**Visual Question Answering（VQA）**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **学号** | **专业班级** | **姓名** | **性别** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  | 3180105546 | 计科1804 | 程川洋 | 女 |

## Project Introduction

* 1. **选题**

本次实验任务为Visual Question Answering，即输入图片I和由n个单词组成的问题集Q={q1,…,qn}，输出由m个单词组成的答案A={a1,…,am}。该任务广泛的运用在图灵测试，盲人导航等应用领域中。

目前绝大多数的VQA任务都基于神经网络，用CNN提取图像特征，用LSTM编码问题特征，然后再结合两种特征来推断答案。

* 1. **工作简介**

我们参考老师给出的第二种方法《Stacked attention networks for image question answering (CVPR 2016)》。该文作者发现绝大数的问题都需要进行多步推理，比如定位出文字中出现的目标，和问题中的关键概念，不断地精确化的聚焦于可能出现答案的图像区域。因此，文章作者提出SAN网络，以进行多步推理过程，SAN网络可以看做是注意力机制的拓展。

在本组的实验中，我们主要流程是先用VGG网络提取图片的特征向量，再用LSTM提取问题的特征，最后将图片和问题的特征融合，计算出图片与文字的相关性、得到问题的答案。

* 1. **开发环境及系统运行要求**

开发环境：ModelArts Ascend Notebook

软件环境：Python3.7.5、 MindSpore1.1

## Technical Details

1. **理论知识**

SAN的网络结构如下图所示，主要由三部分组成：第一部分是图像模型，即用CNN来踢去图像特征；第二部分是问题模型，即用CNN或者LSTM来提取问题的语义向量；第三部分为Stacked Attention Model，即通过多步推理来获得最终答案。

Diagram

Description automatically generated

1. **Image Model图像模型**

我们可以通过VGGNet来得到一个image I的特征fI。SAN模型仅从最后一个pooling layer提取出来特征，舍弃掉原本VGG图像分类任务中需要的线形层等。

Diagram

Description automatically generated

首先我们先将原始图像进行中心化裁剪变为448\*448的形状，然后再通过VGG网络来得到形状为14\*14\*512的特征层，其中，特征层的每个像素都便对应原图像32\*32的区域，即将图像等分为14\*14个小区域，然后每个小区域得到512维的特征向量。然后我们需要先将图像特征向量的长度增加到和后面的问题特征向量的维度一致，然后再通过一个激活函数tanh的全连接层。



1. **Question Model 问题模型**

在论文中提出，我们既可以用LSTM也可以用CNN对问题进行处理。我们在实验中用的LSTM来进行处理。

Diagram

Description automatically generated

LSTM采用gate mechanism，用input gate，output gate和forget gate来控制模型的效果。针对一个由T个单词组成的问题q=[q1,…,qT]，其中qt表示第t个单词的one-hot编码。首先将qt嵌入到向量空间当中，然后再将嵌入的结果xt送到LSTM当中，得到输出ht。最后，对于一整句话来说，最后该问题的表示vQ=hT

Text

Description automatically generated

1. **Stacked Attention Model 堆叠注意力模型**

通过前两步可以分别得到图像特征vI和问题特征vQ，然后到了实验的关键部分。通过采取逐步推理过程，可以避免直接用全局图像特征预测导致的噪声干扰。

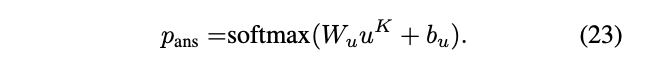
首先将图像和问题特征进行矩阵变化，得到一个注意力分布PI。

Text

Description automatically generated

然后基于注意力分布PI，我们可以重新计算图像向量的加权和，然后再将图像和问题结合起来。每一层Attention机制都按上述步骤完成，最后得到的可以推断答案的注意力Pans。

Diagram, schematic

Description automatically generated

通过不断使用attention机制，我们得到的新的特征向量可以更好的捕捉图像中与文本相关的关键特征信息，最后选出分数最好的答案作为结果。

1. **实现技术过程**

备注：这部分可以不用解释整体意义（上面已经写了），关键是贴代码和代码含义/代码结构等。【！！！！最后记得删掉这行】

1. **数据处理**

本次实验使用的数据集是VQA2.0的一个子集[Train: 44375；validation: 21435；test: 21435]

1. **图像处理**
2. **问题文字序列处理**
3. **Dataset的生成**
4. **模型训练**
5. **VGG**
6. **LSTM**
7. **Stacks Attention Model**

## Experiment Results

用图文并茂的形式给出实验结果，如系统界面、操作说明、运行结果等，并对实验结果进行总结和说明。

## References:

[1] S. Antol et al., "VQA: Visual Question Answering," 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2425-2433, doi: 10.1109/ICCV.2015.279.

[2] Z. Yang, X. He, J. Gao, L. Deng and A. Smola, "Stacked Attention Networks for Image Question Answering," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 21-29, doi: 10.1109/CVPR.2016.10.