是否涉密	

桂林电子科技大学

研究生学位论文开题报告表

（专业学位硕士）

学　　号：20032303051

姓　　名：康博

论 文 题 目：融合情感的图像描述生成算法研究

校外导师姓名：李俊

校内导师姓名：文益民

学 位 类 别：工程硕士

领　　域：计算机视觉

所 属 学 院：计算机与信息安全学院

研究生院制表

2020 年 12 月 10 日填

填 表 说 明

1.凡所列栏目内容填写不够的，可以另加附页。

2.本表须保持原格式不变，不够可增加附页，纸张限用 A4，装订要整齐。

页面设置（页边距为上、下：2.5cm, 左为2.6cm,右为2.1cm；字体为宋体小四，行间距为18磅。）

3.无校外导师者不用填写“校外导师姓名”和“校外导师审核意见”。

4.会计硕士和翻译硕士不用填写领域。

一、学位论文研究内容

学号：20032303051	姓名：康博	研究方向：图像描述
拟定学位论文题目	融合情感的图像描述生成算法研究	

学位论文的课题来源（请在相应栏目打 √）：

□973、863 项目	□国家社科规划、基金项目	□教育部人文、社会科学研究项目
□国家自然科学基金	□中央国家其他部门	□国际合作研究项目
□与港澳台合作研究项目	□国防项目	□企事业单位委托项目
☑省（自治区、直辖市）项目	□外资项目	□学校自选项目
□非立项	□其他	

学位论文研究主要内容

在心理学上已经证明，在情感图像中对象、场景都可以作为情感刺激，不同对象与对象、场景和对象之间都可引起不同的情感，所以挖掘和分析对象与对象之间的关联程度，对于图像情感识别至关重要。此外，情感类别不是由某个特定的对象决定，而是由多个有意义的对象共同决定，但每个对象对于图像情感的所占权重不同。如何得到每个对象所占权重，进而根据其权重去寻找相对应的情感词与对象语义概念相匹配，并融合到图像描述中，是一个值得研究的问题，对于融合情感知识情的图像描述，可以将其与相应的图片结合，不仅可以促进多模式情感识别，也可以应用于图文结合的旅游推荐系统。综上所述，研究图像情感识别和融合情感知识的图像描述算法十分重要。

一、研究如何利用图像中对象与对象间的关联程度，进而识别图像情感的问题？

图像情感识别本质上是分类任务，但它与目标检测不同的是虽然对象语义概念相同，但所表达的情感却可能不同，而且一个对象和其他不同对象相互作用，可以引起不同的情感，此外场景特征的提取对于图像情感分析也尤为重要。据我们了解，目前图像情感识别的公开数据集都是图像级别的标注，那么我们如何在图像级别的标注下找到关键的、与情感相关的对象区域，而且图像中多个对象间的关联程度是不一样的，那么我们怎么得到对象间的关联程度，并利用其识别图像情感呢？要解决上述问题，主要有以下三个难点：

难点 1：在没有对象间关联程度标注的情况下，如何挖掘它们之间的关联程度？

难点 2：如何利用对象间的关联程度，生成对象的情感特征，进而得到整张图像的情感特征？

难点 3：如何找到图像中主导情感的对象区域呢？

二、研究如何找到与图像情感、对象语义概念相匹配的情感词？

目前风格化图像描述的做法是将整个语句以一种积极或消极（浪漫或幽默）的方式表达出来，而并没有找到图像中的哪些局部对象区域主导整个图像的情感，这显然是不合理的。那么我们要做的是：针对图像情感，找到情感词，并将其放在图像描述合适的位置。但是图像中对象的个数不止一个，我们应该将情感词放在图像描述的哪个位置呢，此外与情感相关的形容词有很多，怎么选择最合适的情感词和对象语义概念相匹配呢？我们首先在语料库中找到与情感相关的形容词，然后建立一个记忆模块，包含一系列的嵌入向量，确定情感词的情感类别，进而根据对象区域对于图像情感所占的比重，利用语料库中情感词出现在对象语义概念上下文中的概率找到可能性

	最大的情感词，并将情感词与相应的对象语义概念相匹配，最后融合到图像描述。要解决上述问题，主要有以下难点： 难点 1：情感词有很多，怎么选择最合适的情感词和对象语义概念相匹配呢？

二、学位论文研究依据

学位论文的选题依据和研究意义，以及国内外研究现状和发展态势，附主要参考文献

一、选题依据与研究意义

随着互联网中旅游图像的数量与日俱增，如何有效的管理、有组织的对旅游图像进行分类，以使用户在海量的图像库中找到符合自己情感需求的旅游图像成为新的研究问题，基于情感的图像推荐系统，可以让用户根据情感找到对应的旅游场景图像，众所周知，一图胜千言，图像也可传达丰富的语义信息，所以图像情感的语义分类越来越受到关注。图像描述是利用自然语言描述图像视觉内容的任务，结合视觉信息和语言模型，生成有意义、表达准确的语句，在此基础上，为了避免生成显而易见的描述，并且以一种更吸引用户注意力的方式生成描述，风格化图像描述旨在通过考虑用户的先验知识、常用词汇和书写风格来满足风格化的需求，但是目前风格化描述的做法是将整个语句以一种积极或消极（浪漫或幽默）的方式表达出来，而并没有找到是图像中的哪些局部对象区域主导整个图像的情感，这显然是不合理的，那么如何找到图像情感对应的对象区域，并将表达合适情感强烈程度的形容词匹配到相应的对象语义概念就成为了一个新的研究问题。

针对旅游照片生成带有情感的描述，不仅对经济社会具有重大意义，而且也可以促进科学技术领域的发展。从以下两方面说明为旅游照片生成带有情感的描述的意义：1、针对旅游照片做情感分析和挖掘，既可以更好地解决旅游场景的分类问题，进而生成与用户所需情感更贴近的旅游意图，促进旅游业的发展；又可以提高旅游图像检索的准确率，因为基于情感的图像检索系统最重要的任务就是设计和训练相应的情感模型来划分情感类别，用以定量的描述图像情感信息，以此来映射图像所表达情感与图像内容之间的关系。2、对于带有情感的描述，可以将其与相应的图片结合，这样不仅可以促进多模式情感识别，也可以应用于图文结合的旅游推荐系统，根据旅游景点的宣传海报等信息，为用户推荐更符合其情感需求的旅游地点。

综上所述，研究融合情感知识的图像描述生成算法，是非常有研究价值和研究意义的工作。

二、国内外研究现状和发展态势

1. 图像情感识别国内外研究现状和发展态势

过去二十年里，研究人员提出大量关于图像情感识别的算法，根据算法基于不同的心理模型，即 CES（Categorical Emotion States ）和 DES[1]（ Dimensional Emotion Space），可以分为分类任务和分布式学习任务，我们的工作聚焦在情绪分类任务。Wang[2]等人回顾与总结了图片情感分析的历史与现状，从传统的视觉情感分析方法和深度学习两个方向对图片情感分析相关研究的技术方法进行梳理并评述。早期的图像情感识别主要以设计手工特征来挖掘和分析图像情感，受心理学和艺术学的影响，大多数工作以颜色、纹理、构图和内容等要素构成图像情感特征，用于图像情感识别，Brothet[3]等人引入形容词名词对（ANPs），提出 SentiBank 视觉概念检测器，从语义层面选择与情感密切相关的视觉概念。随着深度学习的快速发展，越来越多的研究人员使用卷积神经网络（CNN）提取图像视觉特征替代手工提取特征，并取得重大的进展。目标检测是计算机视觉中的一个核心问题，作为预处理已经被应用到各种相关任务，包括图像描述、视觉推理和场景图等。根据是否生成区域建议，目标检测任务可分为两大类：一是单阶段目标检测，比如 YOLO[4]、SSD[5]，二是双阶段目标检测，例 R-CNN[6]、Fast R-CNN[7]、Faster R-CNN[8]。R-CNN 是最早应用

CNN 的两阶段目标检测方法，极大的提高了模型检测精度。在此基础上 Faster R-CNN 为了减少生成区域建议所耗费的时间和提取特征所产生的冗余计算，引入了区域提议网络（RPN）来替代 R-CNN 中耗时较长的选择性搜索算法生成区域建议，不仅实现了端到端的训练，而且很大程度上提高了检测精度。由于 Faster R-CNN 的准确性和速度相较于其它目标检测算法有很大的优势，从而被广泛应用于各种相关任务的预处理。

最早将 CNN 用于图像情感识别的工作是 Chen[9]等人在 SentiBank 的基础上，利用深度神经网络构建 DeepSentiBank 的图像情感分类方法，相较于手工提取视觉特征的传统方法，取得了较大的进展。通过利用从网站获取的 50 万张图像和相应的标签，You[10]等人提出渐进式卷积神经网络架构（PCNN）来预测图像情感。单级视觉特征已不能满足我们对图像特征的需求，于是 Rao[11]等人构建了多级深度表征网络，通过从图像语义、美学和低级视觉特征中提取图像的情感特征。为了充分利用图像的多尺度特征，Zhu[12]等人结合 CNN 和循环神经网络（RNN）结构，利用 CNN 提取不同层次的特征，RNN 捕获它们之间的关系。虽然提取了图像的多级特征，但忽略了局部特征也可以引起图像情感这一要素，所以 You[13]等人利用注意力机制去发现与情感相关的局部区域，是第一个通过聚焦局部区域识别图像情感的算法。Yang[14]等人进一步提出了弱监督耦合网络（WSCNet），该网络通过注意力机制发现情感相关的区域，并利用全局特征和局部特征去预测图像情感。此外通过使用深度度量学习结合多任务深度框架可解决图像情感识别的检索和分类任务。Zhang[15]等人提出了一种新的 CNN 模型，来提取和整合视觉内容信息、风格信息来预测图像情感类别。考虑到不同的情感刺激种类，Yang[16]等人提出一个刺激感知图像情感识别网络。

现有的方法大多数是直接利用图像的全局特征或局部特征预测图像情感，由于人类情感涉及高度复杂和抽象的认知过程[17]，很难直接从情感图像的整体或者局部特征中推断出情感，在心理学上已经证明场景和物体都可以作为情感图像中的情感刺激，情感状态不是关于一个特定对象，而是多个有情感意义的对象的感知[18]。SOLVER 算法利用图像中对象和对象、场景和对象间的情感关系做图像情感分析，首先依据目标检测网络得到的对象语义概念及其视觉特征建立情感图，接着用图神经网络去处理构建的情感图，得到各个对象的情感增强向量，然后利用基于场景特征的注意力机制去指导各个区域特征相融合，再结合场景特征一起送到情感分类器中，最后得到情感分类。

虽然关于图像情感识别已经做了大量的工作，但仍有一些挑战和难点存在于这个领域。首先是情感鸿沟的问题，即视觉特征与观众通过感知图像所获得的情感不一致的问题。对于这个问题，我们可以提取能够更好地区分情感的视觉特征也可以结合可用的上下文信息[19]，因为同一对象在不同的语境下会表达出不同的情感，比如我们看到一个人在哭泣，我们可能会感到很悲伤，但如果对这个场景有一个描述，这个人终于实现了他的梦想，他激动的落泪，那么我们就会感受到不一样的情感。其次不同于目标检测的客观性，图像情感识别是一个主观的问题，对于那些很喜欢自然景象的人来说，看到电闪雷鸣的景象也许他们会很兴奋，而对于害怕打雷的人来说，则会感到恐惧。最后则是标签噪声和缺失的问题，在某些情况下，数据集的标注必须由专业人士来完成，例如艺术品只有专家能够提供可靠的标注。在真实世界中，可能只有少量甚至没有标记的数据，那么应该如果应对这种情况呢？无监督和弱监督可能是两个较好的选择。

2. 风格化图像描述国内外研究现状和发展态势

图像描述[20]是用自然语言描述图像的视觉内容，使用视觉系统和语言模型，能够生成有意义和句法正确的句子。近年来风格化图像描述越来越受到人们的关注，因为常规的图像描述生成不具有感情色彩的描述且没有与用户交互的描述事实的句子，所以研究人员为了避免生成常规的、显而易见的描述提出风格化描述，风格化描述旨在通过生成考虑到用户的先验知识、活跃词汇量和写作风格的描述来满足这一要求。Mathews[21]等人于 2016 年首先提出了 switching RNN，用来生成具有消极或积极情感的图像描述。Chen[22]等人在 2018 年提出了 style-factual LSTM 和对抗性训练方法来训练风格化图像描述模型。Yang[23]等人使用 NIC 作为基线模型，针对数据稀疏问题，改变了基线模型中的文本 one-hot 表示，使用 word2vec 对文本进行映射，为了防止过拟合，在模型中加入了正则项和使用 Dropout 技术，并在词序记忆方面取得创新，引入联想记忆单元 GRU，用于文本生成。Tang[24]等人基于逐层优化的多目标优化及多层概率融合的 LSTM。上述方法都较依赖于风格化句子和图像相配对的数据来进行训练。为了减少过度依赖数据的标注，Gan[25]等人提出了 StyleNet 模型，在描述生成的过程中控制风格，以生成所需要的风格化描述。之后 Sinnetal[26]等人提出一种新的方法，在另一个 CNN 的帮助下，将情感术语纳入图像描述中去，对 LSTM 中的权重矩阵进行了分解，以同时对事实句子和风格化句子进行建模。Chen[27]等人提出用领域层规范生成风格化的图像描述，使得生成各种风格化的描述成为可能。Guo[28]等人在多个辅助模块的帮助下，通过在未配对的风格化语料库上训练单一的模型来生成多种风格的图像描述。以前的方法都侧重于设计语言模型或训练算法来捕捉风格因素，而 Zhao[29]等人通过构建记忆模块来显示编码从大型语料库中学到的风格知识。目前风格化描述的做法是将整个语句以一种积极或消极（浪漫或幽默）的方式表达出来，而并没有找到是图像中的哪些局部对象区域主导整个图像的情感，这显然是不合理的，那么如何找到图像情感对应的对象区域[30]，并将表达合适情感强烈程度的形容词匹配到相应的对象语义概念[31-32]就成为了一个新的研究问题。

三、主要参考文献

[1] Zhao S,Yao X, Yang J, et al. Affective Image Content Analysis: Two Decades Review and New Perspectives[J]. 2021.

[2]王仁武，孟现茹．图片情感分析研究综述[J]．图书情报知识，2020(3):9.

[3] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, “Large-scale visual sentiment ontology and detectors using adjective noun pairs,” in ACM MM, 2013, pp. 223–232.

[4] Redmon J， Divvala S， Girshick R , et al. You Only Look Once: Unified, Real-Time Object Detection[J]. IEEE, 2016.

[5] Liu W， Anguelov D， Erhan D , et al. SSD: Single Shot MultiBox Detector[J]. Springer, Cham, 2016.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” inProc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.

[7] R. Girshick, “Fast R-CNN,” inProc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.

[8] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” inProc. NeurIPS, 2015, pp. 91–99.

[9] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, “Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks,” arXiv:1410.8586, 2014.

[10] Q. You, J. Luo, H. Jin, and J. Yang, “Robust image sentiment analysis using progressively trained and domain transferred deep networks.” in AAAI, 2015, pp. 381–388.
[11] T. Rao, M. Xu, and D. Xu, “Learning multi-level deep representations for image emotion classification,” NPL, vol. 51, no. 3, pp. 2043–2061, 2020.
[12] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu, “Dependency exploitation: a unified cnn-rnn approach for visual emotion recognition,” in IJCAI, 2017, pp. 3595–3601.
[13] Q. You, H. Jin, and J. Luo, “Visual sentiment analysis by attending on local image regions,” in AAAI, 2017, pp. 231–237.
[14] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, “Visual sentiment prediction based on automatic discovery of affective regions,” IEEE TMM, vol. 20, no. 9, pp. 2513–2525, 2018.
[15] W. Zhang, X. He, and W. Lu, “Exploring discriminative representations for image emotion recognition with CNNs,” IEEE Trans. Multimedia, vol. 22, no. 2, pp. 515–523, Feb. 2020.
[16] J. Yang, J. Li, X. Wang, Y. Ding, and X. Gao, “Stimuli-aware visual emotion analysis,” IEEE Trans. Image Process., vol. 30, pp. 7432–7445, 2021.
[17] N. H. Frijda, “Emotion experience and its varieties,”Emotion Rev., vol. 1, no. 3, pp. 264–271, Jul. 2009.
[18] M. Bar, “Visual objects in context,”Nature Rev. Neurosci., vol. 5, no. 8, pp. 617–629, 2004.
[19] Anderson, P .; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6077–6086.
[20] Stefanini M， Cornia M， Baraldi L , et al. From Show to Tell: A Survey on Image Captioning[J]. 2021.
[21] Mathews, A. P .; Xie, L.; and He, X. 2016. Senticap: Generating image descriptions with sentiments. InThirtieth AAAI conference on artificial intelligence.
[22] Chen, T.; Zhang, Z.; Y ou, Q.; Fang, C.; Wang, Z.; Jin, H.; and Luo, J. 2018. “factual”or“emotional”: Stylized image captioning with adaptive learning and attention. InProceedings of the European Conference on Computer Vision (EC-CV), 519–535.
[23]杨楠, 南琳, 张丁一,等. 基于深度学习的图像描述研究[J]. 红外与激光工程, 2018, 47(2):8.
[24]汤鹏杰, 王瀚漓, 许恺晟. LSTM 逐层多目标优化及多层概率融合的图像描述[J]. 自动化学报, 2018, 44(7).
[25] Gan, C.; Gan, Z.; He, X.; Gao, J.; and Deng, L. 2017. Stylenet: Generating attractive visual captions with styles. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3137–3146.
[26] Shin, A.; Ushiku, Y .; and Harada, T. 2016. Image captioning with sentiment terms via weakly-supervised sentiment dataset. InBritish Machine Vision Conference.
[27] Chen, C.-K.; Pan, Z.; Liu, M.-Y .; and Sun, M. 2019. Unsupervised stylish image description generation via domain layer norm. InProceedings of the AAAI Conference on Artificial Intelligence, volume 33, 8151–8158.
[28] Guo, L.; Liu, J.; Y ao, P .; Li, J.; and Lu, H. 2019. Mscap: Multi-style image captioning with unpaired stylized text. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4204–4213.
[29] W. Zhao, X. Wu, and X. Zhang, “MemCap: Memorizing style knowledge for image captioning,”

inAAAI, 2020.
[30] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.
[31] Pennington J , Socher R , Manning C. Glove: Global Vectors for Word Representation[C]// Conference on Empirical Methods in Natural Language Processing. 2014.
[32] Mikolov T , Chen K , Corrado G , et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

三、学位论文研究方案及研究计划

1. 研究方案（包括有关理论、方法、技术路线、创新手段、关键技术等）

（1）利用图像中对象与对象间的关联程度，进而识别图像情感的解决方案

难点 1：在没有对象间关联程度标注的情况下，如何挖掘它们之间的关联程度？

解决方案：我们考虑可以通过在语料库中求得对象语义概念的共现次数，共现次数越多，它们之间的相关度也就越高，关联程度也就越强；或者利用对象语义概念的共现次数作为关联程度的伪标签，通过$\gamma ij=(wi\ vi\ )T(wj\ vj\ )$, 训练两个参数向量$wi\ ,wj\ ,\gamma ij$表示对象间的相关度，$\gamma ij$的值越大表示对象 i，j 之间的关联程度越强。若以两个对象语义概念的共现次数作为对象间关联程度的判断因素，我们考虑会不会出现这种情况：在语料库中，两个对象的语义概念共现次数很多，表示对象间的关联程度很强烈，但实际在图像中两个对象间的关联程度并不强。针对这个问题，我们考虑结合对象语义概念的共现次数和对象视觉特征之间的相关度，表示对象之间的关联程度是否强烈。

难点 2：如何利用对象间的关联程度，生成对象的情感特征，进而得到整张图像的情感特征呢？

解决方案：针对这个问题我们考虑在目标检测阶段，采用 bottom-up attention model 检测对象及其属性，Faster R-CNN 作为检测网络，以 Resnet-101 为网络结构，用 ImageNet 数据集来做预训练，Visual Genome Dataset 数据集上训练检测网络，识别出来对象及其属性，并以对象+属性作为对象的语义概念，然后用 GloVe 算法生成对象的语义概念的词向量作为全连接图的节点，对象间的关联程度作为全连接图的边。构建的全连接图如图 1 所示。接着用图神经网络推理得到的全连接图，得到对象的情感特征向量，最后结合对象的情感特征，得到整张图像的情感特征。解决方案的流程图如图 2 所示。

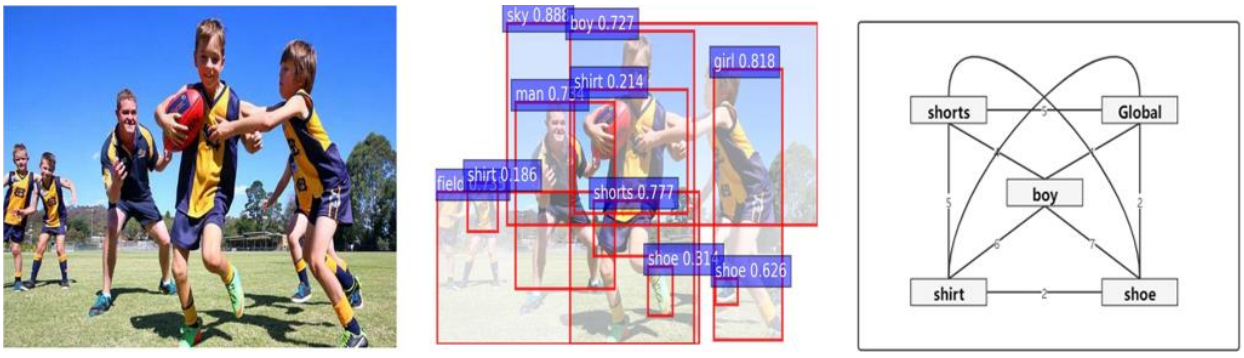

图 1：根据图像中的对象信息构建全连接图

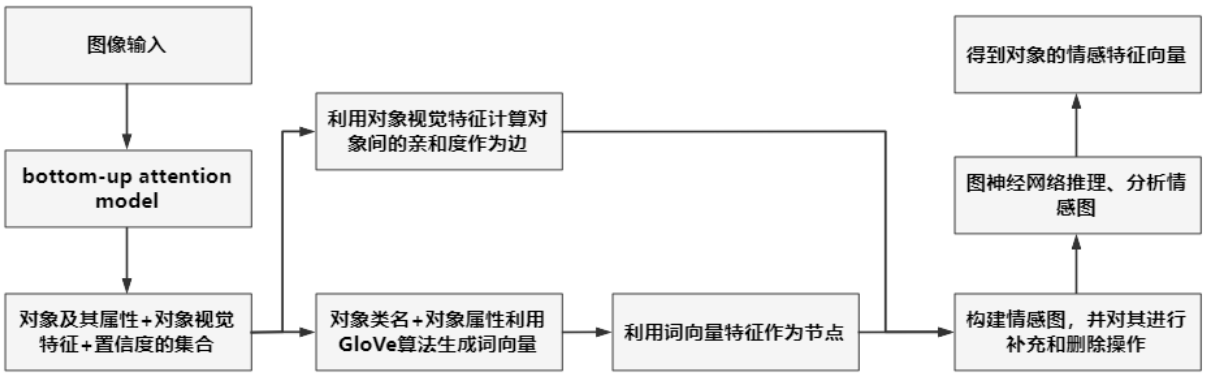

图 2：生成整张图像情感特征的流程图

难点 3：如何找到图像中主导情感的对象区域呢，以及后续我们应该将情感词放在图像描述的哪个位置呢？因为目前的风格化图像描述是将整个句子以一种积极或消极的方式表达出来，而并没有找到是图像中的哪个局部对象（区域）主导整个图像的情感。那么我们要做的是：针对图像情感，找到情感词，并将其放在图像描述合适的位置。但是图像中对象的个数不止一个，我们应该将情感词放在图像描述的哪个位置呢，即我们找到表达情感的形容词后，应该和哪个对象的语义概念相匹配呢？如图 3 所示，这张图像的主体显然是两只长颈鹿，而不是放在后边的树或者建筑上。

解决方案：我们考虑用现有的 Image Caption 模型为图像生成描述，因为图像描述就是建立在理解图像中场景的基础上，所以这里我们结合生成的句子特征和图像的全局特征形成多模式场景特征（语言+视觉），这里句子特征和图像的全局特征属于不同的空间而且维度也不相同，我们怎么结合句子特征和图像的全局特征呢？首先利用 encode-decode 模型对句子中的每个单词分别和图像的全局特征进行编码，通过预测下一单词是否和句子中的下一单词是否相同，作为损失来迭代更新全局特征，最后得到一个融合了句子特征的视觉特征，即为我们需要的多模式场景特征。然后利用基于场景特征的注意力机制去判断哪些对象区域是重要，进而为每个局部特征分配权重，基于每个局部特征被分配的权重，来决定带有情感的形容词匹配到哪个对象语义概念。对于注意力机制的设计，因为我们得到对象的情感特征 Oi 和场景特征 Fs 属于不同的空间，我们首先将两个特征投影到相同的空间中去，然后通过 $f(Oi, Fs) = \text{sigmoid}((wi\ Oi)T(wj\ Fs))$ 训练两个权重矩阵 wi, wj ，最后得到一个 0-1 之间的值，即为对象对于图像情感表达所占权重。基于多模式场景特征的注意力机制模型流程图如图 4 所示：

原图像描述：
Two giraffes are standing outdoors near a building.
想要的描述：
Two *lovely* giraffes are standing outdoors near a building.
对比的描述：
Two giraffes in a *pleasant* park are against *beautiful* trees.(MsCap模型生成的带有积极情感的描述)

图 3：找到图像情感对应的对象区域

图 4：基于多模式场景特征的注意力机制模型流程图

（2）找到与图像情感相匹配的情感词解决方案

难点 4：和情感相关的形容词有很多，怎么选择最合适的情感词和对象语义概念相匹配呢？目前的风格化图像描述是针对于 Positive、Negative 两类去寻找相对应的形容词，虽然都是表达积极情感，显然 happy 和 surprise 所表达的情感强度是不同的。如图 5 所示，图像所表达的情感为 sad，虽然 depressed（沮丧的）和 scary（可怕的）都是带有消极情感的词，但显然这里用 depressed 更合适。

解决方案：我们考虑首先在语料库中通过 MemCap 提到句子分解算法，找到与情感相关的形容词，通过一个记忆模块，包含一系列的嵌入向量，来记忆情感词的类别，然后找到图像中主导情感表达的对象，从语料库中找到情感词出现在对象语义概念上下文的概率Xij和对象语义概念上下文所有单词出现的总次数Xi，$Pij=Xij/Xi$, 得到情感词 j 出现在对象语义概念的上下文中的概率，可以找到单词 i，j 哪一个和对象的语义概念更相关。这样处理感觉耗费的时间和计算量太大，还有另外一种解决办法：就是我们现在语料库中找到大量的形容词名词对存储下来，这样我们每次只用将含有对象语义形容词名词对提取出来，然后我们计算和对象语义概念搭配的情感词出现在其上下文的概率，选择可能性最大的情感词。这种方法会比上一种方法更省时间，但是形容词名词对可能不包括所有的情感词。选择合适情感词模块的流程图如图 6 所示：

原图像描述：

A man sitting on a bench in the park.

想要的描述：

A depressed man sitting on a bench in the park.

对比的描述：

A scary man sitting on a bench in the park.

图 5：找到情感强度合适的形容词

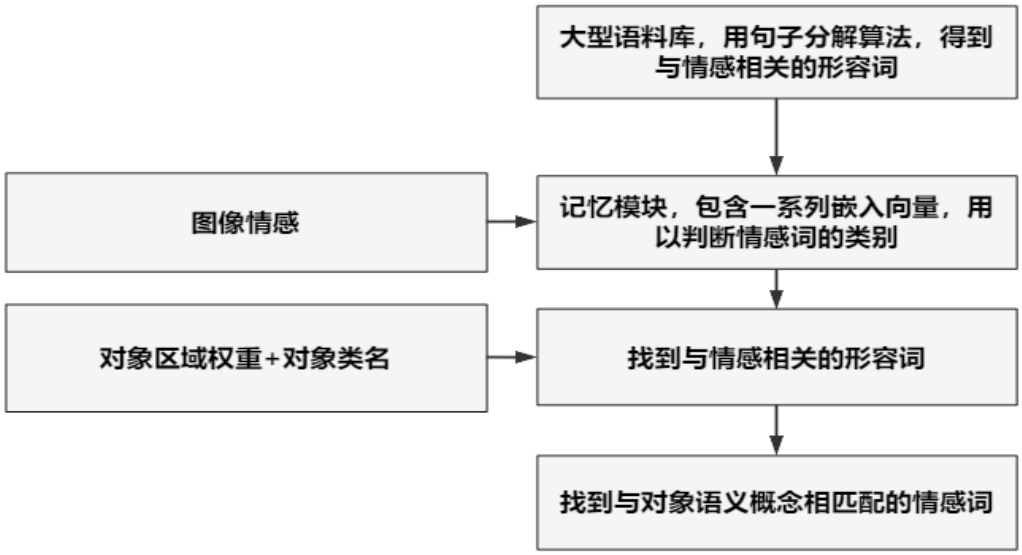

图 6：选择合适情感词模块的流程图

（3）技术路线

图像情感识别的算法流程：

①. 图像输入。

②. 利用 bottom-up attention model 得到对象及其属性+对象视觉特征+置信度集合。

③. GloVe 算法生成（对象类名+对象属性）的词向量作为节点。

④. 利用对象视觉特征计算对象间的关联程度作为边，结合步骤 3 中的节点，构建全连接图，并对其进行一定的筛选或补充操作。

⑥. 由现有的 Image Caption 模型生成图像描述。

⑦. Faster R-CNN 得到图像的全局特征，与图像描述的句子特征形成多模式场景特征。

⑧. 建立基于多模式场景特征的注意力机制为每个对象区域分配相应的权重。

⑨. 根据对象权重，融合各对象区域的情感特征，生成整张图像的情感特征向量。

⑩. 拼接场景特征和图像的情感特征，然后送入情感分类器，得到图像情感类别。

图像情感识别算法流程图如图 7 所示：

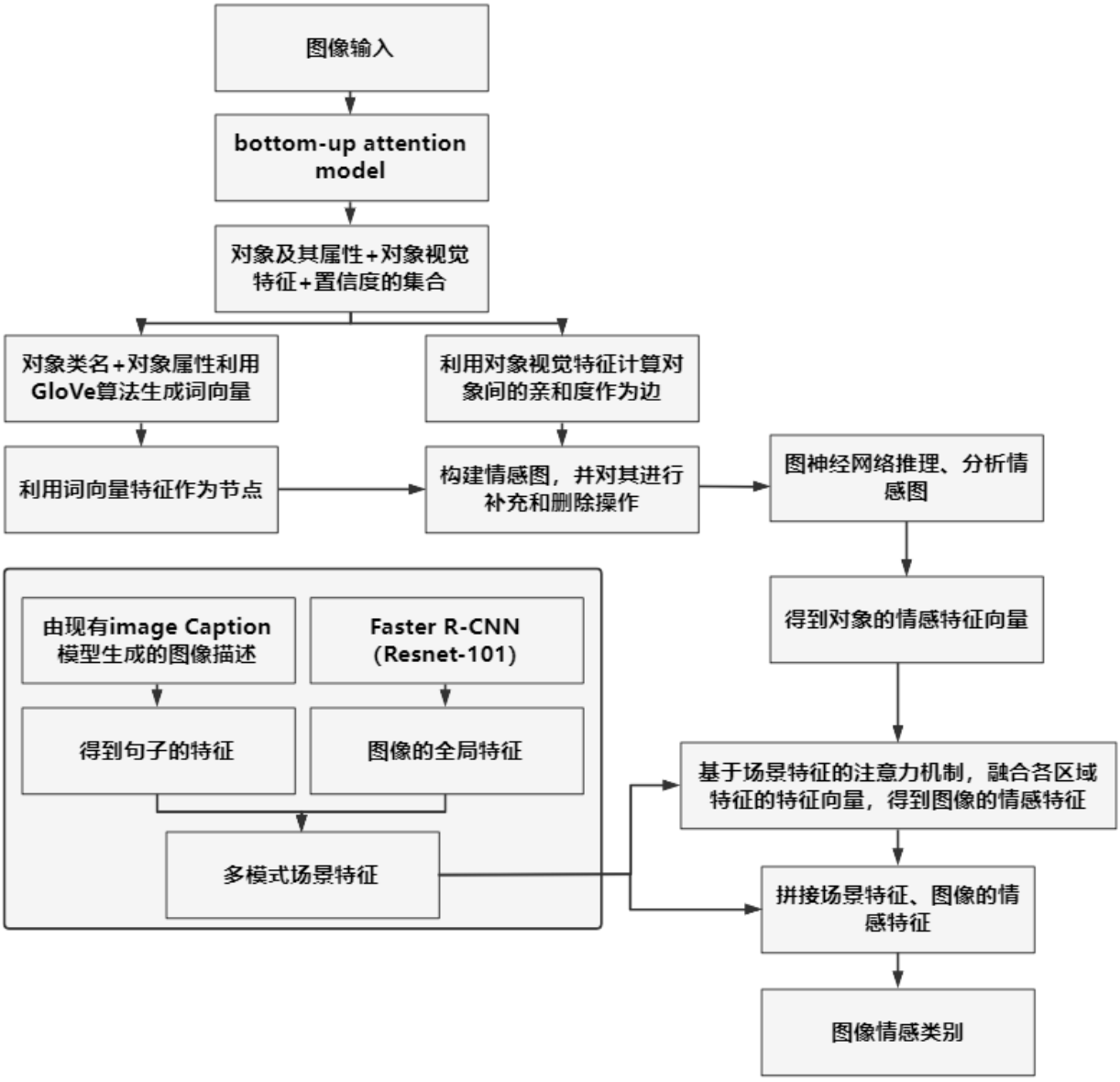

图 7：图像情感识别算法流程图

融合情感的图像描述算法流程：

①. 在语料库中，用句子分解算法得到与情感相关的形容词。

②. 设计记忆模块包含一系列的嵌入向量，用以判断情感词的类别。

③. 根据图像情感，找到与情感相关的形容词集合。

④. 根据区域权重+对象类名，找从步骤③中得到的情感词集合找到与对象语义概念最匹配的情感词。

⑤. 设计句子融合机制，将情感词融合到图像描述。

⑥. 生成带有特定情感的图像描述。

融合情感的图像描述算法流程图如图 8 所示：

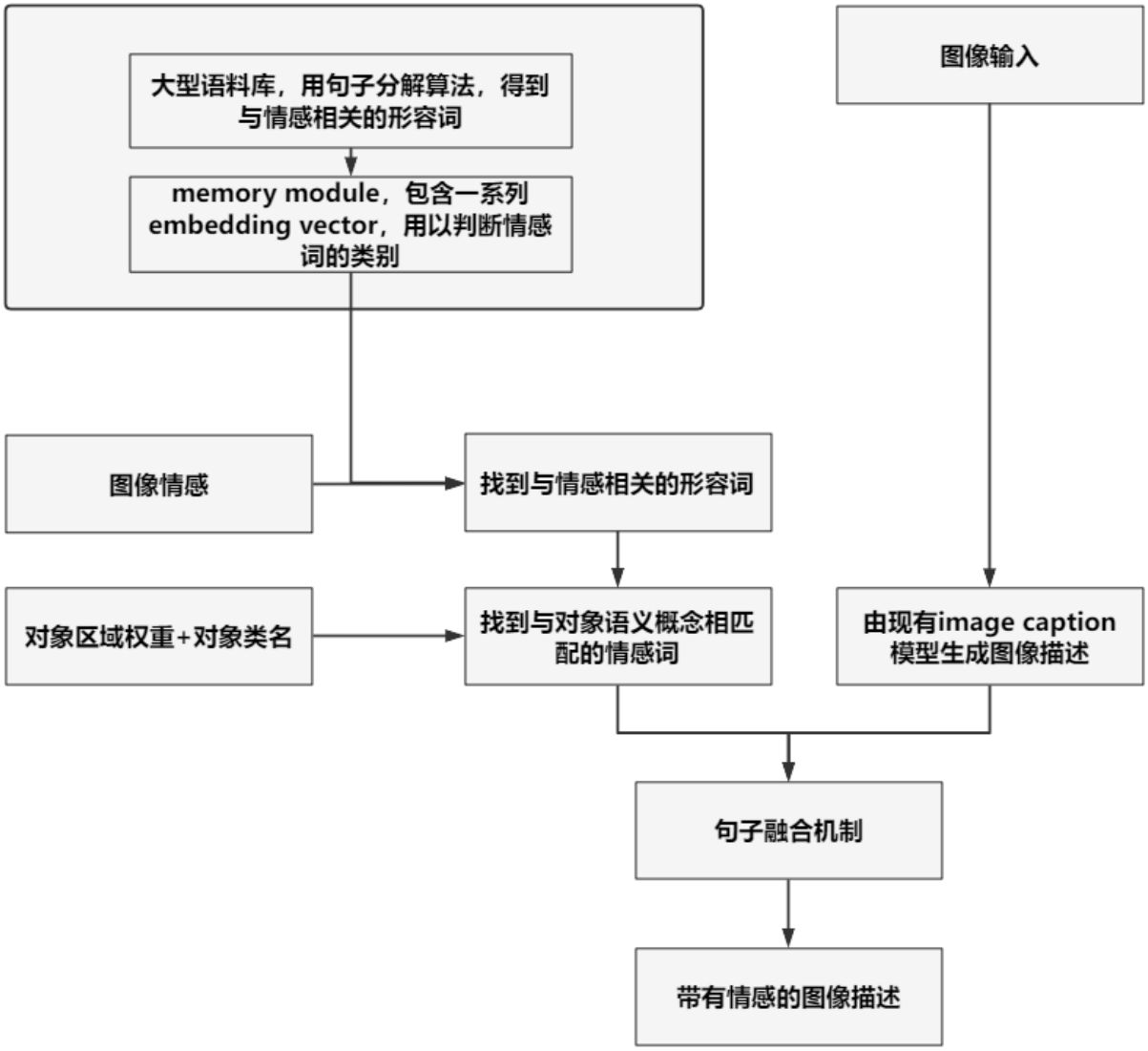

图 8：融合情感的图像描述算法流程图

2、实验条件落实情况，可能存在的问题及解决办法

（1）实验条件落实情况：

软件环境：

算法实现：基于 PyTorch 的 Python 开发环境。

部分数据支持：500-1000 张的旅游照片并附带情感的图像描述。

实验结果可视化：基于 Python 的开发环境。

硬件环境：

一台 DGX-station 深度学习工作站，以及多台已配置好软件开发环境的高性能计算机。

（2）可能存在的问题及解决办法：

问题 1：因为 bottom-up attention model 在 Faster R-CNN 的基础上，又针对每一类执行 NMS（非极大值抑制），最后得到每一类概率最大的候选框，所以我们得到候选框个数基本与图像中对象个数相同，即每幅图像得到候选框的个数不一样。

解决办法：针对这个问题我们考虑两种方法：一是删去 bottom-up attention model 对每类执行 NMS 的步骤，然后用置信度 Top-N 的候选框作为目标检测网络的结果；二是对于图像中对象较多的情况，我们选择置信度 Top-N 的候选框，对于个数较少的，我们考虑用零向量填充或者图像的全局特征填充缺失节点。

问题 2：若以两个对象语义概念的共现次数作为对象间关联程度的判断因素，我们考虑会不会出现这种情况：在语料库中，两个对象的语义概念共现次数很多，表示对象间的关联程度很强烈，但实际在图像中两个对象间的关联程度并不强。

解决办法：针对这个问题，我们考虑结合对象语义概念的共现次数和对象视觉特征之间的相关度，表示对象之间的关联程度是否强烈。

3. 年度研究计划

	起始时间	完　成　内　容
年度研究计划	2021.10~2021.12	阅读相关文献，完成开题报告
	2022.01~2022.06	设计图像情感识别算法，实验分析及小论文撰写
	2022.07~2022.11	设计带情感的图像描述的算法，实验分析及小论文撰写
	2022.12~2023.02	总结前期的工作，撰写大论文初稿
	2023.02~2023.06	完成大论文撰写，修改并定稿

4．学位论文创新和预期研究成果

（1）创新之处

创新点 1：在没有对象间关联程度的标注下，学习到对象间的关联程度，并利用其生成整张图像的情感特征。

创新点 2：利用多模式场景特征的注意力机制，找到图像情感对应的对象区域，并将形容词匹配到相应的对象语义概念。

创新点 3：在没有情感标注的情况下，找到情感相关形容词并确定其表达的情感强度。

（2）预期研究成果

1. 在期刊或学术会议上发表学术论文 1-2 篇。
2. 完成学位论文 1 篇。

四、开题报告审查意见

<table>
<tr><td colspan="3">1. 校外导师对学位论文选题、论文计划可行性等方面意见，是否同意开题。

签名：　　年　　月　　日</td></tr>
<tr><td colspan="3">2. 校内导师对学位论文选题、论文计划可行性等方面意见，是否同意开题。

签名：　　年　　月　　日</td></tr>
<tr><td colspan="3">3. 评审专家意见</td></tr>
<tr><td rowspan="2">开 题
报告会</td><td>时间：　　年　　月　　日</td><td>地点：</td></tr>
<tr><td colspan="2">评审专家（至少 3 位，请专家本人签名）：　答辩主席：_____
答辩委员会成员：_____ _____ _____ _____</td></tr>
<tr><td colspan="3">评审专家组意见：
1. 论文选题的理论意义或实用价值：☐较大　☐一般　☐较小
2. 论文难度：☐偏高　☐适当　☐偏小
3. 论文工作量：☐偏大　☐适当　☐偏小
4. 研究方案的可行性：☐好　☐较好　☐一般　☐不可行
5. 对文献资料及课题的了解程度：☐好　☐较好　☐一般　☐较差
6. 对开题报告中反映出的综合能力和表达能力：☐好　☐较好　☐一般　☐较差
7. 是否同意论文开题报告：☐同意　☐需要重做</td></tr>
<tr><td colspan="3">综合考核结果（请划 √ ）：
☐优秀　☐良好　☐中等　☐及格　☐不合格
组长签名：　　年　　月　　日</td></tr>
<tr><td colspan="3">4. 学院（所、部）意见：

主管领导签名：
单　　位（盖章）　　年　　月　　日</td></tr>
</table>