**贵州大学计算机科学与技术学院**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **《大数据知识工程》课程实验报告** | | | | | |
| **题目** | 基于 PyTorch 的知识图谱表示学习框架。 | **姓名** | 田旭 | **日期** | 2025.3.29 |
| **实验环境：**  虚拟机操作系统 ubuntu-22 | | | | | |
| **实验内容与完成情况：**  **3.1 实验目的**  本次实验旨在：  掌握知识图谱表示学习的工具 OpenKE，熟悉其基本操作和模型实现。  学习使用 TransE、RESCAL 等经典模型进行知识图谱推理。  了解常用数据集（FB15k、WN18、FB15k-237、WN18RR）的特点及其在知识图谱补全中的应用。  理解评价指标（Hit@k、MR、MRR）的意义，并学会评估模型性能。  理解基解基于表示学习知识图谱推理与补全方法。  **3.2 实验平台**  本次实验主要使用的工具和库包括：  **OpenKE**：清华大学开源的知识图谱表示学习框架，支持多种经典模型（如 TransE、RESCAL）和数据集  （如 FB15k、WN18）。  **PyTorch**：深度学习框架，用于实现和训练模型。  **Datasets**：主要使用 FB15k-237 和 WN18RR。  **3.3 实验内容和要求**  请大家酌情调参。另外，请大家选择翻译模型 TransE 或张量分解模型 RESCAL（任选一个模型）分别对常  用数据集（FB15k、WN18、FB15k-237、WN18RR）进行运行并记录相关的结果。  **请大家在自己的电脑编译模型，否则容易出 bug。** | | | | | |
| **三、实验完成情况及实验结果展示**  1.git项目(OpenKE)到本地，python环境配置到本地，激活虚拟环境(虚拟环境名:chen\_home)     1. 执行过程截图      1. 执行相关运行代码为   加载FB15K 数据集进行测试  **测试代码**  from openke.config import Trainer, Tester from openke.data import TrainDataLoader, TestDataLoader from openke.module.model import TransE from openke.module.loss import MarginLoss from openke.module.strategy import NegativeSampling import os *# 1. 加载数据* train\_dataloader = TrainDataLoader(  in\_path = "./benchmarks/ FB15K/",  nbatches = 100,  threads = 8,  sampling\_mode = "normal",  bern\_flag = 0,  filter\_flag = 1,  neg\_ent = 25,  neg\_rel = 0 ) test\_dataloader = TestDataLoader("./benchmarks FB15K/", "link")  *# 2. 定义模型* transe = TransE(  ent\_tot = train\_dataloader.get\_ent\_tot(),  rel\_tot = train\_dataloader.get\_rel\_tot(),  dim = 200,  p\_norm = 1,  norm\_flag = True ) model = NegativeSampling(  model = transe,  loss = MarginLoss(margin = 1.0),  batch\_size = train\_dataloader.get\_batch\_size() )  *# 3. 训练* trainer = Trainer(model=model, data\_loader=train\_dataloader, train\_times=1000, alpha=0.001, use\_gpu=True) trainer.run()  *# 4. 保存模型* os.makedirs('./checkpoint\_tx', exist\_ok=True) transe.save\_checkpoint('./checkpoint\_tx/transe.pth')  *# 5. 测试* transe.load\_checkpoint('./checkpoint\_tx/transe.pth') tester = Tester(model=transe, data\_loader=test\_dataloader, use\_gpu=True) tester.run\_link\_prediction(type\_constrain=False) mrr, mr, hit10, hit3, hit1 = tester.run\_link\_prediction(type\_constrain=False) print("MRR:", mrr) print("MR:", mr) print("Hit@10:", hit10) print("Hit@3:", hit3) print("Hit@1:", hit1)  **指标解释**  **表1 指标说明表**   |  |  |  | | --- | --- | --- | | 指标名称 | 英文全称 | 含义说明 | | MRR | Mean Reciprocal Rank | 平均倒数排名，即正确实体排名倒数的平均值，数值越高越好 | | MR | Mean Rank | 平均排名，越低表示预测越靠前，性能越好 | | Hit@10 | Hits at 10 | 排名前10中包含正确实体的比例，常用于衡量模型对高排名结果的覆盖能力 | | Hit@3 | Hits at 3 | 排名前3中包含正确实体的比例 | | Hit@1 | Hits at 1 | 排名第1就是正确实体的比例，要求最严格，体现模型的精准预测能力 |  1. **FB15K 数据集运行结果：**      1. **WN18数据集测试结果**      1. **FK15K237数据集测试结果**      1. **WN18RR数据集测试结果**     **问题与解决：**  **问题：**  **解决方案：**执行命令bash make.sh 之前需要先执行下列三条命令，不然会报错  apt-get update  apt-get install dos2unix  dos2unix make.sh    **改进的方向**  当前模型性能低下的主要原因是缺少实际训练过程，导致保存和测试的模型参数仍为随机值，评估结果接近随机猜测。为改进这一问题，应首先引入完整的训练流程，利用OpenKE中的Trainer模块配置学习率、训练轮数等参数进行训练。同时，可通过调整TransE模型的超参数（如嵌入维度、边界值、负样本数量和采样策略）进一步提升性能。在评估阶段，建议启用过滤机制并结合类型约束以提高指标的真实性与准确性。此外，还需确保数据集格式完整、ID映射正确，并开启GPU加速训练效率。通过上述改进，模型在WN18上的MRR有望达到0.7以上，Hit@10超过90%，显著优于当前结果。 | | | | | |