面向稠密向量检索的知识蒸馏方法研究与实现

答辩人：张涵

目录
CONTENTS

前期工作与总结

选题背景及依据

研究内容与方法

实验及预期成果

Part 1

前期工作与总结

前期工作与总结

课程学习

- 应用数理统计
- 机器学习

zhanghan9797 Create 11.25周报.md

..

image

10.14周报.md

10.21周报.md

10.28周报.md

10.9周报.md

11.11周报.md

11.17周报.md

11.25周报.md

11.4周报.md

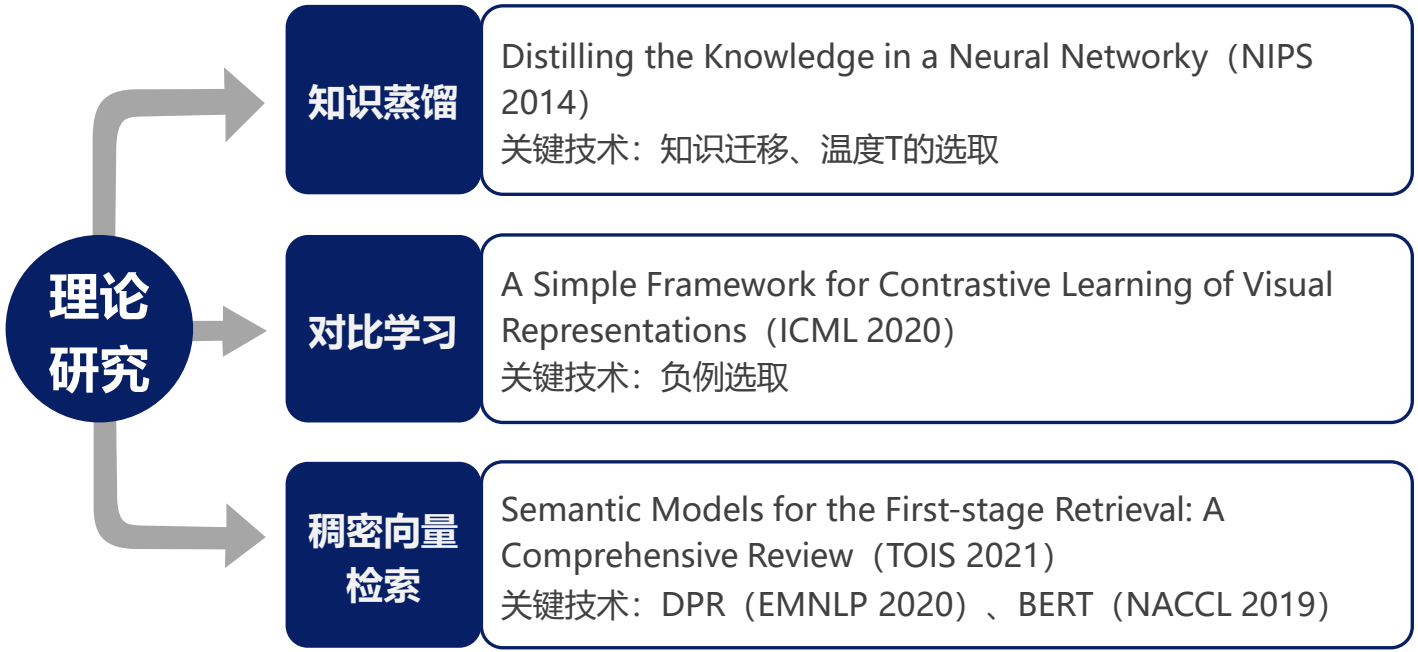
9.23周报.md

9.30周报.md

组会交流

- 每周组会
- 周报制度

理论研究



Karpukhin et al., Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.
Xiong et al., Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. ICLR 2021.

Part 2

选题背景及依据

选题背景及依据

选题背景

- 在数据爆炸式增长的数字时代，各个国家都在加速数字产业布局，网页文档数量达到百亿级别。
- 信息检索的目的是从大规模文档集合中选择文档以满足用户需求。
- 信息检索有很多应用，通常被视为一些 NLP 任务的第一阶段，例如事实验证和问答。

查询和文档之间的相关性

选题背景及依据

- 词汇不匹配问题
  - Q: How many people live in Sydney?
    - ➢ Sydney's population is 4.9 million
      [relevant, but missing 'people' and 'live']
    - ➢ Hundreds of people queueing for live music in Sydney
      [irrelevant, and matching 'people' and 'live']

传统信息检索模型

- 词汇不匹配问题
- 对查询和文档的浅层理解

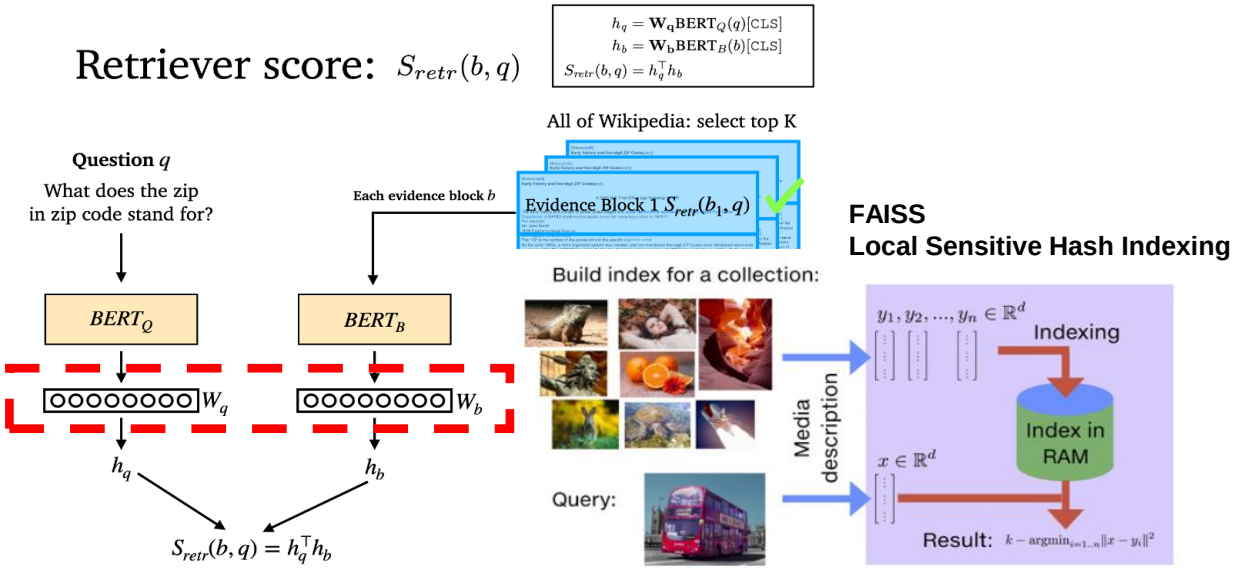
应用神经网络

神经信息检索模型

- 克服传统模型存在的问题
- 预训练的模型进一步提高了模型的性能

选题背景及依据

稠密向量检索



缺点

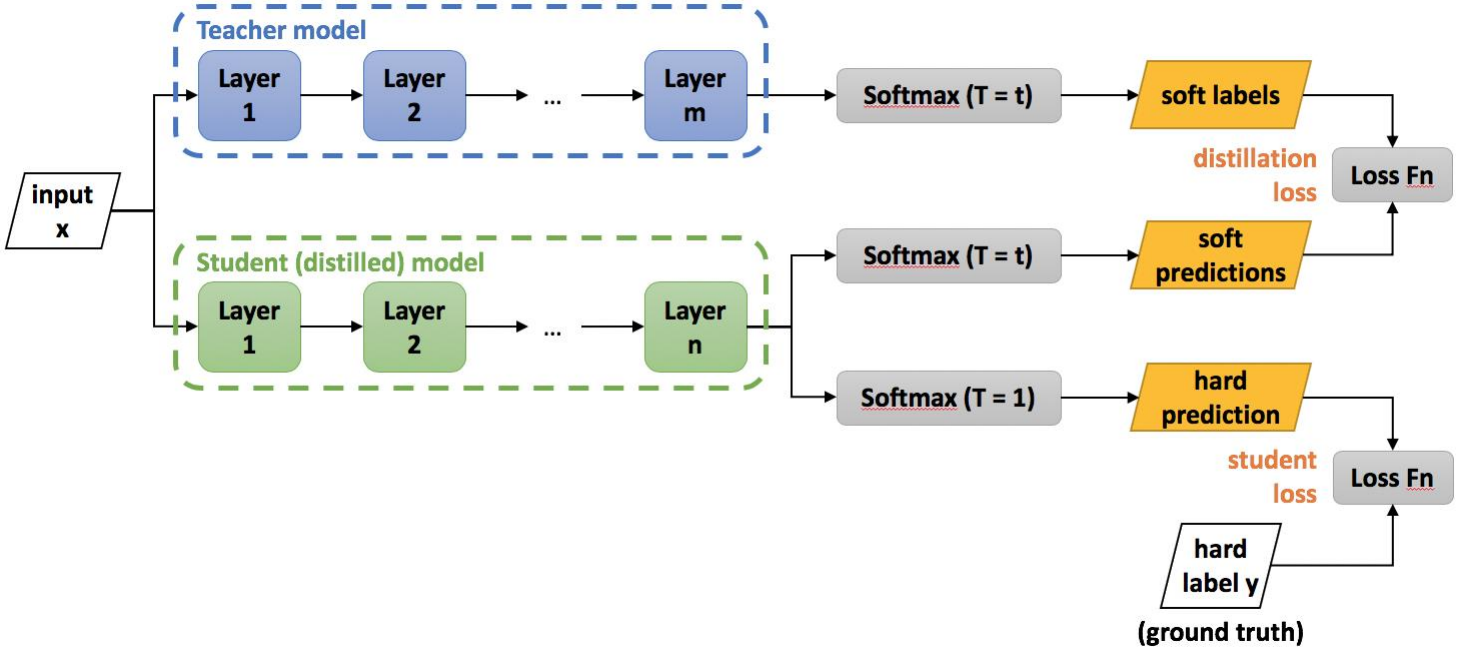
- 稠密向量的编码中存在冗余
- 占用大量不必要的空间和内存

Part 3

研究内容与方法

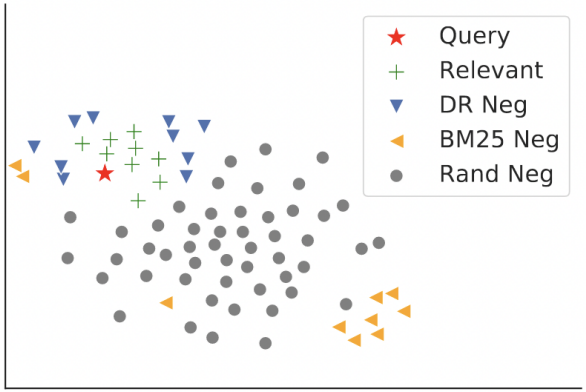
研究内容与方法

1.知识蒸馏方法

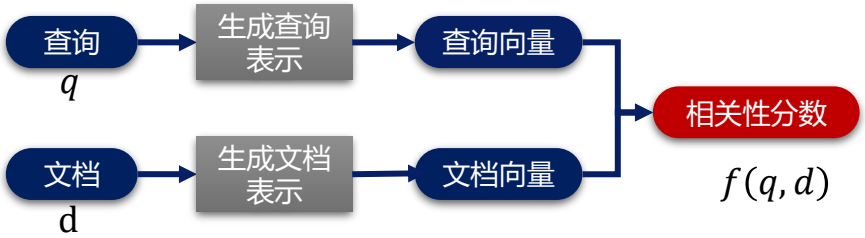


研究内容与方法

2.对比学习

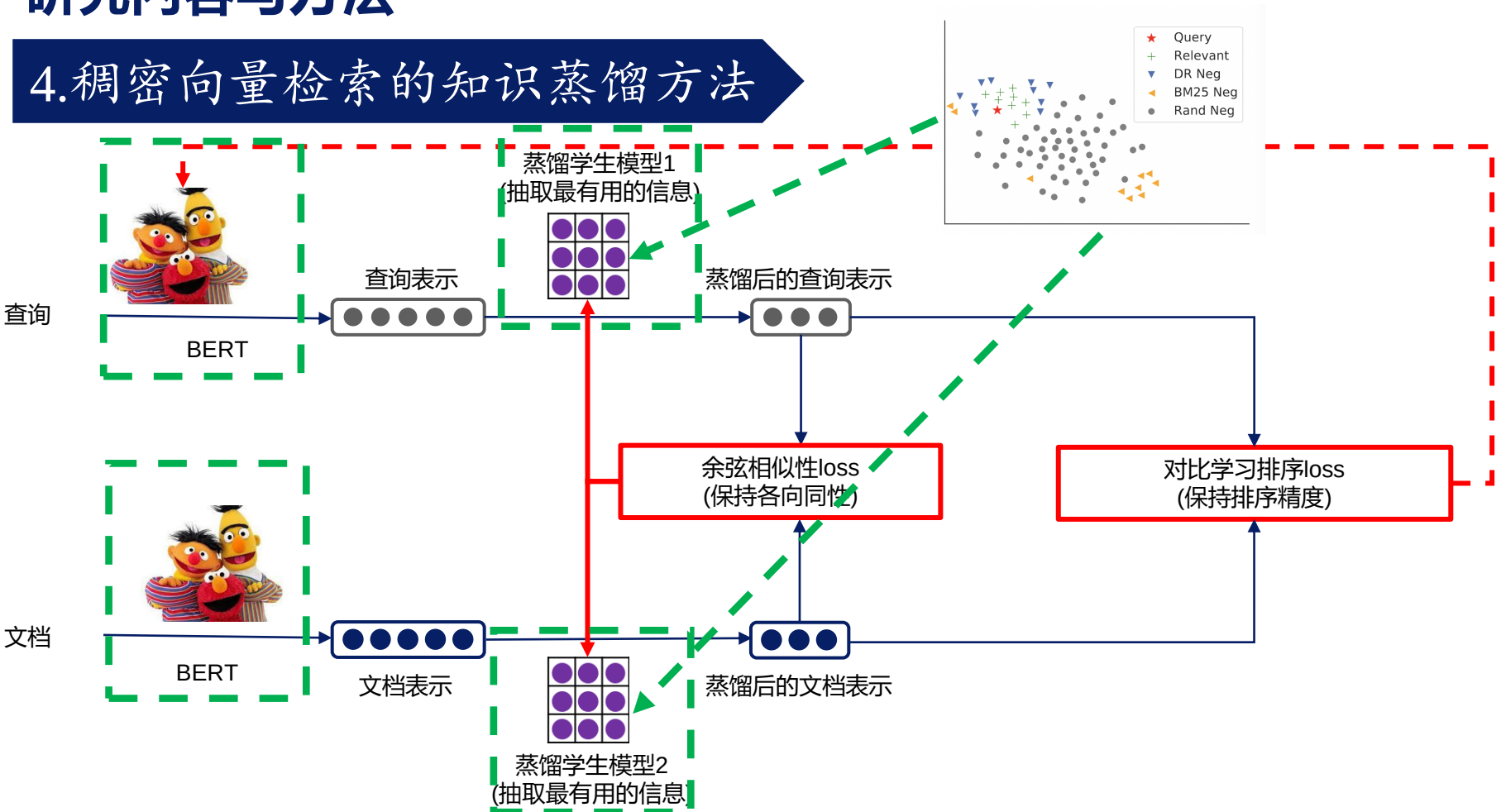


3.稠密向量检索



研究内容与方法

4.稠密向量检索的知识蒸馏方法



Part 4

实验及预期成果

实验

A

数据集

MS MARCO：基于搜索引擎BING构建的大规模英文阅读理解数据集

NQ、TriviaQA、WQ、CuratedTREC、SQuAD：来自维基百科的五个问答数据集

B

基线

DPR（EMNLP 2020）

BERT（NACCL 2019）

C

评估指标

平均倒数排名MRR

召回率Recall

MS MARCO https://microsoft.github.io/msmarco/

Devlin, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NACCL 2019.

Karpukhin et al., Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.

实验

DPR-PCA256-PQ2

```
(pytorch) neulab@omnisky01:/home/hdd/pyserini/DPR_Index_Compression$ bash pyserini.sh
python -m pyserini.eval.evaluate_dpr_retrieval --retrieval /home/hdd/pyserini/DPR_Inde
100%|
Top20 accuracy: 0.7628808864265928
Top100 accuracy: 0.8476454293628809
```

Top 20(paper)	Top 100(paper)
76.2 (74.8)	84.7 (84.1)

DPR-PCA256

```
python -m pyserini.eval.evaluate_dpr_retrieval --r
100%|
Top20 accuracy: 0.7764542936288089
Top100 accuracy: 0.8559556786703602
```

Top 20(paper)	Top 100(paper)
77.6 (77.2)	85.6 (85.5)

复现ANCE结果

```
neulab@omnisky01:/home/hdd/pyserini/indexes$ python -m pyserini.e
.bf.tsv
Downloading https://raw.githubusercontent.com/castorini/anserini-
ssage_eval.py...
/home/neulab/.cache/pyserini/eval/msmarco_passage_eval.py already
Skipping download.
Running command: ['python', '/home/neulab/.cache/pyserini/eval/ms
set.txt', '/home/hdd/pyserini/runs/run.msmarco-passage.ance.bf.ts
Results:
####################
MRR @10: 0.3301573657161042
QueriesRanked: 6980
####################
```

```
$ python -m pyserini.eval.msmarco_passage_eval
######################
MRR @10: 0.3302
QueriesRanked: 6980
######################
```

Ma et al., Simple and Effective Unsupervised Redundancy Elimination to Compress Dense Vectors for Passage Retrieval. EMNLP 2021.

预期成果

创新点

- 利用知识蒸馏的方法完成对稠密向量的降维操作
- 为了优化训练过程，采用对比学习的训练方式对训练负例进行精确选取
- 应用大型预训练语言模型，保证实验结果的精度

预期成果

- 完成顶级会议论文一篇
- 完成发明专利一项

工作进度安排

序号	阶段及内容	工作量估计（时数）	起止日期	阶段研究成果
1	课题方向研究和选择	400	2021.9-2021.10	收集资料，确定研究方向
2	撰写开题报告，明确研究内容	200	2021.11	撰写开题报告，完成开题答辩
3	阅读论文并复现论文中的实验	600	2021.12-2022.2	完成对前人实验工作的整理
4	探索预训练语言模型微调后在信息检索任务上的效果	600	2022.3-2022.6	完成对预训练语言模型的分析报告
5	设计压缩稠密向量的策略和模型	600	2022.7-2022.9	完成对蒸馏模型的分析报告
6	探索利用对比学习进行训练后在信息检索任务上的效果	600	2022.10-2022.12	完成对比学习策略的分析报告
7	总结已使用方法的优点与不足	400	2023.1-2023.2	具体地实验效果分析报告和论文
8	整理之前的工作并指出后继的可改进的地方	200	2023.3	进行一次组会报告
9	对研究工作进行整理并撰写论文	400 合计 4000	2023.4-2023.5	完成毕业论文