本科毕业设计（论文）中期检查报告

专业：计算机科学与技术　　班级：10011805

学号	2018302379	姓名	史宇龙	指导教师	牛凯
报告题目	基于自然语言的跨模态行人重识别研究				
报告日期	2022 年 4 月 25 日		报告地点	在线腾讯会议	

中期报告

一、论文工作任务与进度安排情况

（1）论文工作任务：

搭建一个可以用文本来对行人图像进行检索的深度学习神经网络。本课题采用双向长短记忆网络(Bi-directional Long Short-Term Memory, Bi-LSTM)与深度残差网络(Deep Residual Network, Resnet)分别提取文本和图像的特征，后将两者的全局特征利用卷积核降维到同一维度并放入同一特征空间进行联合嵌入学习，利用跨模态投影匹配损失(Cross-Modal Projection Matching, CMPM)与跨模态投影分类损失(Cross-Modal Projection Classification, CMPC)对两者进行训练，使得正样本对在特征空间中的距离更近，而负样本对距离更远；最终使用余弦距离进行测试，获得以召回率(Recall)为指标的测试结果。

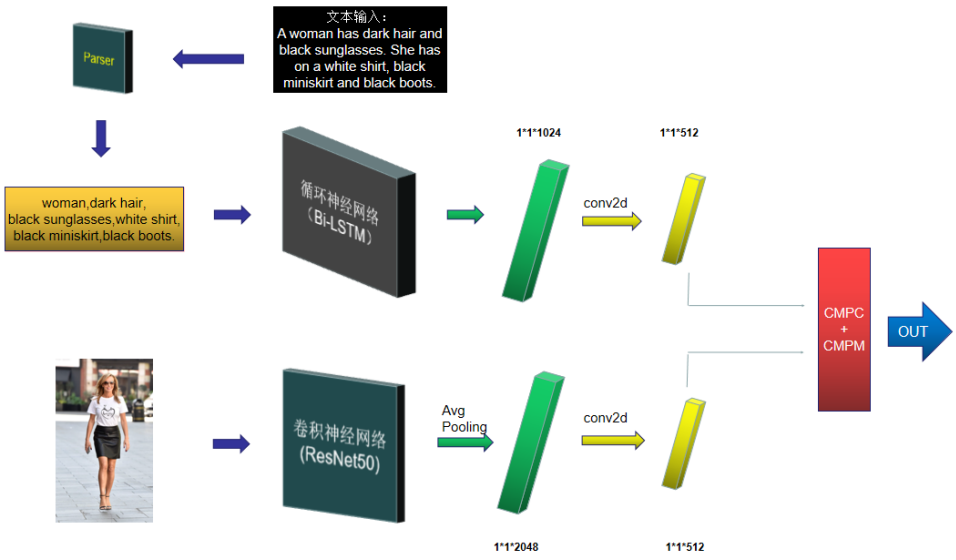

图 1: BiLSTM-ResNet 模型

随后在现有模型的基础上进行优化，将深度残差网络中的最后一层全局平均池化(Global Avg Pooling, GAP)及后续结构删去，替换为更能提取出显著特征的全局最大池化(Global Max Pooling, GMP)，随后加入门控模块(Gated Block, GB)试图消去最大池化带来的噪声，最后将提取的特征送入特征空间进行学习。

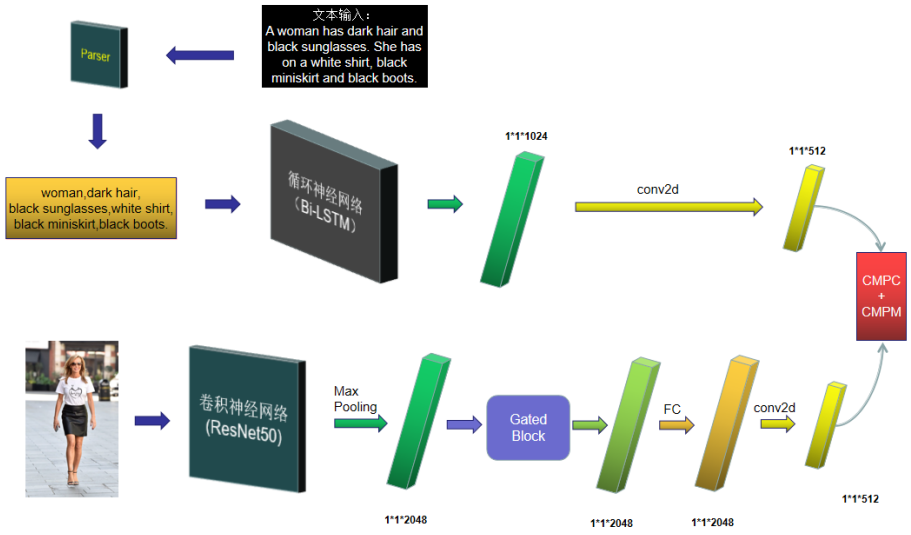

图 2: 改进后模型

（2）进度安排情况：

2021 年 12 月 1 日 - 2021 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。了解行人重识别概念，阅读并深刻理解基于自然语言的行人重识别的各种解决方法；已完成。

2021 年 12 月 16 日 - 2021 年 12 月 31 日，数据获取及预处理。从互联网获取 CUHK-PEDES 数据集，了解其中行人图像，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。

2022 年 1 月 1 日 - 2022 年 1 月 31 日，开题准备。学习基于文本的行人重识别的各种实现方法，并思考对于实现方法的选取，构思出初步的网络结构图；已完成。

2022 年 2 月 1 日 - 2022 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。

2022 年 3 月 1 日 - 2022 年 3 月 31 日，深度学习模型搭建及调试。加入 CMPM 与 CMPC 损失函数，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。

2022 年 4 月 1 日 - 2022 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。

2022 年 4 月 16 日 - 2022 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一组参数；正在进行。

2022 年 5 月 1 日 - 2022 年 5 月 31 日，结果整理，论文写作；未开始。

2022 年 6 月 1 日 - 2022 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 CUHK-PEDES 的具体组成与图像参数，阅读 Li 等人提出的 GNA-RNN 模型[1]。CUHK-PEDES 为第一个大规模行人重识别(Person Re-Identification, Re-ID)数据集，包含来自五个现有行人重识别数据集 CUHK03、MARKET-1501、CUHK-SYSU、VIPER 和 CUHK01 的 13003 人的 40206 张图像，每张图像都带有两个句子描述的注释，因此一共有 80412 个文本描述。

在深度学习网络搭建上，通过对使用 CNN-RNN 结构[1, 2, 3, 4]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文本-视觉匹配中的作用。CNN 用来提取行人图像的特征，而 RNN 用来提取文本特征，试图最小化属于相同身份的文本描述与行人图像之间的特征距离。其中，RNN 部分选用 LSTM，其相对于普通的 RNN 多出了门控装置，可以有选择地存储信息，因此更适用于文本特征提取。

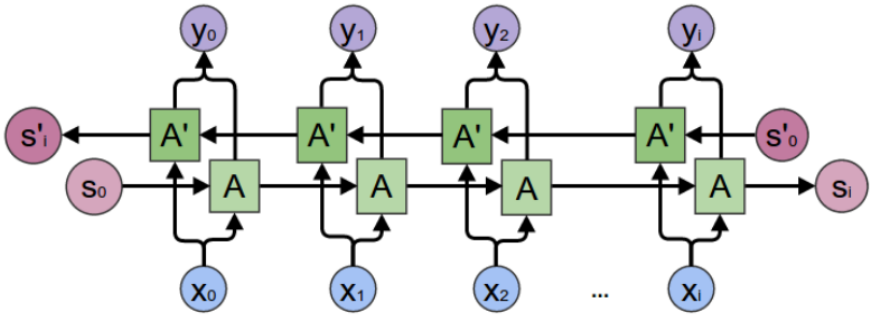

图 3: Bi-LSTM 结构

在阅读语义自对齐网络(Semantically Self-Aligned Network, SSAN)[5]结构时了解到双向长短记忆网络，其在结构上分为 2 个独立的 LSTM，输入序列分别以正序和逆序输入至 2 个 LSTM 神经网络进行特征提取，最后将 2 个输出向量进行拼接后形成的词向量作为该词的最终特征表达。在实际运用中 Bi-LSTM 由于可以编码从后到前的信息，进而能够对句子含义进行更细粒度的分类，因此在文本特征提取中往往会取得更好的效果。

在对 CNN 网络的选择上，经过实验，在使用 CUHK-PEDES 数据集的前提下，MobileNetV1、ResNet50、ResNet101、ResNet152 中 ResNet50 取得的效果最好，故选用 ResNet50 来进行图像特征提取。

在损失函数的选取上，使用跨模态投影匹配损失与跨模态投影分类损失[3]来衡量两种特征嵌入之间的距离。相较以往的排序损失(Ranking Loss)[6]，三联体损失(Triplet Loss)[7]以及跨模态交叉熵(Cross-Modal Cross-Entropy, CMCE)[2]等损失函数，其引入了跨模态特征投影并且不需要选定特定的三联体，在不同大小的批次下具有极大的稳定性，更适用于对图像-文本对的判别学习。

对 ResNet50-BiLSTM 模型进行优化时，将原结构中深度残差网络的最后一层平均池化换为最大池化，意图获得文本的全局信息，随后加入门控模块[8]来抑制最大池化引入的噪声并有选择地保留图像特征。

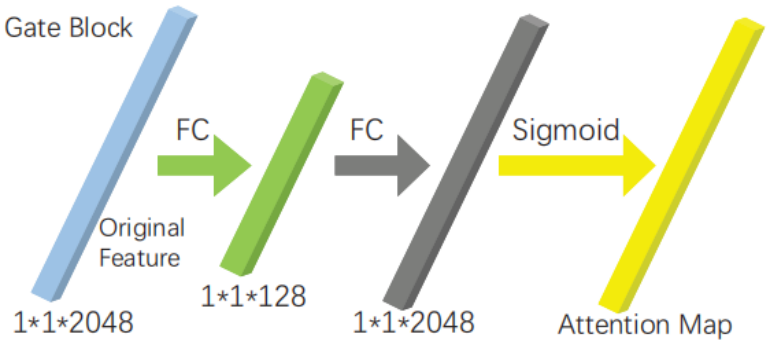

图 4: 门控模块(Gated Block)

三、存在问题及采取的措施

1、对基本模型训练后进行测试，得到的召回率相比同类型模型有较大差距，即使将 CNN 模型更换为参数量更多的 ResNet152，召回率指标也没有体现出明显的提升，随后在适当调整训练参数并进行多 GPU 训练后效果有了明显提升。

2、最初在 Loss 函数的选用上采用三联体损失，但在测试时效果较差，随后更改为 CMCE 并没有明显提升，在查阅相关论文后，最终使用 CMPM 与 CMPC 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，发现自适应矩估计(Adaptive Moment Estimation, Adam)与随机梯度下降(Stochastic Gradient Descent, SGD)都可以取得相对较好的效果，并且收敛速度很快，但是使用自适应梯度算法(Adaptive Gradient, Adagrad)进行优化时，CMPM 最终收敛在 18 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。

四、下一阶段论文工作安排

2022 年 4 月 25 日 - 2022 年 4 月 30 日，对改进后的模型进行训练参数优化，试图获取更好的训练结果。

2022 年 5 月 1 日 - 2022 年 6 月 10 日，对各项实验所得结果进行整理对比，随后开始进行论文写作并进行答辩准备。

五、参考文献

[1] Li S, Xiao T, Li H, et al. Person Search with Natural Language Description[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[2] Li S, Xiao T, Li H, et al. Identity-Aware Textual-Visual Matching with Latent Co-Attention[C]//IEEE International Conference on Computer Vision, 2017.

[3] Zhang Y, Lu H. Deep Cross-Modal Projection Learning for Image-Text Matching[C]//European Conference on Computer Vision, 2018.

[4] Wang Z, Fang Z, Wang J, et al. Vitaa: Visual-Textual Attributes Alignment in Person Search By Natural Language[C]//European Conference on Computer Vision, 2020.

[5] Ding Z, Ding C, Shao Z, et al. Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification[J]. 2107.12666, 2021.

[6] Faghri F, Fleet D J, Kiros J R, et al. Vse++: Improving Visual-Semantic Embeddings with Hard Negatives[J]. 1707.05612, 2017.

[7] Schroff F, Kalenichenko D, Philbin J. Facenet: A Unified Embedding for Face Recognition and Clustering[C]//IEEE Conference on Computer Vision, 2015.

[8] Ma T, Yang M, Rong H, et al. Dual-path CNN with Max Gated Block for Text-Based Person Re-identification[J]. 2021, 111: 104168.