Deep Modular Co-Attention Networks for Visual Question Answering

Zhou Yu　Jun Yu　Yuhao Cui　Dacheng Tao　Qi Tian

摘要

在 VQA 任务中，设计一个有效的“共同注意力”模块来将问题中的关键词和图像中的关键区域联系起来是解决问题的核心。此前，大多数成功的共同注意力学习尝试都是通过使用浅层模型来实现的，而深度共同注意力模型与浅层模型相比几乎没有任何改进。在本文中，作者提出了一种深层模块化共同注意力网络（MCAN），它由多层模块化共同注意力（MCA）层级联组成。每一层 MCA 对问题和图像进行自我注意建模，并使用两个基本注意单元的模块化组合对图像进行问题引导的注意力建模。作者在基准的 VQA-v2 数据集上对 MCAN 进行了定量和定性评估，并进行了广泛的消融研究，以探索 MCAN 有效性背后的原因。实验结果表明，MCAN 明显优于之前的最先进水平。最佳单一模型在测试集上达到了 70.63% 的总体准确率。

关键词：视觉问答 (VQA)；共同注意力；深度模块化共同注意力网络 (MCAN)；模块化共同注意力 (MCA)；自注意 (SA)；引导注意 (GA)

1　引言

多模态学习在计算机视觉和自然语言处理领域受到了广泛关注，并且在视觉语言任务中取得了许多显著的进展，如图像-文本匹配、视觉字幕、视觉基础以及视觉问答 (VQA)。其中，VQA 是一项更具挑战性的任务，要求对图像和问题进行细粒度的语义理解，并通过视觉推理预测答案。注意力机制是深度神经网络研究的重要方面，已被广泛应用于单模态任务和多模态任务。VQA 中，学习图像区域的视觉注意和问题关键词的文本注意是非常重要的。研究表明，同时学习视觉和文本模式的共同注意可以有助于图像和问题的细粒度表征，从而实现更准确的预测。然而，这些共同注意模型学习的是多模态实例的粗糙交互，不能推断出图像区域和问题词之间的相关性，因此存在显著的局限性。

为了克服这些问题，提出了两种密集共注意模型 BAN 和 DCN，以模拟任意图像区域与任意疑问词之间的密集相互作用。然而，这些深度共同注意力模型与浅层模型相比几乎没有改进。这些深度共同注意力模型的瓶颈在于缺乏在各模态内同时建模较为密集的自注意力。

本文提出了一种新的深度模块化协同注意网络 (MCAN)，旨在提高视觉问答 (VQA) 的性能。受到机器翻译中的 Transformer 模型的启发，本文设计了两个注意单元：一个自我注意 (SA) 单元用于模拟模态内密集交互，以及一个引导注意单元 (GA) 用于模拟密集的多模态交互。通过组合 SA 和 GA 单元，得到了深度模块化协同注意网络。实验结果表明，MCAN 模型显著优于现有的共同注意模型，验证了自注意和引导注意在共同注意学习中的协同作用，也表明了深度推理的潜力。此外，研究发现对图像区域的自注意建模可以提高对象计数性能，这对 VQA 来说是一个挑战。

2 相关工作

2.1 视觉问答 (VQA)

近年来，视觉问答 (VQA) 受到了越来越多的关注。最简单的 VQA 解决方案是将图像和问题表示为全局特征，并使用多模式融合模型预测答案[1]。然而，全局特征表示可能会丢失关键信息，因此，最近的研究方法将视觉注意机制引入到了 VQA 中，通过自适应学习给定问题的参与图像特征，然后使用多模态特征融合来获得准确的预测结果。Chen 等人提出了一种问题引导的注意映射技术[2]；Yang 等人提出了一种堆叠注意网络[3]；Fukui 等人[4]，Kim 等人[5]，Yu 等人[6][7]和 Ben 等人[8]利用不同的多模态双线性池化方法，将图像空间网格中的视觉特征与文本特征相结合，预测注意。Anderson 等人[9]引入了一种自底向上和自顶向下的注意机制，来学习对候选对象的注意。

2.2 共同注意模型

共同注意模型是 VQA 的重要研究方向，它不仅要理解图像的视觉内容，还需要完全理解自然语言问题的语义。为了实现这一目标，Lu 等人提出了一种交替学习图像注意和问题注意的共同注意学习框架[10]。Yu 等人简化了共同注意方法为两个步骤：自我注意和问题条件注意[7]。Nam 等人提出了一种基于先前注意记忆的多阶段协同注意学习模型，以加深注意[11]。然而，这些协同注意模型仅学习了每个模态（图像或问题）的独立注意分布，忽略了问题词与图像区域之间的密切交互。因此，理解多模态特征的细节关系成为瓶颈。为解决这一问题，人们提出了密集共同注意模型，该模型建立了问题词与图像区域之间的完整交互[12][13]。与先前的粗略交互的共同注意模型相比，密集交互模型具有更好的 VQA 性能。

3 本文方法

3.1 本文方法概述

整个网络如图 1 所示，分为三个模块, 由于 Deep Co-Attention Learning 有两种方式，对采用 stacking 结果的 L 层 MAC 的模型称为 $MCAN_{sk}$-L，采用 encoder-decoder 结构的模型称为 $MCAN_{ed}$-L

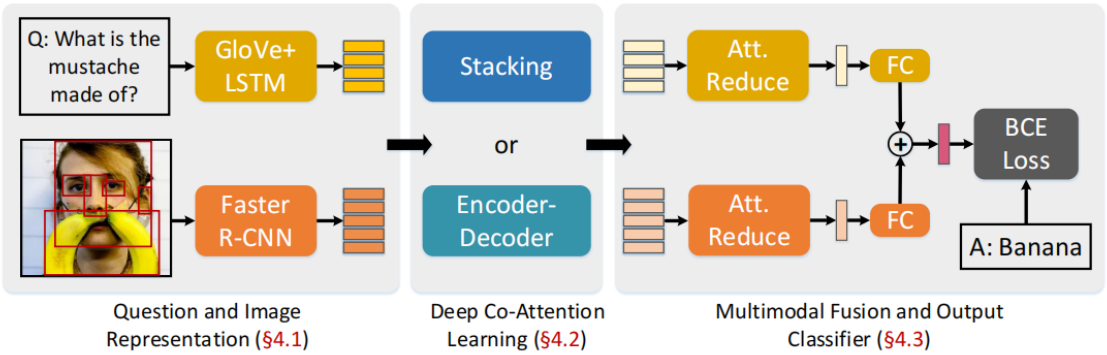
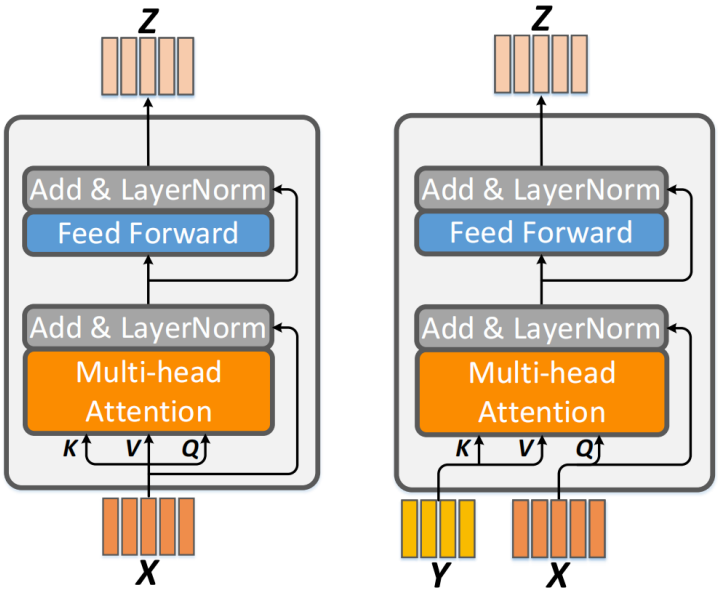


图 1: 整体网络结构图

3.2 MCA 层



(a) Self-Attention (SA) (b) Guided-Attention (GA)

图 2: SA 单元和 GA 单元的模型架构图

根据图 2，MCA 层由 SA 单元和 GA 单元模块化组合构成。在传统的 SA 单元中，把单模态特征 X 转换成查询特征 q、关键字特征 K 和值特征 V，令他们的维度均为 d，从而根据如下公式得到 SA 特征 f。

$$f = A(q, K, V) = \text{softmax}\left(\frac{qK^T}{\sqrt{d}}\right)V \tag{1}$$

为了进一步改进 SA 特征的表示能力，在 SA 单元中引入 multi-head attention，本文得到 multi-head SA 单元，如图 2(a) 所示。在 multi-head SA 单元中，每个 head 是一个 SA 单元，各个 head 之间是并行的，通过计算每个 head 的 SA 特征，获得 multi-head SA 特征 F，如下公式所示。

$$F = MA(q, K, V) = [\,\text{head}_1, \text{head}_2, \cdots, \text{head}_h\,]\,W^{\circ} \tag{2}$$

$$\text{head}_j = A\left(qW_j^Q, KW_j^K, VW_j^V\right) \tag{3}$$

然后将原特征 X 和 F 进行残差和连接并采用全连接进行变换，得到 multi-head SA 特征 F^a，如下公式所示。

$$F^a = \text{LayerNorm}\,(X + F) \tag{4}$$

经过两层全连接层进行线性和非线性变换，得到 multi-head SA 特征 F_Z^a，如下公式所示

$$F_Z^a = FC - RELU - FC\,(F^a) \tag{5}$$

F_Z^a 再与原特征 F^a 进行残差和连接并采用全连接进行变换，得到最终的 multi-head SA 特征 Z，如下公式所示。

$$Z = \text{LayerNorm}\,(F^a + F_Z^a) \tag{6}$$

图 2(b) 中的模态间的 GA 特征计算方法类似于图 1(a) 模态内的 SA 特征。

3.3 MCA 层的变形

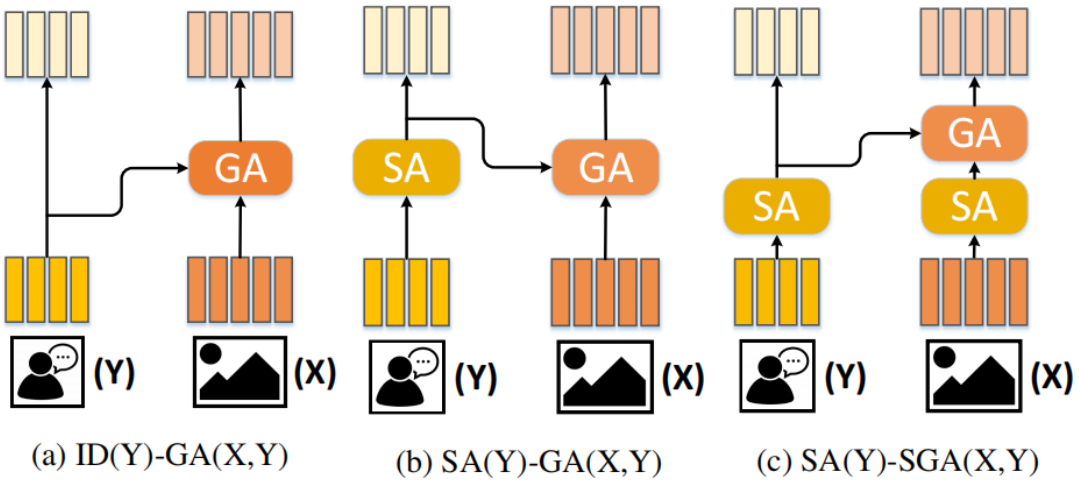


图 3: MCA 层的 3 种变形

根据图 3，通过不同的方式组合 SA 单元和 GA 单元，本文得到 3 种不同的 MCA 层，且它们都能够在深度上级联以得到深层的 MCAN 模型。图 3(a) 中的 MCA 层作为 baseline。在图 2(a) 中，输入端的问题特征一方面通过恒等映射直接传递到输出端作为最终的问题特征，另一方面作为 GA 单元的输入端，指导图片特征生成 dense inter-modal 的图片特征。图 3(b) 在图 3(a) 的基础上增加 SA(Y) 单元，用以建模 dense intra-modal 的问题特征。图 3(c) 在图 3(b) 的基础上增加 SA(X) 单元，用以建模 dense intra-modal 的图片特征。本文通过级联图 3(c) 中的 MCA 层得到深层的 MCAN 模型。

3.4 MCA 层的级联方式

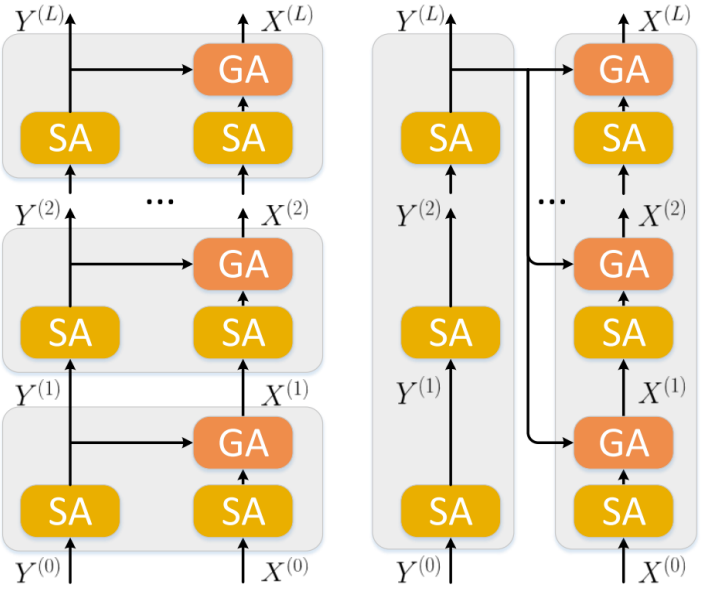


图 4: MCA 层的级联方式

MCA 层在深度级联的方式有 Stacking 和 Encoder-Decoder 两种方式，如图 4 所示。对于 Stacking 的模型，前一层的输出作为下一层的输入，最后一层的输出作为最终的问题特征和图片特征。对于 Encoder-Decoder 的模型，编码器相当于堆叠 SA(Y) 单元得到最终的问题特征，解码器相当于堆叠 SGA(Y,X) 单元，最终得到问题指导的图片特征。

3.5 多模态融合与答案预测

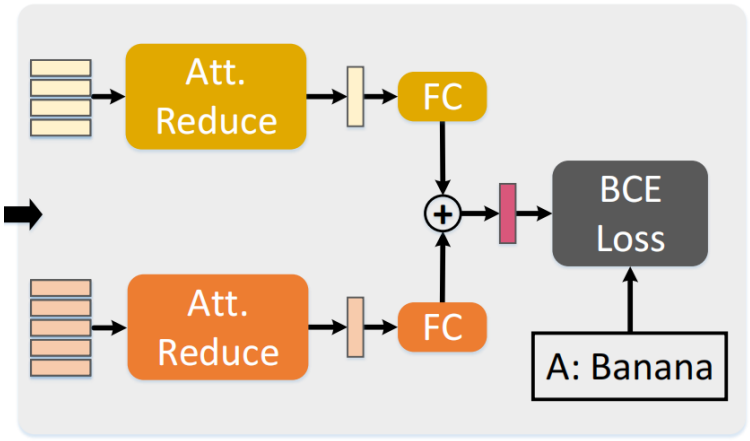


图 5: 多模态融合与答案预测流程图

根据图 5，本文采用深层的 MCAN 模型得到图片特征 X^L 和问题特征 Y^L，然后设计 attentional reduction model 对 X^L 和 Y^L 进行降维处理，得到低维度的图片特征 $\tilde{x}$ 和问题特征 $\tilde{y}$，接着采用和的方式对这两个低维度的特征进行线性多模态融合，得到多模态融合特征 z。对 z 进行非线性变换，得到向量 s，最后采用多分类器对向量 s 分类并获取图片问答的答案。

$$f = A(q, K, V) = \text{softmax}\left(\frac{qK^T}{\sqrt{d}}\right)V \tag{7}$$

4 复现细节

4.1 与已有开源代码对比

实验过程参考作者的源码实现了 MACN 网络结构

首先定义了一个多头注意力机制（MHAtt）的类。该类包括了一个前向传播的函数 forward 和一个多头注意力的函数 att。forward 函数接受 value、key 和 query 三个输入，以及一个 mask，然后将这些输入送入多头注意力函数中进行计算，最后输出经过线性变换后的多头注意力结果。

然后定义 MCAN（Multimodal Compact Bilinear Attention）模型中的 Self-Attention（SA）结构。该结构包含两个子层：多头自注意力层（MHAtt）和前馈神经网络层（FFN）。通过对输入进行多头注意力计算并应用前馈神经网络，该结构能够从图像和问题的输入中提取特征，并输出最终的表示。

具体而言，该模型的实现包括以下几个步骤：

1. 实例化一个多头自注意力层和一个前馈神经网络层

2.定义两个dropout层和两个归一化层

3.在前向传递中，首先对输入张量x应用多头自注意力机制，计算多头注意力输出

4.将多头注意力输出与原始输入x进行加法操作，并对结果进行归一化和dropout处理

5.将上一步的结果应用到前馈神经网络层中，计算FFN输出

6.将FFN输出与上一步的结果进行加法操作，并再次进行归一化和dropout处理

7.返回最终结果

最后实现 MCAN 的 SGA（Second-order Graph Attention）模块，其功能是在一个多模态的场景下，对两个模态的特征进行建模和交互，从而获得更丰富的表征。SGA 模块通过两次多头注意力机制，

实现了两个模态的交互，再通过全连接层进行特征的非线性变换和整合。

4.2 实验环境搭建

首先安装一些必要的软件包:

安装 Python >= 3.5

安装 Cuda >= 9.0 和 cuDNN

使用 CUDA 安装 PyTorch >= 0.4.1（也支持 Pytorch 1.x）安装 SpaCy 并初始化 GloVe

运行以下脚本来设置实验所需的所有配置：

```
sh setup.sh
```

4.3 界面分析与使用说明

以下脚本将使用默认超参数开始训练：

```
python3 run.py –RUN=’train’
```

验证和测试脚本命令：

```
python3 run.py –RUN=’val’ –CKPT_V=str –CKPT_E=int
```

其中，str 取值’small’, ’large’，对应 2 个模型，int 取值 13

5 实验结果分析

5.1 消融实验

5.1.1 不同 MCA 层的实验

Model	All	Y/N	Num	Other
ID(Y)-GA(X,Y)	64.8	82.5	44.7	56.7
SA(Y)-GA(X,Y)	65.2	82.9	44.8	57.1
SA(Y)-SGA(X,Y)	**65.4**	**83.2**	**44.9**	**57.2**

图 6: 不同 MCA 层的实验结果

根据图 6，SA(Y)-GA(X,Y) 的结果全面超过 ID(Y)-GA(X,Y)，表明问题的模态内 attention 能够提高 VQA 的性能；SA(Y)-SGA(X,Y) 的结果全面超过 SA(Y)-GA(X,Y)，表明图片的模态内特征能够提高 VQA 的性能。

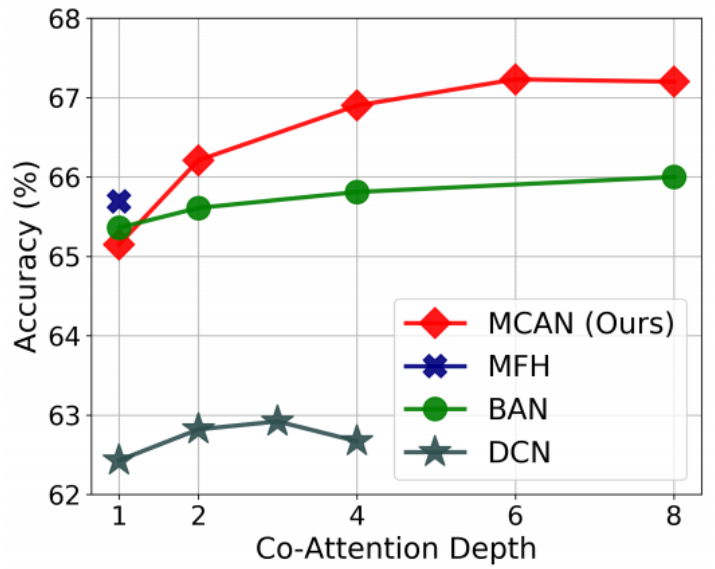


图 7: 不同 MCA 层的实验结果

根据图 7，随着层次加深，ID(Y)-GA(X,Y) 和 SA(Y)-GA(X,Y) 的问答准确率几乎一致，但 SA(Y)-SGA(X,Y) 的准确率显著高于其他两个模型，因此图片的模态内 attention 特征对计数问题起关键作用。

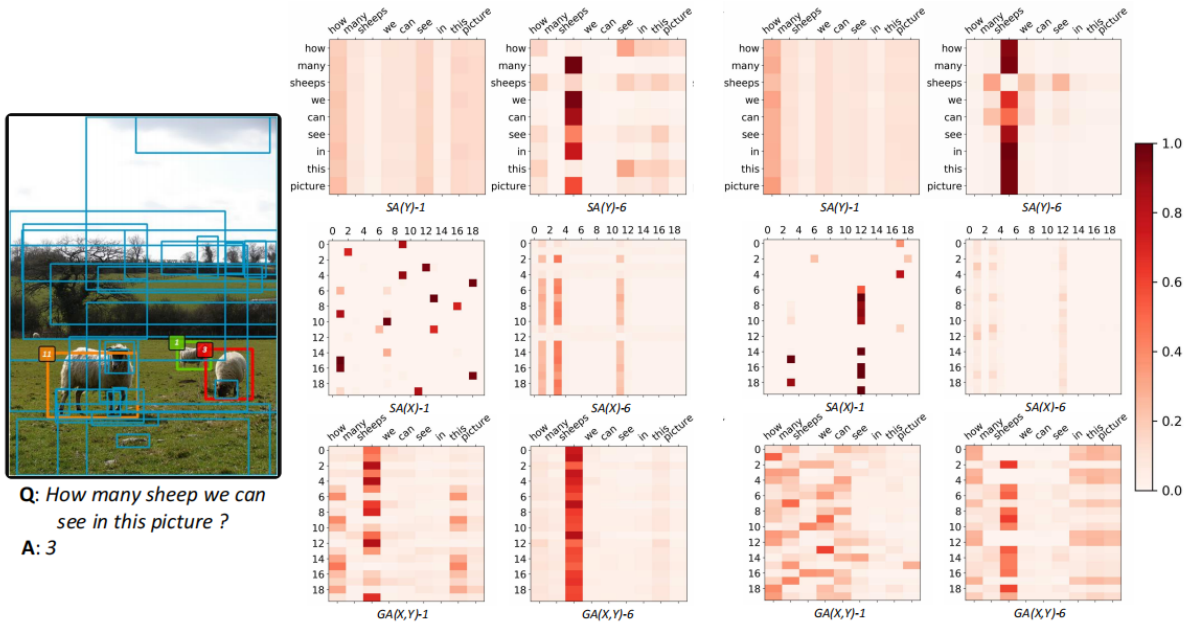
5.1.2 MCA 层级联方式的实验

L	MCAN_{sk}	MCAN_{ed}	Size
2	66.1	66.2	27M
4	66.7	66.9	41M
6	66.8	**67.2**	56M
8	66.8	**67.2**	68M

图 8: 不同 MCA 层的实验结果

根据图 8，Encoder-decoder 级联方式优于 Stacking 方式，在级联的层数为 6 时，两种级联方式都达到最好的性能。

5.2 attention map 的可视化分析



(a) Encoder-Decoder (**P**: 3) (b) Stacking (**P**: 3)

图 9: 不同 MCA 层的实验结果

根据图 9，当层数为 1 时，对于 Encoder-Decoder 方式，问题的模态内 attention 定位 how 和 see，问题的模态间 attention 定位 how、sheep 和 picture，图片的模态内 attention 定位不清晰。当层数为 6 时，对于 Encoder-Decoder 方式，问题的模态内和模态间 attention 都定位到 sheep，图片的模态内 attention 定位 3 个 sheep，故预测答案是 3。同理可以分析 Stacking 方式预测的答案也是 3，但 Stacking 方式定位的效果没有编码器解码器方式定位的效果好，从这也可以看出采用 Encoder-Decoder 方式级联模型能够获得更好的效果。

5.3 MCAN 模型的实验结果

Model	Test-dev				Test-std
	All	Y/N	Num	Other	All
Bottom-Up [28]	65.32	81.82	44.21	56.05	65.67
MFH [33]	68.76	84.27	49.56	59.89	-
BAN [14]	69.52	85.31	50.93	60.26	-
BAN+Counter [14]	70.04	85.42	**54.04**	60.52	70.35
MCAN_{ed}-6	**70.63**	**86.82**	53.26	**60.72**	**70.90**

图 10: 不同 MCA 层的实验结果

根据图 10，本文的模型超过以前的 state-of-the-art 结果。

复现实验结果如图 11 所示

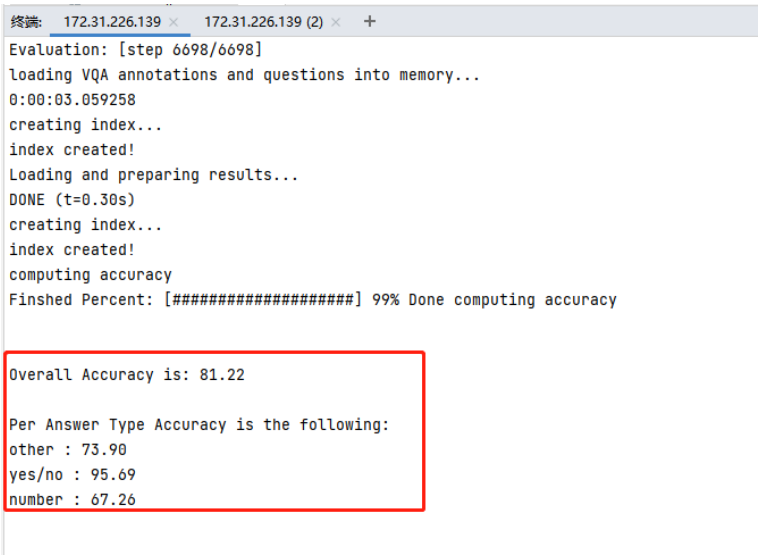


图 11: 复现实验结果

6 总结与展望

该复现论文提出了一种用于多模态推理的深度神经网络模型，该模型可以同时处理图像和文本输入，并在不同的任务上达到最先进的性能。该模型的主要贡献在于提出了多层注意力机制，这些机制旨在对不同的模态和任务进行自适应加权，以提高模型的表示能力。此外，MCAN 还使用了自注意力机制来捕获文本输入中的长期依赖性，并使用共现关系来增强图像和文本之间的交互。该模型在多个视觉和自然语言处理任务上进行了评估，包括视觉问答、图像标注和文本生成等任务，均取得了领先的结果。未来，可以通过进一步改进模型架构和训练方法来进一步提高模型的性能，并将其应用于更广泛的多模态推理任务中。

参考文献

[1] ZHOU B, TIAN Y, SUKHBAATAR S, et al. Simple baseline for visual question answering[J]. arXiv preprint arXiv:1512.02167, 2015.

[2] CHEN K, WANG J, CHEN L C, et al. Abc-cnn: An attention based convolutional neural network for visual question answering[J]. arXiv preprint arXiv:1511.05960, 2015.

[3] YANG Z, HE X, GAO J, et al. Stacked attention networks for image question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 21-29.

[4] FUKUI A, PARK D H, YANG D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[J]. arXiv preprint arXiv:1606.01847, 2016.

[5] KIM J H, ON K W, LIM W, et al. Hadamard product for low-rank bilinear pooling[J]. arXiv preprint arXiv:1610.04325, 2016.

[6] YU Z, YU J, FAN J, et al. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1821-1830.

[7] YU Z, YU J, XIANG C, et al. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering[J]. IEEE transactions on neural networks and learning systems, 2018, 29(12): 5947-5959.

[8] BEN-YOUNES H, CADENE R, CORD M, et al. Mutan: Multimodal tucker fusion for visual question answering[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2612-2620.

[9] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6077-6086.

[10] LU J, YANG J, BATRA D, et al. Hierarchical question-image co-attention for visual question answering [J]. Advances in neural information processing systems, 2016, 29.

[11] NAM H, HA J W, KIM J. Dual attention networks for multimodal reasoning and matching[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 299-307.

[12] NGUYEN D K, OKATANI T. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6087-6096.

[13] KIM J H, JUN J, ZHANG B T. Bilinear attention networks[J]. Advances in neural information processing systems, 2018, 31.