综合图文推理的问答系统：Reading-VQA

周恩贤
2018011438
清华大学计算机系
zex18@mails.tsinghua.edu.cn

陈博涵
2018011441
清华大学计算机系
cbh18@mails.tsinghua.edu.cn

冯卓尔
2017011998
清华大学计算机系
fze17@mails.tsinghua.edu.cn

Abstract

近年来，愈来愈多的研究者投入了视觉问题回答*(Visual Question Answering)*的任务中，且有着不错的成果。视觉问答任务是一项较为综合的任务：利用计算机视觉的知识处理输入图片，透过自然语言处理的方法处理输入问题，并利用注意力等模型综合两个领域，最终生成解答，是一个全新的突破。然而，在实际生活中，我们能同时感知到视觉与听觉，可以同时看见图片也能读到图片的相关描述，再进行综合推理。因此，我们希望能让机器也有阅读思考、处理多模态的能力。我们改良了原有的 *VQA* 模型，提出了两个 *Reading-VQA* 的新模型，并在训练过程中加入图像描述层，利用了图片、描述与问题三个不同维度的信息进行视觉问答的任务。由于此模型类比了人们在生活中”阅读推理”的过程，我们将此模型称为*”Reading-VQA”*。

关键字：视觉问答、多模态、自然语言处理、计算机视觉

1. 引言

几年前火热的图像描述 (Image Caption) 任务涉及了人工智能中自然语言处理 (NLP)、计算机视觉 (CV)、知识表达与推理 (KR) 三大领域，但利用 $n-gram$ 语言模型 [1] 已经能把该问题处理得很好。因此有人发想：如果不仅要求能对图像进行标注描述，而是加上了”回答问题”的限制，是不是就能增加难度呢？也因此，学者开始研究视觉问题回答的任务。由于问题任意性与答案的开放性，目前对于 VQA 任务还没有一个完整确定的解决方式，可说是目前最具挑战性的”AI-complete”问题之一 [2]。

经过这几年的研究，VQA 的任务从各方面得到了改善。模型上的改进如利用 Tucker Fusion 的 MUTAN[3]、使用多模态的 MLB[4] 等；也有学者尝试将视觉问答任务加上已知常识与事实进行推理 (F-VQA)[5] 等等。然而，这些方式都是基于模型的优化。我们好奇，如果在训练 VQA 的过程中给予另一维度的”文字提示”，让 VQA 的问题转为”图文阅读推理”的问题，是否能有效地提升训练结果呢？因此，我们改良了原有的 VQA 模型，在训练时额外输入图像描述，并提出了两种 Reading-VQA 的模型，在 Visual Genome 数据集上进行训练与测试。

2. 相关研究成果

关于 VQA 已有许多研究成果，在此以著名的 MUTAN 介绍，我们的模型也是基于 MUTAN 模型进行改良的。

2.1. MUTAN

MUTAN 在 2017 年提出时曾造成一股轰动，原因在于他拿下了 VQA 挑战的第一名。MUTAN 之所以能成功，关键在于创新性的 Tucker Fusion：以往处理文字与图像的共同问题时，多数模型只使用了简单的 concat、reshape 等方式，无法完整的提取到图片与文字相关信息。

而 MUTAN 定义了一种融合的新技巧–Tucker Fusion，开了五个不同的线性层, 分别处理 Yes/No、What、When、Where、Who 这几种问题，并在融合的同时计算出与各属性的相似度 (Tensor Similarity) 与稀疏度 (Tensor Sparsity)，而进而能够进行分解 (Tucker Decomposition)。这种将图片与文字融合在同一个 Tensor

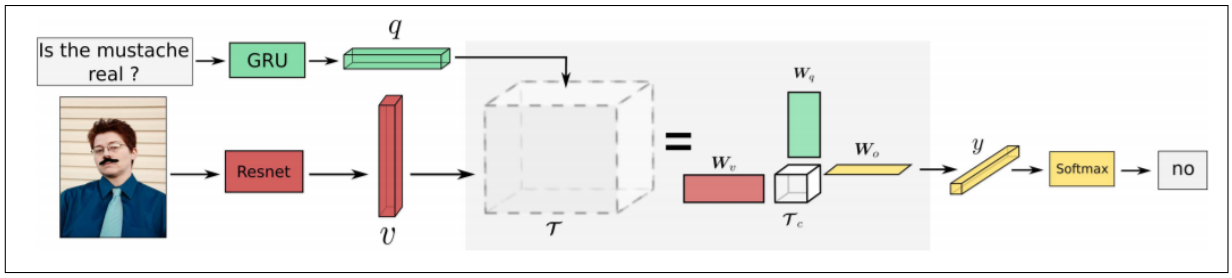

图 1. MUTAN 模型进行 VQA 的过程示意图

中进而进行多模态的分析与对比与分解的就是 Tucker Fusion。

3. 数据集处理

由于 Reading-VQA 的训练不仅需要原本的问题–回答训练集，还需要加入图像描述进行训练，我们使用了涵盖物体识别、区域分析、图片关联性描述、问答数据集的 Visual Genome 进行训练。总计下载了 10.7k 张图片、5M 个图片描述以及 11.5M 个 QA 进行训练。

3.1. 图像描述提取

我们首先建立每张图片对应的描述字典，利用预训练的 BERT 模型 [6] 进行预处理。平均每张图片大概有 10 个描述左右，然而每张图片的描述数量不一样多，为了确保大小与训练资讯的一致性，我们每次训练时仅随机抽取五句图像描述进行训练，每一句话会被 BERT 模型编码成 768 维的向量，再将五句话输入双向 LSTM（维数为 1024），并取 hidden cell（取 LSTM 的长期记忆）的平均值作输出。

3.2. 图片提取

利用 ResNet-152 提取图片的特征，变成 $10 \times 10 \times 2048$ 的图片矩阵。

3.3. 问题提取

利用 BERT 将输入问题提取成 2048 维的向量。

4. 模型提出与训练

我们从两个角度出发，提出了两个改良后的 Reading-VQA 模型。

4.1. Context-Question Concatenated Model

就如同许多简单的文字问题回答 (Text-based question answering) 任务般，答案往往就隐藏在题目之中。因此，若把输入描述并在问题之前，应该能够提升部份正确性与效率。

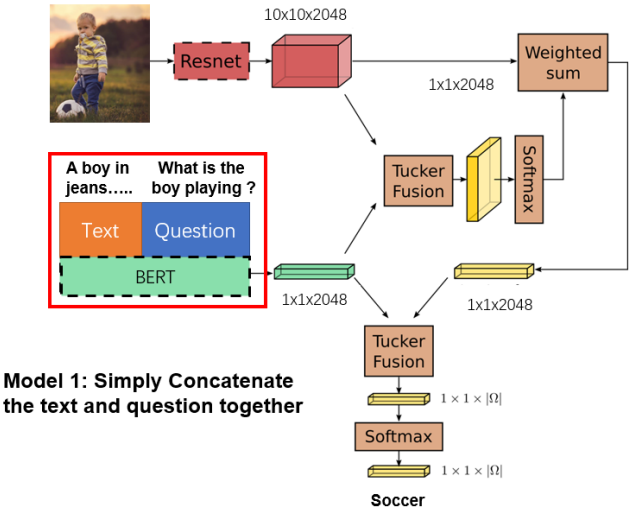

图 2. 将描述视为问题的一部份并将二者相连，作为 MUTAN 模型的问题输入

4.2. Read-Think-Answer Model

第一个模型强制把描述与问题併在一起。然而，经训练、测试后发现其效能并不高；在我们的训练过程中，Context 与 Question 的比例约为 5:1，意味着 BERT 可能提取更多文本信息而忽略了问题的重要性。

事实上，如同生活中所接受到的信息，描述的比例也都相对问题来的多；更重要的是，输入描述与图片的关联性在第一个模型中被遗漏了。因此，一个更好的方法应该是将图片、问题、输入描述三者先分别预处理，

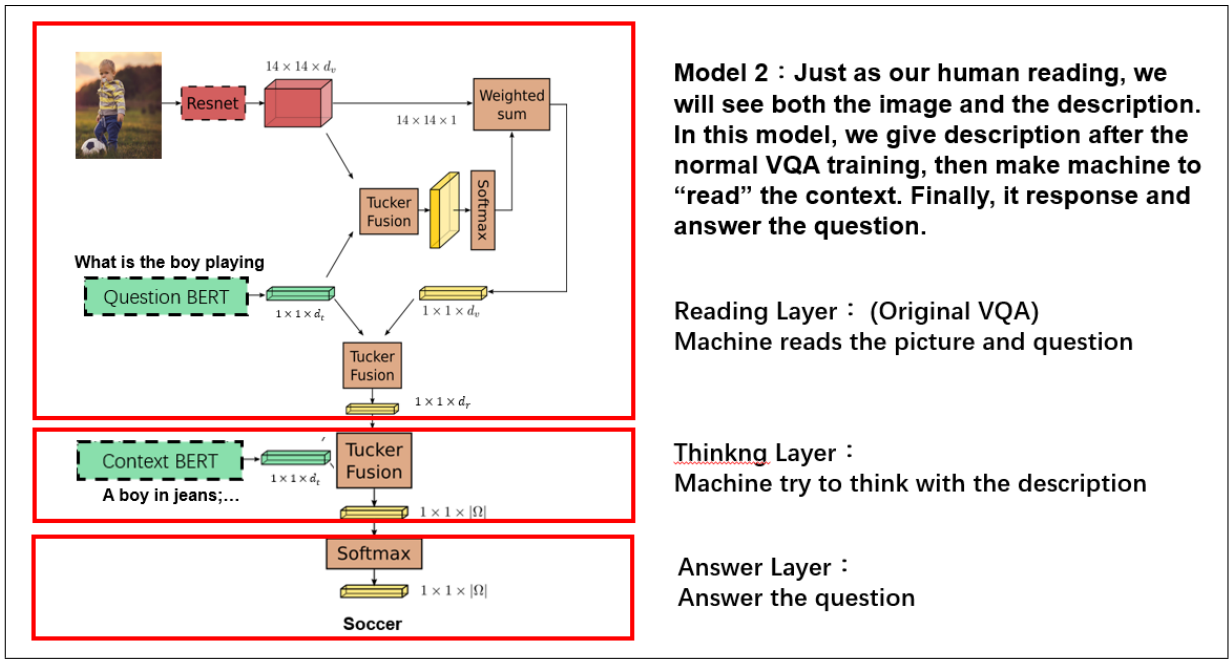

图 3. 读–想–答模型。分别提取问题、文字、描述的特征矩阵，再利用 Tucker Fusion 一层一层的融合起来。

提取相关特征。接着分成" 读、想、答" 三阶段：一开始机器看到输入图片与题目，相当于读到原本 VQA 的任务要求；接着给予一些文本描述, 进而思考图片与问题的关联性；最后才给出解答。在每一阶段的融合时使用了 MUTAN 模型中 Tuker Fusion 进行连接，最后利用 Softmax 函数并连接到输出层中。

5. 第一阶段训练结果

5.1. 与预训练模型对比

在前期的工作中，我们使用了预训练的 MUTAN 模型，并建立了一开始的两种 Reading-VQA 模型（写于 poster 中）。由于设备问题，我们仅在 $\frac{1}{10}$ 的 Visual Genome 上进行训练，并将预测结果上传至 CodaLab 的评估平台中进行测试。

Model	acc(test-dev)
Pretrained MUTAN	41.12
Model 1	38.25
Model 2	39.31

表 1. 训练结果发现，Reading-VQA 比预训练好的 MUTAN 模型还要更差，推测原因为数据集不够大，以及模型的能力受控于预训练 MUTAN 模型参数。

5.2. 模型重构

我们在 poster session 之后接受了老师、助教与同学的建议，重新修正了模型，并做了以下修正:

- 继续使用第二个模型：考量到输入描述与问题长度的比例以及第一阶段的测试结果，我们认为第二个模型的整体表现更好，因此在时间有限的情况下选取了第二个模型。
- 不使用预训练好的 MUTAN 模型, 自己复现并重新开始训练：由于预训练模型是在 COCO 数据集上测试的，且参数多已拟合于原本单纯的 VQA 任务。因此，我们认为自己重新实现并训练有可能会得到更好的结果。
- 使用 BERT 取代 GRU：BERT 作为近期最火热的语言模型，其能力已能远远超过传统的 RNN 模型，因此我们相信使用 BERT 对于 VQA 的任务会有所帮助。
- 在本机选取新的 criterion：由于原本的 MUTAN 模型是上传到 Codalab 进行 test 的，而在本地不太好训练。我们定义了一个粗浅版的损失函数以及准确率，作为判定模型好坏趋势的依据：

$$Loss = CrossEntropyLoss(ans, logits)$$
$$acc = \sum(logits.argmax(1)[0] == ans[0])/totalNum$$

准确度只比对字符串第一个词的原因在于语言生成的多样性：由于 Bert 的字符集 (30000 词) 相对于原本的字符集 (8000 词) 大很多，因此要让输出答案完全匹配是不太可能的。而答案的第一个字往往决定了回答的方向，因此可作为快速判定 acc 的一个标准。

5.3. 训练参数

模型重构后，我们重新进行了训练，训练参数如下：

超参数	相关值
Optimizer	Adam
Betas	0.9, 0.999
Learning Rate	10^{-4}
Batch Size	128
Drop-out	0.5

表 2. 重构后的模型训练结果。表现不够好的原因在于训练只迭代训练了一轮，训练未完全。

5.4. 第二阶段测试结果

由于数据集加大且面临期末考试周，我们没办法完整地训练好我们的新模型。一个只训练过一个 epoch（约三小时）的最好结果如下：

Model	Loss	acc
Rebuild Model	4.23	31.25

表 3. 重构后的模型训练结果。表现不够好的原因在于训练只迭代训练了一轮，训练未完全。

6. 结论

6.1. 结果与分析

藉由本次专题，我们得到了以下成果与结论:

- 两个模型的比较：相对于直接把输入段落并在问题前面，分别处理问题、描述段落与图片后再两两融合所得到的结果更好
- 预训练模型的使用：在 Reading-VQA 模型中使用预训练的 MUTAN 训练速度较快，但受限于训练好的参数，较难无法突破原有模型。同时推测因为第一阶段训练时数据集较小，即使我们使用预训练好的模型也无法重现论文中的 60% 准确度
- 重构新模型：我们尝试从头开始写我们的模型，但由于训练时间过短不完全，最好仅能达到约 30% 的准确度

6.2. 贡献

本次作业中，我们提出了另一种训练 VQA 的方法–结合 Context，并从” 阅读推理” 的角度去分析解决 VQA 任务。我们提出了两个 Reading-VQA 新模型，分别是 Context-Question Concatenated Model 以及 Read-Think-Answer Model 并使用了预训练的 MUTAN 模型构造出基本的 Reading-VQA 模型

我们也尝试去结合 BERT 模型并重构 Read-Think-Answer Model，虽然碍于作业时限原因没有训练完全，但这种结合输入段落以训练 VQA 任务的方式可在之后进行更多的探讨与研究。

参考文献

[1] Cavnar, William B., and John M. Trenkle. ”N-gram-based text categorization.” Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval. Vol. 161175. 1994.

[2] Antol, Stanislaw, et al. ”Vqa: Visual question answering.” Proceedings of the IEEE international conference on computer vision. 2015.

[3] Ben-Younes, Hedi, et al. ”Mutan: Multimodal tucker fusion for visual question answering.” Proceedings of the IEEE international conference on computer vision. 2017.

[4] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha, and B.-T. Zhang.Hadamard Product for Low-rank Bilinear Pooling. In 5th International Conference on Learning Representations, 2017.

[5] Wang, Peng, et al. ”Fvqa: Fact-based visual question answering.” IEEE transactions on pattern analysis and machine intelligence 40.10 (2018): 2413-2427.

[6] Devlin, Jacob, et al. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).