

Bridgeformer：一种基于Bert和ViT中间表示训练的VQA模型及其可解释性

黄奕铭 刘佳琦 贾卓霖 孙海川

2022年11月28日

摘要

目录

1 背景与研究简介	2
1.1 多模态与VQA	2
1.2 深度学习可解释性	3
1.3 VQAv2 数据集	3
1.4 目前研究的问题与我们的解决方法	3
2 相关工作	5
2.1 视觉文本大模型与下流任务	5
2.2 多模态学习中的可解释性	7
3 方法	8
3.1 Bridgeformer的架构	8
3.1.1 token嵌入与预处理	8
3.1.2 模型主体部分	8
3.2 基于Attention Map的可解释性	9
4 实验	10
4.1 超参设置	10
4.2 训练过程	10
4.3 训练结果	10
4.4 推理示例	12

1 背景与研究简介	2
5 可解释性的分析	13
6 结论与展望	14

1 背景与研究简介

1.1 多模态与VQA

人类通过多种感觉器官接触世界，例如眼睛、耳朵、触觉。多模态机器学习(Multimodal Machine Learning)研究包含不同模态数据的机器学习问题。常见的模态包括：视觉、文字、声音。它们通常来自于不同的传感器，数据的形成方式和内部结构有很大的区别，例如，图像是自然届存在的连续空间，而文本是依赖人类知识、语法规则组织的离散空间。多模态数据的异质性(heterogeneity)对如何学习多模态间关联性和互补性提出挑战。

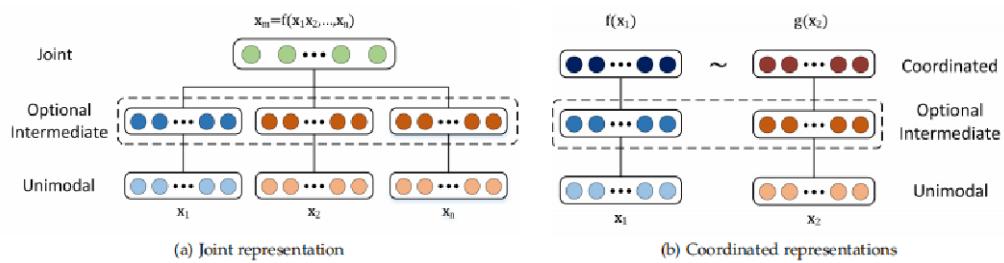


图 1: 多模态中的融合与协作范式[1]

随着互联网技术的快速发展,自然语言处理和计算机视觉的研究推动了人工智能的进步,从而产生了视觉问答(Visual Question Answering,VQA)。VQA是一项旨在将视觉和语言结合起来让计算机理解人类语言的任务,因此研究视觉问答具有重要的意义。

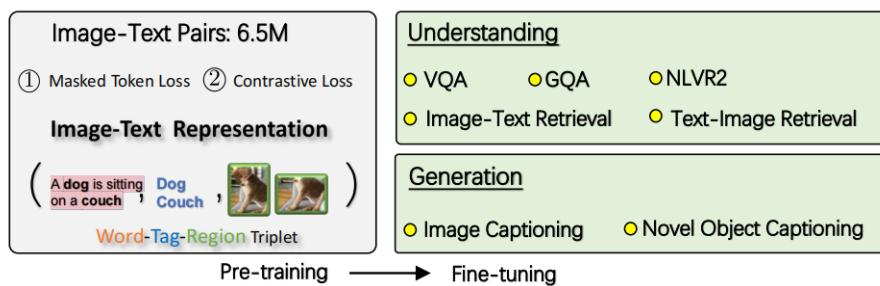


图 2: 视觉文本下游任务

1.2 深度学习可解释性

随着深度学习的广泛应用，人类越来越依赖于大量采用深度学习技术的复杂系统，然而，深度学习模型的黑盒特性对其在关键任务应用中的使用提出了挑战，引发了道德和法律方面的担忧，因此，使它们具有可解释性是使它们令人信服首先要解决的问题。于是，关于可解释的人工智能领域的研究应运而生，主要集中于向人类观察者明确解释模型的决策或行为。因此，为进一步深入研究建立更高效且具有可解释性可以为深度学习模型确立良好的基础。目前研究重点从解释深度学习模型的逻辑规则、决策归因和内部结构表示这三个方面出发介绍了几种可解释性研究的典型模型和算法外还指出了三种常见的内置可解释模型的构建方法；最后，评价方法有忠实度、准确性、鲁棒性和可理解性这四种的评价指标，指明了深度学习可解释性未来可能的发展方向。

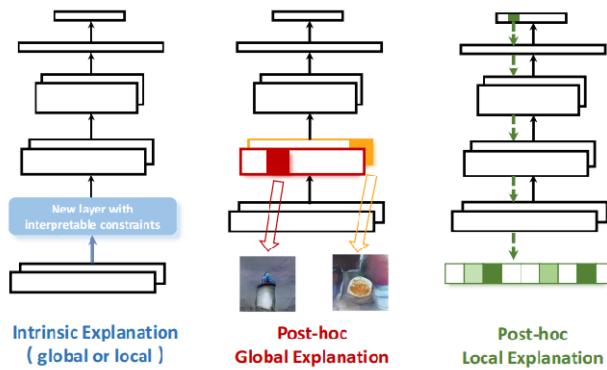


图 3: 深度学习可解释性的范式

1.3 VQAv2 数据集

视觉问答(VQA) 2.0是一个包含关于图像的开放式问题的数据集。这些问题需要对视觉、语言和常识的理解才能回答。它是VQA数据集的第二个版本。

COCO-VQA中以“是否存在一个”为开头的问题,7 9%的答案是“是”Visual于图像整体内容的问题,这可能导致提问中的偏见为了减少数据分布对模型的影响.Goyal等人在2017年提出了VQA 2.0数据集.与VQA 1.0数据集相比,VQA 2.0数据集规模更大,并且主要解决了答案不衡的问题,针对两张不同的图像提问相同的问题,并且尽量使得得到的答案相反但是VQA 2.0数据集仍存在答案分布问题,训练集和测试集的答案分布相似,模型可以利用答案分布带来的偏见得到较高的准确率,降低了模型的泛化性。如论文中所视, 其答案分布如下。

数据集中包含越40万图片文本对用于训练, 20万图片文本对用于验证, 40万图片文本对用于测试, 其中一般的工作比较时会在测试集test-std这个分割进行性能比较。如下图所示

1.4 目前研究的问题与我们的解决方法

目前，针对多模态深度学习的可解释性方法较少，而且这些工作的基线性能都不是很高。

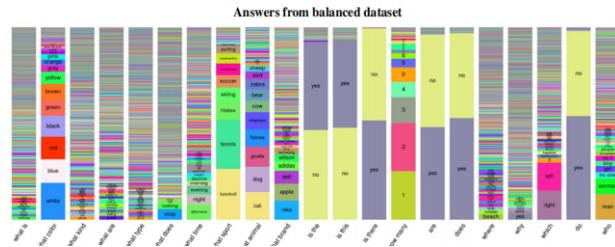


图 4: VQAv2 的分布

Method	VQA	
	test-dev	test-std
VisualBERT [13]	70.80	71.00
VL-BERT [10]	71.16	-
LXMERT [1]	72.42	72.54
12-in-1 [2]	73.15	-
UNITER [2]	72.70	72.91
VL-BART/T5 [54]	-	71.3
VILT [21]	70.94	-
OSCAR [3]	73.16	73.44
VILLA [8]	73.59	73.67
ALBEF (4M)	74.54	74.70
ALBEF (14M)	75.84	76.04

图 5: VQA sota比较

而我们针对这个问题提出面向注意力机制基于实例可视化的可解释方法，并利用预训练Bert和ViT的中间表示训练名为bridgeformer的新模型做为基线模型进行可解释。示意图如下图所示。

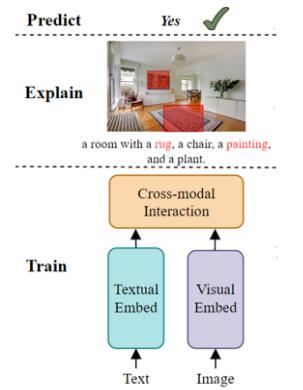


图 6: 研究动机示意

我们的贡献有：1.提出了bridgeformer这个基于预训练模型的架构 2.提出了一种面向多模态模型的可解释方法。

2 相关工作

2.1 视觉文本大模型与下流任务

计算机视觉 (CV) 和自然语言处理 (NLP) 早先是两个较为独立的研究领域。CV 重点关注如何用计算机代替人眼对目标完成识别、跟踪、测量等任务，对图像进行处理；NLP 则研究计算机如何处理、运用自然语言，包括语言生成、问答、对话等任务。近年来，以深度神经网络为代表的机器学习和模式识别技术被广泛应用于 CV 和 NLP 领域，取得了目前最先进的效果[10]。

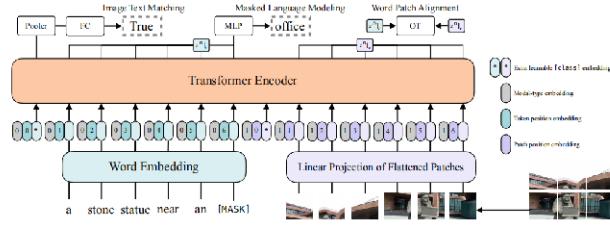


图 7: ViLT[6]

人类可以同时使用视觉和语言这两方面的能力来完成一系列任务，CV 与 NLP 的结合 (V2L) 也成为了人工智能研究领域的重要课题，可以拓展这两个方向的重要应用。例如，将图像理解和语言生成任务结合起来构成了图像描述 (image captioning) 任务；将图像分类、目标检测、图像分割、目标技术、颜色分析等 CV 任务与问答任务结合起来就构成了视觉问答任务；将图像理解和对话任务结合起来就构成了视觉对话任务。

图像说明的目标是为给定的图像生成“标题”，即用一句话总结图像内容。标题通常包含感兴趣的对象、对象的行为以及对象之间的位置关系。

给定一个图像-问题对，视觉问答要求根据图像回答一个问题。大多数研究都将视觉问答视为一个基于预定义答案集的分类问题。例如，VQA v2 有大约2K个预定义答案。

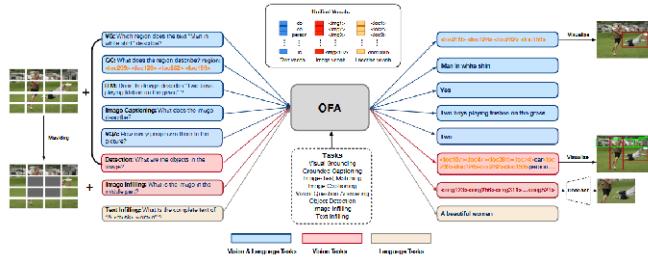


图 8: OFA

图像-文本匹配 (ITM)，或说图文检索，是视觉领域的基本课题之一。给定一个特定模态 (视觉或语言) 的query，它的目标是从另一个模态中找到语义上最接近的目标。根据query和目标模态，它包含两个子任务：图像-文本检索和文本-图像检索。

文本-图像生成: 给定一段文本, 生成包含该文本内容的图像。视觉对话: 给定一个图像, 一段对话历史, 和一个关于图像的问题, 回答这个问题。视觉推理: 与要求回答有关输入图像问题的VQA任务类似, 视觉推理要求进一步理解图像的能力。视觉推理任务通常包含足够的关于图像中的对象、问题结构等的注释。视觉蕴涵: 给定一幅图像和一篇文本, 判断该图像在语义上是否包含输入文本。短语基础和参考表达式理解: 这两个任务需要一个模型来输出与文本对应的边界框。对短语基础而言, 文本是一组短语; 对于引用表达理解而言, 文本是一种表达。

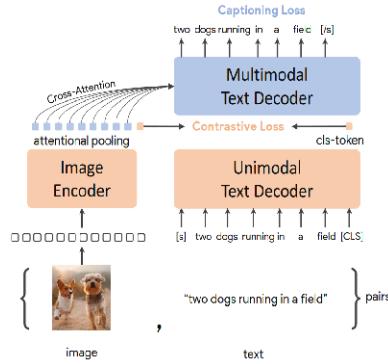


图 9: COCA[11]

谈及模型, 相对于文本预训练语言模型, 多模态预训练模型可以更好地对细颗粒度的多模态语义单元(词或者目标)间的相关性进行建模。例如, 基于语言上下文, 被掩码的词可以被预测为符合语法规则的词“under”等。但这与关联的图片场景“猫在车顶”不符。通过多模态预训练, 模型从图像中捕获“汽车猫”之间的空间关系, 从而可以准确地预测出掩码词。大部分的多模态预训练模型是在视觉-语言对齐数据上进行的。例, 使用图像和文本对齐数据集(MSCOCO, Conceptual Captions, Visual Genome 和 SBU Captions 等)训练的跨模态预训练模型LXMERT、OscarVL.BE 使用视频和文本对齐数据集训练的VideoBERT, ABERT等, Li等人最近还发布了视觉、文本、音三模态预训练模型OPT。

为了充分利用单模态预训练模型, VLP 模型可以将视觉或文本特征发送到 Transformer 编码器。具体来说, VLP 模型利用具有随机初始化的标准 Transformer 编码器来生成视觉或文本表示。此外, VLP 模型也可以利用预训练的视觉 Transformer 对 ViT-PF 进行编码, 例如 ViT 和 DeiT。VLP 模型可以使用预训练的文本转换器对文本特征进行编码, 例如 BERT。从两个不同的角度介绍 VLP 模型的架构: (1) 从多模态融合的角度来看单流与双流, 以及 (2) 单编码器与编码器解码器整体架构设计视角。

Single-stream Architecture: 单流架构是指将文本和视觉特征连接在一起, 然后嵌入单个 transformer 块, 如上图所示。单流结构利用合并注意力来融合多模式输入。单流架构的参数效率更高, 因为两种模式都使用相同的参数集。

Dual-stream Architecture: 双流架构是指文本和视觉特征没有连接在一起, 而是独立地发送到两个不同的 transformer 块, 如上图所示。这两个 transformer 块不共享参数。为了获得更高的性能, 交叉注意力用于实现跨模态交互。为了获得更高的效率, 视觉转换器和文本转换器块之间也可以没有交叉注意力。

许多 VLP 模型采用 Encoder-only 架构, 其中跨模态表示直接嵌入输出层以生成最终输出。相比之下, 其他 VLP 模型使用 transformer 编码器-解码器架构, 其中跨模态表示首先嵌入解码器, 然后嵌入输出层[9]。

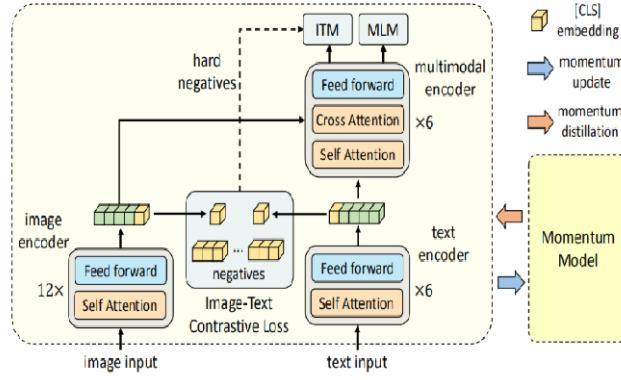


图 10: ALBEF[7]

2.2 多模态学习中的可解释性

有的工作探究在上下文表示中的具体的名词可以在多大程度上与视觉表示进行对应。通过设计探针实验发现文本表示可以为图片patch检索提供一个非常强的信号。在上下文表示中的具体的名词可以在多大程度上与视觉表示进行对应。通过设计探针实验发现文本表示可以为图片patch检索提供一个非常强的信号。如下图所示。

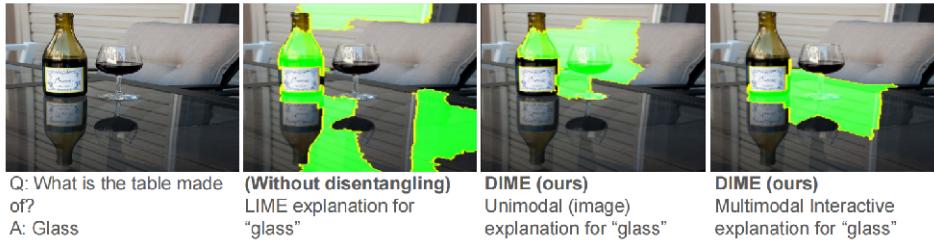


图 11: 基于probe探针的方法[8]

有的工作通过消融某个模态的输入来探究模型的另一模态上的性能，如果模型学会使用多种模态信息来构建多模态表示，那么当某种模态信息缺失时，模型性能应当下降。通过实验发现当视觉信息缺失时预测文本要比文本信息缺失时预测视觉要更加困难，说明多模态模型是不对称的。

总体来说[2]，大部分工作基中于probe的想法，语言探针发现：VL模型的句法理解能力要比语言模型差，原因可能是预训练使用的caption数据集句法变化较小。视觉探针发现：基于region特征的模型表现差于grid，由于它们很难从表示中抽取出正确的视觉信息使其更依靠文本信息。更好的目标检测模型会带来更好的下游任务结果。多模态探针发现：预训练多模态模型可以捕获到一些多模态信息。比如UNITER获得更好的结果，ViLT要差一点，原因是没有设计目标预测预训练任务，导致比较差的目标语义理解能力。但是，对于概念的理解，还是依赖文本信息，也没有能力从多模态层次去理解目标之间的位置关系。fine-tune的局限性多模态的表现依赖于文本数据

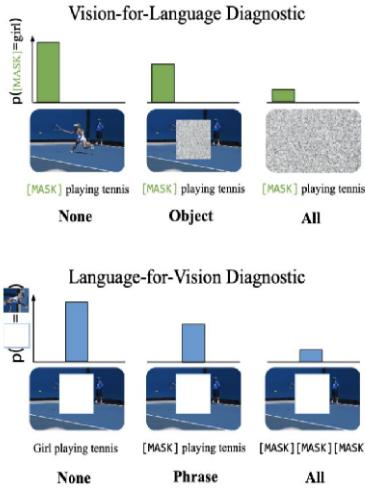


图 12: 基于MASK消融的方法[5]

3 方法

3.1 Bridgeformer的架构

我们基于ViT[4]和Bert[3]的中间表示训练了该模型。

对于ViT，图像将被它打成patch，变换为seqeunce，然后送入Transformer block中，

3.1.1 token嵌入与预处理

$$\begin{aligned}
 \text{Image} &\in \mathbf{R}^{n \times c \times w \times h}, \text{Text} \in \mathbf{R}^{n \times L} \\
 \mathbf{I} &:= \text{ConvProj}(\text{Image}) \in \mathbf{R}^{n \times (p \cdot p) \times d} \\
 \mathbf{T} &:= \text{Embedding}(\text{Text}) \in \mathbf{R}^{n \times L \times d} \\
 \mathbf{I} &:= \text{ViT}(\mathbf{I}), \mathbf{T} := \text{Bert}(\mathbf{T})
 \end{aligned}$$

3.1.2 模型主体部分

对于图像或者文本打成的序列，其在Transformer块中经历着这样的变换：

$$\begin{aligned}
 \mathbf{A} &:= \mathbf{W}_Q \mathbf{I} \cdot \mathbf{W}_K \mathbf{I} \\
 \mathbf{I} &:= (\text{Softmax}(\mathbf{A}) \mathbf{W}_V \mathbf{I}) \mathbf{W}_O \\
 \mathbf{I} &:= \text{FFN}(\mathbf{I}).
 \end{aligned}$$

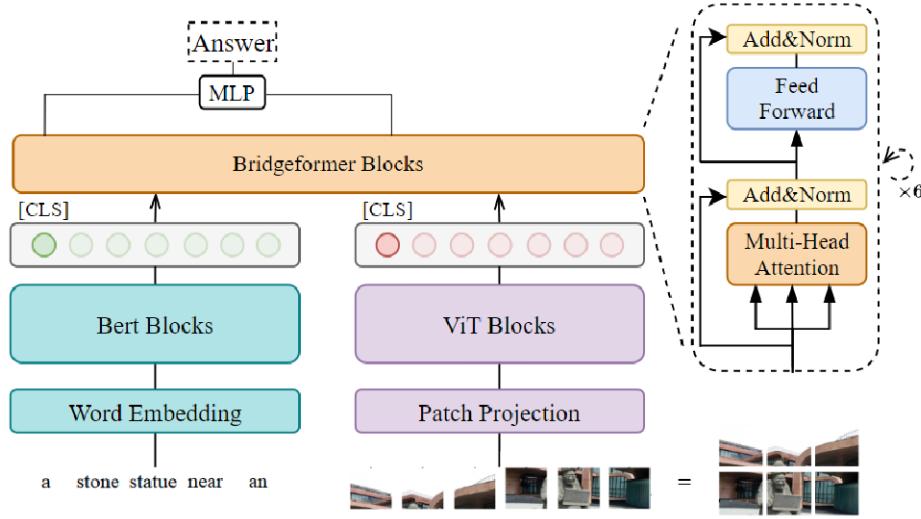


图 13: 模型结构

在模态融合阶段:

$$\mathbf{M} := \text{Concat}(\mathbf{I}, \mathbf{T})$$

$$\mathbf{M} := \text{Bridge}(\mathbf{M})$$

$$\mathbf{M}_{cls} := \text{Concat}(\mathbf{M}_0, \mathbf{M}_{50}) \in \mathbf{R}^{n \times 2d}$$

$$\text{Answer} := \text{MLP}(\mathbf{M}_{cls})$$

3.2 基于Attention Map的可解释性

考虑到解释的直观性，基于实例的可视化能有较好的直观性，如简单的attention map 可视化。

$$A = \text{Softmax}\left(\frac{QW_Q \cdot (KW_K)^T}{\sqrt{dim}}\right)$$

我们的目的是使被大多数token所关注的token在图像中显示出来，这因此工作的重点在于将不同图像区域和问本单词的差异度体现出来。

而联想到Attention机制中Attention Map的含义，即所有一个token对所有其他token的关注度，显然符合我们的需求。对于不同的transfomer层，我们使用效果累加的方式求得。公式如下。

首先我们有attention 注意力图

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^\top}{\sqrt{d_h}}\right)$$

$$\mathbf{O} = \mathbf{A} \cdot \mathbf{V}$$

其次我们初始化相关矩阵

$$\mathbf{R}^{ii} = \mathbf{1}^{i \times i}$$

当遇到同质的Attention时

$$\mathbf{R}^{ss} = \mathbf{R}^{ss} + \mathbf{A} \cdot \mathbf{R}^{ss}$$

遇到异质的Attention时

$$\mathbf{R}^{ss} = \mathbf{R}^{ss} + \mathbf{A} \cdot \mathbf{R}^{qs}$$

最后，我们对相关矩阵进行求和

$$\text{score}_i := \text{Sum}(\mathbf{R}_i)$$

4 实验

4.1 超参设置

我们使用的词嵌入是Bert预训练的词向量，大概长3W，是ULM这个子词嵌入的方法做的。而对于图像的处理则是进行统一的224 resize。

表 1: 训练配置

配置	实际选择
优化器	AdamW
平均移动参数(β_1, β_2)	(0.9, 0.999)
学习率	1e-5
批次大小	128
训练轮数	10
训练步数	80000
损失函数	BCE损失

4.2 训练过程

以下为训练集上的loss下降与准确率提升随时间变化折线图

可以看到最终loss下降到了接近0的值，而在训练集上的准确率达到了82% 左右

4.3 训练结果

可以看到在测试集上的准确率为47%。参考:目前sota的准确率大概在63% 左右.

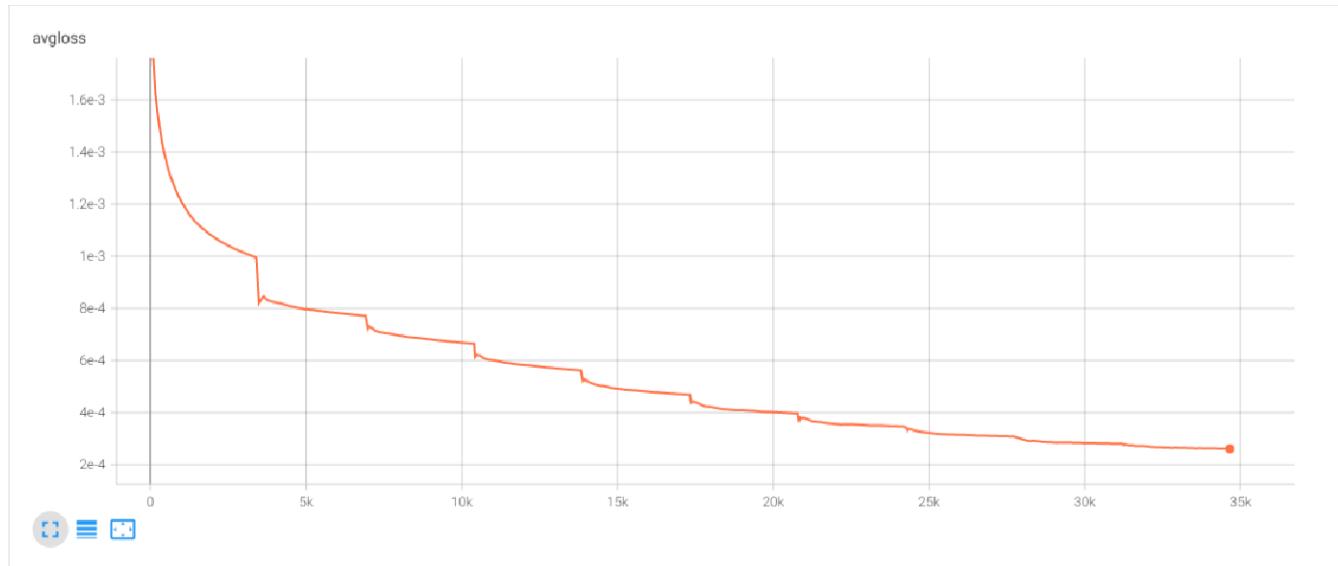


图 14: Loss随轮数的变化

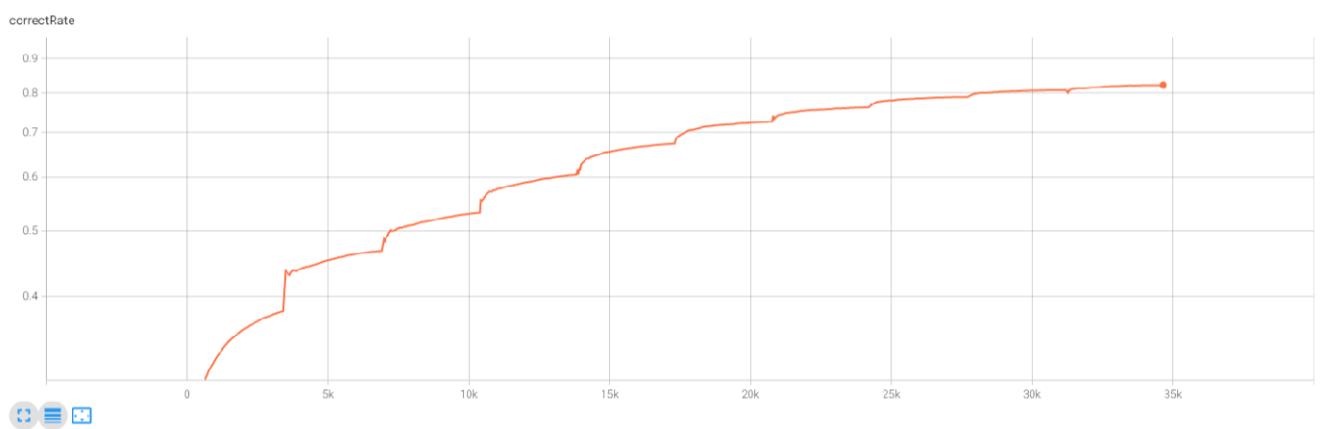


图 15: 训练集上准确率

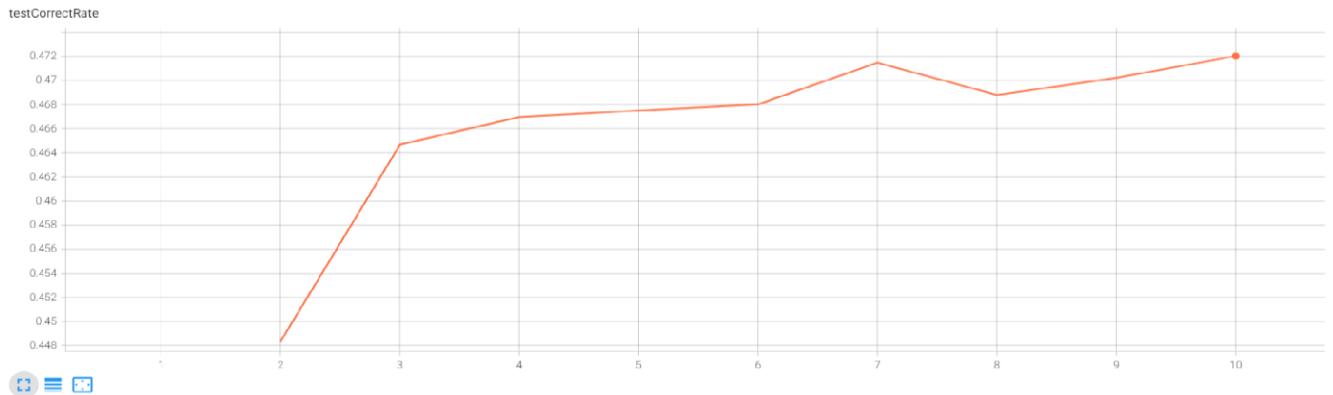


图 16: 测试集上准确率

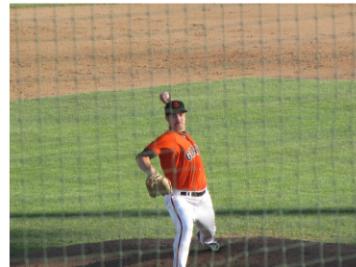
4.4 推理示例

我们从训练集中随机挑出12个图片文本对模型进行测试，可以看到推理结果大致准确。

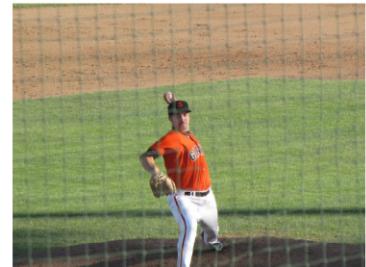
Question: What is this photo taken looking through?
Model Predict: net
Ground Truth: net



Question: What position is this man playing?
Model Predict: pitcher
Ground Truth: pitcher



Question: What color is the players shirt?
Model Predict: orange
Ground Truth: orange



Question: What color is the players shirt?
Model Predict: orange
Ground Truth: orange



Question: Is this man a professional baseball player?
Model Predict: yes
Ground Truth: yes



Question: What color is the snow?
Model Predict: white
Ground Truth: white

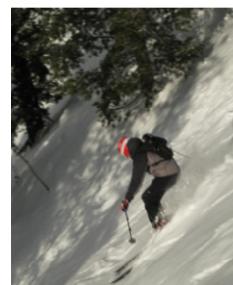


图 17: 推理示意

5 可解释性的分析

我们首先可视化了bert中的attention map。

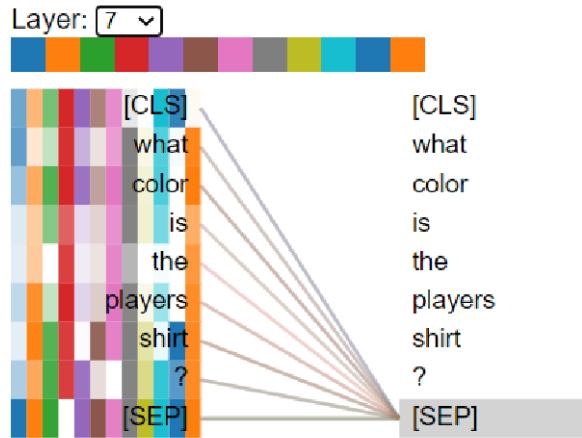


图 18: bert第11层attention map

以下是分层分head的情况。

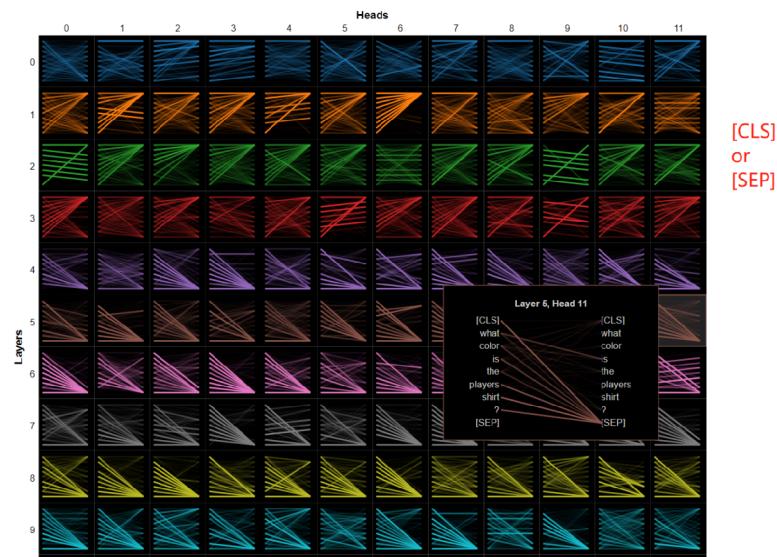


图 19: 全部的attention map

按照我们的可解释性理论，我们对视觉中重要的注意力图进行了相关图的更新进行的可视化，发现如下结果。可以看到模型关注了符合我们常识推断，第三个例子我们的方法并没有关注到红色这个概念。

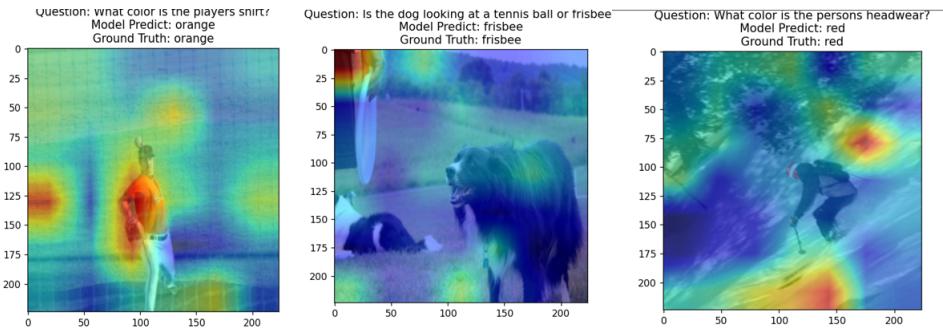


图 20: 重点视觉的可解释性

6 结论与展望

我们通过基于视觉与文本的典型预训练模型，提出了新的VQA基线模型，并进行了可解释性研究。而未来的工作中我们会做出更好理解，更鲁棒的可解释方法。

参考文献

- [1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.
- [2] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [5] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? On cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*. Association for Computational Linguistics, nov 2021.
- [6] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

- [7] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.
- [8] Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. *arXiv preprint arXiv:2203.02013*, 2022.
- [9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [10] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.
- [11] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.