http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2024.0656

基于改进 Transformer 的视觉语言模型

金泽辉[1]，陈凯[1,*]，唐敦兵[1]，方国宇[1]，温菲霞[2]

(1. 南京航空航天大学 机电学院，南京 210016；2. 中国电子科技集团公司第二十八研究所，南京 210007)

摘 要：图像字幕是一项结合计算机视觉和自然语言处理的跨模态任务，模型需要为图像生成恰当的描述。为了应对这一挑战，提出一种带有残差注意力模块和门控线性单元的增强Transformer——残差门控 Transformer (RGFormer)。基于 CLIP 和 RGFormer，提出一种生成图像字幕的模型——CRM。CRM 利用 CLIP 的图像编码器提取特征作为图像字幕的前缀，该前缀通过映射网络 RGFormer 被投影到文本空间，然后被输入 GPT-2 语言模型中生成图像字幕。CLIP 是在一个大型的图像-文本对数据集上进行训练的，包含丰富的视觉和语义信息，非常适合完成视觉-语言跨模态任务。CRM 的核心理念是使用 RGFormer 映射网络来缩小视觉与文本表示之间的差距，从而完成图像字幕任务。CRM 可以在短时间内为大规模的数据集生成有意义的描述，且无需额外监督。定量评估实验表明，CRM 在 COCO Caption 数据集上的表现可与一些先进模型相媲美，且效率更高。

关 键 词：图像字幕；CLIP；Transformer；门控线性单元；残差多头注意力

中图分类号：V221$^+$.3; TB553

文献标志码：A **文章编号**：1001-5965(2026)11-4019-10

图像字幕是一项利用自然语言描述视觉图像的任务。该任务不仅涉及检测和识别图像中的主要对象，还需要用自然语言表达其之间的关系，最终生成贴合图像内容的完整描述。图像字幕生成是一个跨学科、多模态的交叉研究方向，可以应用于许多场景，如自动驾驶、多模态检索、智能监控等，在信息时代具有广阔的应用前景。

目前已有许多成熟的图像字幕生成方法[1-7]，其主要遵循图像编码和文本解码的泛式。具体来说，这些方法通常使用图像编码器提取视觉特征，然后使用文本解码器生成相应的字幕。

然而，由于视觉特征和语义特征之间的差异，这些模型的训练通常需要大量的计算资源。其需要更大规模的数据集、更复杂的网络结构和更多的可训练参数。有时，为了获得更好的性能，还需要目标边界框和对象类别等额外监督标签。这些缺点限制了其在实际场景中的应用。因此，研究轻量级模型以解决图像字幕生成任务具有重要意义。

为了解决这个问题，本文采用强大的预训练模型构建了更轻量级的网络。具体而言，利用在大规模图像文本对数据集上预训练的 CLIP 编码器[8]，从输入图像中提取富含视觉和语义信息的特征，随后设计了残差门控 Transformer(residual gated Transformer, RGFormer)映射网络将其投影至语言空间，作为前缀引导语言模型生成图像字幕。该方法无需大量训练数据或额外监督标签，利用视觉和语义信息之间的强相关性，降低了模型的训练成本和数据需求。

本文在 COCO Caption[9-10] 数据集上进行实验，整个训练过程在一张 Nvidia GTX 3090Ti GPU 上耗时 4 h。实验结果表明，CRM(CLIP-RGFormer model)具备良好的泛化能力。定量评估结果显示，CRM

收稿日期：2024-09-05；**录用日期**：2025-02-14；**网络出版时间**：2025-03-19 11:16

网络出版地址：link.cnki.net/urlid/11.2625.V.20250318.1741.006

基金项目：国家自然科学基金 (52202417)；中国博士后科学基金 (2022TQ0155,2022M721605)；虚拟现实技术与系统全国重点实验室 (北京航空航天大学) 开放课题基金 (VRLAB2023A02)；中国科协青年科技人才托举工程 (2023QNRC001)；中央高校基本科研业务费专项资金 (NS2024030)；江苏省科协青年科技人才托举工程 (JSTJ-2023-XH032)

* **通信作者**. E-mail：chen_kai@nuaa.edu.cn

引用格式：金泽辉，陈凯，唐敦兵，等. 基于改进 Transformer 的视觉语言模型 [J]. 北京航空航天大学学报，2026，52（11）：4019-4028.
Jin Z H，Chen K，Tang D B，et al. Visual language model based on improved Transformer [J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52（11）：4019-4028 （in Chinese）.

在较短的训练时间内达到了与一些先进方法相媲美的性能指标。综上所述,本文的创新点如下:

1)提出一种基于CLIP和RGFormer的轻量级图像字幕生成模型CRM。该模型能够在较短的训练时间内为图像生成贴合内容的字幕。

2)提出一种基于残差多头注意力机制和门控线性单元的改进Transformer——RGFormer,在图像字幕生成任务中表现出优于原生Transformer的性能。

1 相关工作

1.1 图像编码器

图像字幕生成模型需要将输入图像编码为特征向量序列,再输入文本解码器生成单词序列。早期研究使用来自分类任务的预训练卷积神经网络(convolutional neural network, CNN)(如VGG16[11]和ResNet101[12])作为编码器来获得网格级图像特征。然而,Anderson等[1]指出,该方法未能有效确定图像中的关键区域。为了生成更符合人类视觉习惯的字幕,他们使用Faster R-CNN[13]检测图像中的突出部分并提取区域级特征,进而提出了自下而上和自上而下注意力(bottom-up and top-down attention, BUTD)图像字幕生成模型,在公开数据集上取得了更优异的性能。后续的许多研究沿用了BUTD的思想。例如,Zhou等[14]提出的视觉-语言预训练(vision-language pre-training, VLP)模型同样利用预训练的目标检测器检测关键物体并提取图像区域级特征。

然而,这种依赖目标检测网络的方法需要边界框和物体类别等额外的标签进行训练,且特征提取过程非常耗时,导致模型训练和推理时间变长,限制了其在实际应用中的效率和可扩展性。

本文使用CLIP作为图像编码器,取代了传统的目标检测网络。CLIP是一种基于文本和图像的多模态模型,在大规模图像文本对数据集上进行了无监督对比学习,旨在将图像和文本映射到同一特征空间。其核心原理是通过对比学习,最大化图像和文本的相似性,同时最小化不相关的配对。引入CLIP编码器后,无需依赖复杂的目标检测网络即可直接从图像中提取富含视觉和语义信息的特征。已有大量研究表明,CLIP在多种视觉任务中表现优异,包括图像分类[15-17]和目标检测[18-20]等。相比传统方法,CLIP通过广泛的预训练,在缺乏标注数据时仍能提供良好的性能,并且具有一定的零样本推理能力。

1.2 文本解码器

根据文本解码器的类型,现有的图像字幕生成模型可分为两类:基于卷积神经网络-长短期记忆(convolutional neural network-long short-term memory, CNN-LSTM)的模型[1,6-7,21-22]和基于CNN-Transformer的模型[3,5,23-24]。前者使用LSTM[25]作为解码器,能够有效捕捉长期依赖关系,并缓解网络反向传播过程中的梯度消失和爆炸问题。例如,BUTD将底层提取的图像特征输入LSTM解码,实现对图像内容的准确描述。后者使用Transformer[26]作为解码器,凭借其强大的全局表征能力,模型能够更有效地结合上下文信息,从而生成更合理的字幕。近年来,应用Transformer最有影响力的工作之一是BERT[27],其为后续研究大语言模型引入了基本框架:在大规模数据集上进行预训练,然后在较小的数据集上针对下游任务进行微调。Vilbert[28]使用基于BERT的预训练语言模型,结合交叉注意力机制,有效地融合并处理图像和文本2种模态的数据。VLP[14]通过将视觉特征转化为语义前缀,进一步引导BERT生成语义连贯的图像字幕。

然而,前述2种方法均存在一定的局限性。LSTM在训练效率和表达能力方面存在瓶颈。首先,LSTM依赖时间步计算,无法充分利用现代GPU的并行计算能力,导致训练速度较慢。其次,LSTM主要依赖局部信息进行预测,难以整合全局上下文信息,导致生成的描述在连贯性和整体性上有所欠缺。相比之下,Transformer通过自注意力机制实现了全局信息的有效整合,并且其高度并行化的计算方式显著提升了训练效率。然而,基于Transformer的BERT语言模型的微调过程需要耗费大量算力,且对样本规模和质量有较高的要求,难以保证微调后的模型性能。

本文使用GPT-2[29]作为文本解码器,以增强模型全局表达能力。GPT-2是一种基于Transformer架构的大规模预训练语言模型,核心原理是通过自注意力机制捕捉文本中的长距离依赖关系,并利用无监督学习从海量文本数据中学习语言表示。为解决语言模型微调的难题,本文提出以训练轻量级映射网络来替代传统的微调过程。建立视觉和文本共享空间[14,28,30]的研究与本文工作密切相关。VLP对整个模型进行预训练,以获取视觉和语言联合表征。而本文使用CLIP能够自然地获取这种联合表征,且仅需训练轻量级映射网络即可在下游任务中实现良好性能。

2 图像字幕生成方法介绍

给定一个图像文本对数据集$\{x^i,y^i\}_{i=1}^N$,N表示样本数量。图像字幕生成是为每张图像x^i生成最贴合

内容的字幕 y^i。每个字幕 y^i 都可以用一个 token 序列 $\boldsymbol{y}_1^i,\boldsymbol{y}_2^i,\cdots,\boldsymbol{y}_l^i$ 来表示，其中，l 是数据集中真实 token 序列的最大长度。训练过程中每个 token 序列都被填充到长度 l，训练目标如下：

$$\max_{\theta}\sum_{i=1}^{N}\ln p_\theta(\boldsymbol{y}_1^i,\boldsymbol{y}_2^i,\cdots,\boldsymbol{y}_l^i|x^i) \tag{1}$$

式中：θ 为训练参数，p 为生成真实 token 序列 $\boldsymbol{y}_1^i,\boldsymbol{y}_2^i,\cdots,\boldsymbol{y}_l^i$ 的概率。

本文将 CLIP 编码得到的嵌入转化为语义前缀，使用自回归语言模型逐个生成单词。具体来说，在已知输入 x^i 和前 $j-1$ 个 token 的情况下，最大化 $\boldsymbol{y}_j^i$ 的概率：

$$\max_{\theta}\sum_{i=1}^{N}\sum_{j=1}^{l}\ln p_\theta(\boldsymbol{y}_j^i|x^i,\boldsymbol{y}_1^i,\cdots,\boldsymbol{y}_{j-1}^i) \tag{2}$$

2.1　总体框架

受 Li 等[31]的启发，本文采用一种使用前缀引导语言模型生成图像字幕的方法，总体框架如图 1 所示。首先，将图像输入 CLIP 编码器，即 Vision Transformer[32] 网络中，获得一组嵌入。然后，将其与一组可学习常量嵌入拼接起来，并通过映射网络投影到语言空间生成字幕前缀。本文使用 Transformer 作为映射网络，并在此基础上提出了一种增强架构——RGFormer。每个 RGFormer 模块主要由一个残差多头注意力模块（residual multi-head attention，RMHA）和一个带有门控线性单元（gated linear unit，GLU）的前馈层组成。最后，将生成的字幕前缀与真实 token 嵌入进行拼接，并输入 GPT-2 中生成图像字幕。在训练期间，保持编码器和解码器冻结，只对映射网络进行训练，从而实现模型轻量化。

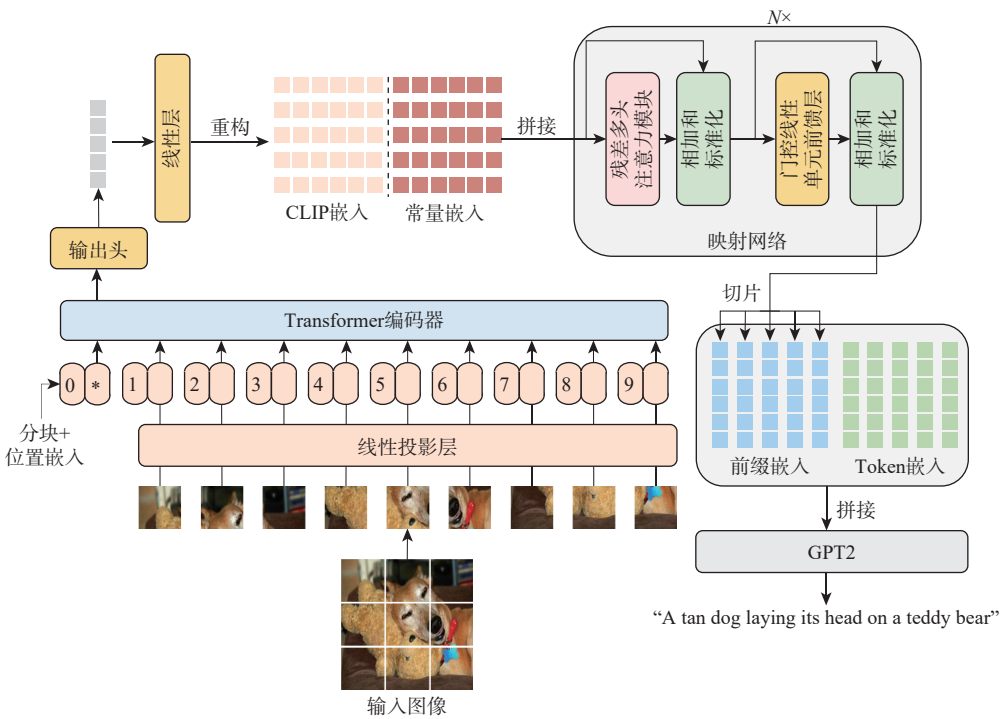

图 1　CRM 网络总体框架

Fig. 1　Overall structure of CRM

本文使用 CLIP 编码器从图像 x^i 中提取特征，得到一组嵌入向量 $\boldsymbol{e}_1^i,\boldsymbol{e}_2^i,\cdots,\boldsymbol{e}_k^i$。随后，将这组嵌入与一组可学习常量嵌入 $\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_k$ 进行拼接，并传递给轻量级映射网络 F，以获得前缀嵌入 $\boldsymbol{p}_1^i,\boldsymbol{p}_2^i,\cdots,\boldsymbol{p}_k^i$：

$$\boldsymbol{p}_1^i,\boldsymbol{p}_2^i,\cdots,\boldsymbol{p}_k^i=F(\mathrm{Concat}(\boldsymbol{e}_1^i,\boldsymbol{e}_2^i,\cdots,\boldsymbol{e}_k^i,\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_k)) \tag{3}$$

式中：k 表示前缀长度。本文使用 GPT-2 分词器对真实文本标签进行分词，获得一组 token 序列 $\boldsymbol{y}_1^i,\boldsymbol{y}_2^i,\cdots,\boldsymbol{y}_l^i$，随后将这些 token 编码为词嵌入向量：

$$\boldsymbol{t}_1^i,\boldsymbol{t}_2^i,\cdots,\boldsymbol{t}_l^i=\mathrm{TextEncoder}(\boldsymbol{y}_1^i,\boldsymbol{y}_2^i,\cdots,\boldsymbol{y}_l^i) \tag{4}$$

式中：TextEncoder 表示 GPT 文本编码器。每个 token 嵌入 $\boldsymbol{t}_x^i(x\in[1,l])$ 的维度与前缀嵌入 $\boldsymbol{p}_x^i(x\in[1,k])$ 相同，本文将其排列在一起，并输入 GPT-2 中。给定前缀，训练目标是使用自回归方法预测下一个 token。本文保持 CLIP 和 GPT-2 冻结，并使用简单有效的交叉熵损失来训练映射网络 F：

$$\mathcal{L}_x=-\sum_{i=1}^{N}\sum_{j=1}^{l}\ln p_\theta(\boldsymbol{t}_j^i|\boldsymbol{p}_1^i,\boldsymbol{p}_2^i,\cdots,\boldsymbol{p}_k^i,\boldsymbol{t}_1^i,\boldsymbol{t}_2^i,\cdots,\boldsymbol{t}_{j-1}^i) \tag{5}$$

2.2　图像编码器结构

CRM 编码器主要包括 3 个模块，如图 2(a)所示。首先，线性投影层将输入图像分割成一组块，并

将每个块展平为向量。在每个向量上添加位置编码后，将其输入 Transformer 编码器中。Transformer 编码器由多个堆叠的基础模块组成，如图 2(b)所示。每个模块主要包含一个多头注意力层和一个前馈层，其中，Lx 表示有 2 个模块堆叠。多头注意力机制使模型能够评估不同输入部分的重要性，前馈网络应用线性组合和非线性激活来增强模型的表达能力。最后，本文将输出头的结果通过线性变换重构为一组一维向量，每个向量的维度与 GPT-2 的词嵌入向量维度相匹配，以满足输入要求。

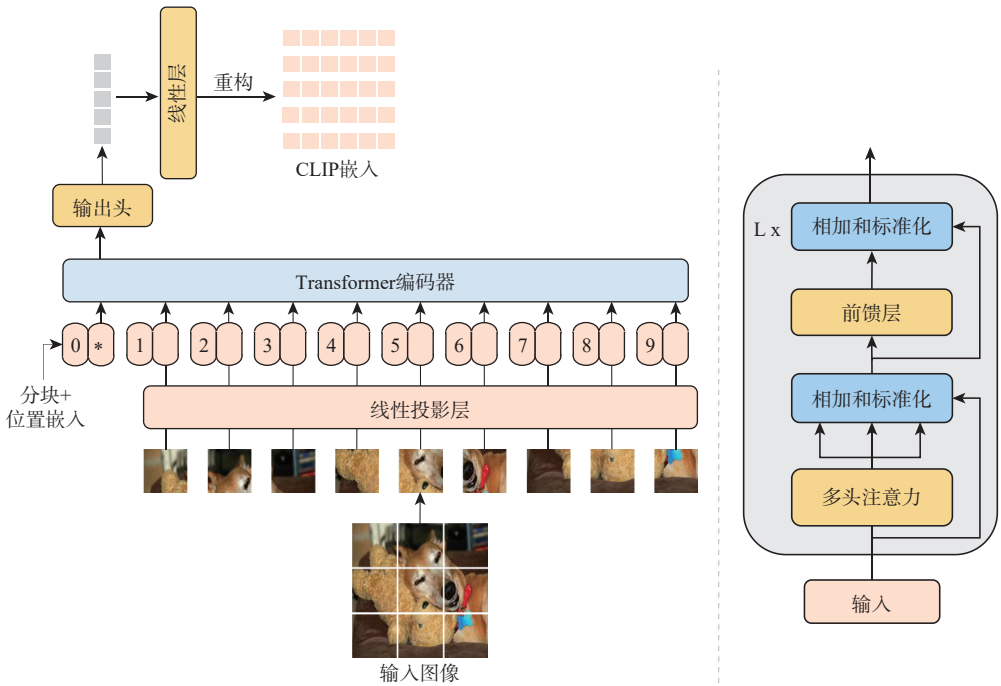

图 2 CRM 编码器和 Transformer 编码器

Fig. 2 Image encoder and Transformer encoder

CRM 编码器的前向传播过程如式(6)～式(10)所示。x_p^n表示图像块，$\boldsymbol{E}$表示线性投影层的权重矩阵，$\boldsymbol{E}_{\text{pos}}$表示位置编码，$z_l$表示 Transformer 编码器的输出，$l$表示堆叠深度，$\boldsymbol{y}$表示输出头结果。MHA表示多头注意力，MLP表示前馈层，LN表示线性变换，Reshape表示张量重塑。编码器的最终输出结果是一组嵌入向量$\boldsymbol{cp}_1^i,\boldsymbol{cp}_2^i,\cdots,\boldsymbol{cp}_k^i$。

$$\boldsymbol{z}_0=[x_{\text{prefix}};x_p^1\boldsymbol{E};x_p^2\boldsymbol{E};\cdots;x_p^N\boldsymbol{E}]+\boldsymbol{E}_{\text{pos}} \tag{6}$$

$$\boldsymbol{z}'_l=\text{MHA}(\boldsymbol{z}'_{l-1})+\boldsymbol{z}'_{l-1} \tag{7}$$

$$\boldsymbol{z}_l=\text{MLP}(\boldsymbol{z}'_l)+\boldsymbol{z}'_l \tag{8}$$

$$\boldsymbol{y}=\text{LN}(\boldsymbol{z}_l) \tag{9}$$

$$\boldsymbol{cp}_1^i,\boldsymbol{cp}_2^i,\cdots,\boldsymbol{cp}_k^i=\text{Reshape}(\text{LN}(\boldsymbol{y})) \tag{10}$$

2.3 映射网络结构

映射网络是本研究的重点，是实现模型跨模态能力的关键组件。该网络构建了视觉-语言共享空间，将提取的图像特征投影并转化为语义前缀，实现了模态间的转换。为了实现模型轻量化，同时避免微调语言模型，本文保持 CLIP 和 GPT-2 参数冻结，并采用 Transformer 作为映射网络。

Transformer 在文本序列任务上表现出色，但仍存在一定的局限性。Transformer 依靠注意力机制关注上下文信息，但堆叠的各个基础模块之间的关系相对独立。具体而言，每个模块计算得到的注意力分数仅受前一模块输出的影响，而更早期的原始注意力分数未得到有效保留。这种设计限制了模块之间的信息流动，可能导致重要信息的丢失。此外，由多个基础模块堆叠而成的深度神经网络可能存在梯度消失和爆炸问题，导致训练不稳定和收敛困难。

2.3.1 RGFormer 整体结构

为了解决上述问题，本文首先提出了 RMHA，用于替代 Transformer 原始注意力模块。RMHA 维护了一条注意力分数传播路径，以保留更多的原始信息。此外，将前馈层中的第 1 个线性层替换为 GLU，利用门控机制缓解梯度消失问题。基于这 2 个模块，本文提出了一种改进的 Transformer 架构——RGFormer，如图 3 所示。

2.3.2 残差多头注意力

多头注意力机制将注意力分散到各个子空间中，使模型可以学习多维特征信息。最后，将各子

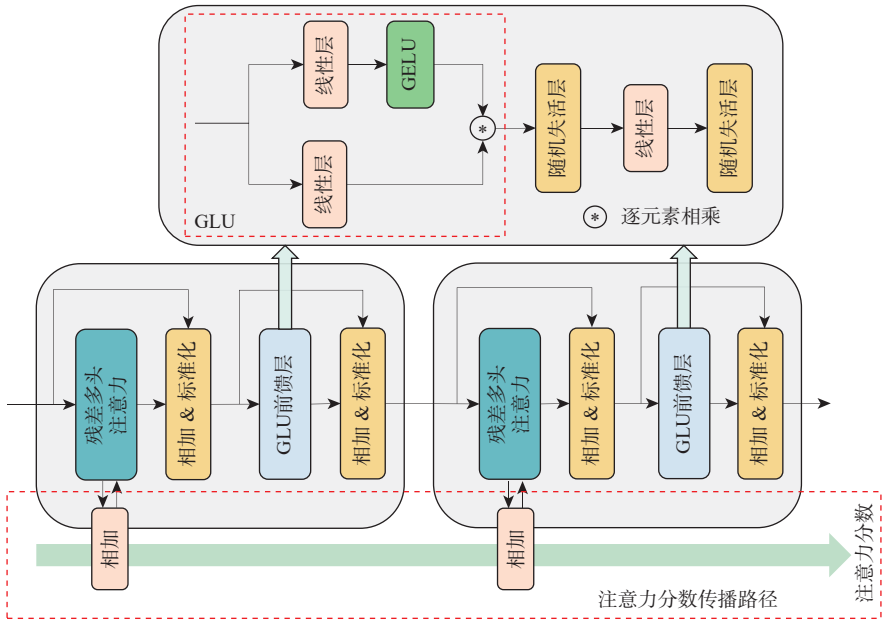

图 3　RGFormer 模块结构

Fig. 3　Structure of RGFormer module

空间的输出进行合并，具体计算过程如下：

$$\begin{cases} \boldsymbol{Q}_i = \boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}}, \boldsymbol{K}_i = \boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}}, \boldsymbol{V}_i = \boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}} \\ h_i = \text{Attention}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) \\ \text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(h_1, h_2, \cdots, h_n)\boldsymbol{W}^{\text{O}} \end{cases} \tag{11}$$

式中：$\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 为维度为 d_k 的矩阵；$\boldsymbol{V}$ 为维度为 d_v 的矩阵；i表示子空间的索引；$\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$、$\boldsymbol{W}_i^{\boldsymbol{V}}$分别为将 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 投影到子空间的矩阵；$\boldsymbol{W}^{\text{O}}$为对注意力头输出进行合并的权重矩阵。注意力函数采用使用缩放点积来实现，如下所示：

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{12}$$

本文提出了一种改进的注意力模块 RMHA，旨在保留更多的原始特征和缓解梯度消失问题。RMHA 的核心思想是创建一条“直接”路径来传播注意力分数。具体来说，每个 RMHA 模块将传播注意力分数作为残差项添加到当前层的注意力分数矩阵中，并计算得到最终的注意力权重，如图 4 所示，其中，Block$N$ 表示堆叠的第 N 个模块。RMHA 具体计算过程如下所示：

$$\begin{cases} \text{RMHA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \text{Prev}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}} + s\right)\boldsymbol{V} \\ s = s + \frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}} \end{cases} \tag{13}$$

式中：s表示传播注意力分数。s通过迭代更新，使每个 RMHA 模块能够利用全局所有注意力分数。此外，RMHA 的残差结构能够有效缓解梯度消失问题，从而提升模型训练的稳定性。

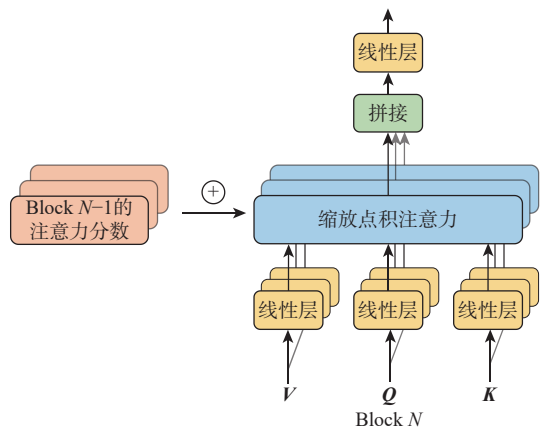

图 4　RMHA 模块结构

Fig. 4　Structure of RMHA module

2.3.3　门控线性单元前馈层

原始 Transformer 前馈层包含一个隐藏线性层，本文使用 GLU 进行替代。GLU 的核心思想是引入门控机制来增强网络的表达能力，其前向传播过程如下：

$$\text{GLU}(x, W_1, W_2, b_1, b_2) = \sigma(xW_1 + b_1) \otimes (xW_2 + b_2) \tag{14}$$

式中：σ表示激活函数；$\otimes$表示逐元素乘法；W和b表示权重和偏置。

本文使用 GELU 函数，相比于常用的 ReLU 等激活函数，GELU 更为连续和平滑，有助于梯度的稳定传播。GLU 由 2 个线性层“并联”组成，其中仅一个线性层包含激活函数。在前向传播过程中，输入向量进入 2 个分支，得到线性和非线性结果，随后逐元素相乘。GLU 通过门控机制选择性地激活输入特征，这使模型能够更有效地捕捉输入数据

中的关键信息，同时忽略不相关的部分。此外，GLU同时保留了线性和非线性路径，在梯度的反向传播过程中有利于保持梯度的稳定，从而缓解梯度消失问题。

2.4 CRM推理过程

GPT-2使用Softmax来确定当前时间步生成词汇表中每个单词的概率。基于此，本文使用集束搜索算法，根据概率在每个时间步中选择若干候选词，并维持一个固定大小的候选集。具体而言，设定集束宽度为k，在每个时间步，算法根据当前候选序列的累积概率扩展新的候选词，并筛选出概率最高的k个序列继续迭代。最终，选择具有最高累积概率的词序列作为生成的图像字幕。图5展示了集束宽度为2的字幕生成过程，蓝色标注表示当前时间步累积概率最高的候选字符。集束搜索算法不仅提升了字幕的连贯性和语义一致性，还能够有效避免生成过程中的局部最优问题。

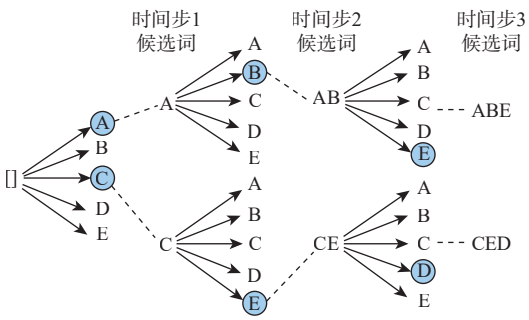

图5 集束搜索算法

Fig. 5 Beam search algorithm

3 实验结果

3.1 数据集和对比模型

本文在COCO Caption数据集上进行实验，并按照Karpathy等[33]的方法对数据集进行划分。训练集包含120 000幅图像，每幅图像对应5个标题。验证集和测试集各由5 000幅图像组成。

本文将CRM与BUTD、VLP、VL-T5[34]、VLKD[35]和Oscar[3]等方法进行比较。其中，BUTD、VLP和Oscar均使用目标检测器提取图像区域级特征，并输入解码器以生成文本序列。BUTD结合顶层注意力机制聚焦于检测到的物体，以生成细节丰富的字幕。VLP的训练方法类似于BERT，在数百万个图像文本对上进行了预训练以获取视觉-语言联合表征。Oscar在预训练过程中利用检测到的对象标签作为锚点，将视觉特征与文本中的相应语义对齐，促进多模态信息的融合。VL-T5使用基于Transformer的T5[36]模型，提出了一种多任务学习框架。通过预训练的联合编码器理解图片和文本之间的复杂关系，然后在下游任务上进行微调。VLKD将CLIP文本编码器替换为预训练的BART[37]，并通过知识蒸馏的方式让BART学习CLIP的跨模态知识，将文本特征映射到共享空间中。

3.2 评估指标和实验结果

3.2.1 评估指标

本文使用图像字幕生成任务中的常见指标，包括SPICE[21]、BLEU@4[38]、METEOR[39]和CIDEr[40]，其中，SPICE为语义命题型图像字幕评价指标，BLEU@4为4元BLEU，METEOR为显序翻译评估指标，CIDEr为基于共识的图像描述评价指标。在COCO Caption数据集上验证CRM的性能。此外，本文研究的目标是设计一种轻量化的图像字幕生成方法。因此，本文对比了CRM和基线模型的可学习参数规模和训练时间，以验证其轻量化效果。

3.2.2 定量评估

表1提供了CRM与各对比模型在COCO Caption数据集上的定量评估结果，表中加粗项表示对应列指标的最优结果。实验结果表明，CRM在SIPCE指标上仅次于VL-T5和Oscar，这是因为SPICE指标侧重于评估生成字幕在语义层次上的相似性。CRM使用预训练的CLIP编码器提取图像特征，并结合增强的RGFormer映射网络实现跨模态信息转化，强调图像和文本之间的语义对齐，生成的描述具有较强的语义一致性和全局关联性。相比之下，BUTD和VLP等模型可能更注重局部细节，容易忽略全局语义表达。在BLEU@4、METEOR和CIDEr指标上，CRM略低于部分模型。一方面是因为其他模型具有更多的可学习参数，对复杂特征的表达能力更强；另一方面上述指

表1 COCO Caption数据集定量评估结果

Table 1 Quantitative evaluation on COCO caption dataset

模型	BLEU@4	METEOR	CIDEr	SPICE	参数规模	训练时间/h	设备型号
BUTD[1]	36.2	27.0	113.5	20.3	5.20×10^7	960	M40
VLP[14]	36.5	28.4	117.7	21.3	1.15×10^8	48	V100
VL-T5[34]	34.5	28.7	116.5	21.9	2.70×10^8		
VLKD[35]	34.1	27.5	114.3	21.0			
Oscar[3]	**36.58**	**30.4**	**124.12**	**23.17**	1.35×10^8	74	V100
CRM	33.61	27.85	114.60	21.67	**4.60×10^7**	**4**	3090 Ti

标更注重词汇匹配及语法结构的准确性。而CRM更强调轻量化和在不同场景中的泛化能力，生成的字幕通常具有更高的自由度和灵活性，可能不完全符合某些特定的语法结构。总的来说，CRM以更小的参数规模和更短的训练时间，达到了媲美主流方法的性能，在保证模型性能的前提下实现了轻量化。

3.2.3　定性评估

图6展示了CRM为从COCO Caption测试集中随机选择的图像生成的未经筛选的字幕。从结果来看，这些字幕都较准确地描述了图像的内容，值得注意的是，CRM能够识别一些不常见的物体。例如，CRM识别出了第4张图片的“double decker bus”，Ground Truth中并没有相应的描述，这归因于CLIP图像编码器在多样化大规模数据集上进行了预训练，具备一定的零样本推理能力。此外，CRM展现出一定的联想能力。在第1张图片中，通过识别“toilet”关联到“in a bathroom”，使字幕内容更为丰富。然而，结果表明模型也存在一定的不足，偶尔无法识别图像中的某些次要物体，如第1张图中的“glasses”或“hat”。相比之下，VLP在这方面的表现更优秀，因为VLP使用了Faster R-CNN对象检测器，并结合对象标签，从图像中提取区域级特征，更擅长捕捉图中的对象。

原图					
Ground Truth	Man in glasses and a cap sitting on the toilet.	A big cat laying down in a chair on a porch.	A small dog is curled up on top of the shoes.	Two red busses are traversing next to a tall brick building.	Three cows walking along the street in a town.
BUTD	A smiling man sitting on a toilet.	A cat siting in a blue chair.	A dog sleeping with some shoes.	Red buses driving down a street.	Two cows walking along the street.
VLP	A man in a hat sitting on a toilet.	A cat sitting on top of a chair.	A dog sleeping in the shoes.	Two red busses are driving on a street.	A few cows walking down the street in a town.
CRM	A man sitting on a toilet in a bathroom.	A cat is siting outside on a chair.	A dog laying on top of the shoes.	A red double decker bus driving down a street.	Black and white photo-graph of cows walking down a street.

图6　COCO Caption测试集中随机选择图像的未筛选结果

Fig. 6　Unfiltered results of randomly selected images from COCO Caption test set

为了更直观地展示RMHA模块的效果，本文选择图6中的第2和第3张图片进行注意力可视化，如图7所示，其中，蓝色到红色的渐变表示权重逐渐升高。实验结果表明，RMHA能够准确地聚焦于图像中的关键区域，如主体“cat”和“dog”，从而有效促进信息流动并增强模型的表达能力。

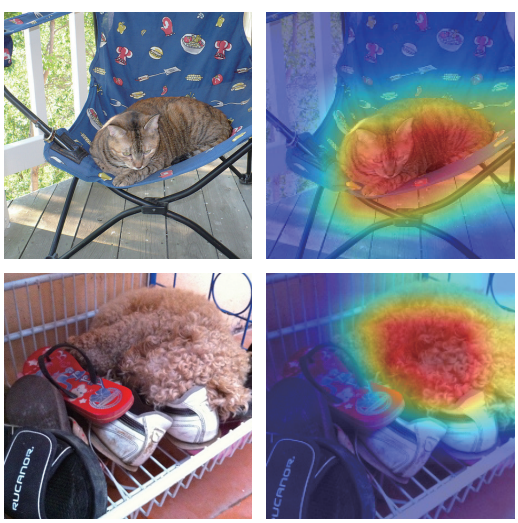

图7　注意力可视化结果

Fig. 7　Attention visualization results

定性实验结果表明，CRM在COCO Caption测试集上表现出了良好的性能。为了进一步评估模型的泛化能力，本文从网络上随机收集了一些图像并用CRM生成字幕，如图8所示。可以观察到，面对COCO Caption数据集之外的任意网络图片，CRM仍能够生成准确且有意义的字幕，表现出了优秀的泛化能力。

3.2.4　前缀长度实验

Li等[31]指出，增加前缀长度可以提升模型在下游任务中的表现。在前缀长度达到特定值后，模型性能将趋于稳定，达到的饱和值因下游任务而异。对于图像字幕生成任务，本文研究了不同前缀长度对模型性能的影响，在COCO Caption数据集上对具有不同前缀长度的CRM分别进行了4轮训练，并使用BLEU@4和CIDEr分数在训练集和测试集上进行了定量评估。

如图9所示，当前缀长度过小时，模型无法有效表达完整信息，性能较差。当前缀长度增加到约40时，模型在训练集和测试集上的表现趋于稳定，



BUTD: A group of people are riding skateboards
VLP: A group of kids riding skateboards on a wooden ramp
CRM: A group of young skateboarders pose for a picture



BUTD: There are some penguins and a man in a hat
VLP: A group of penguins are standing in the snow
CRM: A man standing in the snow with penguins around



BUTD: A street with many stores and a bicycle
VLP: A person walking down a street at night
CRM: A person walking down a narrow alley at night



BUTD: A man in a black hat swinging a bat
VLP: A baseball player swinging a bat at a ball
CRM: A baseball player swinging a bat during a game

图 8 CRM 为网络图片生成的字幕

Fig. 8 Captions for online pictures generated by CRM

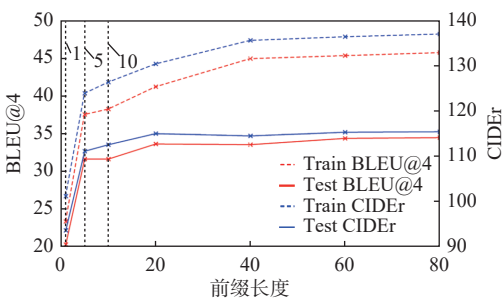

图 9 CRM 在不同前缀长度下的性能

Fig. 9 Performance of CRM under different prefix lengths

并且进一步增加前缀长度无法显著提升性能，反而会造成计算资源的浪费。考虑到 Transformer 在计算注意力权重时会产生较大的内存开销，本文在实验中将最大前缀长度设置为 80。

3.2.5 消融实验

本文进行了消融实验来验证 RGFormer 模块的有效性，结果如表 2 所示，表中，加粗项表示对应指标的最优结果。以原始 Transformer 作为映射网络的模型称为 CTM（CLIP transformer model）。实验结果表明，RMHA 模块和 GLU 模块均有助于提升 CTM 的性能，且 RMHA 模块的提升效果更明显。这是因为注意力机制是 Transformer 架构最核心的部分，直接决定了模型对输入数据的理解和处理能力。RMHA 能够有效地保留并传递重要的注意力信息。相比之下，前馈层更多作用在特征转换和非线性表示上。引入 GLU 能够提升非线性表达能力，但对模型整体性能的影响相对有限。当将 RMHA 和 GLU 组合成 RGFormer 作为映射网络时，模型表现出优于单独使用任一模块的性能指标。

表 2 RGFormer 消融实验结果

Table 2 Results of ablation experiment of RGFormer

模型	BLEU@4	METEOR	CIDEr	SPICE
CTM	33.53	27.45	113.08	21.05
CTM+RMHA	33.59	27.73	114.27	21.56
CTM+GLU	33.50	27.62	113.49	21.23
CTM+RGFormer(CRM)	**33.61**	**27.85**	**114.60**	**21.67**

4 结 论

1）CRM 在图像字幕生成任务中以更轻量化的结构取得了良好的性能，在 COCO Caption 测试集上的 BLEU@4、METEOR、CIDEr、SPICE 指标分数分别为 33.61、27.85、114.60、21.67，可与 BUTD 和 VLP 等先进方法相媲美，且参数和训练时间大大减少；

2）RGFormer 相较于原生 Transformer 提升了性能，其中，RMHA 模块通过残差结构连接前后注意力模块，保留了更多的原始注意力信息，GLU 前馈层通过门控机制选择性激活输入特征，增强了模型的表达能力。

3）脱离了使用目标检测器提取图像区域级特征的传统形式。CRM 无需额外的监督标签或进行耗时的目标检测过程，实现了轻量化。

4）CRM 使用 RGFormer 为映射网络，有效地将 CLIP 提取的图像特征映射到文本空间中，解决了图像字幕生成任务跨模态的难点。

5）CRM 聚焦于更轻量化的结构，以更低的训练成本实现了良好的性能。但由于舍弃了目标检测器，在一些图像中可能无法识别到次要物体，存在一定局限性。

参考文献（References）

[1] Anderson P, He X D, Buehler C, et al. Bottom-up and top-down at-

tention for image captioning and visual question answering [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway Palo Alto: IEEE Press, 2018: 6077-6086.

[2] Chen X L, Zitnick C L. Learning a recurrent visual representation for image caption generation[EB/OL]. (2014-11-20)[2024-12-18]. https://doi.org/10.48550/arXiv.1411.5654.

[3] Li X J, Yin X, Li C Y, et al. Oscar: object-semantics aligned pre-training for vision-language tasks[C]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 121-137.

[4] Stefanini M, Cornia M, Baraldi L, et al. From show to tell: a survey on deep learning-based image captioning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 539-559.

[5] Tan H, Bansal M. LXMERT: learning cross-modality encoder representations from transformers[EB/OL]. (2019-12-03)[2024-12-20]. https://doi.org/10.48550/arXiv.1908.07490.

[6] Xu K, Ba J L, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention[C]//International Conference on Machine Learning. New York: ACM, 2015: 2048-2057.

[7] Yao T, Pan Y W, Li Y H, et al. Exploring visual relationship for image captioning[C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 711-727.

[8] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. Virtual Event: PMLR, 2021: 8748-8763.

[9] Chen X L, Fang H, Lin T Y, et al. Microsoft COCO captions: data collection and evaluation server[EB/OL]. (2015-04-03)[2024-12-20]. https://doi.org/10.48550/arXiv.1504.00325.

[10] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[C]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2024-12-20]. https://arxiv.org/abs/1409.1556.

[12] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[13] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[14] Zhou L, Palangi H, Zhang L, et al. Unified vision-language pre-training for image captioning and VQA[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence Palo Alto: AAAI Press, 2020: 13041-13049.

[15] Conde M V, Turgutlu K. CLIP-art: contrastive pre-training for fine-grained art classification[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2021: 3951-3955.

[16] Zhang R R, Zhang W, Fang R Y, et al. Tip-adapter: training-free adaption of CLIP for few-shot classification[C]//Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022: 493-510.

[17] Liu H D, Xu S Y, Fu J M, et al. CMA-CLIP: cross-modality attention CLIP for image-text classification[EB/OL]. (2021-12-09)[2024-12-21]. https://arxiv.org/abs/2112.03562.

[18] Vidit V, Engilberge M, Salzmann M. CLIP the gap: a single domain generalization approach for object detection[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 3219-3229.

[19] Teng Z, Duan Y N, Liu Y, et al. Global to local: clip-LSTM-based object detection from remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5603113.

[20] Gu X Y, Lin T Y, Kuo W C, et al. Open-vocabulary object detection via vision and language knowledge distillation[EB/OL]. (2022-05-12)[2024-12-21]. https://doi.org/10.48550/arXiv.2104.13921.

[21] Anderson P, Fernando B, Johnson M, et al. SPICE: semantic propositional image caption evaluation[C]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 382-398.

[22] Yong C, Fei H, Lian Z, et al. A hierarchical multimodal attention-based neural network for image captioning[J]. ACM SIGIR Forum, 2017, 51(CD): 889-892.

[23] Ji J Y, Luo Y P, Sun X S, et al. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021, 35(2): 1655-1663.

[24] Luo Y P, Ji J Y, Sun X S, et al. Dual-level collaborative transformer for image captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021, 35(3): 2270-2277.

[25] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Long Beach: Curran Associates, Inc. , 2017, 30: 5998-6008.

[27] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24)[2024-12-24]. https://doi.org/10.48550/arXiv.1810.04805.

[28] Lu J, Batra D, Parikh D, et al. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J]. Advances in neural information processing systems, 2019, 32.

[29] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[30] Zhang P C, Li X J, Hu X W, et al. VinVL: revisiting visual representations in vision-language models[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 5575-5584.

[31] Li X L, Liang P. Prefix-tuning: optimizing continuous prompts for generation[EB/OL]. (2021-01-01)[2024-12-25]. https://doi.org/10.48550/arXiv.2101.00190.

[32] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. V2. (2021-06-03)[2024-12-25]. https://doi.org/10.48550/arXiv.2010.11929.

[33] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 3128-3137.

[34] Cho J, Lei J, Tan H, et al. Unifying vision-and-language tasks via text generation[C]//International Conference on Machine Learning. Virtual Event: PMLR, 2021: 1931-1942.

[35] Sun L, Li Y W, Dong Y M. Learning from expert: vision-language knowledge distillation for unsupervised cross-modal hashing retrieval[C]//Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. New York: ACM, 2023: 499-507.

[36] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21(140): 1-67.

[37] Lewis M, Liu Y H, Goyal N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[EB/OL]. (2019-10-29)[2024-12-25]. https://doi.org/10.48550/arXiv.1910.13461.

[38] 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: Association for Computational Linguistics, 2002: 311-318.

[39] Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language[C]//Proceedings of the Ninth Workshop on Statistical Machine Translation. Philadelphia: Association for Computational Linguistics, 2014: 376-380.

[40] Vedantam R, Zitnick C L, Parikh D. CIDEr: consensus-based image description evaluation[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 4566-4575.

Visual language model based on improved Transformer

JIN Zehui[1], CHEN Kai[1,*], TANG Dunbing[1], FANG Guoyu[1], WEN Feixia[2]

(1. College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China;
2. Nanjing Research Institute of Electronic Engineering, Nanjing 210007, China)

Abstract: Natural language processing and computer vision are combined in the cross-modal job of image captioning. The model is required to generate an appropriate caption for the given image. To address this challenge, we proposed residual gated Transformer (RGFormer) as an enhancement of the Transformer architecture. The model, based on CLIP and RGFormer, CRM, is then proposed for image captioning. CRM utilizes the encoder of CLIP to extract image features as a prefix to the caption. RGFormer, a lightweight mapping network, projects the prefix into language space before feeding it into GPT-2 to produce captions. CLIP was trained on an extensive dataset comprising image-text pairs, which contains rich visual and semantic information and is exceptionally well-suited for vision-language tasks. The core idea of CRM is to reduce the disparity between visual and textual representations by using RGFormer to accomplish the cross-modal task. CRM could generate meaningful captions for large-scale datasets in a short training time without additional annotations or pre-training. Quantitative evaluation experiments show that CRM achieves results comparable to some advanced models on the COCO Caption dataset more efficiently.

Keywords: image caption; CLIP; Transformer; gated linear unit; residual multi-head attention

Received: 2024-09-05; **Accepted**: 2025-02-14; **Published Online**: 2025-03-19 11:16

URL: link.cnki.net/urlid/11.2625.V.20250318.1741.006

Foundation items: National Natural Science Foundation of China (52202417); China Postdoctoral Science Foundation (2022TQ0155,2022M721605); Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (VRLAB2023A02), Young Elite Scientists Sponsorship Program by CAST (2023QNRC001); The Fundamental Research Funds for the Central Universities (NS2024030); Young Elite Scientists Sponsorship Program by JSTJ (JSTJ-2023-XH032)

* **Corresponding author**. E-mail: chen_kai@nuaa.edu.cn