Computer Science and Application 计算机科学与应用, 2023, 13(12), 2432-2446
Published Online December 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1312243

Hans汉斯

PDID：视觉离散化智能问答模型

——基于图像像素离散化和图像语义离散化的VQA模型

陈页名，张思禹，孙杳如

同济大学电子与信息工程学院，上海

收稿日期：2023年11月26日；录用日期：2023年12月21日；发布日期：2023年12月29日

摘　要

视觉问答是一项具有挑战性的多模态任务，它连接了计算机视觉和自然语言处理两个领域。在这项任务中，模型需要根据给定的图片和相关问题，有效地提取信息并给出正确答案。然而，由于图像和文本属于不同的模态，存在着严重的语义差异，因此如何有效地将不同模态的信息对齐并减少语义差异，是当前视觉问答领域的重点关注问题。本文针对当前视觉问答方法在多模态对齐阶段图像和文本信息颗粒度的巨大差异，提出了基于视觉离散化(PDID: Pixel Discretization and Instance Discretization)的智能问答模型并辅助以模态注意力机制完成跨模态信息和语义对齐。图像以像素为最小单位的特征数据与文本以单词为最小单位的特征数据，它们在数据的信息颗粒度上存在巨大的差异，即语言通过至多数万单词即可完成整个文本语义空间的构建，而图像则是通过亿级的RGB三原色数组构建而成。这说明了直接建模以像素为单位的图像是很难和文本做好对齐的。本文通过了多种图像离散化的方式，一方面通过离散化图像像素，以颜色离散化、强度离散化、纹理离散化、空间离散化四种形式将图像像素完成离散化，在数量级上逼近文本特征的最小基元数量；另一方面通过图像语义特征的软编码，离散化图像深层次的语义特征，将图像的语义特征与文本的单词语义对齐，在语义层面上逼近文本特征的单词语义信息量。除此以外，本文提出了一种新型的视觉关系融合模块，视觉关系融合模块用来捕获同种模态内离散化特征和连续特征的交互信息，为模型提供丰富的视觉特征。本文先使用自注意力方法提取模态内特征之间的相关性，即提取视觉全局关系，再使用通道空间分离注意力进行跨模态结合，为局部引导的全局特征提供更大的表示空间和更多的补充信息。为了验证本方法的有效性，在VQA-v2，COCO-QA，VQA-CP v2数据集上进行了广泛实验，充分验证了该方法在视觉问答任务中的基于离散机制的视觉问答研究有效性。同时也体现了该模型在其他跨模态任务(图像文本匹配、指示表达)中仍有很强的泛化能力。

关键词

VQA，像素离散化，语义离散化，自注意力，跨模态融合

PDID: Visual Discretization Intelligent Question Answering Model

—VQA Model Based on Image Pixel Discretization and Image Semantic Discretization

文章引用: 陈页名, 张思禹, 孙杳如. PDID: 视觉离散化智能问答模型[J]. 计算机科学与应用, 2023, 13(12): 2432-2446.
DOI: 10.12677/csa.2023.1312243

Yeming Chen, Siyu Zhang, Yaoru Sun

College of Electronic and Information Engineering, Tongji University, Shanghai

Received: Nov. 26th, 2023; accepted: Dec. 21st, 2023; published: Dec. 29th, 2023

Abstract

Visual question answering is a challenging multimodal task that bridges the fields of computer vision and natural language processing. In this task, the model needs to effectively extract information and give the correct answer based on the given picture and related questions. However, since images and texts belong to different modalities, there are serious semantic differences. Therefore, how to effectively align information from different modalities and reduce semantic differences is a key concern in the current field of visual question answering. In view of the huge difference in the granularity of image and text information in the multi-modal alignment stage of current visual question answering methods, this paper proposes an intelligent question answering model based on visual discretization (PDID: Pixel Discretization and Instance Discretization) and is assisted by a modal attention mechanism, cross-modal information and semantic alignment. There is a huge difference in the information granularity of the feature data of images with pixels as the smallest unit and the feature data of text with words as the smallest unit. That is, language can complete the construction of the entire text semantic space with up to tens of thousands of words, and the image is constructed from a billion-level RGB three primary color array. This shows that it is difficult to align the image with the text by directly modeling the image in pixels. This article adopts a variety of image discretization methods. On the one hand, it discretizes image pixels and discretizes image pixels in four forms: color discretization, intensity discretization, texture discretization, and space discretization, approaching text in an order of magnitude. The minimum number of primitives of the feature; on the other hand, through soft coding of image semantic features, the deep-level semantic features of the image are discretized, the semantic features of the image are aligned with the word semantics of the text, and the word semantic information of the text features is approximated at the semantic level quantity. In addition, this paper proposes a new type of visual relationship fusion module. The visual relationship fusion module is used to capture the interactive information of discrete features and continuous features within the same modality, providing rich visual features for the model. This paper first uses the self-attention method to extract the correlation between features within the modality, that is, extracts the visual global relationship, and then uses the channel space separation attention for cross-modal combination to provide a larger representation space and locally guided global features and more supplementary information. In order to verify the effectiveness of this method, extensive experiments were conducted on the VQA-v2, COCO-QA, and VQA-CP v2 data sets, which fully verified the effectiveness of this method in visual question answering research based on discrete mechanisms in visual question answering tasks. At the same time, it also reflects that the model still has strong generalization ability in other cross-modal tasks (image text matching, instruction expression).

Keywords

VQA, Pixel Discretization, Semantic Discretization, Self-Attention, Cross-Modal Fusion

1. 引言

视觉问答(Visual Question Answering, VQA)代表着跨学科研究的前沿，融合了计算机视觉和自然语言处理领域。VQA 系统旨在解释和回答关于数字图像的问题，这是一项需要对视觉内容和语言语义进行精准理解的任务。将图像离散化集成到 VQA 系统中已成为该领域的重要发展。图像离散化涉及将连续的图像数据转换为离散形式，从而促进更高效和准确的分析。这一方法显著增强了 VQA 系统处理复杂视觉数据的能力，带来了更好的性能和可靠性。

本文旨在深入探讨图像离散化在 VQA 中的应用。本文探讨这一方法如何将原始图像数据转换为更适合进行分析和与自然语言组件交互的格式。本文将从像素信息离散化和图像语义离散化两个角度分别建模，讨论将涵盖图像离散化的理论基础、在 VQA 中的应用、其带来的益处以及所面临的挑战。

现实世界的图像通常多变而复杂，包含各种颜色、纹理和物体。图像离散化有助于标准化这些数据，简化对多样化视觉输入进行分析和解释的任务。在图像质量、光照或构图方面不理想的场景中，这种标准化特别有益。图像离散化还提高了 VQA 系统的效率。处理连续的视觉数据可能会在计算上具有较大的消耗且耗时。通过将图像转换为离散格式，这些系统能够更快速地处理视觉信息，实现更快的响应时间。这种效率对于需要及时响应的应用非常关键。此外，图像离散化促进了先进的机器学习和深度学习模型在 VQA 中的整合。这些模型，尤其是卷积神经网络(CNNs)，在分析离散化的图像方面非常有效。它们能够从这些图像中提取复杂的模式和特征，这对于回答复杂而详细的问题至关重要。图像离散化与深度学习之间的协同作用显著拓展了 VQA 系统的能力。

2. 背景

2.1. VQA 概述

一个 VQA 模型需要多种技术、工具协作完成，根据这些技术、工具的作用不同可以将 VQA 模型分为四个阶段，分别为：特征提取阶段(提取图像特征和文本特征)；多模态特征对齐阶段；多模态特征融合阶段以及答案预测阶段。提取图像特征需要用到卷积神经网络(Convolutional Neural Network, CNN)，如 ResNet [1]、Faster R-CNN [2]等；提取文本特征需要用到 Glove 词嵌入[3]，循环神经网络(Recurrent Neural Network, RNN) [4]，如长短期记忆网络(Long Short-Term Memory, LSTM) [5]、门控循环网络(Gate Recurrent Unit, GRU) [6]等；在多模态特征对齐阶段，为了充分的挖掘图像特征和文本特征的关系需要用到注意力机制(Attention Mechanism, AM) [7]；对于多模态特征的融合需要用到加法、拼接、相乘、多模态低秩双线性池化(Multimodal Low-Rank Bilinear, MLB) [8]等融合方法；在答案预测阶段选择多层感知机(Multi-Layer Perception, MLP) [9]进行答案预测。上述的几个阶段中，目前的研究难点在于如何进行多模态特征对齐，也即如何充分的挖掘不同模态特征间的关系，最大程度减少不同模态信息间的语义鸿沟，因此如何搭建一个具有高效跨模态信息对齐能力的 VQA 模型仍然是人工智能领域的一个热点问题。如图 1 所示，目前的基线模型在预测上仍然存在很大问题，即模型的预测分布极大程度依赖于问题，导致 answer 的分布非常容易向着高频率答案偏移。

2.2. 图像特征提取概述

目前绝大多数的方法，除去在融合部分的注意力机制的添加，都可以归纳为联合嵌入模型。基于联合嵌入的方法是指：将输入的图像和问题映射到相同的子空间进行答案的预测。对于该方法而言，如何将不同模态信息映射到相同的子空间变得十分重要。早期对不同模态特征大多采用线性融合方法，例如：相加、相乘、拼接等方式。Malinowski 等人[10]提出了一种名为“Neural-Image-QA”的方法，该方法首

先使用 CNN [11]提取图像特征，然后将提取出的图像特征与问题中每一个单词的词向量进行拼接产生联合特征，再将该联合特征依次送入到 LSTM [5]中预测答案。Ren 等人[12]提出的“VIS + LSTM”模型使用了与上述相同的方法，区别在于特征拼接的时候将图像特征看作一个视觉单词，仅与第一个单词的词向量进行拼接，作为第 0 个时间步的特征输入到 LSTM 中，而不是每个时间步都进行拼接。大多数的视觉问答模型使用 RNN 处理文本信息，Ma 等人[13]针对该特点进行了新的尝试，提出了一种完全基于 CNN 的视觉问答模型，该模型可以分为 3 个 CNN 模块，分别用于提取文本特征、提取图像特征以及多模态特征融合。虽然该方法在主流 VQA 数据集上表现不是很好，但是为视觉问答模型搭建提供了一种新的思路。

特征提取上，通过 grid-feature 和通过 bounding-box 各自的问题。grid-feature 会导致图像的语义丢失，而 bounding-box 的问题则更为隐蔽，如下图 2 所示，其实图像以 bounding-box 为基元进行分类会出现一个弊端，即图像的 bounding-box 存在大量的互相遮掩的情形。

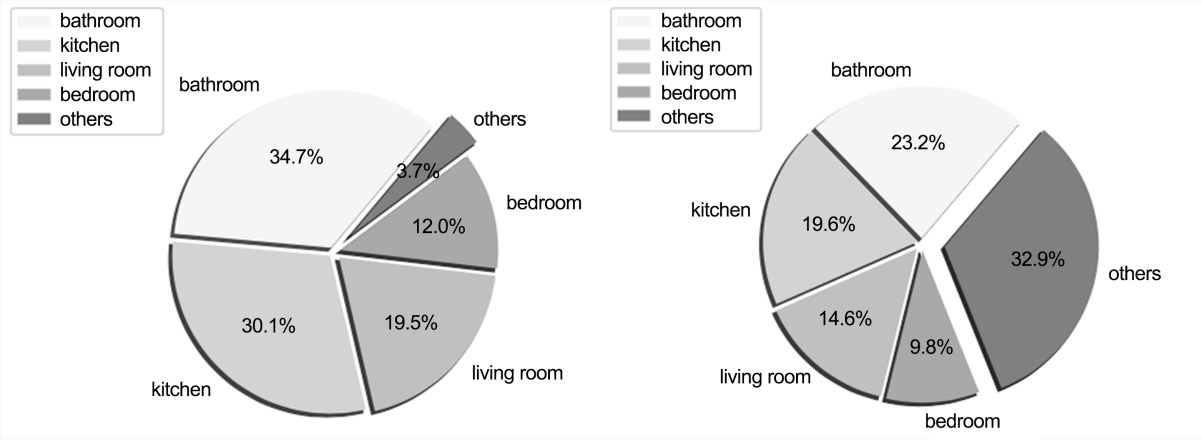

Figure 1. Baseline model prediction and labeling for location issues
图 1. Baseline 模型对于地点问题的预测和标签

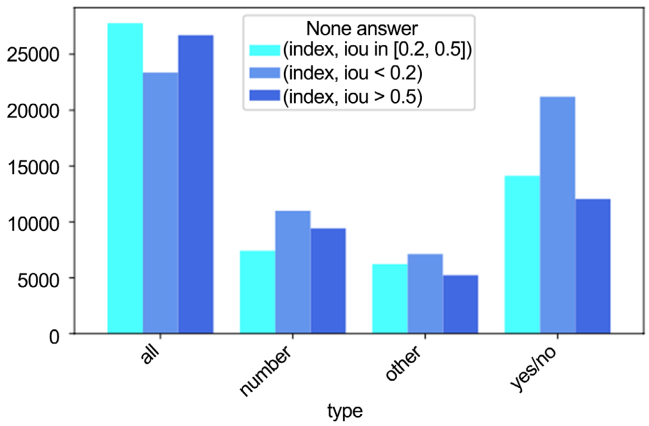

Figure 2. Statistical chart of bounding-box label overlap in Coco dataset
图 2. Coco 数据集 bounding-box 标签重叠情况统计图

但是仅仅通过连续神经网络映射辅助以注意力机制去做信息颗粒度的对齐，仍然是困难的。图像离散化，尽管在概念上是一个简单的过程，却在这一领域带来了转变性的变革。通过将连续的视觉数据转换为离散的、可量化的单元，图像离散化使得对视觉元素进行更细粒度和精确的分析成为可能。这种离

散化过程在数字图像处理和计算机图形等领域是基础性的，而且在 VQA 中同样至关重要。

图像离散化的基本前提在于将图像分割成不同的区域或像素，每个区域代表特定的视觉特征。这种分割有助于更有结构的图像分析，使得 VQA 系统能够以更高的精度解释和回答复杂的视觉问题。这种方法起源于传统的图像处理技术，但随着计算能力和算法效率的进步而有了显著的发展。

Figure 3. Similar semantic image combination
图 3. 相似图像语义组合

如上六张图所示(图 3)，六张图依次为原图，亮度 1.5 倍，亮度 0.5 倍，轻微旋转，加高斯噪声，加椒盐噪声的情况。考虑实际图像的获取情况，随着一天时间的光照变化、摄像头的平稳程度，摄像头的清晰程度会依次出现上述这些情况。而这些图像输入进一个连续神经网络的编码器后获取到的特征向量是存在巨大差异的，而从实际的最终 VQA 模型的推理角度出发，这些图像所包含的语义是十分接近的。这就要求了我们需要以一个全新的方式去看待图像，即怎么能够将这些在像素层面上数值相差巨大的像素基元建模成相同的输入，或者在特征编码阶段完成特征聚类和抽象。

3. 视觉特征离散化

图像离散化建立在涉及数学、计算机视觉和机器学习等多个领域的理论基础之上。其核心概念涉及将图像从其连续形式(与人类视觉感知相一致的表示)转换为更适合计算处理的离散格式。在 VQA 系统中，这种转换对于对视觉数据的高效分析和解释至关重要。

在 VQA 的背景下，图像离散化通常涉及将图像分解为像素或段，每个表示不同的属性，如颜色、纹理或强度。这些离散元素成为特征提取的构建块，是解释图像内容的关键步骤。特别是基于卷积神经网络(CNNs)的特征提取算法利用这些离散元素来识别和分类与提出的问题相关的视觉模式。

3.1. 像素离散化

本文使用了基于颜色的离散化、基于纹理的离散化、基于强度的离散化和基于空间的离散化。

3.2.1. 颜色离散化

对 RGB 颜色进行量化可以使用以下公式：

$$C_q = \frac{C_{原始}}{Q} \times Q \tag{1}$$

其中，C_q 是量化后的颜色，$C_{原始}$ 是原始颜色，Q 是量化级别。

3.2.2. 纹理离散化

当涉及到局部二值模式(Local Binary Pattrns，简称 LBP)时，其计算过程可以用一下的方式表示：

假设我们考虑一个图像中的一个特定像素点 P，并以其为中心，选择一个半径为 R 的圆形邻域。在这个邻域内，选择 P 周围的 N 个像素点(通常均匀分布在圆周上)，然后比较这些像素的灰度值与中心像素 P 的灰度值。

1) 对于这些邻域像素点中的每一个像素 P_i，其中 $i = 0,1,2,\cdots,N-1$；

2) 计算 P_i 的灰度值与中心像素 P 的灰度值的差异。

3) 如果 P_i 的灰度值大于或等于中心像素 P 的灰度值，则将 P_i 对应的位置记为 1，否则记为 0。

4) 将所有的 0 和 1 连接成一个二进制数串(可能的长度是 N)，形成一个二进制模式。

5) 将这个二进制模式转换为十进制数值，作为中心像素 P 的新值。

这个过程可以用一个数学公式表示为：

$$LBP_{P,R} = \sum_{i=0}^{N-1} s(i) \times 2^i \tag{2}$$

其中 $s(i)$ 是一个函数，表示与中心像素 P 相邻的像素点的灰度值与中心像素的比对结果。如果 P_i 的灰度值大于或等于中心像素 P 的贵读职，则 $s(i)=1$，否则 $s(i)=0$。

P 是中心像素，R 是邻域半径，N 是选取的邻域像素点数。

LBP 的公式描述了如何从图像中提取局部纹理特征，并将其编码为具有不同模式的整数值，这些特征可以用于图像识别、纹理分析以及其他计算机视觉任务。

3.2.3. 强度离散化

对图像强度进行二值化(Binarization)是一种基本的离散话方法。一种简单的二值化公式是：

$$B(x,y) = \begin{cases} 1, & \text{if } I(x,y) > T \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

其中，$B(x,y)$ 是二值化后的图像，$I(x,y)$ 是原始图像的强度，T 是阈值。

3.2.4. 空间离散化

本文中，使用卷积函数辅助硬编码池化层完成空间离散化。

3.3. 语义离散化

卷积神经网络(CNN)特征提取：

$$F_{i,j,k} = \sigma\left(\sum_{u,v} I_{i+u,j+u,k} \cdot W_{u,v,k}\right) \tag{4}$$

其中，$F_{i,j,k}$ 是卷积后的特种图中的一个元素，$I_{i+u,j+u,k}$ 是输入图像中的像素，$W_{u,v,k}$ 是卷积核中的权重，σ 是区别于常规 sigmoid 函数以外的硬编码激活函数，用于离散化。

在 VQA 中，有时需要将图像和问题的信息联合表示。一种方式是使用注意力机制，其中图像特质和问题特质被赋予不同的权重。注意力机制的公式可以表示为：

$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)} \tag{5}$$

其中，α_i 是第 i 个特征的注意力权重，e_i 是与该特征相关的能量分数。

4. 基于图像离散化的 VQA 模型

将上述图像离散化技术以并行方式融合到本文的 VQA 模型中，模型整体结构如下流程图 4 所示。

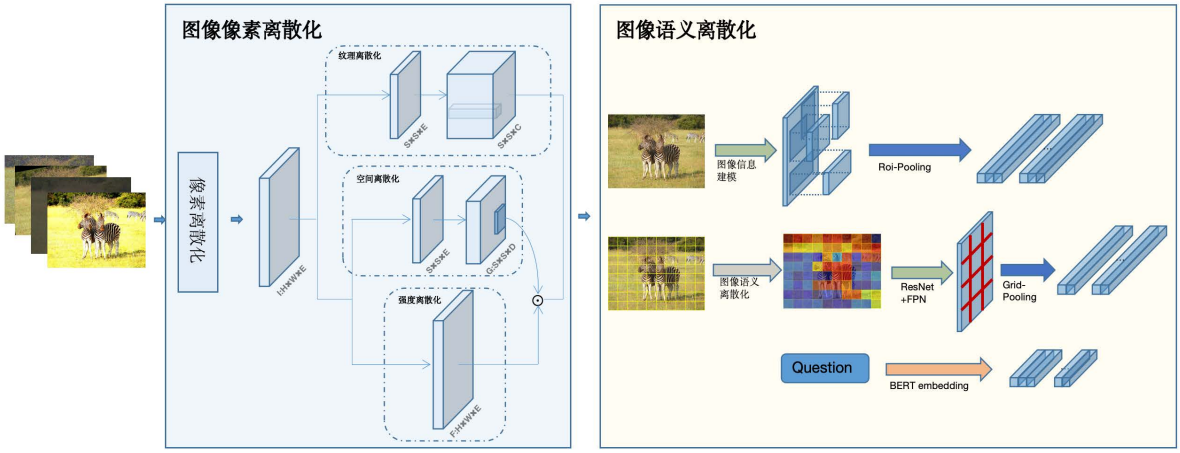

Figure 4. PDID model flow chart
图 4. PDID 模型流程图

本文输入来自 COCO 数据集的图像进行编码，首先经过图像像素离散化模块，在该模块，PDID 模型会并行的进行空间离散化、纹理离散化、强度离散化。空间离散化本质来说是将临近像素进行聚类，强度离散化则是让图像在整张图的层面上分离出光照度等亮度信息，而纹理离散化则是侧重于图像共性的局部细节并作保留。随后联合编码得到的特征向量会输入进图像语义离散化模块，在该模块，图像一方面通过常规的卷积神经网络和 RPN 网络完成特征提取，另一方面通过 grid-feature 的形式对图像的语义进行权重的再分配，最终通过模态内的注意力融合以上两种视觉特征，再通过跨模态注意力融合图像和文本特征，最终完成输出。模型的模态内和模态间注意力结构如图 5 所示。

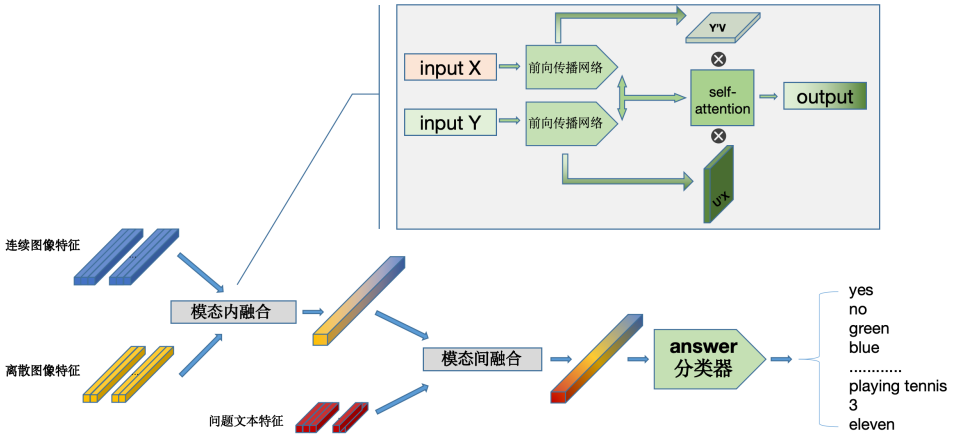

Figure 5. Structure diagram of intra-modal attention and inter-modal attention
图 5. 模态内注意力和模态间注意力的结构图

注意力机制的结构如上所示，我们定义了一种双线性融合的方式，该注意力能够很好的平衡所有特征向量之间的关系，这种双向的权重再加权保证了最终每一个特征都能融合尽可能多的不同层次的信息和不同模态的信息。

5. 实验

5.1. 实验数据集

VQA v1 是第一个广泛采用的 VQA 数据集。它包含 248,349 个训练问题、244,302 个测试问题和 121,512 个验证问题。从 Microsoft COCO 数据集中获取了 204,721 张自然图像。VQA v1 中的问答对是基于人类注释者进行注释的，包含三种类型的对应答案，即“是/否”、“数字”和“其他”。VQA v1 数据集由开放式和多项选择两个子任务组成。VQA v2 是 VQA v1 的扩展版本，其重点是通过平衡对减少数据集偏差。整个数据集包含 443,757 个训练问题、214,354 个验证问题和 447,793 个测试标准问题。每个问题都会由人工注释者生成 10 个自由回答答案。提供基于准确率的评估指标来预测。回答评分根据以下公式给出：

$$\text{Accuracy}(A)=\min\left(\frac{\#\text{humans that said}(A)}{3},1\right) \tag{6}$$

其中 A 是不同注释者提供的答案的数量。如果预测答案由三个以上注释者给出，则相应得分为“1”。COCO-QA 是根据 Microsoft COCO 数据集创建的，该数据集小于 VQA v1 和 VQA v2。该数据集包含 66.9% 的训练集和 33.1%的测试集。整个数据集包含 69,172 张图像、92,396 个问题和 435 个答案。问题分为四种类型：“物体(69.84%)”、“数字(16.59%)”、“颜色(7.47%)”和“位置(6.10%)”。此外，采用 Wu-Palmer 相似度(WUPS)作为附加指标来测量真实答案与预测答案之间的语义距离。WUPS 分数设置为 0.0 和 0.9。VQA-CP v2 是从 VQA v2 重组而来的，它改变了答案的先验分布，以减少训练和测试分割中面向问题的偏差。具体来说，它包含 121k 个图像、438k 个问题和 440 万个训练集答案。VQA-CP v2 的测试集有 98k 张图像、220k 个问题和 220 万个答案。

5.2. 实验参数

本文模型使用 Resnet-101 作为 backbone 来初始化整个网络，该网络加载来自 ImageNet 上预训练后的模型参数辅助训练。对于单词特征，我们将所有问题用 0 填充到最大长度 14，并通过 BERT 提取 14 × 768 作为问题嵌入 Q。BERT 模型在 Wikipedia (2500M)和 Books Corpus (800M Words)上进行了预训练。融合视觉和文字特征后，我们通过全连接层将它们转换为 1024 维。多头注意力的数量设置为 8。对于优化模型，所有全连接层使用 0.5 的 dropout 率。所有梯度均被裁剪为 0.5，批量大小固定为 128。所有模型均训练 200 个 epoch，初始学习率为 lr = 0.001，并且预热比率设置为 1/3。选择随机梯度下降(SGD)作为优化器，动量为 0.8，权重衰减为 10−3。我们的操作是通过 PyTorch 1.11.2 和 Tensorflow 2.4.0 实现的。所有代码都在四个 NVIDIA A100 GPU 上运行。训练网格单元时每张图像花费 0.026 秒。每个图像的像素离散化计算时间为 0.012 秒。在整个训练过程中，每一步需要 0.29 秒，30 个 epoch 总共需要 11.3 小时。

5.3. 实验结果

模型的训练准确率和损失函数值如图 6 所示，模型在 30 个 epoch 后逐渐平稳，证明模型收敛。

表 1、表 2、表 3 是模型在 VQA-v2.0，COCO-QA，VQA-CP v2 数据集上的表现结果，模型基本都取得了 sota 的效果。

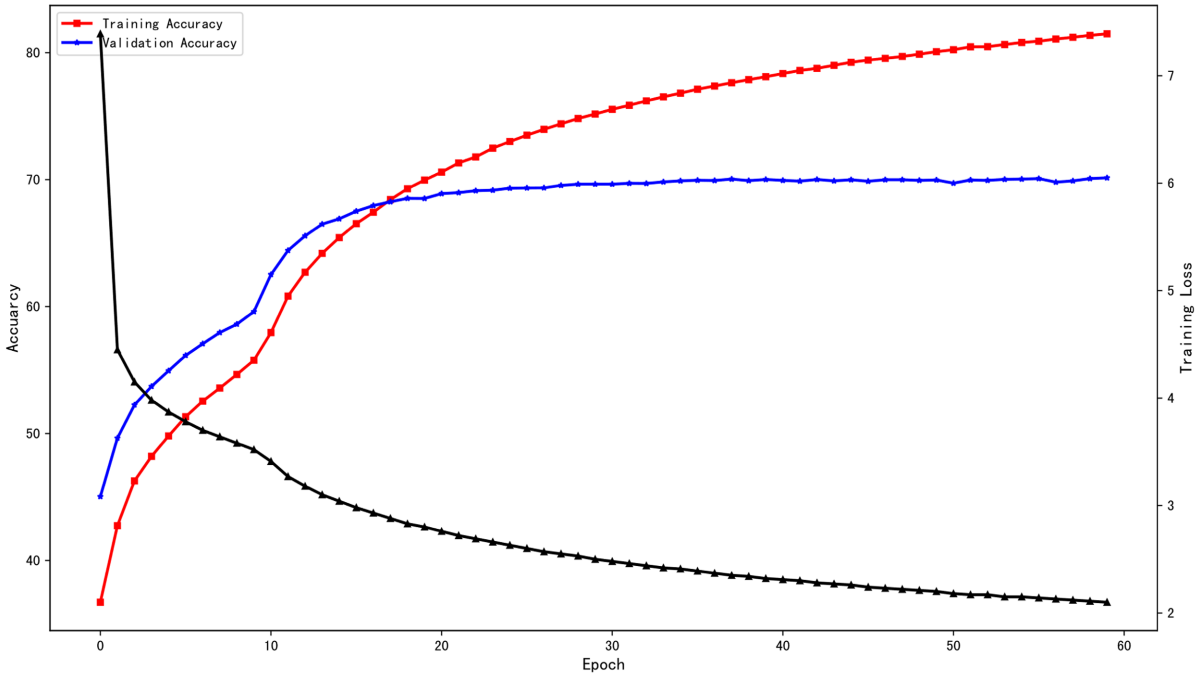

Figure 6. Model training accuracy and loss value chart
图 6. 模型训练准确率和损失值图

Table 1. Comparative experimental results of PDID model on VQA-v2.0 data set
表 1. PDID 模型在 VQA-v2.0 数据集上的对比实验结果

Method	Faster RCNN	Test-dev (%)				Test-std (%)			
		Y/N	Num	Other	Overall	Y/N	Num	Other	Overall
SLA [14]	×	79.95	40.35	55.86	63.89	80.01	40.63	55.82	64.06
MLB [15]	×	-	-	-	-	84.02	37.90	54.77	65.07
Bottom-up [16]	√	81.82	44.21	56.05	65.32	82.20	43.90	56.26	65.67
CGN [17]	√	-	-	-	-	82.91	47.13	56.22	66.18
v-VRANet [18]	√	83.31	45.51	58.41	67.20	83.39	44.96	58.49	67.34
Multi-grained [19]	√	83.60	47.02	58.24	67.41	83.88	46.60	58.50	67.73
VCTREE-HL [20]	√	84.28	47.78	59.11	68.19	84.55	47.36	59.34	68.49
CRA-Net [21]	√	84.87	49.46	59.08	68.61	85.21	48.43	59.42	68.92
Cap-Aid [22]	√	-	-	-	-	86.15	47.41	60.41	69.66
MLIN [23]	√	85.96	52.93	60.40	70.18	-	-	-	70.28
DFAF [24]	√	86.09	53.32	60.49	70.22	-	-	-	70.34
MCAN [25]	√	86.82	53.26	60.72	70.63	-	-	-	70.90
CAM [26]	√	85.18	47.35	59.76	68.82	85.22	46.98	59.91	68.99
MRA-Net [27]	√	85.58	48.92	59.46	69.02	85.83	49.22	59.86	69.46
ALSA [28]	√	85.73	48.98	59.17	69.21	-	-	-	-
PDID (ours)	×	86.81	53.36	60.33	71.02	86.99	53.09	65.28	71.23

Table 2. Comparative experimental results of PDID model on COCO-QA data set
表 2. PDID 模型在 COCO-QA 数据集上的对比实验结果

Method	Faster RCNN	Object	Number	Color	Location	WUPS0.9	WUPS0.0	Overall
SAN [29]	×	65.40	48.60	57.90	54.00	71.60	90.90	61.60
HieCoAtt [30]	×	68.00	51.00	62.90	58.80	75.10	92.00	65.40
Dual-MFA [31]	√	68.86	51.32	65.89	58.92	76.15	92.29	66.49
CVA [32]	√	69.55	50.76	68.96	59.93	76.70	92.41	67.51
ODA [33]	√	70.48	54.70	74.17	60.90	78.29	93.02	69.33
CoR-3 [34]	√	70.42	55.83	74.13	60.57	78.10	92.86	69.38
MRA-Net [35]	√	71.40	56.42	74.69	60.62	79.03	93.21	70.27
CAM [36]	√	70.32	55.26	77.10	59.28	78.53	92.97	69.68
PDID (ours)	×	71.46	55.04	74.88	61.01	78.89	92.94	70.11

Table 3. Comparative experimental results of PDID model on VQA-CP v2 data set
表 3. PDID 模型在 VQA-CP v2 数据集上的对比实验结果

Method	Faster RCNN	Y/N	Num	Other	Overall
SAN [29]	×	38.35	11.14	21.74	24.96
RAMEN [30]	√	-	-	-	39.21
MulRel [31]	×	42.85	13.17	45.04	39.54
Bottom-up [16]	×	42.27	11.93	46.05	39.74
ReGAT [32]	√	-	-	-	40.42
CAM [36]	√	43.29	12.31	45.41	39.75
MRA-Net [35]	√	44.53	13.05	45.83	40.45
PDID (ours)	×	50.82	13.24	40.39	42.09

5.4. 消融实验

5.4.1. 离散模块消融实验

本文对 PDID 做了三组消融实验，分别验证了 backbone 对模型结果的影响，注意力头数的影响以及离散化模块有效性的验证。

5.4.2. backbone 结构消融实验

数量问题的解决没有能够得到很好的处理，因为对于 vqa-cpv2 数据集，数量问题的准确率一向很低，而本文的离散化模块，并不会对数量问题造成太大的影响，所以本文认为对于数量问题是需要一些更新颖的架构，仅仅通过对实例背景的分割等途径对于数量问题依然不是一个很好的解决思路。根据表 4 可知，resnet101 比 resnet50，能够更有效的提取图像的深层特征，所以效果都有不同程度的细小的上升。本文的方法在 vqa1.0 上的提升最小，原因可能是 vqa1.0 本身的数据集的回答就比较依赖文本特征，对于图像特征的提取要求不算高。而后续的新的数据集如 vqa2.0 vqa-cp 等，都是对数据集分布做过调整的，需要更好的提取并融合图像特征。

5.4.3. 注意力头数量测试

如图 7 所示，注意力的头数越多，模型表现越好，但是同时需要综合考虑模型训练的参数量和训练时间。

Table 4. Test-dev and test-std indicators of the PDID model under different backbones on four data sets
表 4. 不同 backbone 下的 PDID 模型在四个数据集上的 test-dev 和 test-std 指标

Method	Data Set	Faster RCNN	Test-dev (%)				Test-std (%)			
			Y/N	Num	Other	Overall	Y/N	Num	Other	Overall
PDID (Resnet-18)	VQA v2.0	×	75.57	34.68	40.99	54.13	72.96	36.05	40.95	54.33
PDID (Resnet-50)	VQA v2.0	√	83.77	49.12	57.13	68.81	81.82	42.29	55.99	64.38
PDID (Resnet-101)	VQA v2.0	√	86.81	53.36	60.33	71.02	86.99	53.09	65.28	71.23
PDID (Resnet-18)	VQA v1.0	×	75.01	29.69	39.87	52.02	75.22	29.91	40.26	53.95
PDID (Resnet-50)	VQA v1.0	√	82.59	44.24	56.24	67.93	81.93	43.95	56.03	68.12
PDID (Resnet-101)	VQA v1.0	√	83.88	44.85	57.33	68.99	83.01	44.17	56.36	69.01
PDID (Resnet-18)	VQA-CP v2	×	-	-	-	-	40.55	11.05	30.79	29.32
PDID (Resnet-50)	VQA-CP v2	√	-	-	-	-	46.84	12.72	35.97	36.39
PDID (Resnet-101)	VQA-CP v2	√	-	-	-	-	47.25	13.21	36.98	37.75
PDID (Resnet-18)	COCO-QA	×	-	-	-	-	-	-	-	60.24
PDID (Resnet-50)	COCO-QA	√	-	-	-	-	-	-	-	68.68
PDID (Resnet-101)	COCO-QA	√	-	-	-	-	-	-	-	68.91
PDID (Original)	VQA v2.0	×	86.91	53.34	60.45	71.08	87.99	54.08	65.37	71.87
PDID (Original)	VQA v1.0	×	86.72	45.24	60.45	70.11	86.99	45.24	60.37	69.68
PDID (Original)	VQA-CP v2	×	-	-	-	-	50.82	13.24	40.39	42.09
PDID (Original)	COCO-QA	×	-	-	-	-	-	-	-	70.94

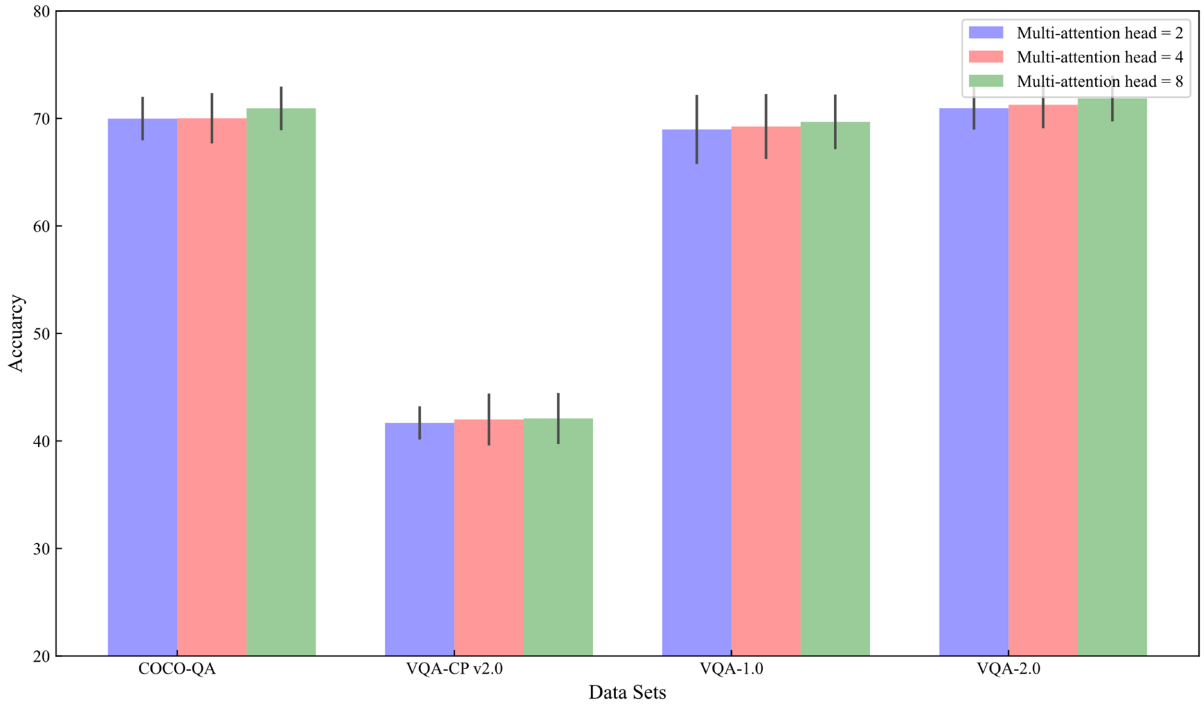

Figure 7. Number of attention heads on model performance on different datasets
图 7. 注意力头数在不同数据集上对模型表现的影响

5.4.4. 离散模块的消融测试

Table 5. Test-dev and test-std indicators of the PDID model with PD or ID on four data sets
表 5. 有无离散模块的 PDID 模型在四个数据集上的 test-dev 和 test-std 指标

Method	Data Set	Faster RCNN	Test-dev (%)				Test-std (%)			
			Y/N	Num	Other	Overall	Y/N	Num	Other	Overall
PD	VQA v2.0	×	74.37	34.68	40.99	52.23	70.37	35.00	40.25	53.59
ID	VQA v2.0	√	82.87	49.15	56.13	67.83	80.22	41.20	55.69	64.39
PDID	VQA v2.0	√	83.95	49.35	57.25	68.98	80.90	42.81	58.27	66.69
PD	VQA v1.0	×	74.71	29.35	37.97	51.62	72.22	29.91	41.28	52.78
ID	VQA v1.0	√	81.69	44.22	54.14	68.03	80.93	43.95	58.03	68.62
PDID	VQA v1.0	√	82.98	44.86	57.23	68.99	82.01	44.17	58.39	68.06
PD	VQA-CP v2	×	-	-	-	-	40.55	11.05	34.69	25.38
ID	VQA-CP v2	√	-	-	-	-	46.84	12.72	37.62	33.39
PDID	VQA-CP v2	√	-	-	-	-	47.25	13.21	36.9	35.75
PD	COCO-QA	×	-	-	-	-	-	-	-	62.24
ID	COCO-QA	√	-	-	-	-	-	-	-	68.72
PDID	COCO-QA	√	-	-	-	-	-	-	-	68.91

PD 模型指的是只有 pixel-discretization 的模型，而 ID 模型相应的指的是只有 instance-discretization 的模型，而 PDID 则是本文的原模型，我们通过消融实验表 5 来观察像素离散化模块和语义离散化模块的有效性。最终发现像素离散化是十分重要的，如果没有像素离散化，只有语义离散化模型表现极差，这是因为语义离散化依赖于像素离散化模块得到的输出，离散的端到端输入输出是重要的。而没有语义离散化的模块表现相较来说差异不大，这说明对于语义的离散化仍然需要更精细的模型来建模。

5. 总结与展望

在视觉问答(Visual Question Answering, VQA)系统中实施图像离散化带来了几项显著的优势。首先，它提高了 VQA 系统的准确性。另一个优势是对复杂视觉数据的改进处理。现实世界的图像通常多变而复杂，包含各种颜色、纹理和物体。图像离散化有助于标准化这些数据。

尽管存在这些优势，但在 VQA 系统中实施图像离散化也面临挑战。其中一个主要挑战是平衡离散化过程中保留的细节水平。过多的细节可能导致计算效率低下，而过少则可能导致丢失对于准确解释所必要的关键信息。此外，将多样且复杂的现实世界图像离散化为标准化格式的过程可能具有挑战性，需要精密的算法和谨慎的校准。

另一个挑战是在多样化和不受控制的环境中准确解释离散化图像。现实世界中的图像在质量、光照、透视和构图方面可能有很大的变化。确保离散化过程始终捕捉到这些不同图像的关键特征是一项复杂的任务。这种变异性可能影响系统的泛化能力，使其在不同类型的图像上准确回答问题的能力受到影响。

图像离散化与 VQA 系统的其他组件(如自然语言处理和深度学习算法)的整合也带来挑战。确保这些组件之间的无缝交互对系统的整体性能至关重要。视觉和文本数据表示的不一致可能导致误解和错误答案。实现有效地将视觉分析与语言解释相结合的和谐整合仍然是研究和发展的一个关键领域。

总的来说，图像离散化显著提升了 VQA 系统的性能。它提高了准确性，更有效地处理复杂的视觉数据，提高了计算效率，并使先进的机器学习技术得以整合。这些优势使图像离散化成为 VQA 技术持续发

展和完善中的关键组成部分。

致　谢

在完成本论文的过程中，我要衷心感谢所有给予我支持和帮助的人。首先，我要感谢我的导师孙杳如教授，他给予了我宝贵的建议、专业的指导和无私的支持，使得我能够顺利完成这篇论文。他们的耐心指导和鼓励对我在研究过程中起到了至关重要的作用。我还要感谢实验室的同事和其他教授/同学们，在研究和讨论中给予了我很多启发和帮助，使我的研究更加丰富和深入。此外，我要感谢我的家人和朋友，他们在我写作过程中给予了我精神上的支持和鼓励，让我能够专注于论文的完成。最后，我要感谢所有曾经为本论文提供帮助的人，尽管我无法一一列举，但你们的支持对我而言是非常宝贵的。感谢你们在我学术道路上的陪伴和支持。

参考文献

[1] Ardila, A., Bernal, B. and Rosselli, M. (2015) Language and Visual Perception Associations: Meta-Analytic Connectivity Modeling of Brodmann Area 37. *Behavioural Neurology*, **2015**, Article ID: 565871. https://doi.org/10.1155/2015/565871

[2] Antol, S., Agrawal, A., Lu, J., Mitchell, M. and Parikh, D. (2015) VQA: Visual Question Answering. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 2425-2433. https://doi.org/10.1109/ICCV.2015.279

[3] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[4] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv: 1409.1556.

[5] Peng, Z., Yash, G., Douglas, S.S., Dhruv, B. and Devi, P. (2016) Balancing and Answering Binary Visual Questions. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 5014-5022.

[6] Shih, K.J., Singh, S. and Hoiem, D. (2016) Where to Look: Focus Regions for Visual Question Answering. *IEEE Conference on Computer Vision and Pattern Recognition* 2016, Las Vegas, 27-30 June 2016, 4613-4621. https://doi.org/10.1109/CVPR.2016.499

[7] Alex, K., Ilya, S. and Geoffrey, E.H. (2017) Imagenet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. https://doi.org/10.1145/3065386

[8] Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A. and Fergus, B. (2015) Simple Baseline for Visual Question Answering. arXiv: 1512.02167.

[9] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., *et al*. (2015) Going Deeper with Convolutions. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[10] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T. and Rohrbach, M. (2016) Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *Proceedings of the* 2016 *Conference on Empirical Methods in Natural Language Processing*, Austin, November 2016, 457-468. https://doi.org/10.18653/v1/D16-1044

[11] Charikar, M., Chen, K.C. and Colton, M.F. (2002) Finding Frequent Items in Data Streams. In: Widmayer, P., Eidenbenz, S., Triguero, F., Morales, R., Conejo, R. and Hennessy, M., Eds., *Automata, Languages and Programming*, Springer, Berlin, 693-703. https://doi.org/10.1007/3-540-45465-9_59

[12] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., *et al*. (2015) Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the* 32*nd International Conference on Machine Learning*, Lille, 6-11 July 2015, 2048-2057.

[13] Yang, Z., He, X., Gao, J., Deng, L. and Smola, A. (2015) Stacked Attention Networks for Image Question Answering. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 21-29. https://doi.org/10.1109/CVPR.2016.10

[14] Lu, P., Li, H., Zhang, W., Wang, J. and Wang, X. (2018) Co-Attending Freeform Regions and Detections with Multi-Modal Multiplicative Feature Embedding for Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 7218-7225. https://doi.org/10.1609/aaai.v32i1.12240

[15] Teney, D., Anderson, P., He, X. and Van Den Hengel, A. (2018) Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4223-4232. https://doi.org/10.1109/CVPR.2018.00444

[16] Wu, C., Liu, J., Wang, X. and Li, R. (2019) Differential Networks for Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 8997-9004. https://doi.org/10.1609/aaai.v33i01.33018997

[17] Zhang, L., *et al*. (2021) Rich Visual Knowledge-Based Augmentation Network for Visual Question Answering. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 4362-4373. https://doi.org/10.1109/TNNLS.2020.3017530

[18] Xie, E., *et al*. (2020) PolarMask: Single Shot Instance Segmentation with Polar Representation. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 12190-12199. https://doi.org/10.1109/CVPR42600.2020.01221

[19] Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y. and Yan, Y. (2020) BlendMask:Top-Down Meets Bottom-Up for Instance Segmentation. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 8570-8578. https://doi.org/10.1109/CVPR42600.2020.00860

[20] Gao, P., *et al*. (2019) Dynamic Fusion with Intra- and Inter-Modality Attention Flow for Visual Question Answering. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 6632-6641. https://doi.org/10.1109/CVPR.2019.00680

[21] Stefanini, M., Cornia, M., Baraldi, L. and Cucchiara, R. (2021) A Novel Attentionbased Aggregation Function to Combine Vision and Language. 2020 25*th International Conference on Pattern Recognition* (*ICPR*), Milan, 10-15 January 2021, 1212-1219. https://doi.org/10.1109/ICPR48806.2021.9413269

[22] Wu, C., Liu, J., Wang, X. and Dong, X. (2018) Object-Difference Attention: Asimple Relational Attention for Visual Question Answering. *Proceedings of the* 26*th ACM International Conference on Multimedia*, Seoul, 22-26 October 2018, 519-527. https://doi.org/10.1145/3240508.3240513

[23] Peng, L., Yang, Y., Wang, Z., Wu, X. and Huang, Z. (2019) CRA-Net: Composed Relation Attention Network for Visual Question Answering. *Proceedings of the* 27*th ACM International Conference on Multimedia*, Nice France, 21-25 October 2019, 1202-1210. https://doi.org/10.1145/3343031.3350925

[24] Yang, Z., He, X., Gao, J., Deng, L. and Smola, A. (2016) Stacked Attention Networks for Image Question Answering. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 21-29. https://doi.org/10.1109/CVPR.2016.10

[25] Li, L., Gan, Z., Cheng, Y. and Liu, J. (2019) Relation-Aware Graph Attention Network for Visual Question Answering. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 10312-10321. https://doi.org/10.1109/ICCV.2019.01041

[26] Osman, A. and Samek, W. (2019) DRAU: Dual Recurrent Attention Units for Visual Question Answering. *Computer Vision and Image Understanding*, **185**, 24-30. https://doi.org/10.1016/j.cviu.2019.05.001

[27] Peng, L., *et al*. (2019) Word-to-Region Attention Network for Visual Question Answering. *Multimedia Tools and Applications*, **78**, 3843-3858. https://doi.org/10.1007/s11042-018-6389-3

[28] Liu, Y., Zhang, X., Zhao, Z., Zhang, B., Cheng, L. and Li, Z. (2022) ALSA: Adversarial Learning of Supervised Attentions for Visual Question Answering. *IEEE Transactions on Cybernetics*, **52**, 4520-4533. https://doi.org/10.1109/TCYB.2020.3029423

[29] Zhong, H., Chen, J., Shen, C., Zhang, H., Huang, J. and Hua, X.S. (2021) Selfadaptive Neural Module Transformer for Visual Question Answering. *IEEE Transactions on Multimedia*, **23**, 1264-1273. https://doi.org/10.1109/TMM.2020.2995278

[30] Peng, L., Yang, Y., Wang, Z., Huang, Z. and Shen, H.T. (2022) MRA-Net: Improving VQA via Multi-Modal Relation Attention Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 318-329. https://doi.org/10.1109/TPAMI.2020.3004830

[31] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805. http://arxiv.org/abs/1810.04805

[32] Wang, X., Zhang, R., Kong, T., Li, L. and Shen, C. (2020) SOLOv2: Dynamic and Fast Instance Segmentation. 34*th Conference on Neural Information Processing Systems* (*NeurIPS* 2020), Vancouver, 6-12 December 2020, 17721-17732.

[33] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X. and Huang, T.S. (2018) Generative Image Inpainting with Contextual Attention. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 5505-5514. https://doi.org/10.1109/CVPR.2018.00577

[34] Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W. and Zhang, B.T. (2017) Hadamard Product for Low-Rank Bilinear Pooling. arXiv: 1610.04325. http://arxiv.org/abs/1610.04325

[35] Bai, Y., Fu, J., Zhao, T. and Mei, T. (2018) Deep Attention Neural Tensor Network for Visual Question Answering. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *ECCV* 2018: *Computer Vision—ECCV* 2018, Springer, Cham, 21-37. https://doi.org/10.1007/978-3-030-01258-8_2

[36] Wu, C., Liu, J., Wang, X. and Dong, X. (2018) Chain of Reasoning for Visual Question Answering. 32*nd Conference on Neural Information Processing Systems* (*NeurIPS* 2018), Montréal, 3-8 December 2018, 275-285.