本科毕业设计（论文）中期检查报告

专业：人工智能　　班级：10051901

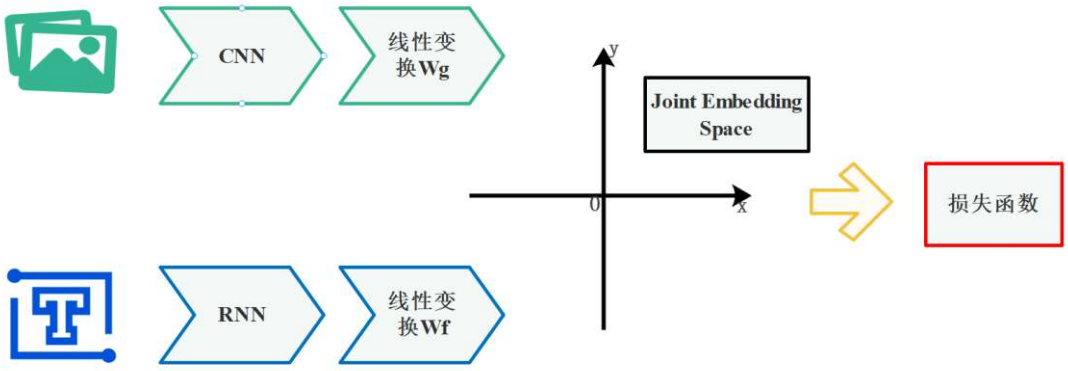
学号	2019302973	姓名	牛远卓	指导教师	牛凯
报告题目	基于深度学习的跨模态图像文本匹配研究				
报告日期	2023 年 4 月 18 日		报告地点		线上报告

中期报告（不少于 1500 字）

一、论文工作任务与进度安排情况

（1）论文工作任务

搭建一个可以用来图文匹配的深度学习神经网络。本课题采用卷积神经网络（例如 Resnet152 网络）与循环神经网络（例如 Gated Recurrent Unit，GRU 网络）分别提取图像和文本的特征，将两者的特征通过线性变换并归一化转换到同一语义空间，利用内积计算一对图文对的相似度。最终使用带有最负样本（Hard Negatives）的基于铰链的三联体排名损失（Triplet Ranking Loss）作为损失函数进行训练。测试时采取召回率（Recall）作为评测指标，具体流程框架如下图所示。



目前主要在现有模型的基础上进行优化改进，尝试接入不同神经网络提取图文特征，并且调整超参数以提高搜索准确率。

（2）进度安排情况

2022 年 12 月 1 日 - 2022 年 12 月 15 日，文献调研。学习深度学习基础

知识，了解卷积神经网络（Convolutional Neural Network, CNN），循环神经网络（Recurrent Neural Network, RNN）基本结构与作用。了解图文匹配概念，阅读并深刻理解图文匹配各种解决方法；已完成。

2022 年 12 月 16 日 - 2022 年 12 月 31 日，数据获取及预处理。从互联网获取 Flickr30k 和 MSCOCO 数据集，了解其中图像、文本以及标签的对应关系，随后通过编码解码的方式对数据进行预处理；已完成。

2023 年 2 月 1 日 - 2023 年 2 月 28 日，深度学习模型搭建及调试。配置深度学习模型运行所需环境，搭建对应 CNN 与 RNN 模型并测试；已完成。

2023 年 3 月 1 日 - 2023 年 3 月 31 日，深度学习模型搭建及调试。选用 Resnet152 代替 VGG19，对完整网络进行试运行并调整训练参数以获取更好的测试结果；已完成。

2023 年 4 月 1 日 - 2023 年 4 月 15 日，检索性能评估，模型改进。思考对现有模型的改进方向，通过查找文献了解各种方法的具体思路并进行优劣对比，决定出最佳优化方法并进行代码实现；已完成。

2023 年 4 月 16 日 - 2023 年 4 月 30 日，参数优化。对改进后的模型进行调参并对比不同参数下的召回率指标，选取具有最佳召回率的参数；正在进行。

2023 年 5 月 1 日 - 2023 年 5 月 31 日，结果整理，论文写作；未开始。

2023 年 6 月 1 日 - 2023 年 6 月 10 日，论文写作，答辩准备；未开始。

二、论文相关资料查阅情况

在数据集的选用与准备上，为了了解训练与测试所用数据集的具体组成与数据参数，阅读 Young 等人提出的 Flickr30k[1]数据集和 Lin 等人提出的 MSCOCO[2]数据集。Young 等人提出使用语义表征的视觉指称定义指称相似性矩阵。为了计算这些指称相似性，他们构建了一个指称图：基于 30K 图像和 150K 描述性标题的大型语料库的成分及其指称的层次结构。Lin 等人为了将目标检测扩展到场景理解，推出了一个简单的 91 种、328,000 张图片，并带有 250 万个标记的实例的数据集。

在深度学习网络搭建上，通过对使用 CNN-RNN 结构[3, 4, 5, 6]进行特征提取的论文进行阅读，了解到卷积神经网络以及循环神经网络在图文匹配中的作用。其中，CNN 用来提取图像的特征，而 RNN 用来提取文本特征，损失函数最小化

正确的图文对之间的距离，同时增大错误图文对之间的距离。在 CNN 的选择上，经过多次实验，在使用 Flickr30k 数据集的前提下，VGG19、ResNet50、ResNet101、ResNet152 中 ResNet152 取得的平均效果最好，故选用 ResNet152 来进行图像特征提取。RNN 部分则选用 GRU 结构。

在损失函数的选取上，正确的图文对应该相较于其他在语义空间中更近，因此使用三联体损失（Triplet Loss），这与学习排名问题[7]和最大边际结构预测[8，9]类似。当然，还有其他可以考虑的损失函数。其中之一是成对的铰链损失（Pairwise Hinge Loss）。它鼓励正确的图文对间的距离位于语义空间中半径为 ρ1 的超球内，而错误的对不应接近于 ρ2，其中 ρ2-ρ1>k>0，其中 k 为裕量值。

三、存在问题及采取的措施

1、对不同 CNN 进行测试，选用 VGG19 并调优后的召回率较为有限，将 CNN 模型更换为参数量更多的 ResNet152 并调优，召回率指标体现出明显的提升。因此，选择 ResNet152 并进行多轮 GPU 训练后效果有了明显提升。

2、最初在损失函数的选用上采用带有负样本之和的三联体损失，但测试时效果有限，随后更改为带有最负样本的三联体损失的方式使得召回率有了明显提升。

3、在采用不同优化器进行训练时，发现使用自适应梯度算法（Adaptive Gradient，Adagrad）进行优化时，在使用 Flickr30k 的情况下 R@1 仅为 2%左右，在学习了 Adagrad 的具体原理之后，推测是其将学习率调整为接近 0 的值，使得梯度下降不能进一步进行。于是，在相同数据集与超参数的情况下，选用收敛效果更好，速度更快的自适应矩估计（Adaptive Moment Estimation, Adam）优化方法，获得性能提升。

四、下一阶段论文工作安排

2023 年 4 月 22 日 - 2023 年 4 月 30 日，对模型进行训练参数优化，试图获取更好的训练结果。

2023 年 5 月 1 日 - 2023 年 6 月 10 日，对各项实验所得结果进行整理对比分析，随后开始进行论文写作并进行答辩准备。

五、参考文献

[1] Peter Young, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL 2014.

[2] Lin, et al. Microsoft COCO: Common Objects in Context. ECCV 2014.

[3] Li S, et al. Person Search with Natural Language Description. CVPR 2017.

[4] Li S, et al. Identity-Aware Textual-Visual Matching with Latent Co-Attention. ICCV 2017.

[5] Zhang Y, et al. Deep Cross-Modal Projection Learning for Image-Text Matching. ECCV 2018.

[6] Wang Z, et al. Vitaa: Visual-Textual Attributes Alignment in Person Search By Natural Language. ECCV 2020.

[7] Hang Li. Learning to rank for information retrieval and natural language processing. SLHLT 2014.

[8] Olivier Chapelle, et al. Large margin optimization of ranking measures. NIPS 2007.

[9] Quoc Le, et al. Direct optimization of ranking measures. arXiv 2007.

<table>
<tr><td>指导教师意见：

同意通过中期检查。

签名： 
2023 年 4 月 18 日</td></tr>
<tr><td>中期检查小组成员：

齐召帅 刘婷 牛凯</td></tr>
<tr><td>中期检查小组意见： 1. 进度: ☐ 较快；☑ 正常；☐ 较慢；☐ 太慢。
2. 质量: ☑ 较好；☐ 一般；☐ 较差；☐ 太差。
3. 结论: ☑ 合格；☐ 基本合格；☐ 不合格
（在相应的方块内作记号“ √ ”）</td></tr>
<tr><td>评 议
结 论</td><td>中期检查是否合格：☑ 合格；☐ 不合格
（在相应的方块内作记号“ √ ”）

同意通过中期检查。

评议小组组长签名：齐召帅 2023 年 4 月 18 日</td></tr>
</table>