朱明航,冯杰,马汉杰,等. 基于昇腾平台的图像描述算法的部署与优化[J]. 智能计算机与应用,2024,14(11):52-58. DOI:10.20169/j.issn.2095-2163.241107

基于昇腾平台的图像描述算法的部署与优化

朱明航[1], 冯 杰[1], 马汉杰[1], 邵蒙悦[2], 刘新天[1], 张海翔[1]
(1 浙江理工大学 计算机科学与技术学院(人工智能学院), 杭州 310018; 2 浙江理工大学 信息科学与工程学院, 杭州 310018)

摘 要: 图像描述是一种通过文字来解释和呈现图像内容的技术,在计算机视觉、图像识别和人工智能等领域中具有重要的应用。针对华为昇腾平台,提出一个可部署的高性能图像描述模型。首先从高精确度的图像描述模型出发,通过算子可行性、精确度与计算量两个方面进行综合分析,得到可部署且高效的方案,并对其进行一些算法上的优化,最终得到一个高性能的昇腾离线模型。在华为昇腾平台上,使用生成的模型对多张图像进行描述并分析描述结果,所生成的模型各个指标均有提升,其中 CIDEr 指标提升 9%,每张图片推理时间为 210.43 ms。

关键词: 昇腾; 图像描述; 强化学习; 束算法; 神经网络处理器
中图分类号: TP311.1 **文献标志码**: A **文章编号**: 2095-2163(2024)11-0052-07

Deployment and optimization of image captioning algorithm based on Ascend platform

ZHU Minghang[1], FENG Jie[1], MA Hanjie[1], SHAO Mengyue[2], LIU Xintian[1], ZHANG Haixiang[1]
(1 School of Computer Science and Technology(School of Artificial Intelligence), Zhejiang Sci-Tech University, Hangzhou 310018, China; 2 School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Image captioning is a technique for explaining and presenting image content through texts. It has important applications in the fields of computer vision, image recognition and artificial intelligence. A deployable high-performance image description model is proposed for the Huawei Ascend platform. Firstly, starting from the high-precision image description model, through the comprehensive analysis of operator feasibility, accuracy and calculation amount, the deployable and efficient solution is obtained, and some algorithm optimizations are performed on it. Finally the solution, namely the high-performance Ascend offline model, is obtained. On the Huawei Ascend platform, the generated model is used to describe multiple images and the description results are analyzed. All indicators of the generated model are improved, among which the CIDEr indicator is increased by 9%, and the inference time for each image is 210.43 milliseconds.

Key words: Ascend; image captioning; reinforcement learning; beam search; NPU

0 引 言

图像描述技术的发展前景相当广阔,可以涵盖多个领域。例如帮助视障人士理解图像内容,从而提高生活质量;通过自动生成对图像内容的描述,更加精确地进行图像检索,提高图像检索的准确率;帮助自动标注图片,从而减轻人工标注的负担,提高标注效率等。图像描述的关键步骤包括图像特征提取、语义理解和句子生成。在特征提取阶段,计算机会从图像中提取出有用的特征向量,通常使用卷积神经网络进行处理。接下来,在语义理解阶段,计算机会将特征向量与语义信息进行关联,以便理解图像中的对象、场景和关系。最后,在句子生成阶段,计算机利用这些信息生成自然语言的描述。然而在这些步骤中使用的算子复杂,模型适配性差,导致其部署困难且性能较低。针对这一问题,本文提出一

基金项目: 浙江省科技计划项目(2021C01163)。

作者简介: 朱明航(1999—),男,硕士研究生,主要研究方向:图像描述,语音合成; 马汉杰(1982—),男,博士,副教授,主要研究方向:视频图像传输与处理,机器视觉,情感计算等; 邵蒙悦(1998—),女,硕士研究生,主要研究方向:图像描述; 刘新天(1998—),男,硕士研究生,主要研究方向:文字检测与识别; 张海翔(1973—),男,博士,副教授,主要研究方向:三维数字人,三维重建。
通信作者: 冯 杰(1980—),男,博士,讲师,主要研究方向:自动推理与文字检测与识别,视频分析与处理。Email: arlose@zstu.edu.cn。

收稿日期: 2023-06-25

个高性能且易于部署的 ResNet[1] 与 X-LAN[2] 的组合模型方案，并将其成功部署至带有神经网络处理器（Neural-network Processing Unit，NPU）的昇腾平台设备。

1　相关技术

1.1　图像描述算法

2015 年受到机器翻译领域新的研究成果的启发，Vinyals 等学者[3] 首次采用了编码器-解码器结构。其中，编码器为特征提取模块，使用预先训练好的 CNN 模型，例如 ResNet。解码器为描述生成模块，使用循环神经网络（Recurrent Neural Network，RNN）将图像信息转换为文本信息。2018 年，Anderson 等学者[4] 首先采用 Bottom-Up 模型提取图像特征，该模型基于目标检测 Faster R-CNN[5] 网络改进。深度学习方法具有较好的泛化能力，在大规模数据集上进行训练，经过不断的学习与更新，可以处理未知的语言结构和语义，应用于更丰富的场景。郝义铭[6]、乔玉聪[7] 皆采用 CNN-RNN 结构来实现图像描述目的。Deng 等学者[8] 尝试使用特征金字塔网络（Feature Pyramid Networks，FPN）[9] 模型提取多层特征，使模型在不增加参数的情况下更有效地检测图像中不同尺度的物体。Nejatishahidin 等学者[10] 尝试使用图像分割网络提取对象掩码和中层表示特征图，以利于在有限的训练数据场景下获得具有竞争力的性能。Yao 等学者[11] 设计了 5 种 LSTM 的变体，分析并得到具有竞争力的结果。Huang 等学者[12] 提出了注意力上的注意（Attention on Attention，AoA）模块。AoA 通过 2 条线扩展标题模型的注意力，有利于不同模式（文本和图像）的信息建模和融合。Guo 等学者[13] 提出了拥有几何感知的自注意力模块，该方法通过考虑物体的成对几何关系和内容信息来改进自注意力机制模块，从而有助于对视觉信息进行推理。Pan 等学者[2] 提出了 X-Linear 注意力模块，将 X-Linear 注意块集成到描述生成模型中得到 X-线性注意模型（X-Linear Attention Networks，X-LAN）。该模型可以捕获高阶内模态和多模态交互信息，旨在增强视觉信息，并对图像标题进行复杂的多模态推理。

1.2　昇腾平台

昇腾计算产业是基于昇腾系列（HUAWEI Ascend）处理器和基础软件构建的全栈 AI 计算基础设施、行业应用及服务，包括昇腾系列处理器、系列硬件、异构计算架构（Compute Architecture for Neural Networks，CANN）、AI 计算框架、应用使能、开发工具链、管理运维工具、行业应用及服务等全产业链。本文具体使用的昇腾平台设备是华为 Atlas200DK 开发者套件。Atlas200DK 开发者套件 Atlas200Developer Kit（简称 Atlas200DK）是以昇腾 310 AI 处理器为核心的一个开发者板形态产品，具有至少 8 *TOPS*（Tera Operations Per Second）的 AI 算力。CANN 是华为针对 AI 场景推出的异构计算架构，对上支持多种 AI 框架，对下服务 AI 处理器与编程，发挥承上启下的关键作用，是提升昇腾 AI 处理器计算效率的关键平台。同时针对多样化应用场景，提供高效易用的编程接口，支持用户快速构建基于昇腾平台的 AI 应用和业务。

综上所述，图像描述主要分为 2 个部分：特征提取和描述生成。其中，特征提取部分的算法主要基于 CNN 网络，用于区域特征和全局特征提取并融合。描述生成部分的算法则主要基于 RNN 或者 Transformer。本文主要针对特征提取模块和描述生成模块的部署方案进行研究，并对其进行性能上的优化。最终在昇腾平台上进行实验并对实验结果做出分析。

2　部署方案选择

图像描述模型分为特征提取和描述生成 2 部分，接下来将从算子可行性、精确度与计算量 2 个方面对这 2 部分进行分析。本文在 ResNet101 基本网络结构的基础上，去除最后一层全连接层，并且修改自适应最大池化层的输出，将输出大小从 1 × 1 修改为 14 × 14。在对图像进行编码后，将所得的特征输入到修改过后的空间自适应最大池化层，会得到 14 × 14 × 2 048 的特征。这样既增加网络的感受野又能保留更多的有用特征信息，从而提高后续描述网络的表达能力。本文为了区分标准的 ResNet101，将上述描述的 ResNet 特征提取网络的模型命名为 RExtract 模型。Anderson 等学者[4] 提出的 Bottom-Up 特征提取算法来提取区域特征。该算法在保留 Faster R-CNN 的基本结构的基础上，将平均池卷积特征与真实对象类的信息连接起来，并将其输入到一个额外的输出层，定义每个属性类和“无属性”类的 Softmax 分布，从而预测区域目标的属性。本文将以 RExtract、Bottom-Up 等特征提取模型和 LSTM-A[14]、RFNet[15]、SCST[16]、X-LAN 等描述生成模型为候选方案并对其评估，最终选择出最优的部署方案。

2.1 算子可行性评估

模型部署的首要任务是将深度学习框架的模型,例如 PyTorch 模型,转换为昇腾离线模型。昇腾离线模型适配要求模型中不包含动态计算形状的算子,例如 Range 算子与 Max 算子的组合,以及一些控制类的算子,例如 Loop、If 算子。在 RExtract 和 X-LAN 网络中,并不包含控制类算子,以及不存在动态计算形状的算子。表 1 说明了特征提取模型和描述生成模型的控制类算子和动态计算形状算子的存在情况。由表 1 可知,在候选的特征提取模型中,Bottom-Up 模型无法满足适配需求,所以选择 RExtract 作为特征提取模型。在描述生成模型中,LSTM-A、X-LAN、SCST、RFNet 均满足算子适配要求,所以将 LSTM-A、X-LAN、SCST、RFNet 作为候选的描述生成模型。

表 1 所有候选模型的算子适配情况

Table 1 Operator adaptation of all candidate models

模型	无控制类算子	无动态计算形状算子
RExtract	√	√
Bottom-Up	√	×
LSTM-A	×	×
X-LAN	√	√
SCST	√	√
RFNet	√	√

2.2 精确度与计算量评估

初始方案的描述生成模型为 X-LAN,下面将从精确度与计算量方面对其进行评估。对于昇腾离线模型而言,少量的精度损失与较多的速度提升是部署的最优方案。因此本文将结合精确度和计算量进行综合评估。本文所使用的评价指标包括 BLEU[17]、METEOR[18]、ROUGE-L[19]、CIDEr[20] 和 SPICE[21]。其中,BLEU 指标有 BLEU@1-4。这 8 个指标是常用于图像描述的指标,能从 4 个方面全面评估模型生成的文本的精确性。

本文对 X-LAN 与 RExtract 的组合模型进行精确度与计算量分析,结果见表 2。根据 LSTM-A 和 RFNet 论文内容描述,其复杂度都高于 SCST 模型。

表 2 RExtract 与不同的图像描述网络的组合模型的性能比较

Table 2 Performance comparison of different image caption models and RExtract

Model	*Flops*/ G	*B*@1	*B*@2	*B*@3	*B*@4	*M*	*R*	*C*	*S*
SCST	10.1	-	-	-	30.0	25.9	53.4	99.4	-
LSTM-A	-	75.4	-	-	35.2	26.9	55.8	108.8	20.0
RFNet	-	76.4	60.4	46.6	35.8	27.4	**56.8**	112.5	20.5
X-LAN	14.9	**76.4**	**60.5**	**46.9**	**36.1**	**28.1**	56.6	**116.5**	**21.1**

表 2 的结果表明,对于其他同样使用 RExtract 特征提取模型的图像描述网络而言,X-LAN 模型的效果更优,CIDEr 指标显著提升。虽然使用 X-LAN 计算量有一定的提高,但这些计算量的浮动,仍在可控的范围内。根据研究得到的结果,RExtract + X-LAN 组合的模型在昇腾平台,模型推理大约需要 180 ms(不包括预处理的时间)。因此本文选择 X-LAN 模型作为描述生成模型。

综上所述,本文从算子可行性、精确度与计算量等方面对模型方案进行综合评估,最终使用 RExtract 模型与 X-LAN 模型为可部署的最高效方案。RExtract 提取特征的流程类似于 Faster-RCNN,最后拼接了一系列卷积用于得到细粒度特征。将细粒度特征输入到 X-LAN 描述生成模型中,经过 X-LAN 的编码器和解码器,即可得到最后的图像描述结果,如图 1 所示。为了方便后续描述,本文将 RExtract + X-LAN 组合模型简称为 R-XLAN 模型。

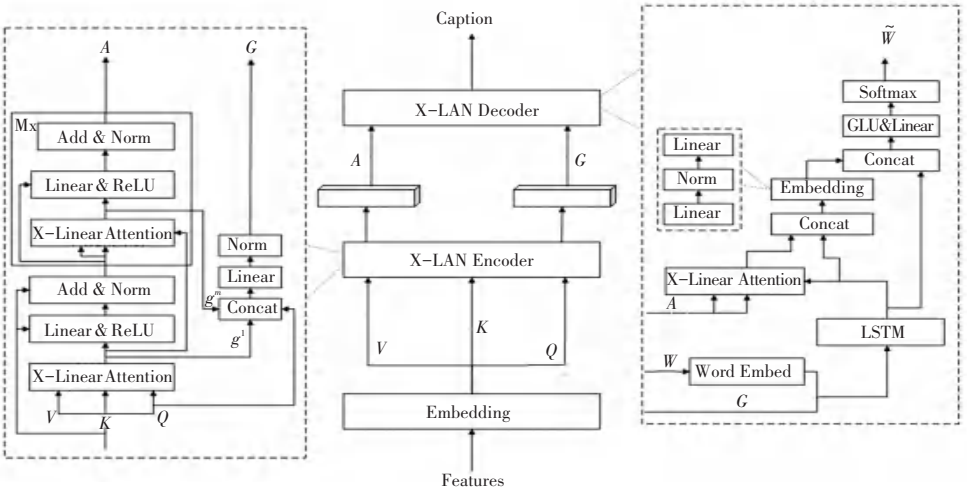

图 1 图像描述模型的网络结构图

Fig. 1 Network structure diagram of image caption model

3　部署模型的优化

根据第 2 节的综合分析，本文最终选定的部署模型是 R-XLAN 模型，并且对这个部署模型进行优化。本文在 2 个过程中对模型进行优化。一个是训练的过程，一个模型转换的过程。在训练的过程中，本文采用强化学习方法对 R-XLAN 模型进行优化；在模型转换的过程中，本文根据何时使用束算法以及开启昇腾的 AI 预处理（Artificial Intelligence Pre-Processing，AIPP）配置对 R-XLAN 模型进行优化。

3.1　强化学习优化

强化学习（Reinforcement Learning）是一种机器学习方法，后期也被广泛用于深度学习算法中。该方法的原理是首先让一个智能体（Agent）根据环境（Environment），不断学习更新自身状态（State）以及下一步动作（Action），并得到相应的奖励（Reward），之后再根据所得的奖励继续调整状态和动作，最终得到一个最优策略，使其在未来的决策过程中获得最大化的收益。

本文采用 SCST 中阐述的强化学习方法自我评估（Self-Critical）对 R-XLAN 网络进行训练，提升网络性能。Self-Critical 基于 Reinforce 算法[22]，通过最大化图像描述的评估指标来优化网络。Reinforce 算法是一种无模型的策略优化算法，不需要对环境建立模型，只需要与环境进行交互即可，适用于连续动作空间和非线性策略的优化问题。本文将 Self-Critical 方法融入到 R-XLAN 网络中，以 CIDEr 指标的得分作为奖励，在最小化奖励的负期望值的同时，不断更新模型参数，提升模型性能。本文对 R-XLAN 网络进行强化学习训练，并将其结果与只在交叉熵损失函数下训练的结果进行对比。性能对比结果见表 3。其中，Cross-Entropy 表示只在交叉熵损失函数下，没有进行强化学习的训练结果，CIDEr 表示强化学习下的训练结果。实验结果表明，通过强化学习，各个指标的分数都有显著提升，其中 CIDEr 指标的分数提升幅度最大，可以达到 9% 左右。因此，本文选择使用强化学习训练的模型作为部署模型。

表 3　不同训练方式下的性能比较

Table 3　Performance comparison under different training methods

Method	*B*@1	*B*@2	*B*@3	*B*@4	*M*	*R*	*C*	*S*
Cross-Entropy	76.4	60.5	46.9	36.1	28.1	56.6	116.5	21.1
CIDEr	**79.4**	**63.8**	**49.6**	**37.8**	**28.6**	**57.7**	**125.3**	**22.5**

3.2　束算法优化

束算法（Beam Search）是一种用于在离散搜索空间中寻找最优解的启发式搜索算法。该算法在搜索空间中找出当前最优的 b 个解，每次从所有候选解中选择分数最高的 k 个解（k 通常小于等于 b），对这些解执行扩展（即向后面的解移动），并对每个新解重复生成、排序、选择步骤，直到找到目标解或达到预定的停止条件。Beam Search 算法可以平衡搜索时间和搜索精度，且实现起来较简单。

在自然语言处理领域，束算法常用于解码某些基于概率的模型生成的序列，如机器翻译中，解码模型生成的序列，得到每个位置的单词。对于基于深度学习的语言模型而言，每个单词的生成都是一个条件概率问题。给定前面的 N 个单词，模型需要预测下一个最有可能出现的单词。在这个生成过程中，Beam Search 算法可以通过保留最有可能的若干个候选单词序列，以减少搜索空间并快速达到最优解。R-XLAN 模型在描述生成部分类似于自然语言处理领域里的语言模型，输出基于概率的序列，并且在得到该序列的过程中，需要不断地预测下一个单词中。因此也常使用了 Beam Search 算法获得最优的描述文本。

然而束算法在 PyTorch 模型转换为 ONNX（Open Neural Network Exchange）模型时存在结果固定的问题，因此需要将 Beam Search 算法改为贪心算法（Greedy）。然而改用贪心算法，CIDEr 指标将下降 1%。经测试发现，模型转换时，只要转换过程用的代码为 Greedy 算法，无论 PyTorch 模型是否使用 Beam Search 算法都能成功转换。因此本文选择训练时使用束算法，转换模型时用贪心算法。实验证明这样的方法优于训练和模型转换都使用 Beam Search 的方法（见表 4）。R-XLAN-1 模型为训练和模型转换时都使用 Greedy 方法；R-XLAN-2 模型为训练时使用 Beam Search 方法，模型转换时使用 Greedy 方法；R-XLAN-3 模型为训练和模型转换时都使用 Beam Search 方法；greedy 表示使用贪心算法；beam 表示使用 Beam Search 算法。实验结果表明，与先前训练转换都用 Beam Search 的方法（R-

XLAN-3)相比,本文方法的结果虽然有下降,但是下降幅度较小,CIDEr 指标分数只下降 0.2%,优于训练转换都用 Greedy 的方法(R-XLAN-1)。因此,本文选择 R-XLAN-2 为部署模型。

表 4 训练和转换时选择不同搜索方法的结果对比

Table 4 Comparison of the results of different search methods during training and conversion

Model	Train	ONNX	*B*@1	*B*@2	*B*@3	*B*@4	*M*	*R*	*C*	*S*
R-XLAN-1	greedy	greedy	78.8	60.1	48.9	37.2	28.7	57.7	124.3	22.5
R-XLAN-2	beam	greedy	79.2	63.5	49.3	37.5	28.6	57.7	125.1	22.5
R-XLAN-3	beam	beam	79.4	63.8	49.6	37.8	28.6	57.7	125.3	22.5

3.3 AIPP 配置优化

本文采取配置不同的输入形式的方法来优化模型的推理速度。昇腾张量编译器(Ascend Tensor Compiler,ATC)工具可以配置 3 种形式的输入类型:Float32、Float16 以及 Uint8。开启 Uint8 输入类型,就是默认启动 AIPP 算子。AIPP 算子是图像预处理算法,可以在模型中直接对图像进行裁剪、填充、色域转换、归一化等操作,因此不需要在 C 代码中进行图像预处理操作。

为了得到速度最快的模型转换方式,本文实验对比 4 种不同输入形式的方法。首先,是不开启 AIPP 配置,这个条件下有 3 种输入形式:Float32、Float16。在这个情况下,需要先将读入的图像进行裁剪、归一化等操作。其次,是开启 AIPP 配置,此时的输入不需要预处理操作,实验结果见表 5。值得注意的是,这里的时间并非只是模型推理的时间,还包括图像读取和预处理。

表 5 不同部署方式的时间对比

Table 5 Time comparison of different deployment methods

Method	Resize shape	*Time*/ ms
Float32	256×256	232.80
Float16	256×256	215.03
AIPP	256×256	210.43

根据表 5 可知,输入精度为 Float32 时,模型推理加图像预处理的时间为 232.80 ms,输入精度为 Float16 时,时间为 215.03 ms。输入精度为 Float32 的速度远远大于输入精度为 Float16 的速度,这因为模型内部运算常使用 Float16 的精度,当输入为 Float32 时会生成一个 cast 算子,将 Float32 的精度转为 Float16 的精度,这个算子会大大增加计算量。后续的计算方式一致,因此两者之间速度的差距主要来自于 cast 算子。开启 AIPP 配置时,模型推理时间为 210.43 ms,在这 3 个方法中,速度最快,比 Float16 快了近 5 ms。开启 AIPP 配置,使用的输入精度为 Uint8,但是在经过一系列的预处理操作后,得到的精度也为 Float16。因此,本质上开启 AIPP 模型整体的计算方式和 Float16 一致,区别就在于一个在模型内部进行图像预处理,另一个在模型外部进行图像预处理。因此可以看出使用 AIPP 算子预处理图像的速度快于在 C 代码中使用 OpenCV 预处理图像。

最终本文选择开启 AIPP 配置,并且实验得到,完成从图像输入到描述语句输出整个过程,开启 AIPP 配置的模型需要花费 210.43 ms。

综合 3.1 节到 3.3 节,本文选择 R-XLAN-2 模型进行部署,该模型通过强化训练,并在训练时开启束算法,转为 ONNX 的时候使用贪心算法。最后,在 ONNX 转为昇腾离线模型的时候开启 AIPP 配置。

4 实验结果分析

本文完成模型的部署开发以及优化后,使用 R-XLAN-2 的昇腾离线模型得到对图像的描述结果,具体结果见图 2。

从结果看出模型可以将图片内的大致内容准确地表述出来,并且对于图片内含有单个目标的物体。除了对目标进行描述,R-XLAN-2 模型还能关注到一些周边信息,例如图 2(a)中的“in a field”;(b)中的“on top of an airport tarmac”;(d)中的“on the side of the road”。

然而,R-XLAN-2 模型的描述还存在缺陷,对于包含多个不同类型物体的图片,该模型的描述并不丰富,且存在一些误差。比如,图 2(c)中描述为“a group of women”,但是该图中并不全是女人。发生这样的偏差是由于目前的模型为了实现部署,放弃了区域特征,因此缺少对类别进行分辨的能力,使得模型以单一的类别特征来描述图像中所有相似类别。

总而言之,模型对于图像的描述是较为准确的。

但是如果图片中存在多个物体，模型只能对其进行简单的描述，通常会错误地使用“a group of”这样的词组，并且对个体的描述可能存在偏差。而对于单一目标而言，描述结果准确且用词丰富。

(a) 输出结果 1　(b) 输出结果 2　(c) 输出结果 3　(d) 输出结果 4　(e) 输出结果 5

图 2　R-XLAN-2 模型输出结果展示图

Fig. 2　Model output result display of R-XLAN-2

5　结束语

本文首先从最初的方案出发，并与常使用的特征提取算法和描述生成算法从算子可行性、精确度与计算量等方面进行综合评估，得到 R-XLAN 模型方案。R-XLAN 模型分为 2 部分：使用 REXtract 模型提取图像特征和使用 X-LAN 模型描述图像内容。此后本文通过 3 个方式优化模型：强化学习优化、束算法优化以及 AIPP 配置优化，从而得到 R-XLAN-2 的昇腾离线模型。最后，本文测试对比了昇腾离线模型得到的结果与参考文本，测试结果表明，该昇腾离线模型可以较为准确地描述图像内的信息，但是对于多目标的图像，描述的信息较为简单，并有一定误差，而且仍有很多关注不到的地方，比起人类描述图像的能力仍显不足。为解决这个问题，未来将寻找更优、且易部署的特征提取模型以及更轻量的描述生成模型。

参考文献

[1] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778.

[2] PAN Yingwei, YAO Ting, LI Yehao, et al. X-linear attention networks for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10971-10980.

[3] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3156-3164.

[4] ANDERSON P, HE Xiaodong, 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. Piscataway, NJ: IEEE, 2018: 6077-6086.

[5] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems. Montreal, Canada: NIPS Foundation, 2015: 91-99.

[6] 郝义铭. 基于深度学习的图像自动描述算法研究[D]. 济南：山东大学, 2022.

[7] 乔玉聪. 基于深度学习的图像描述研究[D]. 西安：西安邮电大学, 2022.

[8] DENG Zelin, ZHOU Bo, HE Pei, et al. A position-aware transformer for image captioning[J]. Computers, Materials and Continua, 2021, 70(1): 2005-2021.

[9] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2117-2125.

[10] NEJATISHAHIDIN N, FAYYAZSANAVI P, KOŠECKA J. Object pose estimation using mid-level visual representations[C]//Porceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, NJ: IEEE, 2022: 13105-13111.

[11] YAO Ting, PAN Yingwei, LI Yehao, et al. Boosting image captioning with attributes[C]//Proceedings of the IEEE

International Conference on Computer Vision. Piscataway, NJ: IEEE,2017: 4894-4902.
[12] HUANG L, WANG Wenmin, CHEN Jie, et al. Attention on attention for image captioning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ:IEEE, 2019: 4634-4643.
[13] GUO Longteng, LIU Jing, ZHU Xinin, et al. Normalized and geometry-aware self-attention network for image captioning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10327-10336.
[14] YAO Ting, PAN Yingwei, LI Yehao, et al. Boosting image captioning with attributes [C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 4894-4902.
[15] JIANG Wenhao, MA Lin, JIANG Yugang, et al. Recurrent fusion network for image captioning[C]//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 499-515.
[16] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ:IEEE, 2017: 7008-7024.
[17] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: ACL, 2002: 311-318.
[18] DENKOWSKI M, LAVIE A. Meteor universal: Language specific translation evaluation for any target language [C]// Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, USA:ACL, 2014: 376-380.
[19] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Text Summarization Branches Out. Barcelona, Spain:ACL, 2004: 74-81.
[20] VEDANTAM R, LAWRENCE Z C, PARIKH D. Cider: Consensus-based image description evaluation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ:IEEE,2015: 4566-4575.
[21] ANDERSON P, FERNANDO B, JOHNSON M, et al. Spice: Semantic propositional image caption evaluation [C]// Proceedings of the 14th European Conference on Computer Vision. Cham:Springer, 2016: 382-398.
[22] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning [J]. Reinforcement Learning, 1992,8: 5-32.