



为谋 (/users/403881) 2020-09-08 20:28:10 (最初创作于: 2020-09-08 20:24:19) 发表于: 蚂蚁机器学习技术团队 (/teams/617)>>GraphML (/teams/617?cid=2778)

165 阅读

知识体系: 机器学习/深度学习 (/articles/?kid=941) 算法框架/工具 (/articles/?kid=948) 人工智能 (/articles/?kid=951) 修改知识体系

文章标签: NLP (/search?q=NLP&type=ARTICLE) ALPS (/search?q=ALPS&type=ARTICLE) 多模态 (/search?q=多模态&type=ARTICLE) 图学习 (/search?q=图学习&type=ARTICLE) GraphML (/search?q=GraphML&type=ARTICLE) 修改标签 标签历史 (/articles/180451/tags/history)

ALPS-GraphML多模态初探：GNN + NLP多模应用

背景

随着Graph Neural Network（GNN）的火热发展，GNN在表达深度上不断突破，在应用广度上也在尝试“破圈”。近年来，GNN与CV、NLP、生物化学、强化学习等领域[1]都摩擦出了火花，相关领域的论文与工业应用层出不穷。作为ALPS-GraphML团队，本文将尝试从GNN的角度探秘GNN与NLP的结合模式，对业界的多模应用进行调研，对公司内的业务场景抽丝剥茧，希望能够从中总结出一些统一的多模模式。其中，对于NLP的演进和理解难免管中窥豹，欢迎对该方向的感兴趣的同学一起讨论。

GNN的现状

业界现状

CNN是CV领域的大法宝，它通过引入参数化的卷积核大大提升了从邻域中提取特征的能力。然而，相比图像场景中排列整齐的矩阵，图（Graph）场景中的邻域则是不规则的，是非欧几里得空间。2016年Thomas N. Kipf提出的GCN[2]，从拉普拉斯矩阵出发首次将卷积过程引入非欧几里得空间，从此GNN进入了一个快速发展的阶段。Structure2Vec[3]、GraphSAGE[4]、GAT[5]等方法不断涌现出来。

Structure2Vec对邻居聚合的模式进行了修改，同时也引入边特征的聚合。

$$H^{(t+1)} = \sigma(D^{-1}AH^tW^t + XV^t)$$

GraphSAGE基于这些方法的共性，进一步提出了一个多层AGG(aggregator)的框架，支持了mean、max、LSTM等多种AGG函数。

$$h^{t+1} = \sigma(W^t \cdot \text{CONCAT}[\text{AGG}(\{h_j, \forall j \in \mathcal{N}(i)\}), h_i^t])$$

1. mean: $\text{AGG} = \sum_{j \in i} \frac{h_j^t}{|\mathcal{N}(i)|}$

2. max: $\text{AGG} = \max(\{W^t h_j^t, \forall j \in \mathcal{N}(i)\})$

3. LSTM: $\text{AGG} = \text{LSTM}(\{W^t h_j^t, \forall j \in \mathcal{N}(i)\})$

GAT[5]则结合了当时NLP领域内大红大紫的Attention技术，在聚合过程中引入权重，一方面间接的实现了邻居的采样，另一方面也有效的解决了随着网络层数加深，卷积结果过平滑的问题。

ALPS-GraphML



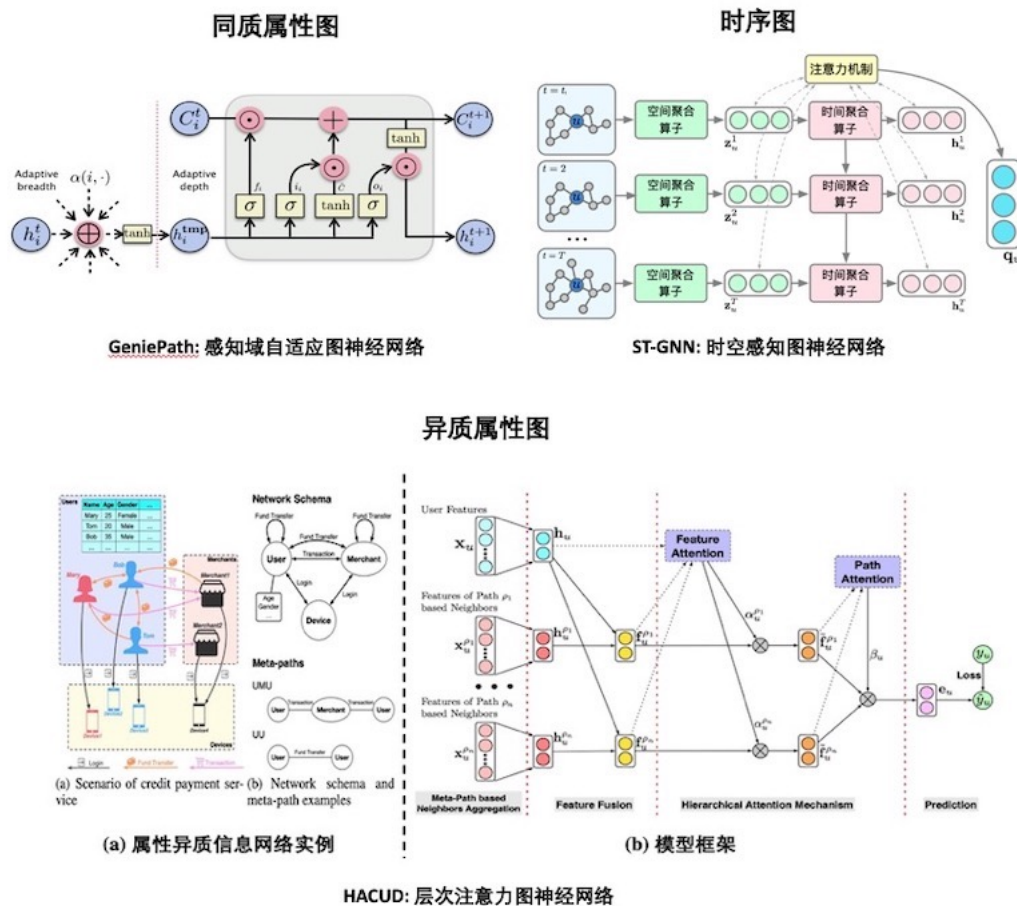
17年初，ALPS-GraphML团队成立，尝试将学术界的GNN算法落地在蚂蚁内拥有大规模图数据的金融场景中。经过几年的成长，平台除了沉淀了上述业界主流算法之外，也在支付、信贷、保险、财富、风控等金融核心业务的智能化建设中遇到了一些特色问题，对于图的表达也提出了更高的要求：

- 动则节点上亿，边几百亿的图数据规模如何学习、预测？
- 不同团队、不同业务需求，构图过程和点边特征完全不同，海量数据如何存储？
- 业务图数据中有大量噪声，有用的信息可能隐藏在更深层的链接中，如何正确的学习到深层的网络关系？
- 用户/商户/评论等不同语义的节点，转账/交易/评论等不同的类型的边，这些共同构建的一张大图，如何有效地利用这些图中异质信息？
- 随着时间的变化，图信息随之演变，这些随时间演变的信息如何有效的利用在图学习过程中呢？



针对上述1、2问题，GraphML工程团队设计了GraphFlat[6]方案，一方面针对海量图数据场景，打通了“数据预处理-图学习-图预测”的全链路；另一方面，该方案基于海量廉价的磁盘空间支持了多图的数据存储，训练过程数据分离，支持多任务的并行训练。

针对3、4、5三个问题，GraphML算法团队分别设计“Geniepath：感知域自适应图神经网络算法”[7]、“HACUD(HeGNN)：面向异质属性图的层次注意力图神经网络”[8]和“ST-GNN：面向时序图的时空感知图神经网络”[9]。在信息聚合时，能自适应的过滤噪声，传递更深层的信息；在图表达上，支持异构和时序等更加丰富的图数据。



NLP+GNN的多模探索

随着深度学习的发展，NLP预训练领域近年来也实现了从简单模型word2vec到GPT2、Bert等复杂模型的进化[10]，Bert的成功催生了后续大量的改进版。其中，Bert与知识的结合是众多扩展领域中备受关注的的一个[11]，ERNIE-baidu、BRNIE-stinghua、K-Bert、KnowBERT、SemBERT等算法都尝试将知识引入Bert的学习过程，从将单词作为mask，到引入知识谱图，再到利用图谱里面的图结构、实体/边特征信息，NLP预训练过程不断在向与Graph结合的方向演进。

另一方面，在端到端的NLP任务中，GNN的方法也在不断引入学习过程[1, 12]，典型的任务包括：

- 文本分类：基于单词共现构图后，利用GNN将图信息编码到单词的表达中，进而用于文本分类。
- 机器翻译：基于句子的语法依存树构图，将GCN中的邻接矩阵变化为句法依存的权重转化矩阵进行训练



- 关系抽取（实体发现和分类）：同样基于单词共现进行构图，利用GNN对单词进行编码，后接其他网络进行关系预测
- 事实验证：利用GAT的实现的的不同线索之间的信息聚合和推理能力
- 其他：语义角色标注、文本生成、阅读理解。

Area	Application	Algorithm	Deep Learning Model	References
Text	Text classification	GCN	Graph Convolutional Network	[1], [23], [48] [2], [22], [46]
		GAT	Graph Attention Network	[68]
		DGCNN	Graph Convolutional Network	[106]
		Text GCN	Graph Convolutional Network	[107]
		Sentence LSTM	Graph LSTM	[62]
	Sequence Labeling (POS, NER)	Sentence LSTM	Graph LSTM	[62]
	Sentiment classification	Tree LSTM	Graph LSTM	[60]
	Semantic role labeling	Syntactic GCN	Graph Convolutional Network	[108]
	Neural machine translation	Syntactic GCN	Graph Convolutional Network	[109], [110]
		GGNN	Gated Graph Neural Network	[38]
	Relation extraction	Tree LSTM	Graph LSTM	[111]
		Graph LSTM	Graph LSTM	[44], [112]
	Event extraction	GCN	Graph Convolutional Network	[113]
		Syntactic GCN	Graph Convolutional Network	[114], [115]
	AMR to text generation	Sentence LSTM	Graph LSTM	[116]
		GGNN	Gated Graph Neural Network	[38]
	Multi-hop reading comprehension	Sentence LSTM	Graph LSTM	[117]
	Relational reasoning	RN	MLP	[96]
		Recurrent RN	Recurrent Neural Network	[118]
		IN	Graph Neural Network	[4]



NLP+Alps-GraphML在蚂蚁的尝试

鉴于业务落地的实际需求和复杂模型训练的难度，我们所知的蚂蚁业务在使用NLP+Graph多模态的时候，大多采用的是模型解耦的二阶段训练模式，即NLP模型先训练得到的Embedding，随后在构图中将Embedding作为节点或边的特征，随后利用GNN的编码能力再进行任务相关的训练。另一种解耦的模式是，NLP进行预训练，随后利用GNN和任务去Fine-tune整体的模型。

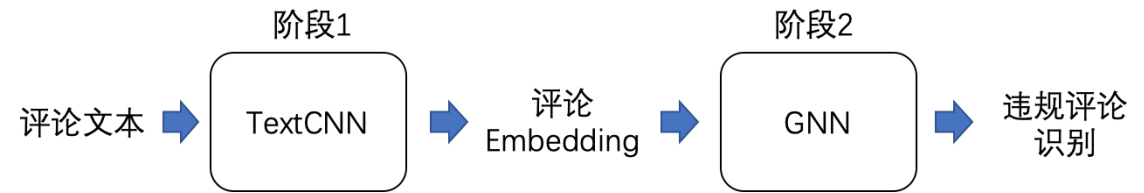
违规留言识别

背景：在口碑商家详情页面的用户评论中，存在一些描述违规广告的违规评论，识别并剔除这部分评论是口碑安全保障中的重要环节。安全风控同学在分析了打标数据之后，发现违规评论除了评论本身存在语义特征之外，还存在以下几个特点：1. 违规评论在部分行业的商家评论中存在聚集性；2. 违规评论与发布用户存在多对一的关系；3. 二级评论（留言回复）往往是违规评论的重灾区。从上面三个特点分析，可以看出如何利用商家与评论的关联，用户与评论的关联，评论与评论的关联是提高识别准确率的关键，而这些关系最自然的表达便是考虑到节点属性的差别的异构图。

方案简述：

评论识别的全链路采用二阶段模型解耦的训练模式：

1. 基于TextCNN将原始评论文本转化为评论Embedding；
2. 利用评论、二级评论、评论用户、商家之间的关联关系构图，并将评论Embedding作为评论/二级评论的节点特征，同时构图过程还需保留4种节点类别特征，随后采用异构图GNN算法HACUD进行建模，训练样本。



模型拆分后，GNN的模型训练过程非常纯粹，POC到上线阶段直接复用了Alps-GraphML提供的全链路能力，具体包括：

1. 数据预处理：基于GraphFlat预处理方案计算得到图特征
2. 模型训练：采用HACUD(HeGNN)进行端到端训练
3. 模型打分：



- i. 离线：使用ALPS提供的ODPS大规模离线打分方案
- ii. 在线：HBase导入H+1图特征后，基于Arks进行在线打分

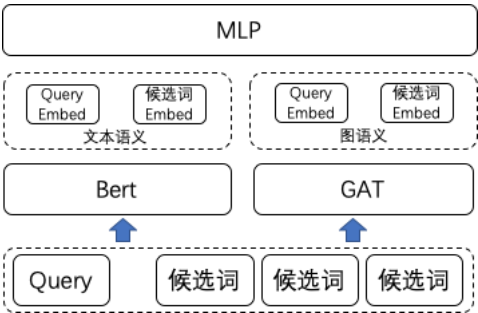
短文本匹配

背景：搜索词的短文本匹配是支付宝搜索推荐的重要一环，如何根据用户输入的优先信息，准确地召回和推荐支付宝提供的服务是这里要解决的核心问题。全流程一般包含2个阶段：1. 粗匹配：基于粗匹配算法得到圈选候选集；2. 精匹配：将搜索词与候选集输入精匹配模型，获得最佳匹配结果。在精匹配阶段，搜索词与候选集之间往往无法直接构建关联，单纯的分析语义信息以不足以得到很好的匹配结果，而刻画这种间接关系最优先的便是通过图结构和GNN提供的传播能力。



方案简述：NLP结合Graph的过程依然采用一种两阶段的训练模式，不同在于，在第二阶段，NLP预模型与GNN模型联合训练，进行Fine-tune。下面仅简述NLP与GNN联合训练阶段：

- 1. NLP模型部分选择加载预训练的Bert进行初始化。将搜索词与匹配项分别通过Bert编码得到文本语义Embedding
- 2. 搜索词与匹配项从知识图谱中获取邻居信息，基于GAT编码图语义Embedding
- 3. 将文本语义Embedding和图语义Embedding结合，后接MLP编码任务标签信息（是否匹配）



上述方案还在POC阶段，在第二阶段的训练过程中，ALPS-NLP和ALPS-Graph两大模块通过模型桥接代码组合，极大地简化建模流程，同时可充分复用已有能力。

结语

本文对于NLP与Graph的结合探索还在初级阶段，相信在蚂蚁丰富的业务矩阵中还存在更多NLP+GNN结合的场景，其训练模式和结合模式也有广阔的探索空间。

BERT的预训练过程是否能够使用GNN更有效利用知识？

特定目标的任务中，二阶段的NLP+Graph模型转化为端到端模型是否会有更好的效果？

如果你对NLP+GNN如何使用还不确定，不如我们一起来聊聊~

参考文献

[1] 清华大学整理的GNN必读论文及应用分类：<https://github.com/thunlp/GNNPapers>
(<https://github.com/thunlp/GNNPapers>)

[2] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2016 Sep 9.

[3] Dai H, Dai B, Song L. Discriminative embeddings of latent variable models for structured data. InInternational Conference on Machine Learning 2016 Jun 11 (pp. 2702-2711).

[4] GraphSAGE: <http://snap.stanford.edu/graphsage/> (<http://snap.stanford.edu/graphsage/>)

[5] GAT: <https://github.com/PetarV-/GAT> (<https://github.com/PetarV-/GAT>)

[6] GraphFlat ata: <https://www.atatech.org/articles/149482?spm=ata.13269325.0.0.f79549faHY8DLu>
(<https://www.atatech.org/articles/149482?spm=ata.13269325.0.0.f79549faHY8DLu>)



[7] Geniepath ata: <https://www.atatech.org/articles/116737?spm=ata.13261165.0.0.366f79carP3KcT>
(<https://www.atatech.org/articles/116737?spm=ata.13261165.0.0.366f79carP3KcT>)

[8] HeGNN ata: <https://www.atatech.org/articles/149555?spm=ata.13269325.0.0.4c7849fa11rYbz>
(<https://www.atatech.org/articles/149555?spm=ata.13269325.0.0.4c7849fa11rYbz>)




[9] ST-GNN ata: <https://www.atatech.org/articles/166667?spm=ata.13269325.0.0.6b6d49faLboD0U>
(<https://www.atatech.org/articles/166667?spm=ata.13269325.0.0.6b6d49faLboD0U>)

[10] NLP的发展历程: <https://zhuanlan.zhihu.com/p/49271699> (<https://zhuanlan.zhihu.com/p/49271699?spm=ata.13261165.0.0.6a355ceeHbFB0L>)

[11] 知识赋能NLP: <https://developer.aliyun.com/article/741285?spm=a2c6h.12873581.0.0.5f43187aUgqwiz>
(<https://developer.aliyun.com/article/741285?spm=a2c6h.12873581.0.0.5f43187aUgqwiz>)

[12]图神经网络在NLP中的应用: http://nlp.csai.tsinghua.edu.cn/~yangcheng/publications/SMP2019_yc.pdf
(http://nlp.csai.tsinghua.edu.cn/~yangcheng/publications/SMP2019_yc.pdf)



评论文章 (2)  7 (/articles/180451/voteup)  0  4 收藏 (/articles/180451/mark/)

他们赞过该文章


轩与 (/users/24012) 溪亭 (/users/675331) 玄若 (/users/449667) 百策 (/users/296029) 清砚 (/users/118846) 石铎 (/users/557009)
皇羲 (/users/272809)

上一篇: XGRep: 基于随机游走的分布式图表达框架 | AI特...

- 1F 皇羲 (/users/272809)

2020-09-08 20:25:29

赞!

 0 (/comments/304614/voteup)  0
- 2F 石铎 (/users/557009)

2020-09-08 20:27:10

赞!

 0 (/comments/304615/voteup)  0

写下你的评论...



评论

