基于 Transformer 解决 VQA 中语言先验问题的尝试

摘要

大部分视觉问答（Visual Question Answering，VQA）模型都存在由固有数据偏差导致的语言先验问题。具体来说，现有 VQA 模型倾向于忽略图片内容而根据答案频率回答问题。本文尝试基于 SSL 框架将 Transformer 作为 VQA 模型以解决 VQA 中语言先验的问题。实验结果表明，最终基于 Transformer 的 VQA 模型在参数量减少到原来六分之一的情况下仍能取得与基线模型相当的性能，证明了 Transformer 在解决语言先验问题上的有效性。

1. 研究背景

VQA（Visual Question Answering）作为多模态方向的下游任务之一，近些年来得到了越来越多研究者的关注。VQA 需要模型根据图片内容回答给定问题，因为要求模型同时理解文本和图像两种模态的信息。然而之前的一些工作(Jabri et al., 2016; Agrawal et al., 2016; Zhang et al., 2016; Goyal et al., 2017)表明部分的 VQA 系统存在语言先验（Language Priors）的问题，具体表现为 VQA 系统总是根据给定问题的类型输出该类型问题最为常见的答案而忽略图片信息。例如当 VQA 系统遇到测试集中的“how many”类型问题时总会给出答案 2，因为这类问题在训练集中的答案通常都是 2。上述现象表明了现存的 VQA 系统并没有真正的理解图片和文本信息。

为了解决语言先验的问题，各种方法随之被提出，其大致可以分为两类：（1）通过模型设计来减少语言偏置；归为这类方法的大多数工作都是基于集成（Ensemble）的模型(Ramakrishnan et al., 2018; Grand and Belinkov, et al., 2019; Cadene et al., 2019; Clark et al., 2019; Mahabadi and Henderson, 2019)。这类方法中最为经典的莫过于 LMH 模型(Clark et al., 2019)，它通过惩罚不根据图片内容回答问题的样本来减少语言偏置。（2）利用数据增强来减少语言偏置：这类方法(Zhang et al., 2016; Goyal et al., 2017; Agrawal et al., 2018)的主要思想是构建更加平衡的数据集来克服语言先验，例如 SSL 方法(Zhu et al., 2020)首先自动生成一些平衡的问题-图片对，之后利用这些数据并引入一个辅助任务将分类任务转化为多任务学习，最终达到效果的提升。

2. 系统实现

本文主要基于 SSL 框架对 Transformer(Vaswani et al., 2017)模型进行修改，并使用修改后的 Transformer 模型替换之前所使用的 Updn(Anderson et al., 2018)模型，以此试图解决 VQA 任务中语言先验的问题。下面对本文所使用的技术进行介绍。

2.1 问题定义

在 VQA 任务中，模型需要同时接受一个问题和图片对作为输入。要求 VQA 系统能够根据提供图片 I 的内容回答给定问题 Q。VQA 系统的目标是从答案集合 A 中得到满足下列公式的最优答案$\bar{a}$:

$$\bar{a} = \arg\max_{a \in A} P(a \mid I, Q)$$

其中 P(a| I, Q)表示在给定 I，Q 的条件下，生成答案 a∈A 的条件概率。

2.2 SSL 框架

本文基于 SSL 框架进行实验，该框架结构如图 2 所示。该框架类似于多任务学习，由两个任务组成既包括 VQA 任务，又引入了一个名叫 Question-Image Correlation Estimation(QICE)的辅助任务，该任务是一个二分类任务用于 VQA 系统在回答问题之前判断问题与图片是否相关。该框架利用 QICE 任务将整体分为两个分支，因为该框架的两个分支有完全相同的输入和类似的输出，所以两个 VQA 模型可以共享神经网络参数，即使用同一个 VQA 模型；第一个分支是将问题与图片相关的问题-图片对(Q, I)输入到 VQA 模型中；第二个分支则是将问题与图片不相关的问题-图片对(Q,I')输入到 VQA 模型中。不相关的问题-图片对是通过在一个 mini batch 中随机挑选另一张图片的特征向量输入到模型中实现。

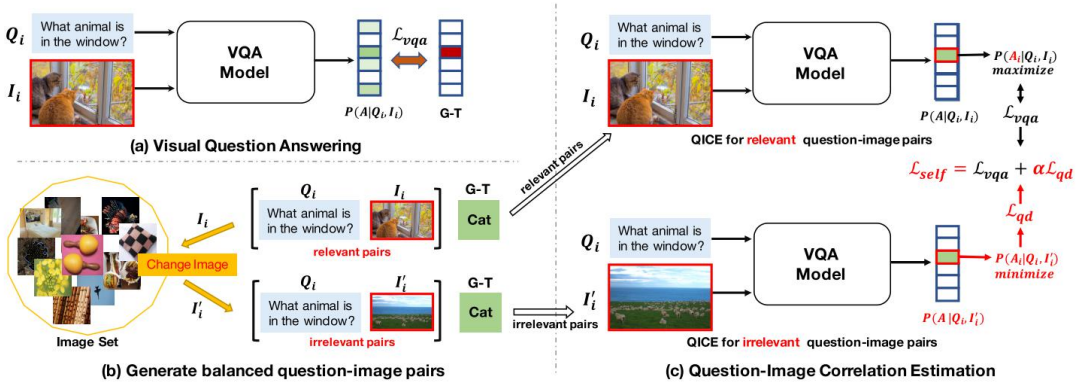

图 2：SSL 模型架构图

如图 2(a)中的 VQA 模型所示，它接收相关的问题-图片对作为输入来预测答案的概率分布 $P(a_i|Q_i, I_i)$。VQA 模型通过最小化如下损失函数进行优化

$$L_{vqa} = -\frac{1}{N}\sum_{i}^{N}[t_i \log(\delta(P(A|Q_i, I_i))) + (1 - t_i)\log(1 - \delta(P(A|Q_i, I_i)))]$$

其中δ(·)代表 sigmoid 函数，t_i是第 i 个问题对应每个答案的软目标分数(soft target score)，可以表示为$\frac{k}{n}$，n 是第 i 个问题的有效答案数目，k 则是第 i 个问题人工标记的答案数目。

如图 2(c)当给定不相关的问题图片对时，直观上可以通过 VQA 模型对正确答案的置信度来衡量 VQA 模型的问题依赖性，置信度越大则依赖性越强。因此考虑通过最小化该置信

度来防止 VQA 模型被语言先验过度取顶，这里将它命名为问题依赖损失L_{qd}：

$$L_{qd} = -\frac{1}{N}\sum_{i}^{N} log\ (1 - P(A_i|Q_i, I_i^{'}))$$

数学上最小化L_{qd}中的$-\log(1-P(A|Q,I^{'}))$等价于最小化$P(A|Q,I^{'})$。从实验中发现在训练时最小化$P(A|Q,I^{'})$比最小化$-\log(1-P(A|Q,I^{'}))$更稳定，原因是$P(A|Q,I^{'})$的梯度比$-\log(1-P(A|Q,I^{'}))$更稳定。因此选择优化$P(A|Q,I^{'})$)，而问题依赖损失L_{qd}可以修改为：

$$L_{qd} = \frac{1}{N}\sum_{i}^{N} P(A_i|Q_i, I_i))$$

最终损失函数写为：

$$L_{self} = L_{vqa} + \alpha L_{qd}$$

其中L_{vqa}是 VQA 损失函数，α是超参数。当$\alpha = 0$时，L_{self}可以被视作单纯的 VQA 损失函数，这意味着问题依赖损失L_{qd}实际上是充当一个正则项来防止模型记忆语言先验并且强迫它更好的理解图片。L_{self}提供了平衡问答问题和减少语言先验之间的灵活性，并且做到了在不使用外部监督的情况下以自监督的方式减轻语言先验。

2.3 多模态 Transformer

VQA 任务要求模型同时接受文本和图片两种模态的输入，然而 Transformer 设计之初便被用于机器翻译任务，只能接收文本一种模态的信息。受 MCAN(Yu et al., 2019)模型启发，本文对原生 Transformer 进行部分修改，模型的整体结构如下图所示。下面将对模型的细节进行介绍。

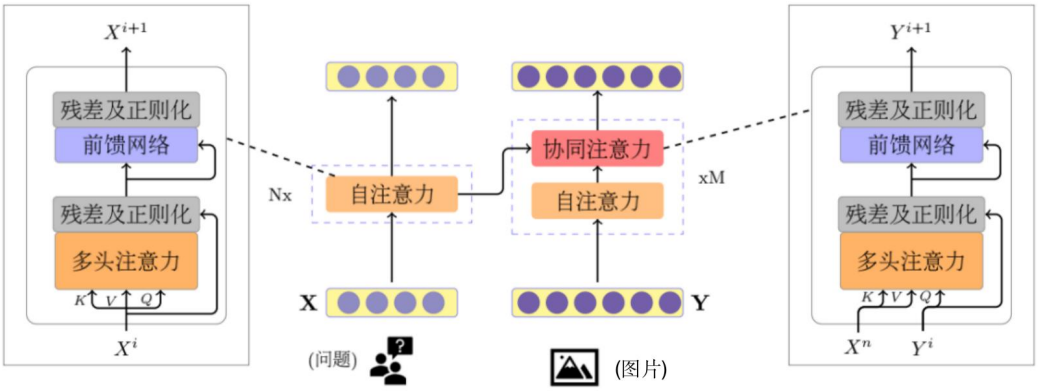

图 3 多模态 Transformer 结构

2.3.1 问题和图片表示

问题表示:

本文得到问题表示的具体做法是首先将输入的问题进行分词，然后类似于(Anderson et al.,2018)中的做法将其修剪为最多 14 个词。之后再使用 300 维的 Glove 对问题中的每个单词作嵌入使得每个问题都被转化为一个 $n \times 300$ 的二维矩阵表示，其中 n 是问题中单词的个数，随后再将这些词嵌入传进一个一层的长短时记忆网络（Long Short-Term Memory, LSTM)中进一步生成包含上下文信息的特征向量，最终将得到的特征向量 $X \in \mathbb{R}^{14 \times d}$作为给定问题的表示也就是模型的输入。每个问题 $Q = \{w_i | i = 1,2,\ldots,n\}$经过上述转换过程得到特征向量 X 的形式化表示如下：

$$X = LSTM(Glove(Q))$$

其中 n 为问题中单词的个数。

图片表示:

传统的 Transformer 并不接收图片作为输入，因此需要找到一个对图片合理的处理方式。针对图片本文采用了“自下而上”的区域编码方法(Anderson et al, 2018):利用 Visual Genome 数据集(Krishna et al., 2017)预训练得到 Faster R-CNN 模型(Ren et al., 2015)，基于该模型对每张图片进行目标检测，每张图片被编码为一组包含 36 个区域框的集合，每个区域框都对应一个对象，最终每个对象都被表示为一个 2048 维的特征向量，因而每张图片则对应一个 $Y \in \mathbb{R}^{36 \times 2048}$的图片表示。图片 I 经过上述转换过程得到特征向量 Y 的形式化表示如下：

$$Y = FasterRCNN(I)$$

2.3.2 编码器和解码器

编码器:

多模态 Transformer 的编码器（Encoder）与传统 Transformer 中的编码器组成相同，同样包含多头注意力层、正则化、前馈神经网络层和残差连接。多模态编码器接收问题特征 $X \in \mathbb{R}^{14 \times d}$作为输入，经过矩阵映射后得到相应的查询矩阵 $Q \in \mathbb{R}^{n \times d_{query}}$,键值矩阵 $K \in \mathbb{R}^{n \times d_{key}}$和实值矩阵 $V \in \mathbb{R}^{n \times d_{value}}$。这里$d_{query} = d_{key} = d_{value} = d$。注意力采用缩放点积(Scaled dot-product)运算，其计算方法如下所示：

$$Q = XW^Q, K = XW^K, V = XW^V$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V$$

为了进一步提高表示能力，模型采用多头注意力(Multi-head Attention)机制(Vaswani et al., 2017)。多头注意力包含 h 个注意力运算，每个注意力运算对应了缩放点积运算，将运算结果拼接成为多头注意力层的输出表示：

$$M\ Attention(Q, K, V) = Concat(Attention_1, Attention_2, \ldots, Attention_h)W^o$$

这里 $W^o \in R^{(d \times h) \times d}$是可训练参数。多头注意力层的输出 $f \in \mathbb{R}^{n \times d}$再经过残差连接与层正则化以防止梯度消失和加速模型收敛：

$$F = LayerNorm(X + M\ Attention(Q, K, V))$$

经过前向层后得到注意力模块的最终输出为：

$$X^* = LayerNorm(f + FFN(f))$$

解码器:

多模态 Transformer 中的解码器（Decoder）与编码器不同之处主要在于查询矩阵、键值矩阵和实值矩阵的生成方式。为了实现问题特征与图片特征两种模态的交互，本文选择将经过编码器计算后输出的问题特征X^*进行矩阵映射作为查询矩阵，而图片特征 Y 进行矩阵映射作为键值矩阵和实值矩阵传入注意力层：

$$Q = X^*W^Q, K = YW^K, V = YW^V$$

之后的运算与自注意力模块相同。

2.3.3 输出层

经过模型的学习，多模态编码器输出的问题特征X^*与解码器输出的多模态特征Y^*已经包含了足够多的注意力信息，因此需要对其作进一步的融合。本文首先使用一个两层的 MLP(FC(d)-ReLu-Dropout(0.1)-FC(1)对X^*(或者Y^*)进行降维得到x(或者$\tilde{y}$)。以问题特征X^*为例说明得到$\tilde{x}$的过程：

$$\alpha = \text{softmax}(\text{MLP}(X^*))$$

$$\tilde{x} = \sum_{i=1}^{m} \alpha_i x_i$$

其中$\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_m] \in \mathbb{R}^m$是所学习到的注意力权重。$\tilde{y}$采用同样的方式使用一个独立的 MLP 计算得到。

之后再对得到的x和y进行融合，融合的具体方式如下：

$$z = \text{LayerNorm}(W_x^T x + W_y^T y)$$

其中W_x，$W_y \in \mathbb{R}^{d \times d_z}$是两个线性映射矩阵，$d_z$是融合特征的维度，并且使用层正则化稳定训练。融合后得到的特征 z 通过 sigmoid 函数映射成$s \in \mathbb{R}^N$的向量，其中 N 为训练集中答案的个数。随后模型使用二分类交叉熵(Binary Cross-Entropy, BCE)作为损失函数训练 N 分类器，最终选择分数最高的分类结果作为给定问题的答案返回。

3. 实验

3.1 数据集和基线模型

数据集：

本文使用相关论文中最为常用的 VQA-CP v2[Agrawalet al., 2018]数据集来对模型进行评估。VQA-CP v2 数据集是由 VQA v2[Goyal et al., 2017]数据集重新组织训练集和验证集构成，其 QA 对在训练集与验证集有不同的数据分布，因此该数据集非常适合评测 VQA 系统的泛化能力。VQA-CP v2 数据集由训练集和测试集两部分组成，其中训练集包括大约 121 千张图片和 448 千个问题，而测试集大约包括 98 千张图片和 220 千个问题。

基线模型：

本文将所使用的方法与下列基线模型进行对比：Updn(Anderson et al., 2018)，AReg(Ramakrishnan et al., 2018), RUBi(Cadene et al., 2019)，LMH(Clark et al., 2019)，SSL(Zhu et al., 2020)和 LXMERT(Tan and Bansal, 2019)。上面列举的大部分模型设计之初便是为了解决语言先验的问题，而 LXMERT 模型则代表最近一系列多模态预训练模型 Oscar(Li et al., 2020)、Visualbert(Li et al., 2019)等，这些预训练模型在多模态任务上通常有非常好的效果包括在 VQA v2 数据集上。

3.2 实验设置

本文使用与基线模型相同的预训练好的 Faster R-CNN 来抽取图片特征。每张图片被编码为一组包含 36 个对象的集合，每个对象使用一个 2048 维的向量来进行表示，再经过一个线性层转化为 128 维。所有的问题都被填充为相同的长度，长度为 14。问题中的每个单词都通过 300 维 Glove 对其进行初始化，之后又被送入 LSTM 最终得到 128 维句子级别的表示。

本文先使用 VQA 损失函数预训练模型 10 个 epoch，之后再引入辅助任务的损失函数作为模型整体的损失函数微调模型 20 个 epoch。本文使用的 batch size 大小为 64。训练模型过程中所使用的不相关图片随机选自与同一 mini-batch 中的其他图片特征。

3.3 实验结果与分析

3.3.1 主要结果

模型	VQA-CP v2 test(%)			
	Overall	Yes/No	Num	Other
Updn(Anderson et al., 2018)	39.74	42.27	11.93	46.05
Areg(Ramakrishnan et al., 2018)	41.17	65.49	15.48	35.48
RUBI(Cadene et al., 2019)	47.11	68.65	20.28	43.18
LMH(Clark et al., 2020)	52.45	69.81	44.46	45.54
LXMERT(Tan and Bansal, 2019)	46.23	42.84	18.91	55.51
Mcan(Yu et al., 2019)	43.13	42.69	14.87	51.11
SSL(Zhu et al.,2020)	57.59	86.53	29.87	50.03
SSL+Transformer	57.34			

表 1：模型在 VQA-CP v2 数据集上的结果

模型	Overall	参数量
SSL(Zhu et al., 2020)	57.59	36M
SSL+Transformer	57.34	6M

表 2：模型参数量的对比

本文受 Mcan(Yu et al., 2019)模型启发, 尝试将 Transformer 模型移植到 SSL(Zhu et al., 2020)框架上。表 1 此外还汇总了包括 Updn(Anderson et al., 2018)、Areg(Ramakrishnan et

al., 2018)、RUBI(Cadene et al., 2019)等相关工作。结合表 1 和表 2 可以看出，本文参数量减少为基线模型的六分之一，总数为 6M 而性能也与基线模型相当。

3.3.2 Pretrain epoch 位置的影响

为了分析 Pretrain epoch 位置对系统性能的影响，我尝试了在不同的 epoch 位置进行 pretrain。对比结果如下表所示：

Layer	pretrain_epoch	Overall
4	7	48.58
4	12	52.97
4	15	53.65
4	18	52.57
1	9	42.5
1	10	57.34
1	11	56.96
1	12	56.79
1	13	56.37

表 3：Pretrain epoch 位置的结果对比

从结果可以看出不同层数的 Transformer 的最佳 pretrain epoch 位置有所不同，对于 4 层的 Transformer，pretrain_epoch=15 时 VQA 系统性能最好，而对于 1 层的 Transformer，pretrain_epoch=10 更为合适。

3.3.3 Hidden size 的影响

我还通过发现 Hidden size 对系统的性能也影响巨大，实验对比了文本和图片不同维度对 VQA 系统的影响，实验发现维度越小 VQA 系统性能更好。

Hidden size	Overall	参数量
128	56.79	6M
256	56.16	9M
512	52.35	19M
1280	51.19	58M

表 4：Hidden size 对模型性能的影响

3.3.4 Loss 权重的影响

两个分支的损失函数的权重对 VQA 性能同样存在影响，我们通过设置不同的权重大小进行对比实验并且在不同层数的 Transformer 上进行实验，实验结果显示虽然 Transformer 层数不同，但是都在权重取为 6 的时候性能最好。

层数	loss 权重	Overall
4	1.5	48.94
4	3	51.41
4	6	52.97

4	9	51.1
4	12	49.5
2	3	50.08
2	6	51.83
2	9	48.03
1	5	55.52
1	6	57.34
1	7	56.9

表 5： Loss 权重的影响

参考文献

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In European conference on computer vision, pages 727-739. Springer.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In EMNLP.

Peng Zhang, Yash Goyal, Douglas Summer-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5014-5022.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6904-6913.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In NeurIPS.

Gabriel Grand and Yonatan Belinkov. 2019. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. NAACL HLT 2019, page 1.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. Advances in Neural Information Processing Systems, 32:841–852.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4060–4073.

Rabeeh Karimi Mahabadi and James Henderson. 2019. Simple but effective techniques to reduce biases. arXiv preprint arXiv:1909.06321, 2(3):5

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4971–4980.

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. Overcoming language priors with self-supervised learning for visual question answering.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pages 6077-6086.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv: 1602.07332.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (NIPS), pages 91–99.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, Qi Tian. 2019. Deep Modular Co-attention Networks for Visual Question Answering. In Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6281-6290.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natu- ral Language Processing (EMNLP-IJCNLP), pages 5103–5114.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In European Conference on Computer Vision, pages 121 – 137. Springer.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.