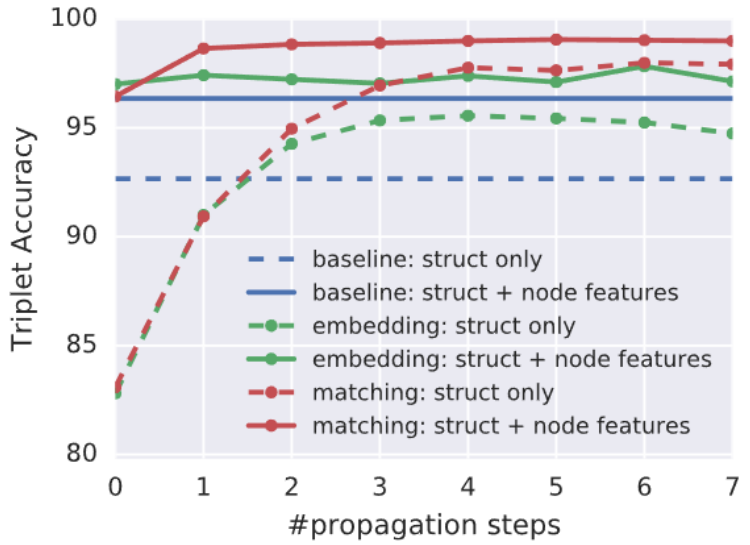
- 与图嵌入模型相比，匹配模型能够根据与之比较的其他图改变图的嵌入结果



GMN 实验结果

- 评价指标：AUC、准确率
- 图匹配（红色）> 图嵌入（绿色）> 传统方法（蓝色）
- 综合使用结构+节点特征（实线） > 仅使用结构特征（虚线）
- 孪生网络 vs 图匹配
  - 提前提取跨图关联信息 > 在网络末端进行信息融合



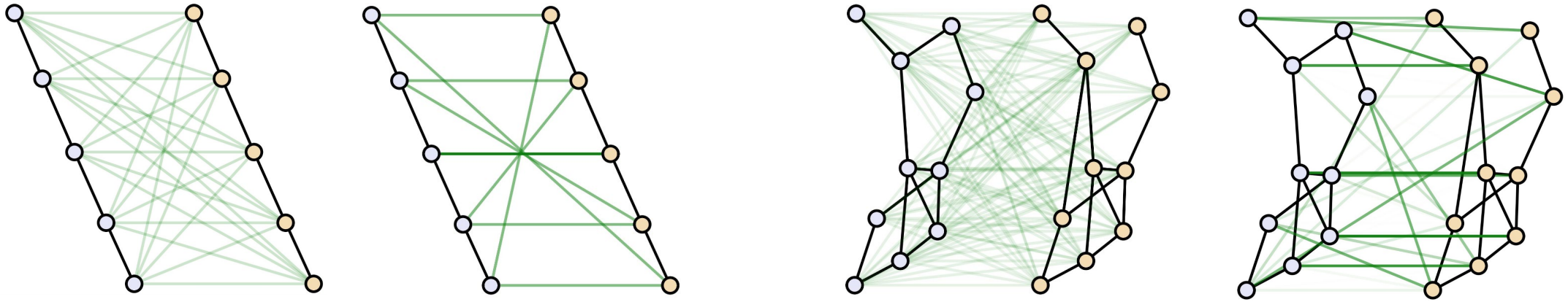


Model	Pair AUC	Triplet Acc
Baseline	96.09	96.35
GCN	96.67	96.57
Siamese-GCN	97.54	97.51
GNN	97.71	97.83
Siamese-GNN	97.76	97.58
GMN	**99.28**	**99.18**

Function Similarity Search

GMN 实验&总结

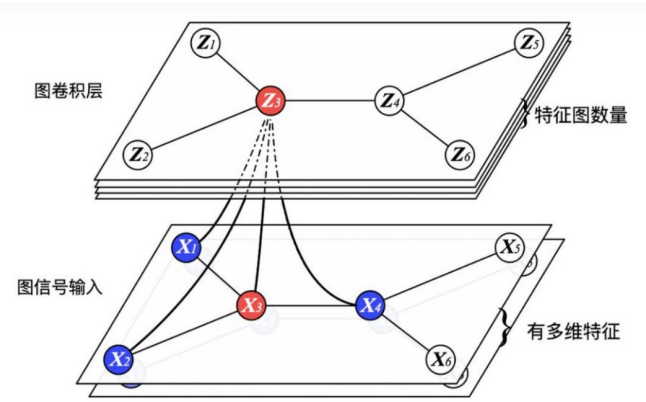
- 跨图注意力可视化
  - 当图完全相同时，注意力迭代结果与节点匹配结果相近
  - 当图存在差异时，跨图注意力倾向于关注“度”较高的节点
- 存在问题
  - 跨图匹配带来的算力开销随图大小和迭代轮次指数级上升
  - 由于图嵌入结果可变，方法仅支持一对一相似度分析，不支持相似图搜索



T	目标	计算图之间的相似性
I	输入	待比较相似度的一对图
P	处理	1. 使用GCN计算节点嵌入向量 2. 计算跨图节点关联信息 3. 聚合跨图节点关联信息向量，生成跨图关系向量 4. 聚合各节点向量，生成图嵌入向量 5. 拼接跨图关系向量和图嵌入向量，计算相似性分数
O	输出	图之间的相似性分数

P	问题	GMN模型算力开销大
C	条件	有监督训练任务，训练时需提供图间相似/不相似的标签
D	难点	1. 跨图节点关联信息提取与迭代轮次解耦 2. 综合利用节点-图和图-图关联信息
L	水平	2023 SCI 1区期刊（论文完成于2021年）

- **GMN问题1：算力开销随图大小和迭代轮次指数级上升**
  - **将跨图节点关联信息提取与迭代轮次解耦**
  - **先进行节点嵌入向量提取，在此基础上再进行跨图分析**
- **使用GCN+孪生网络计算节点嵌入向量**

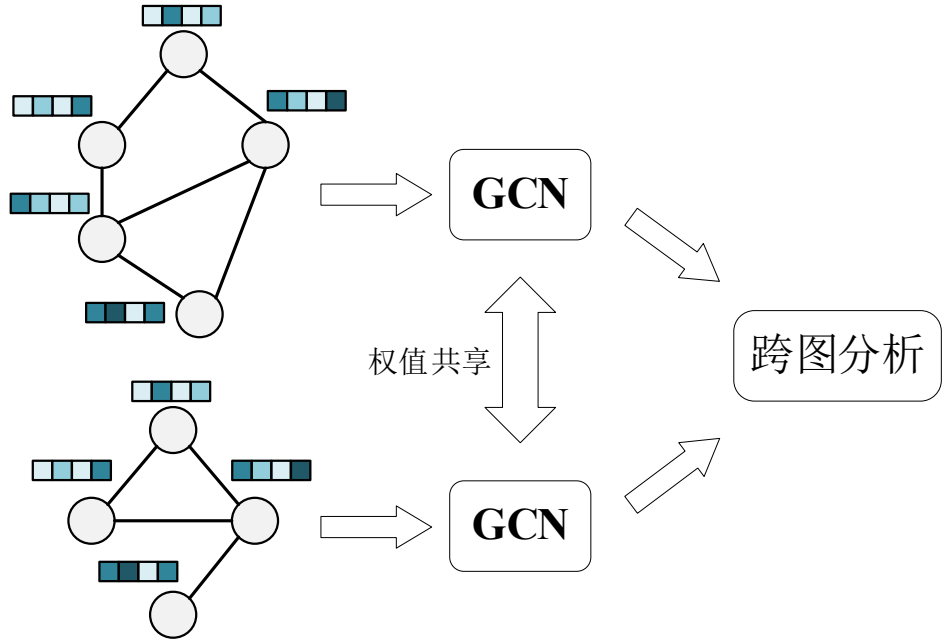


后一层特征　邻接矩阵　前一层特征　权重

$$H^{l+1} = \sigma(AH^lW^l)$$

$$H^{l+1} = \sigma(D^{\frac{1}{2}}\hat{A}D^{-\frac{1}{2}}H^lW^l)$$

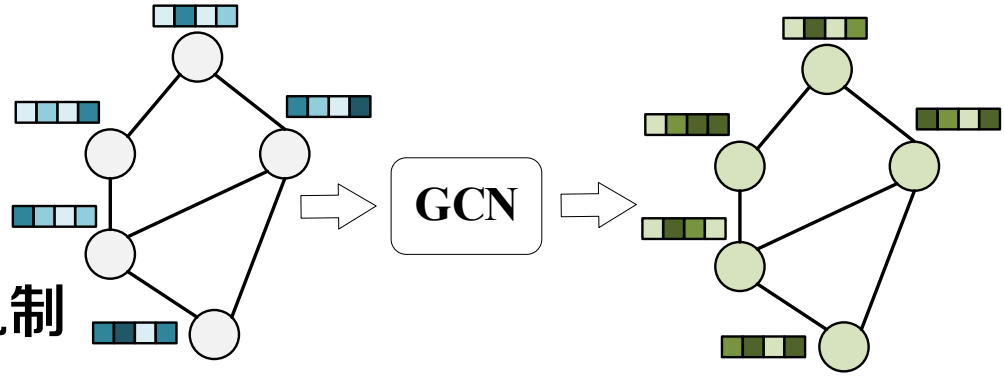
Model	FFmpeg		
	[3, 200]	[20, 200]	[50, 200]
NGMN-GCN (Our)	97.73±0.11	98.29±0.21	96.81±0.96
NGMN-GraphSAGE	97.31±0.56	98.21±0.13	97.88±0.15
NGMN-GIN	97.97±0.08	98.06±0.22	94.66±4.01
NGMN-GGNN	**98.42±0.41**	**99.77±0.07**	**97.93±1.18**



解耦后图之间的差异不再具有传播特性?

MGMN 节点-图匹配

- 问题：无法通过迭代传播捕获节点间差异
- 解决方案
  - 直接计算节点-图差异关系
  - 借助孪生网络训练权重矩阵，替代GMN中注意力机制



各节点嵌入向量的加权平均

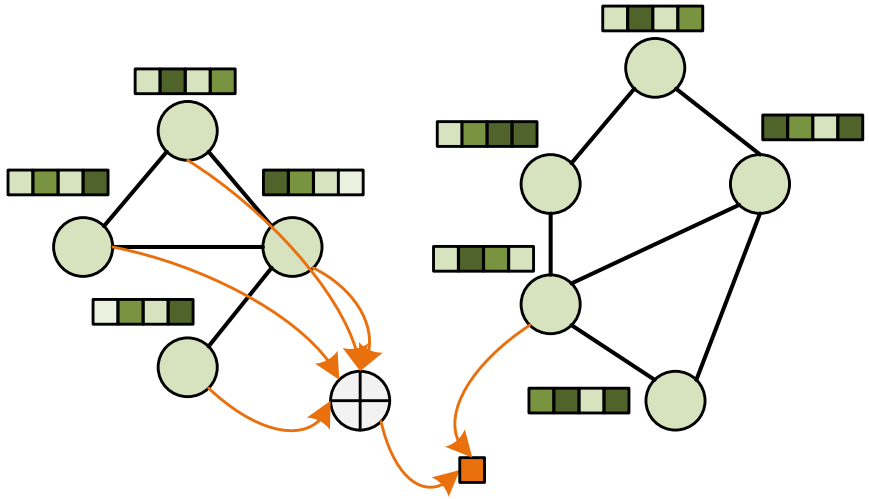
$$\boldsymbol{h}_{G,avg} = \sum a_{i,j}\boldsymbol{h}_j$$

节点嵌入向量间的差异度

$$a_{i,j} = f_s(\boldsymbol{h}_i, \boldsymbol{h}_j) = cosine(\boldsymbol{h}_i, \boldsymbol{h}_j)$$

$$\widetilde{\boldsymbol{h}_i} = f_m(\boldsymbol{h}_i, \boldsymbol{h}_{G,avg}, \boldsymbol{W}_m) = cosine(\boldsymbol{h}_i \odot \boldsymbol{W}_m, \boldsymbol{h}_{G,avg} \odot \boldsymbol{W}_m)$$

节点-图差异关系　　矩阵点乘　　多视角权重矩阵



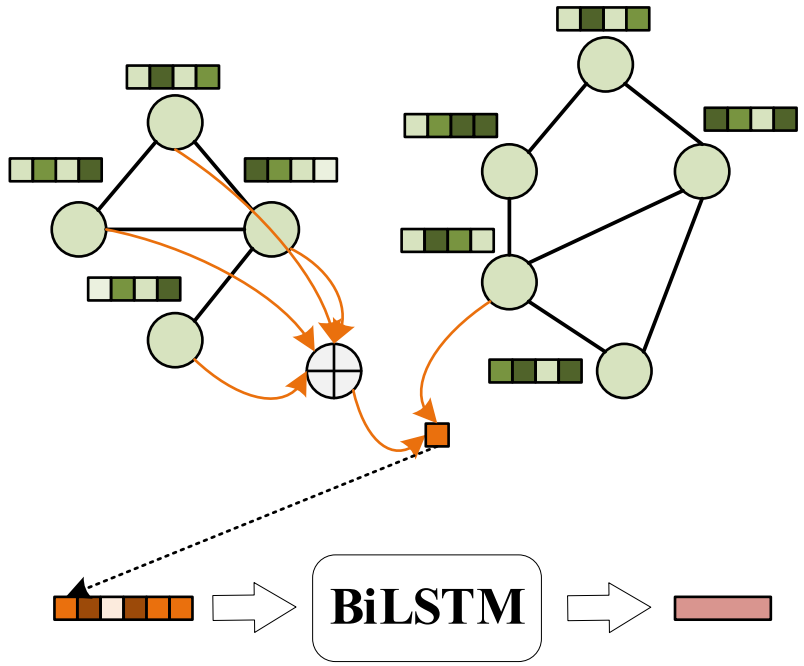
$$[v_1 \quad v_2 \quad v_3]^T \odot \begin{bmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \end{bmatrix}^T = \begin{bmatrix} v_1w_1 & v_2w_2 & v_3w_3 \\ v_1w_4 & v_2w_5 & v_3w_6 \end{bmatrix}^T$$

MGMN 特征向量聚合

- 使用BiLSTM聚合各节点差异向量

$$\widetilde{\boldsymbol{h}_G} = \mathrm{BiLSTM}\left(\{\widetilde{\boldsymbol{h}_i}\}^{\{N,M\}}\right)$$

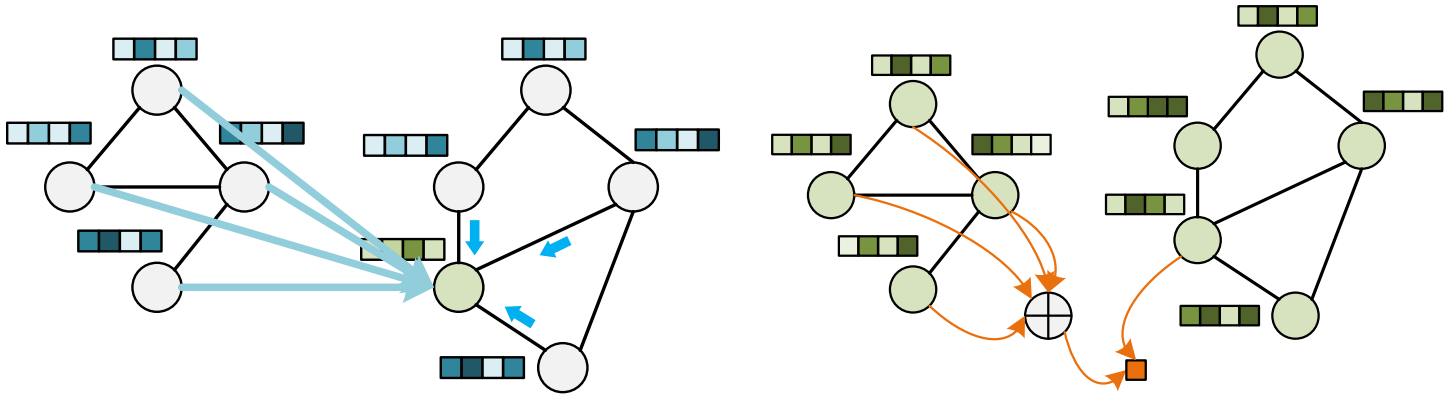
- 为何使用BiLSTM?
  - 使用长短期记忆网络进行聚合已有先例
  - 为了减少顺序对于最终结果的影响，随机打乱特征向量顺序后输入网络
  - 实验证明BiLSTM拥有最佳性能



Model	FFmpeg			OpenSSL		
	[3, 200]	[20, 200]	[50, 200]	[3, 200]	[20, 200]	[50, 200]
NGMN (Max)	73.74±8.30	73.85±1.76	77.72±2.07	67.14±2.70	63.31±3.29	63.02±2.77
NGMN (FCMax)	97.28±0.08	96.61±0.17	96.65±0.30	95.37±0.19	96.08±0.48	**95.90±0.73**
NGMN (BiLSTM)	97.73±0.11	98.29±0.21	96.81±0.96	96.56±0.12	**97.60±0.29**	92.89±1.31

MGMN 图-图匹配

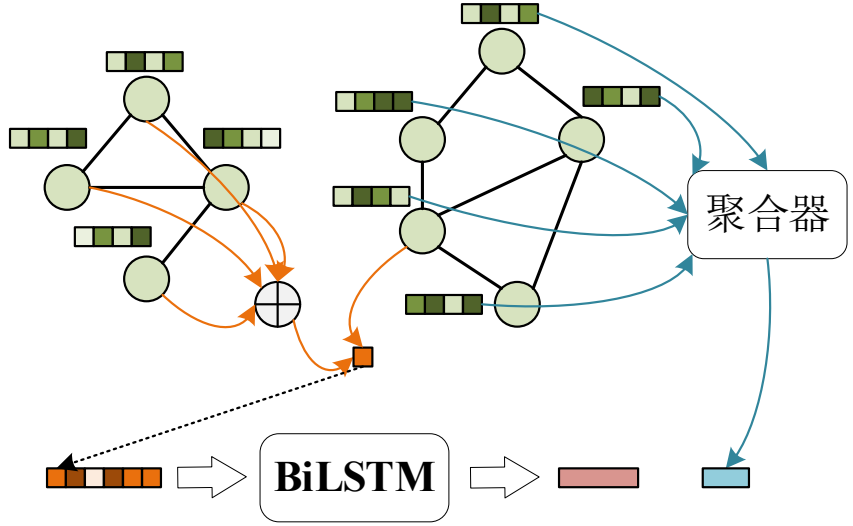
- 解决方案
  - 直接利用GCN输出的节点嵌入向量
  - 聚合节点嵌入向量得到图嵌入向量
  - 拼接差异向量与图嵌入向量为最终结果



GMN　　MGMN

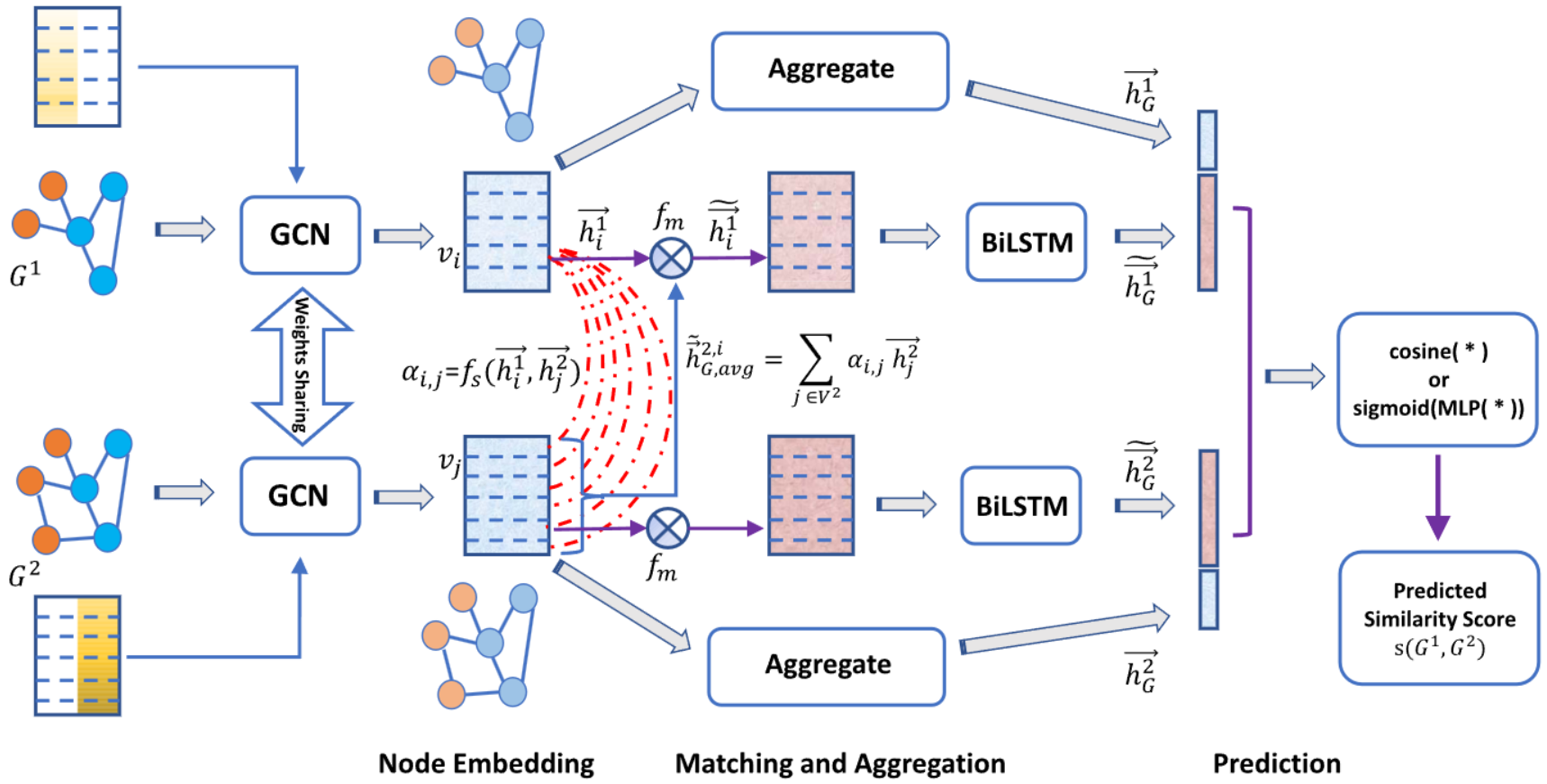
Model	FFmpeg		
	[3, 200]	[20, 200]	[50, 200]
MGMN (Max + BiLSTM)	97.44±0.32	97.84±0.40	97.22±0.36
MGMN (FCMax + BiLSTM)	**98.07±0.06**	**98.29±0.10**	**97.83±0.11**
MGMN (BiLSTM + BiLSTM)	97.56±0.38	98.12±0.04	97.16±0.53

仅考虑跨图匹配信息，
忽略图本身信息？



MGMN 整体框架

- 节点-图匹配/图-图匹配共享GCN输出结果
- 结合孪生网络+图匹配优势



MGMN 实验结果

- 分类任务
  - 判断输入的两个图是否相似
  - 评价指标：AUC

Model	FFmpeg			OpenSSL		
	[3, 200]	[20, 200]	[50, 200]	[3, 200]	[20, 200]	[50, 200]
SimGNN	95.38±0.76	94.31±1.01	93.45±0.54	95.96±0.31	93.58±0.82	94.25±0.85
GMN	94.15±0.62	95.92±1.38	94.76±0.45	96.43±0.61	93.03±3.81	93.91±1.65
GraphSim	97.46±0.30	96.49±0.28	94.48±0.73	96.84±0.54	94.97±0.98	93.66±1.84
SGNN (Max)	93.92±0.07	93.82±0.28	85.15±1.39	91.07±0.10	88.94±0.47	82.10±0.51
SGNN (FCMax) 仅使用图-图	95.37±0.04	96.29±0.14	95.98±0.32	92.64±0.15	93.79±0.17	93.21±0.82
SGNN (BiLSTM)	96.92±0.13	97.62±0.13	96.35±0.33	95.24±0.06	96.30±0.27	93.99±0.62
NGMN (Max)	73.74±8.30	73.85±1.76	77.72±2.07	67.14±2.70	63.31±3.29	63.02±2.77
NGMN (FCMax) 仅使用节点-图	97.28±0.08	96.61±0.17	96.65±0.30	95.37±0.19	96.08±0.48	**95.90±0.73**
NGMN (BiLSTM)	97.73±0.11	98.29±0.21	96.81±0.96	96.56±0.12	**97.60±0.29**	92.89±1.31
MGMN (Max + BiLSTM)	97.44±0.32	97.84±0.40	97.22±0.36	94.77±1.80	97.44±0.26	94.06±1.60
MGMN (FCMax + BiLSTM)	**98.07±0.06**	**98.29±0.10**	**97.83±0.11**	96.87±0.24	97.59±0.24	95.58±1.13
MGMN (BiLSTM + BiLSTM)	97.56±0.38	98.12±0.04	97.16±0.53	**96.90±0.10**	97.31±1.07	95.87±0.88

MGMN 实验结果

- 回归任务
  - 计算输入两个图的相似性分数
  - 从候选集中选择与被查询图最相似的图
  - 评价指标：均方误差、前k个结果的正确率

Datasets	Model	mse (10^{-3})	ρ	τ	p@10	p@20
LINUX1000	SimGNN	2.479±1.038	0.912±0.031	0.791±0.046	0.635±0.328	0.650±0.283
	GMN	2.571±0.519	0.906±0.023	0.763±0.035	0.888±0.036	0.856±0.040
	GraphSim	0.471±0.043	0.976±0.001	**0.931±0.003**	**0.956±0.006**	0.942±0.007
	SGNN (Max)	11.832±0.698	0.566±0.022	0.404±0.017	0.226±0.106	0.492±0.190
	SGNN (FCMax) 仅使用图-图	17.795±0.406	0.362±0.021	0.252±0.015	0.239±0.000	0.241±0.000
	SGNN (BiLSTM)	2.140±1.668	0.935±0.050	0.825±0.100	0.878±0.012	0.865±0.007
	NGMN (Max)*	16.921±0.000	-	-	-	-
	NGMN (FCMax) 仅使用节点-图	4.793±0.262	0.829±0.006	0.665±0.011	0.764±0.170	0.767±0.166
	NGMN (BiLSTM)	1.561±0.020	0.945±0.002	0.814±0.003	0.743±0.085	0.741±0.086
	MGMN (Max + BiLSTM)	1.054±0.086	0.962±0.003	0.850±0.008	0.877±0.054	0.883±0.047
	MGMN (FCMax + BiLSTM)	1.575±0.627	0.946±0.019	0.817±0.034	0.807±0.117	0.784±0.108
	MGMN (BiLSTM + BiLSTM)	**0.439±0.143**	**0.985±0.005**	0.919±0.016	0.955±0.011	**0.943±0.014**

MGMN 实验结果

- GCN卷积层数选择

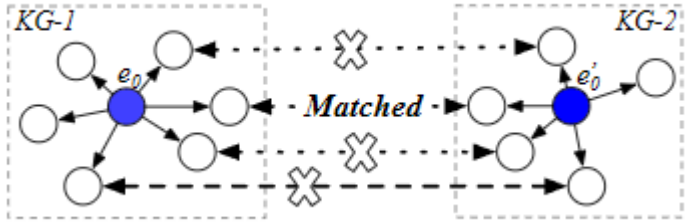
Model	FFmpeg			OpenSSL		
	[3, 200]	[20, 200]	[50, 200]	[3, 200]	[20, 200]	[50, 200]
NGMN-(1 layer)	97.84±0.08	71.05±2.98	75.05±17.20	97.51±0.24	88.87±4.79	77.72±7.00
NGMN-(2 layers)	**98.03±0.15**	84.72±12.60	90.58±10.12	**97.65±0.10**	95.78±3.46	86.39±8.16
NGMN-(3 layers)	97.73±0.11	**98.29±0.21**	96.81±0.96	96.56±0.12	97.60±0.29	92.89±1.31
NGMN-(4 layers)	97.96±0.22	98.06±0.13	**97.94±0.15**	96.79±0.21	**98.21±0.31**	**93.40±1.78**

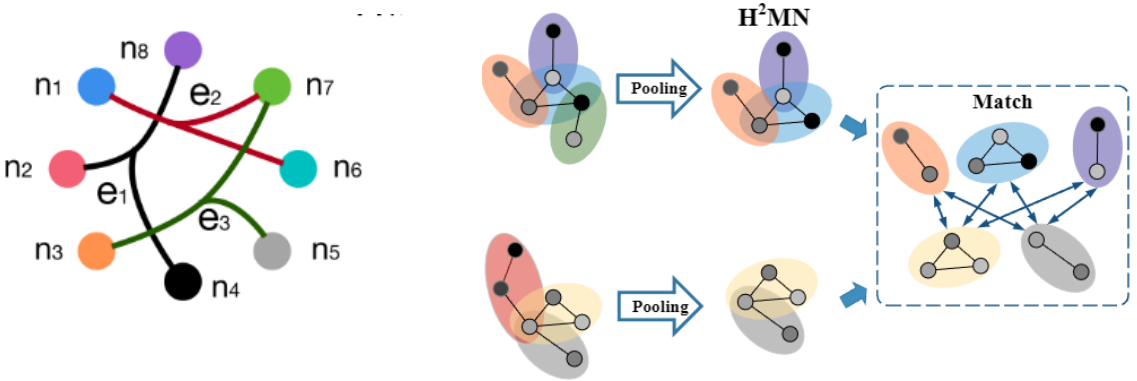
- 视角数选择

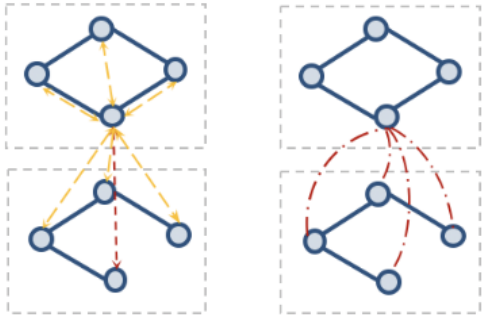
Model	FFmpeg			OpenSSL		
	[3, 200]	[20, 200]	[50, 200]	[3, 200]	[20, 200]	[50, 200]
NGMN-($\tilde{d}$ = 50)	98.11±0.14	97.76±0.14	96.93±0.5	**97.38±0.11**	97.03±0.84	93.38±3.03
NGMN-($\tilde{d}$ = 75)	97.99±0.09	97.94±0.14	97.41±0.05	97.09±0.25	98.66±0.11	92.10±4.37
NGMN-($\tilde{d}$ = 100)	97.73±0.11	**98.29±0.21**	96.81±0.96	96.56±0.12	97.60±0.29	92.89±1.31
NGMN-($\tilde{d}$ = 125)	98.10±0.03	98.06±0.08	97.26±0.36	96.73±0.33	**98.67±0.11**	96.03±2.08
NGMN-($\tilde{d}$ = 150)	**98.32±0.05**	98.11±0.07	**97.92±0.09**	96.50±0.31	98.04±0.03	**97.13±0.36**

前沿发展

- 任务细化
  - 明确节点间一对一匹配关系
  - 狭义图匹配问题
- 扩展关系模型
  - 超图匹配网络
  - 超图：一条边支持连接多个节点
  - 论文引用关系、社交网络等
- 引入前沿技术
  - 对比学习+图匹配网络
  - 支持无监督训练



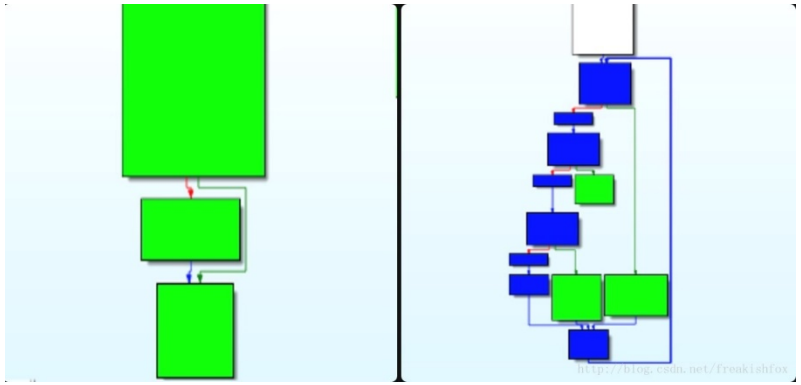


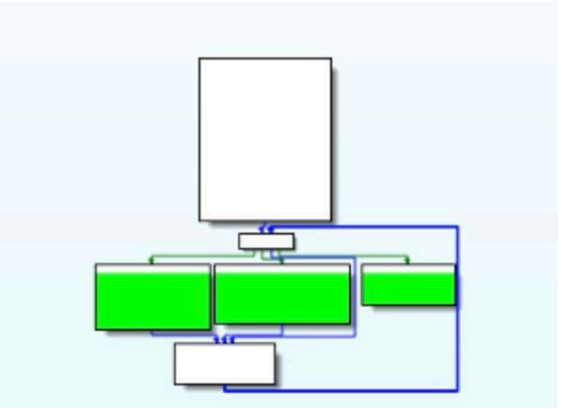


(a) Contrastive learning. (b) Cross-view interaction.

应用示例

- 混淆条件下二进制函数相似性检测
  - 混淆：将计算机程序的源代码或机器码，转换为功能上等价，但是难于阅读和理解的形式
  - 常见混淆方法
    - 流程伪造
    - 控制流图扁平化
- 函数控制流图结构遭到破坏
  - 图嵌入结果受图结构影响大，相似性判断出现偏差
- 核心代码块语义未发生变化
  - 利用图匹配网络思想
  - 重点关注核心代码块匹配状况





参考文献

- LI Y, GU C, DULLIEN T, et al. Graph Matching Networks for Learning the Similarity of Graph Structured Objects. Proceedings of the 36th International Conference on Machine Learning. [C] New York: ACM, 2019:3835-3845.
- LING X, WU L, WANG S, et al. Multilevel Graph Matching Networks for Deep Graph Similarity Learning. IEEE Transactions on Neural Networks and Learning Systems [J].2023，34: 799-813.
- Ma G, Ahmed N K, Willke T L, et al. Deep graph similarity learning: A survey[J]. Data Mining and Knowledge Discovery, 2021, 35: 688-725.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！