CogDL: 基于图的深度学习的研究工具

知识工程实验室 (KEG)

清华大学，北京

CogDL 是什么

CogDL 是由清华大学计算机系知识工程实验室（KEG）的唐杰教授所主持的基于图的深度学习的研究工具，基于 Python 语言和 Pytorch 库进行开发。CogDL 允许研究人员和开发人员可以轻松地训练和比较基线算法或自定义模型，以进行结点分类，链接预测，图分类，社区发现等基于图结构的任务。它提供了许多流行模型的实现，包括：非图神经网络算法例如 Deepwalk、LINE、Node2vec、NetMF、ProNE、methpath2vec、PTE、graph2vec、Infograph 等；图神经网络算法例如 GCN、GAT、GraphSAGE、FastGCN、GTN、HAN、GIN、DiffPool 等。它也提供了一些下游任务，包括结点分类（具有或不具有结点属性），链接预测（具有或不具有属性，边异构或不异构），图分类（有监督或无监督）以及为这些任务构建各种算法效果的排行榜。

CogDL 具有以下特性:

- 任务导向：CogDL 以图上的任务为主，提供了相关的模型、数据集以及我们得到的排行榜。
- 一键运行：CogDL 支持用户使用多个 GPU 同时运行同一个任务下多个模型在多个数据集上的多组实验。
- 多类任务：CogDL 支持同构/异构网络中的节点分类和链接预测任务以及图分类任务。
- 可扩展性：用户可以基于 CogDL 已有的框架来实现和提交新的数据集、模型和任务。

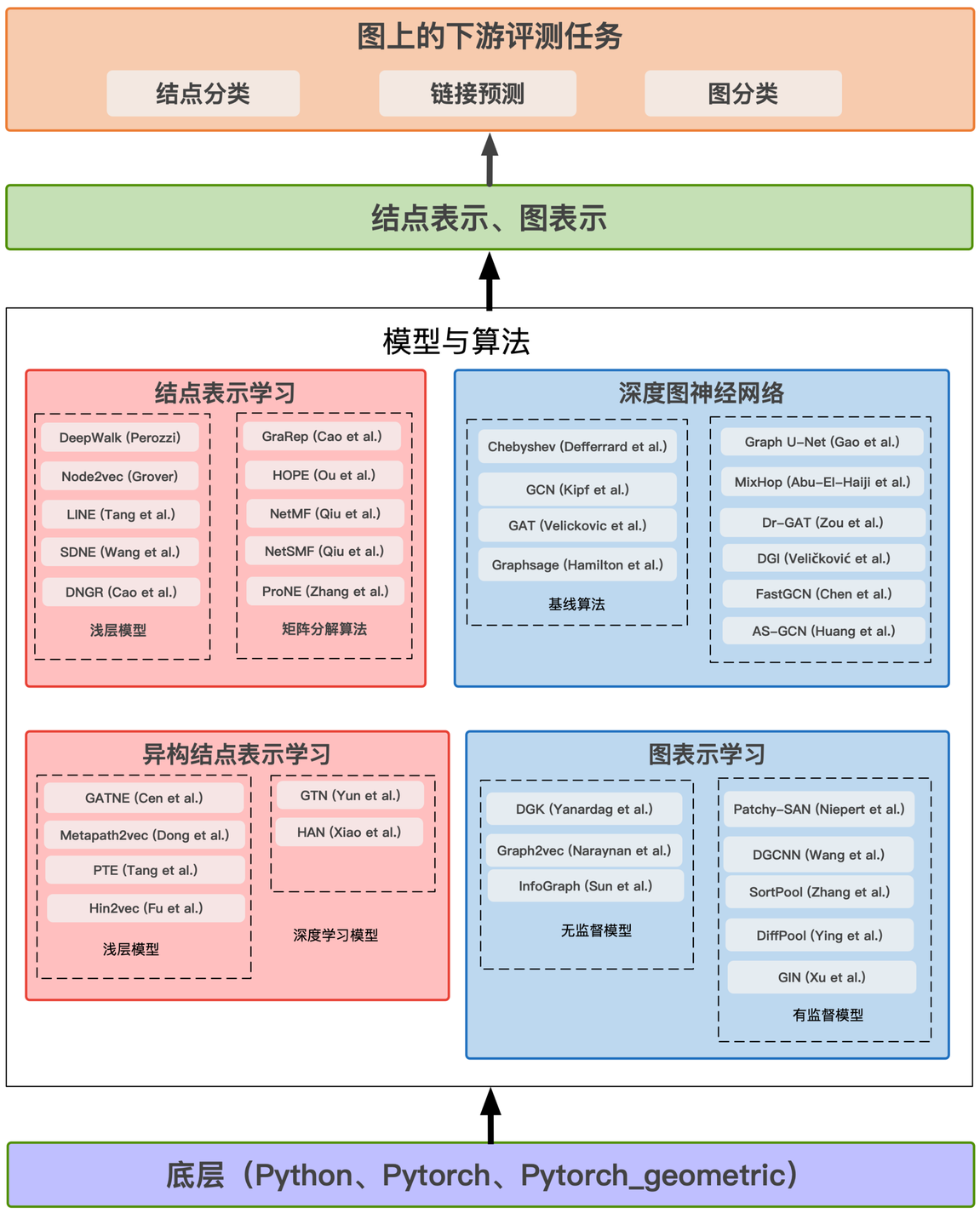

图 1: CogDL 的整体框架

开始使用

基本用法:

python train.py --task example_task --dataset example_dataset --model example_method 来在 example_data 上运行 example_method，并使用 example_task 评测其效果。

- --task, 运行的任务名称，像 node_classification, unsupervised_node_classification, link_prediction 这样来评测表示质量的下游任务。更多任务集在 cogdl/task。
- --dataset, 运行的数据集名称，可以是以空格分隔开的数据集名称的列表, 现在支持的数据集包括 cora, citeseer, pumbed, PPI, wikipedia, blogcatalog, dblp, flickr 等。更多数据集在 cogdl/datasets.
- --model, 运行的模型名称, 可以是个列表，支持的模型包括 gcn, gat, deepwalk, node2vec, hope, grarep, netmf, netsmf, prone 等。更多模型在 cogdl/models.

对于每个算法的具体参数，可以参考 https://github.com/THUDM/cogdl/tree/master/cogdl/models.

自定义数据集或模型

如果你有一个表现良好的算法或者独特的数据集并且想要发布出来，你可以将你的模型实现或数据集通过在我们的仓库中新开一个 github issue 或者在我们的页面上评论来进行提交。提交的内容可以包括：

- 提交你前沿的算法的结果
- 添加你自己的数据集
- 在 CogDL 中实现你自己的模型

LeaderBoard

CogDL 提供了一些下游任务，包括结点分类（具有或不具有结点属性），链接预测（具有或不具有属性，异构或非异构）和图分类（有监督或无监督）任务。我们建立了几个排行榜，这些排行榜列出了各类算法在这些任务上的最新结果。

无监督多标签结点分类

这是一个根据无监督的多标签结点分类设置而构建的排行榜，我们在几个真实的数据集上运行 CogDL 上的无监督表示学习算法，并将输出的表示和 90 %的结点标签作为经 L2 归一化的逻辑回归中的训练数据，使用剩余 10 %的标签作为测试数据，计算并按照 Micro-F1 的大小进行排序。

Rank	Algorithm	PPI	Blogcatalog	Wikipedia
1	ProNE(Zhang et al, IJCAI'19)	**26.32**	**43.63**	57.64
2	NetMF(Qiu et al, WSDM'18)	24.86	43.49	**58.46**
3	Node2vec(Grover et al, KDD'16)	23.86	42.51	53.68
4	NetSMF(Qiu et at, WWW'19)	24.39	43.21	51.42
5	DeepWalk(Perozzi et al, KDD'14)	22.72	42.26	50.42
6	LINE(Tang et al, WWW'15)	23.15	39.29	49.83
7	Hope(Ou et al, KDD'16)	23.24	35.52	52.96
8	SDNE(Wang et al, KDD'16)	20.14	40.32	48.24
9	GraRep(Cao et al, CIKM'15):	20.96	34.35	51.84
10	DNGR(Cao et al, AAAI'16):	16.45	28.54	48.57

半监督有属性的结点分类

下面是几种常见的图神经网络算法在半监督结点分类任务上构建的排行榜。我们在经典的三个数据集 Cora, Citeseer 和 Pubmed 进行了实验，以 Accuracy 指标来评价模型的效果。

Rank	Method	Cora	Citeseer	Pubmed
1	Graph U-Net (Gao et al., 2019)	**84.4 ś 0.6**	**73.2 ś 0.5**	79.6 ś 0.2
2	MixHop (Abu-El-Haija, ICML'19)	81.9 ś 0.4	71.4 ś 0.8	**80.8 ś 0.6**
3	DR-GAT (Zou et al., 2019)	83.6 ś 0.5	72.8 ś 0.8	79.1 ś 0.3
4	GAT (Velikovi et al., ICLR'18)	83.0 ś 0.7	72.5 ś 0.7	79.0 ś 0.3
5	DGI (Velikovi et al., ICLR'19)	82.3 ś 0.6	71.8 ś 0.7	76.8 ś 0.6
6	GCN (Kipf et al., ICLR'17)	81.4 ś 0.5	70.9 ś 0.5	79.0 ś 0.3
7	GraphSAGE (Hamilton, NeurIPS'17)	80.1 ś 0.2	66.2 ś 0.4	76.9 ś 0.7
8	Chebyshev (Defferrard, NeurIPS'16)	79.2 ś 1.4	69.3 ś 1.3	68.5 ś 1.2

CogDL 网站及代码 http://keg.cs.tsinghua.edu.cn/cogdl/ and https://github.com/thudm/cogdl