文章编号: 1000-5641(2020)05-0056-12

通过细粒度的语义特征与 Transformer 丰富图像描述

王俊豪, 罗轶凤

(华东师范大学 数据科学与工程学院, 上海 200062)

摘要: 传统的图像描述模型通常基于使用卷积神经网络 (Convolutional Neural Network, CNN) 和循环神经网络 (Recurrent Neural Network, RNN) 的编码器-解码器结构, 面临着遗失大量图像细节信息以及训练时间成本过高的问题. 提出了一个新颖的模型, 该模型包含紧凑的双线性编码器 (Compact Bilinear Encoder) 和紧凑的多模态解码器 (Compact Multi-modal Decoder), 可通过细粒度的区域目标实体特征来改善图像描述. 在编码器中, 紧凑的双线性池化 (Compact Bilinear Pooling, CBP) 用于编码细粒度的语义图像区域特征, 该模块使用多层 Transformer 编码图像全局语义特征, 并将所有编码的特征通过门结构融合在一起, 作为图像的整体编码特征. 在解码器中, 从细粒度的区域目标实体特征和目标实体类别特征中提取多模态特征, 并将其与整体编码后的特征融合用于解码语义信息生成描述. 该模型在 Microsoft COCO 公开数据集上进行了广泛的实验, 实验结果显示, 与现有的模型相比, 该模型取得了更好的图像描述效果.

关键词: 图像描述; 精细化特征; 多模态特征; Transformer

中图分类号: TP399 **文献标志码**: A **DOI**: 10.3969/j.issn.1000-5641.202091004

Enriching image descriptions by fusing fine-grained semantic features with a transformer

WANG Junhao, LUO Yifeng

(*School of Data Science and Engineering, East China Normal University, Shanghai* 200062, *China*)

Abstract: Modern image captioning models following the encoder-decoder architecture of a convolutional neural network (CNN) or recurrent neural network (RNN) face the issue of dismissing a large amount of detailed information contained in images and the high cost of training time. In this paper, we propose a novel model, consisting of a compact bilinear encoder and a compact multi-modal decoder, to improve image captioning with fine-grained regional object features. In the encoder, compact bilinear pooling (CBP) is used to encode fine-grained semantic features from an image's regional features and transformers are used to encode global semantic features from an image's global bottom-up features; the collective encoded features are subsequently fused using a gate structure to form the overall encoded features of the image. In the decoding process, we extract multi-modal features from fine grained regional object features, and fuse them with overall encoded features to decode semantic information for description generation. Extensive experiments performed on the public Microsoft COCO dataset show that our model achieves state-of-the-art image captioning performance.

Keywords: image captioning; fine-grained features; multi-modal features; transformer

收稿日期: 2020-08-04

基金项目: 国家重点研发计划 (2018YFC0831904)

通信作者: 罗轶凤, 男, 副教授, 硕士生导师, 研究方向为文本数据挖掘与知识图谱. E-mail: yifluo@dase.ecnu.edu.cn

0 引　言

图像描述的主要任务是为图像生成自然语言描述, 并利用所生成的描述帮助应用程序理解图像视觉场景中表达的语义. 例如, 图像描述可以将图像检索转换为文本检索, 用于对图像进行分类并改善图像检索结果. 人们通常只需快速浏览一下即可描述图像视觉场景的细节, 但对计算机而言, 自动为图像添加描述则是一项全面而艰巨的任务, 需要将图像中包含的复杂信息转换为自然语言描述. 与普通的计算机视觉任务相比, 图像描述不仅需要从图像中识别目标实体, 还需要将识别出的目标实体与自然语义相关联以使用自然语言进行描述. 因此, 图像描述需要人们提取图像的深层特征, 与语义特征关联并进行转换以生成描述.

早期图像描述方法大多基于传统机器学习, 倾向于从图像中提取目标实体和属性, 然后将获得的目标实体和属性填充到预定义的句子模板中. 随着深度学习的普及, 近些年的图像描述方法主要遵循编码器-解码器体系结构[1], 这种体系结构通常将 CNN 作为特征提取的编码器, 将 RNN 作为生成描述的解码器. 编码器-解码器体系结构可以生成超出预定义模板的描述语句, 大大提高了所生成语句的多样性.

传统的编码器-解码器图像描述模型通常基于图像中提取的全局特征来生成图像描述. 后续的研究或将注意机制与编码器-解码器体系结构结合在一起, 或从全局特征中提取感兴趣区域特征以关注图像感兴趣区域, 用于提高图像描述的质量, 但是这些现有研究仍然在自然语言生成过程中损失了图像视觉场景中的大量详细信息. 因而, 具有注意力机制的编码器-解码器模型面临以下两个挑战: ①当图像中包含复杂目标实体和属性时, 从全局图像特征图中提取的区域特征不能很好地表示图像实体的语义; ②由于 RNN 的固有顺序性质, 基于 RNN 的模型难以执行并行优化计算, 导致模型训练的时间成本过高.

本文提出了一种新颖的编码器-解码器模型, 通过提取检测到的目标实体的细粒度区域特征丰富图像特征, 利用 Transformer[2] 对图像中包含的语义信息进行编码和解码, 以生成图像描述, 从而提高图像描述的效果. 具体来说, 使用预先训练的深度残差网络[3](Deep Residual Network, ResNet) 来提取图像中检测到的目标实体的区域特征图; 然后, 在编码器中使用 Compact Bilinear Pooling 对图像的区域特征图中的细粒度语义特征进行编码, 使用多层 Transformer encoder 对图像的语义特征编码, 将两部分特征进行融合, 并将所有层次的编码特征与门结构融合, 形成图像整体编码特征. 在解码器中, 从细粒度的区域目标实体图像特征和区域目标实体类别特征中提取多模态特征, 并将它们与编码器输出特征融合, 以解码语义信息从而进行描述生成. 图 1 展示了添加细粒度特征后的描述效果: LSTM-C(Long Short-Term Memory-C) 模型生成的描述效果; 本文模型 (Ours) 生成的描述效果; GT(Ground Truth) 为数据集中真实标注的描述.

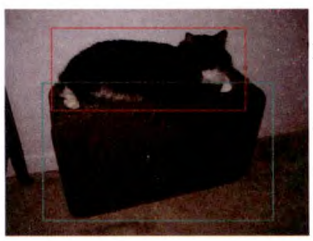

图 1　模型描述生成效果图

Fig. 1　An example of description generation using our model

综上所述, 本文有以下的贡献.

(1) 采用提取精细化的区域目标实体特征的方法以生成具有精细语义的图像描述. 本文是第一篇

提出细粒度图像描述的概念, 并且使用紧凑的双线性池化 (CBP) 模型融合编码特征以进行图像描述的文章.

(2) 引入 Transformer 对语义特征进行编码和解码, 以改善图像描述.

(3) 在 Microsoft COCO 2014 数据集上进行实验, 以评估模型在图像描述任务上的性能. 实验结果表明, 本文的模型达到了优异的效果. 本文还将通过实验进一步验证模型中引入的各种因素的影响.

1 图像描述相关工作

1.1 图像描述

图像描述的任务是给定一张图像, 自动生成其对应的自然语言描述, 随着越来越多的应用需要理解图像的语义, 人们对图像描述进行了大量的研究. 图像描述模型可以分为 3 类: 基于模板填充的模型[4-6]、基于检索的模型[7-9] 和生成式模型[10-12].

基于模板填充的模型的流程为: 首先人为地定义一系列句子模板; 然后提取图像特征以检测图像中的目标实体和属性; 最后用检测到的目标实体和属性填充预先定义的模板以生成描述. 使用基于模板填充的模型生成的句子通常可以清楚地描述图像中包含的实体和属性, 但是缺乏句子和单词的多样性.

基于检索的模型需要维护包含大量图像描述的句子集, 该方法通过比较图像与包含在维护的句子集中的描述句子之间的相似度来获得候选句子集, 选择最相似的句子作为图像的最终描述. 基于检索的模型无法生成新图像的描述, 也无法确保生成的描述的准确性.

生成式模型主要应用深度学习模型进行描述生成. Google 提出了神经图像描述生成器 (Neural Image Caption Generator, NIC)[13], 引入了在机器翻译中被广泛采用的编码器-解码器结构进行图像描述. NIC 模型使用基于 CNN 的 InceptionNet[14] 来提取图像特征, 并使用 RNN 网络作为解码器来解码 CNN 中提取的图像特征, 其中 RNN 网络也可以用长短期记忆网络[15](LSTM) 或门控循环单元[16](Gate Recurrent Unit, GRU) 代替. 斯坦福大学在同一时期提出了神经对话系统[17], 采用 VGGnet[18] 网络作为编码器提取图像特征, RNN 作为解码器.

1.2 注意力机制

注意力机制已被广泛应用于图像描述中. Xu 等[19] 最先将注意力机制应用于图像描述, 他们在解码器中利用硬注意力和软注意力这两种机制来融合从 CNN 层提取的特征. 硬注意力机制将最关注的区域权重设置为 1, 其他权重设置为 0, 以使模型仅专注于一个区域, 而软注意力机制为每个图像区域学习在 0 到 1 之间的权重, 这些权重的总和等于 1, 可以让模型从所有区域中提取加权特征. 软注意力已成为用于图像描述的标准注意力机制. Anderson 等[20] 提出了 bottom-up 和 top-down 两种注意力机制用于图像描述, 使用 Faster R-CNN[21] 检测图像中的感兴趣区域 (Region of Interests, RoIs), 并从 RoIs 中提取图像特征作为 bottom-up 注意力特征, 再使用两个 LSTM 层分别作为 top-down 注意力层和语言模型来生成图像描述. 这两种注意力机制使模型在生成图像描述时能够专注于图像的实体周围区域并忽略一些次要信息.

1.3 双线性池化

Lin 等[22] 提出使用双线性池化 (Bilinear Pooling, BP) 进行细粒度分类, 对于使用 CNN 模型提取的图像特征, 可以使用 BP 将两种特征双线性融合, 以获得更深的隐藏信息从而进行细粒度分类. 由于 BP 中包含大量计算, 后续的研究工作集中在如何减少 BP 中的计算. CBP[23] 通过 Random Maclaurin (RM) 和 Tensor Sketch (TS) 的方法来降低 BP 的维数. Low-rank Bilinear Pooling[24] 提出使用低秩分

解近似分解矩阵以减少计算量并应用于细粒度分类. Grassmann Pooling[25] 通过对矩阵执行奇异值 (Singular Value Decomposition, SVD) 分解, 并使用由前 k 个左奇异向量组成的矩阵逼近原始矩阵, 从而减少计算量.

2　模型实现

2.1　模型概览

常见的基于编码器-解码器体系的图像描述模型利用 CNN 提取深度特征以对图像中包含的隐藏语义进行编码, 然后利用 RNN 从编码的深度特征中解码语义并生成图像描述. 人们倾向于从图像中识别多个感兴趣的区域 (RoIs), 提取 RoIs 区域的图像特征, 并将这些特征送入图像描述模型以生成描述. 在实现中, 可以引入诸如 Faster R-CNN 之类的预训练目标检测模型来检测 RoIs, 并且可以使用诸如 ResNet 之类的预训练模型来提取所识别的 RoIs 的区域图像特征. 由于图像可能包含多个实体, 并且图像中包含的实体产生的 RoIs 对于生成的描述可能有不同的贡献度, 因此, 引入注意力机制可以使图像描述模型在生成包含不同实体的描述时将注意力集中在相关的 RoIs 上.

即使添加了注意力机制, 现有图像描述模型依然无法生成具有足够准确细节的描述, 因为现有 RoIs 特征主要从全局图像特征中提取; 而提取全局图像特征的过程中会有大量的细节信息丢失, 会使提取的区域图像特征缺乏细粒度的语义信息. 由此, 本文提出了一种新颖的编码器-解码器模型, 通过显式提取细粒度的实体特征来丰富图像的详细语义信息, 从而提高图像描述的质量. 本文提出的图像描述模型由 M 个 Compact Bilinear Encoder 和 N 个 Compact Multi-modal Decoder 组成, 模型的总体架构如图 2 所示. 在 Compact Bilinear Encoder 中, Compact Bilinear Pooling 从实体区域特征中提取实体的细粒度特征, 然后将这些细粒度特征与 bottom-up 特征融合在一起, 以对图像的语义信息进行编码. 在 Compact Multi-modal Decoder 中, 在解码语义信息并生成图像描述时会利用多模态信息来进一步增强图像的语义. 本章的剩余部分说明实体区域特征和 bottom-up 特征的提取方法, 以及 Compact Bilinear Encoder 和 Compact Multi-modal Decoder 的实现过程.

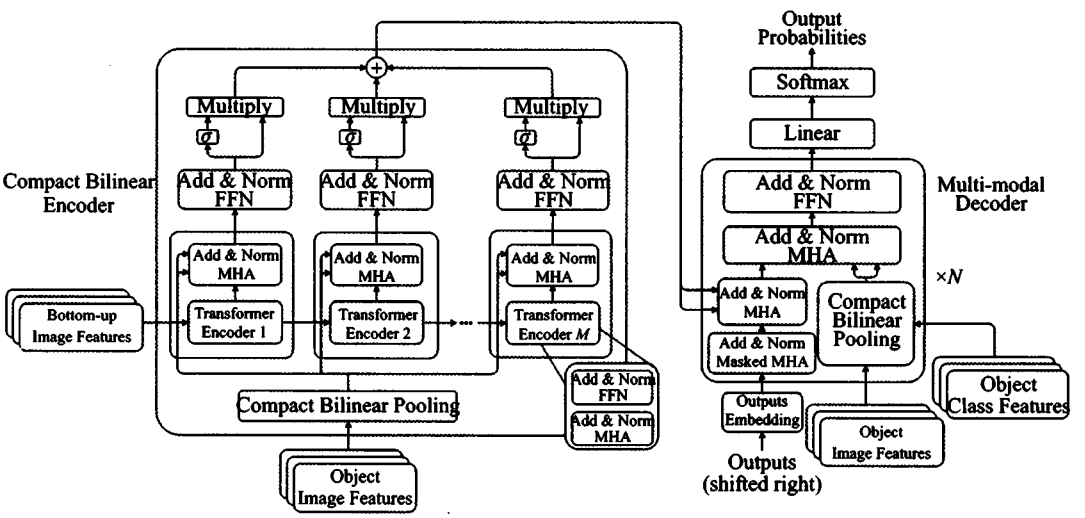

图 2　模型结构概览图

Fig. 2　Model structure overview diagram

2.2　特征提取

为提取图像的实体区域特征, 本文使用经典的两阶段目标检测模型 Faster R-CNN 来检测包含的

实体,使用边界框标注检测到的实体并判断出实体类别,通过预训练的 ResNet 模型提取实体图像特征并进行最大池化,使用词向量表征类别特征,每个实体最终获得 D 维的图像特征和 N 维的类别信息特征. 然后,我们将实体的图像特征和文本类别特征视为区域实体特征. 对于一张检测出了1个实体的图像,它的区域实体特征将被表示为 $F_{\mathrm{o}}=\{f_{1,\mathrm{o}},\cdots,f_{i,\mathrm{o}},\cdots,f_{l,\mathrm{o}}\}$ 和 $F_{\mathrm{c}}=\{f_{1,\mathrm{c}},\cdots,f_{i,\mathrm{c}},\cdots,f_{l,\mathrm{c}}\}$,其中 $f_{i,\mathrm{o}}\in\mathbb{R}^{D}$ 表示第 i 个 D 维实体图像特征, $f_{i,\mathrm{c}}\in\mathbb{R}^{N}$ 表示第 i 个 N 维实体类别特征.

与图像全局特征图相比,bottom-up 特征指的是图像中包含的实体周边相关区域的 RoIs 特征,利用 bottom-up 特征可以使图像描述模型在生成描述时更加关注检测到的目标实体周边的区域,去除一些无关的部分图像特征. 本文遵循 Anderson 等提出的 bottom-up 注意力机制来提取图像的 bottom-up 特征,其中使用 Faster R-CNN 识别实体相关区域,然后将其映射到模型全局特征图上抽取特征,再应用平均池化来生成最终特征. 假定 bottom-up 特征由 k 个 RoIs 组成,则可以将 bottom-up 特征表示成 $F_{\mathrm{b}}=\{f_{1,\mathrm{b}},\cdots,f_{i,\mathrm{b}},\cdots,f_{k,\mathrm{b}}\}$,其中 $f_{i,\mathrm{b}}\in\mathbb{R}^{D_f}$ 表示第 i 个 D_f 维 bottom-up 特征.

2.3 Compact Bilinear 编码器

Compact Bilinear 编码器主要包括 Compact Bilinear Pooling(CBP) 层、多个常规的 Transformer 编码器层、Multi-head Attention 层和前馈神经网络层. 本文利用 Transformer 编码器层对于 bottom-up 特征进行不同级别的语义信息编码,并利用 Compact Bilinear Pooling 层对实体图像特征的细粒度语义信息进行编码;然后,通过 Multi-head Attention 层将不同级别的全局语义特征与细粒度的语义特征融合在一起;最后通过前馈神经网络层生成整体编码器输出.

(1) Compact Bilinear Pooling

Compact Bilinear Pooling 可以融合两个特征,并利用两个向量的外积提取其隐藏的二阶信息. 为了避免计算两个向量外积的成本过高, Compact Bilinear Pooling 使用 Count-Sketch 映射函数[26] 将两个向量投影到较低维空间,然后通过快速傅立叶变换 (Fast Fourier Transform, FFT) 来计算两个向量的外积. 给定两个特征向量 $x_1\in\mathbb{R}^{d_1}$ 和 $x_2\in\mathbb{R}^{d_2}$,它们的融合特征 $x_{\mathrm{out}}\in\mathbb{R}^{d_{\mathrm{out}}}$ 可以通过 $x_{\mathrm{out}}=\mathrm{FFT}^{-1}(\mathrm{FFT}(C_1(x_1))\circ\mathrm{FFT}(C_2(x_2)))$[27] 计算得出,其中 d_1 和 d_2 是输入特征的维度, d_{out} 是输出维度, C_1 和 C_2 是两个 Count-Sketch 映射函数, $\circ$ 是两个向量的点积操作. 详细的 CBP 计算过程由算法 1 给出. 算法 1 中,对输入向量的每个分量进行映射,将分量映射的位置信息存储于 h 中,并给予每个分量一个映射相关系数,存储于 s 中.

算法 1 Compact Bilinear Pooling

输入: $x_1\in\mathbb{R}^{d_1},x_2\in\mathbb{R}^{d_2}$

输出: $x_{\mathrm{out}}\in\mathbb{R}^{d_{\mathrm{out}}}$

for k in [1, 2] **do**
  for $i=1$ to d_k **do**
    $h_k[i]=rand.sample(\{1,2,\cdots,d_{\mathrm{out}}\},1)$
    $s_k[i]=rand.sample(\{-1,+1\},1)$
  end for
  for j=1 to d_{out} **do**
    $O_k[j]=\sum_{i:h_k[i]=j}s_k[i]x_k[i]$
  end for
end for

10: $x_{\text{out}} = \text{FFT}^{-1}\left(\text{FFT}\left(o_1\right) \circ \text{FFT}\left(o_2\right)\right)$, $\circ$ 是两个向量的点积操作

11: **return** x_{out}

对于之前抽取出的实体图像特征 F_{o}, 使用 CBP 层进行二阶特征提取, 获得细粒度特征 Z_{o},公式为

$$Z_{\text{o}} = \text{CBP}(F_{\text{o}}W^{F_{\text{o}}}, F_{\text{o}}W^{F_{\text{o}}}), \tag{1}$$

其中, $Z_{\text{o}} \in \mathbb{R}^{l \times d_{\text{out}}}$, l是检测出的目标实体个数, d_{out}是细粒度图像特征的维度. $W^{F_{\text{o}}} \in \mathbb{R}^{D \times d_{\text{in}}}$是一个参数矩阵, 其中 d_{in}是 CBP 层输入的特征维度.

(2) 全局特征编码层

在本文模型中, Compact Bilinear 编码器中的全局特征编码器由多个全局特征编码层组成, 每一个全局特征编码层包括一个多头注意力 (Multi-head Attention, MHA) 子层和一个全连接的前馈网络 (Feed-Forward Network, FFN) 子层. 将注意力机制进行点积放缩, 即

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d}}\right)V, \tag{2}$$

其中, Q、K和 V分别是 query、 key 和 value 矩阵, $\sqrt{d}$是放缩因子. 多头注意力通过线性映射将 Q、K和 V分为 h个部分, 对每个划分的部分进行按比例缩放的点积注意力运算, 最后结合线性映射的各个部分来生成输出, 过程为

$$\text{MHA}\left(Q, K, V\right) = \text{Concat}(\text{head}_1, \cdots, \text{head}_n)W_{\text{o}}, \tag{3}$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{4}$$

其中, $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$, d_{model}是 MHA 中特征的维度, $d_k = d_{\text{model}}/h$, h是注意力子空间的个数. FFN 子层包含 2 个线性变换和 1 个 ReLU 激活函数, 即

$$\text{FFN}\left(x\right) = \max\left(0, xW_1 + b_1\right)W_2 + b_2. \tag{5}$$

对于图像的 bottom-up 特征 F_{b}, 本文使用全局特征编码层进行特征抽取, 其中 Q、K和 V都是 bottom-up 特征. 每一个 MHA 子层和 FFN 子层都由一个残差 (Residual) 子层和一个标准化 (Normalization, Norm) 子层[28] 连接, 最终的一个全局编码层的输出 Z_{b} 可以定义为

$$\tilde{Z} = \text{Norm}\left(F_{\text{b}} + \text{MHA}\left(F_{\text{b}}, F_{\text{b}}, F_{\text{b}}\right)\right), \tag{6}$$

$$Z_{\text{b}} = \text{Norm}(\tilde{Z} + \text{FFN}(\tilde{Z})), \tag{7}$$

其中 Norm 是一个标准化操作.

(3) 编码层输出

对于图像, 在生成 Compact Bilinear 编码器的整体输出之前, 将细粒度图像特征 Z_{o} (从 CBP 层提取) 与经过 MHA 的每个全局特征编码层的输出 $Z_{i,\text{b}}$融合为

$$\tilde{Z}_i = \text{Norm}\left(Z_{i,\text{b}}W^z + \text{MHA}\left(Z_{i,\text{b}}W^z, Z_{\text{o}}W^{z_{1,\text{o}}}, Z_{\text{o}}W^{z_{2,\text{o}}}\right)\right), \tag{8}$$

$$Z_i = \text{Norm}(\tilde{Z}_i + FFN(\tilde{Z}_i)), \tag{9}$$

其中, $W^z \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$, $W^{z_{1,\text{o}}}, W^{z_{2,\text{o}}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$是线性映射的参数矩阵. 最后, 结合上述不同层次的融合编码特征, 采用门结构来控制每个编码器子层可以提供的信息量, 对应公式为

$$Z_{\text{sum}} = \sum_{i=1}^{M} \text{gate}_i\left(Z_i\right) Z_i, \tag{10}$$

$$\text{gate}_i\left(Z_i\right) = \sigma\left(W^Z Z_i + b^Z\right), \tag{11}$$

其中, M 是编码层的个数, σ 是 sigmoid 激活函数.

2.4 Compact Multi-modal 解码器

Compact Multi-modal 解码器由多个解码器层组成. 它从编码特征以及图像的细粒度多模态特征中提取隐藏语义信息以生成描述语句. 图像的细粒度多模态特征是通过 Compact Multi-modal Bilinear Pooling(CMB) 将实体图像特征和文本类别特征融合在一起形成的, 公式为

$$Z_{\text{CMB}} = \text{CBP}\left(F_{\text{o}} W^{F_{\text{o}}}, F_{\text{c}} W^{F_{\text{c}}}\right), \tag{12}$$

其中, $W^{F_{\text{o}}} \in \mathbb{R}^{D \times d_{\text{in}}}$, $W^{F_{\text{c}}} \in \mathbb{R}^{N \times d_{\text{in}}}$ 是实体图像特征和实体类别特征的映射参数矩阵.

将需要生成的语句序列按词向量进行编码, 并添加其位置向量作为序列输入特征 F_{s}, 使用一个带掩码的 MHA 层 (在进行 Attention 权重计算时使用一个上三角为 1, 其余为 0 的矩阵进行掩码, 防止模型采用当前时序之后的单词进行训练) 对它进行编码得到带权重的序列特征 Z_{s}, 经过多个解码层后, 使用一个 MHA 层和一个 FFN 层将编码输出与最后一层的序列特征融合, 经过一个 softmax 层生成最后的预测概率分布, 相应公式为

$$Z_{\text{s}} = \text{Norm}(F_{\text{s}} + \text{Masked\_MHA}(F_{\text{s}}, F_{\text{s}}, F_{\text{s}})), \tag{13}$$

$$Z_{\text{d}} = \text{Norm}(Z_{\text{s}} + \text{MHA}(Z_{\text{s}}, Z_{\text{sum}}, Z_{\text{sum}})), \tag{14}$$

$$Z_{\text{m}} = \text{Norm}(Z_{\text{d}} + \text{MHA}(Z_{\text{d}}, Z_{\text{CMB}}, Z_{\text{CMB}})), \tag{15}$$

$$\tilde{Z}_{\text{out}} = \text{Norm}(Z_{\text{m}} + \text{FFN}(Z_{\text{m}})), \tag{16}$$

$$Z_{\text{out}} = \text{Softmax}(\tilde{Z}_{\text{out}} W^{\tilde{Z}_{\text{out}}} + b), \tag{17}$$

其中, Z_{sum} 是多层编码器的融合输出, $W^{\tilde{Z}_{\text{out}}} \in \mathbb{R}^{d_{\text{model}} \times C}$, 这里 C 是数据集中所有词组成的词表的大小.

2.5 模型优化方法

正如大多数图像描述模型所做的那样, 首先使用交叉熵损失 (Cross Entropy Loss, CE) 训练模型, 然后使用 self-critical(SCST)[29] 方法再对其进行训练. 给定一个真实数据集的描述单词序列 $Y^* = (y_1^*, y_2^*, \cdots, y_T^*)$, 其中 T 表示单词序列的长度, 最小化交叉熵损失为

$$L_{\text{CE}}(\theta) = -\sum_{t=1}^{T} \log\left(P_\theta(y_t^* | y_1^*, \cdots, y_{t-1}^*)\right), \tag{18}$$

其中, θ 表示模型参数. 用交叉熵损失训练模型后, 按照 SCST 进行强化学习训练模型. 将模型视为一个“agent”, 而文字和图片被视为“environment”. 公式 (18) 中表示为 P_θ 的“agent”策略是本文 Compact Bilinear 编码器和 Compact Multi-modal 解码器的参数. 当生成图像描述时, 生成的序列在生成下一个单词单元时被视为当前时间的“state”. agent 每次采取行动生成新单词时, 都会根据所采取的 state 和 action 来返回 CIDEr-D 分数的奖励. 使用 SCST 训练模型的目标是最大程度地减少负期望. 相应公式为

$$L_{\text{SCST}}(\theta) = -E_{Y_{\text{s}} \sim p_\theta}\left[r\left(Y_{\text{s}}\right)\right], \tag{19}$$

其中, $Y_{\mathrm{s}}=(y_{1,\mathrm{s}},y_{2,\mathrm{s}},\cdots,y_{t,\mathrm{s}},\cdots,y_{T,\mathrm{s}})$的$y_{t,\mathrm{s}}$是在$t$时刻从模型采样的单词, r是 CIDEr-D 分数函数. 公式 (19) 所表示的损失可以进一步近似为

$$\nabla_{\theta}L_{\mathrm{SCST}}(\theta)\approx-\left(r(Y_{\mathrm{s}})-r\left(\hat{Y}\right)\right)\nabla_{\theta}\log p_{\theta}(Y_{\mathrm{s}}),\tag{20}$$

其中, $Y_{\mathrm{s}}=(y_{1,\mathrm{s}},y_{2,\mathrm{s}},\cdots,y_{T,\mathrm{s}})$是利用 Monte-Carlo 的思想根据$p_{\theta}$进行采样得到的, $\hat{Y}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_T)$是根据模型贪心解码得到的。

3　实验评估

3.1　实验设置

本文在公开的 Microsoft COCO 2014 数据集[30]上进行了广泛的实验以评估本文模型的性能. 数据集中每张图像拥有 5 句英语描述标签. 使用 Karpathy 提供的方法[17]将验证集分为 3 个部分: 5 000 张图像用于验证; 5 000 张图像用于测试; 剩余的 30 504 张图像以及训练集中的 82 783 张图像用于模型训练. 将所有描述句子中的字符转换为小写字母后拆分为带有空格的单词, 并丢弃数据集中少于 5 次的单词.

对于模型中的超参数, 本文将 bottom-up 特征的维度D_f设置为 2 048, 1 张图片的 RoIs 数量k固定为 50, 实体图像特征维度 D 设置为 2 048, 类别名称使用 BERT[31]模型抽取, 维度N设置为 1 024, 1 张图片检测出的目标实体数量固定为 5, CBP 层的输入输出维度d_{in}和d_{out}分别设置成 1 024 和 8 192, MHA 层中的特征维度d_{model}设置成 1 024, MHA 的h设置成 8, 每一个 head 的特征维度d_k为 128, 全局特征编码器和解码器的个数固定为 3.

在两步的训练过程中, 首先根据文献 [2] 中提供的学习率策略设置交叉熵损失的学习率, 然后在强化学习阶段将学习率固定为4×10^{-7}. 本文设置 batch 的大小为 64, 使用 Adam 优化器进行优化. 为了评价模型质量, 本文选择了 BLEU[32](B-1、B-4)、METEOR[33](M)、ROUGE-L[34](R) 和 CIDEr-D[35]分数 (C) 作为评价指标, 并以百分比值作为结果.

3.2　消融实验

本文进行了一系列的消融实验来证明本文模型 Ml-Transformer 中的众多改进对模型有增进效果. 首先将传统 Transformer 作为基线模型与增加了门限结构的多维度 Transformer 进行对比; 接着将未添加 CBP 层、仅添加已抽取出的实体图像特征与类别特征的模型, 与添加了 CBP 层抽取细粒度特征的模型加入对比. 表 1 显示了对比结果, 其中 B-1、B-4、M、R 和 C分别表示模型在经过交叉熵损失训练后达到的 BLEU、METEOR、ROUGE-L 和 CIDEr-D 分数. 从表 1 可以看到, 在添加门限结构融合多层特征以及添加细粒度特征后, 本文模型的指标均有所提升, 其中添加了门限结构和 CBP 层的比未添加的在 CIDEr-D 得分上分别提高了 1.5 和 1.4 个百分点.

表 1　添加门限结构和 CBP 的效果

Tab. 1　The effect of introducing gated feature fusion and compact bilinear pooling

添加类型	对比模型	B-1/%	B-4/%	M/%	R/%	C/%
门限结构	Transformer(未添加)	74.6	35.1	26.6	55.3	112.3
	Ml-Transformer(仅添加门限单元)	76.0	36.0	27.8	56.4	113.8
CBP层	Ml-Transformer(仅添加特征)	76.7	36.3	27.5	56.6	115.1
	Ml-Transformer(添加特征与CBP层)	77.0	36.9	28.0	57.3	116.5

本文验证了利用不同特征提取器提取目标实体特征的效果,表 2 显示了实验结果,其中单视图(single-view)和多视图(multi-view)分别表示使用同一个提取器和两个不同特征提取器提取初始图像实体特征的情况,R-101、X-101 和 D-121 分别表示使用 ResNet-101、ResNeXt-101 和 DenseNet-121 提取特征. 当使用两个不同特征提取器时,将提取的特征通过 Compact Bilinear Pooling 融合以形成初始实体特征. 可以看到,使用通过 ResNext 模型和 DenseNet 模型提取的 single-view 特征,可以使模型获得更高的 BLEU 得分,但是会略微降低 METEOR、ROUGE 和 CIDEr-D 分数,使用 multi-view 特征会使模型获得比使用 single-view 特征更高的 BLEU 得分,但是降低了 CIDEr-D 得分. 相对关注 CIDEr-D 得分,本文使用 single-view 的 ResNet 模型作为特征提取器.

表 2 不同特征抽取器抽取效果

Tab. 2 The effect of extracting initial object features with various extractors

视图	特征抽取器	B-1/%	B-4/%	M/%	R/%	C/%
单视图	R-101	77.0	36.9	28.0	57.3	116.5
	X-101	77.2	36.6	27.8	57.0	115.9
	D-121	77.1	37.0	27.6	57.0	116.2
多视图	R-101和X-101	77.5	37.2	27.4	56.8	115.5
	X-101和D-121	77.4	37.2	27.5	56.7	115.4

本文还验证了不同的固定的识别出图像实体的数量 m 对于模型的效果影响,表 3 显示了实验结果. 当 1 张图片识别出的实体数量少于 m 时,使用 0 向量填充;当超过 5 时,直接截取概率最大的不同类别的前 5 个实体. 可以看到模型参数 m 在到达 5 之前,实体数量越多,实体信息越丰富,模型效果越好. 但是由于大部分图片仅包含 3–5 个实体,太大的参数 m 会导致模型效果下降. 因此本文设置超参数 m 为 5.

表 3 不同超参数 m 效果

Tab. 3 The effect of various settings for the hyper-parameter m

参数m	B-1/%	B-4/%	M/%	R/%	C/%
1	76.3	36.2	27.5	56.2	115.4
3	76.5	36.3	27.6	56.4	115.8
5	77.0	36.9	28.0	57.3	116.5
7	76.9	36.7	27.9	57.3	116.3

3.3 模型整体表现

将本文模型的整体表现与以下几个基线模型进行比较.

(1)SCST,该模型将传统的基于 Attention 的 Image Caption 模型与强化学习结合.

(2)LSTM-A[36],该模型使用多实例学习[37]检测图像实体属性,将实体属性添加到 LSTM 解码中进行描述.

(3)Up-Down,该模型通过 Faster-RCNN 抽取图像 bottom-up 特征,使用两个 LSTM 分别作为 top-down attention 和语言模型进行描述.

(4)RFNet[38],该模型融合从不同的 CNN 中抽取的特征进行图像描述.

(5)GCN-LETM[39], 该模型使用图卷积神经网络 (Graph Convolutional Neural Network, RCNN) 抽取图像中两个实体的关系, 并将抽取到的关系作为附加特征进行图像描述.

(6)SGAE[40], 该模型使用自编码的场景图进行图像描述.

表 4 显示了上述模型的最终效果, 其中每个模型都是先进行交叉熵损失训练再进行强化学习训练. 从表 4 中可以看到, 在进行交叉熵损失训练后, 本文模型在除 BLEU-1 之外的所有评估指标上均取得了最佳性能. 虽然 GCN-LSTM 的 BLEU-1 得分略高于本文模型, 而在 CIDEr-D 优化之后, 本文模型在所有评估指标上的表现均优于所有基线模型, 其中 B-1、B-4、METEOR 和 ROUGEL 的得分均以明显的优势领先于在基线模型上的得分, 并且 CIDEr-D 的得分明显高于所有的基线模型. 与最近的两个基线模型 (GCN-LSTM 和 SGAE) 相比, 本文模型的 CIDEr-D 得分提高了 1.5 和 1.3 个百分点; 与其他 4 个基准模型相比, CIDEr-D 得分提高了 7 个百分点以上. 本文模型的性能之所以领先于其他基准模型, 是因为本文模型提取了图像中包含目标实体更多的细粒度特征, 丰富了其语义表达. 因此使用本文模型生成的描述更接近训练数据集中的真实描述.

表 4　整体模型效果对比

Tab. 4　Overall image caption performance using all models

模型	交叉熵损失					CIDEr-D优化				
	B-1/%	B-4/%	*M*/%	*R*/%	*C*/%	B-1/%	B-4/%	*M*/%	*R*/%	*C*/%
SCST		30.0	25.9	53.4	99.4	-	34.2	26.7	55.7	114.0
LSTM-A	75.4	35.2	26.9	55.8	108.8	78.6	35.5	27.3	56.8	118.3
Up-Down	77.2	36.2	27.0	56.4	113.5	79.8	36.3	27.7	56.9	120.1
RFNet	76.4	35.8	27.4	56.8	112.5	79.1	36.5	27.7	57.3	121.9
GCN-LSTM	77.3	36.8	27.9	57.0	116.3	80.5	38.2	28.5	58.3	127.6
SGAE						80.8	38.4	28.4	58.6	127.8
Ml-Transformer(本文模型)	77.0	36.9	28.0	57.3	116.5	80.9	39.0	28.9	58.5	129.1

最后测试了使用 Transformer 进行编码解码的训练效率, 结果见表 5 所示, 其中, 1 个 epoch 为全部训练集训练一轮. 从表 5 可以看出, 相较于使用基于传统 Attention 机制、CNN 作为编码器、RNN 作为解码器的 SCST 模型和使用预先抽取 bottom-up 特征、两层 LSTM 作为解码器的 Up-Down 模型, 本文模型在训练速度和收敛速度上都有较大提升. 本文模型训练 1 个 epoch 的时间较短, 训练需要的 epoch 数 (epoches) 也更少.

表 5　不同编码器/解码器训练效率对比

Tab. 5　Model training efficiencies for various kinds of encoders/decoders

模型	1个epoch/h	epoches
SCST	1.2	40
Up-Down	0.85	40
Ml-Transformer(本文模型)	0.85	15

4　总　　结

当前的图像描述模型主要采用编码器-解码器结构, 采用 CNN 编码器提取深层图像特征, 并使用

RNN 解码器基于提取的特征生成图像描述. 这些模型会丢失图像中包含的大量详细信息, 模型的训练时间成本也会急剧增大. 本文提出了一种新颖的模型, 通过使用 Compact Bilinear Pooling 显式提取检测目标实体的细粒度区域特征, 并利用融合多层信息的 Transformer 对图像中包含的语义信息进行编码和解码, 从而改善图像描述. 此外, 我们还通过融合实体图像特征和文本类别特征来获得多模态信息以增强语义解码. 本文在公开的 Microsoft COCO 数据集上进行了广泛的实验, 实验结果表明, 本文的模型表现出了更优异的性能, 因此, 证明了利用细粒度的目标实体特征可以有效地丰富细节信息以生成图像描述.

[参 考 文 献]

[1] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 652-663.

[2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 6000–6010.

[3] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 770-778.

[4] KULKARNI G, PREMRAJ V, ORDONEZ V, et al. Babytalk: Understanding and generating simple image descriptions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903.

[5] MITCHELL M, HAN X F, DODGE J, et al. Midge: Generating image descriptions from computer vision detections [C]//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. ACM, 2012: 747-756.

[6] YANG Y Z, TEO C L, DAUMÉ H, et al. Corpus-guided sentence generation of natural images [C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. ACM, 2011: 444-454.

[7] DEVLIN J, CHENG H, FANG H, et al. Language models for image captioning: The quirks and what works [EB/OL].(2015-10-14)[2020-06-30]. https://arxiv.org/pdf/1505.01809.pdf.

[8] FARHADI A, HEJRATI M, SADEGHI M A, et al. Every picture tells a story: Generating sentences from images [C]//Computer Vision – ECCV 2010, Lecture Notes in Computer Science, vol 6314. Berlin: Springer, 2010: 15-29.

[9] KARPATHY A, JOULIN A, L F F. Deep fragment embeddings for bidirectional image sentence mapping [C]//Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. Cambridge, MA: MIT Press, 2014: 1889-1897.

[10] MAO J H, XU W, YANG Y, et al. Explain images with multimodal recurrent neural networks [EB/OL]. (2014-10-04)[2020-06-30]. https://arxiv.org/pdf/1410.1090.pdf.

[11] LU J S, XIONG C M, PARIKH D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 3242-3250.

[12] YAO T, PAN Y W, LI Y H, et al. Exploring visual relationship for image captioning [C]//Computer Vision – ECCV 2018, Lecture Notes in Computer Science, vol 11218. Cham: Springer, 2018: 711-727.

[13] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015: 3156-3164.

[14] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015: 1-9.

[15] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[16] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL].(2014-09-03)[2020-06-30]. https://arxiv.org/pdf/1406.1078.pdf.

[17] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664-676.

[18] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10)[2020-6-30]. https://arxiv.org/pdf/1409.1556.pdf.

[19] XU K, BA J L, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention [EB/OL].(2016-04-19)[2020-6-30]. https://arxiv.org/pdf/1502.03044.pdf.

[20] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 6077-6086.

[21] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [C]//Advances in Neural Information Processing Systems 28 (NIPS 2015).[S.l.]: Curran Associates, Inc., 2015: 91-99.

[22] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNN models for fine-grained visual recognition [C]//2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2015: 1449-1457.

[23] GAO Y, BEIJBOM O, ZHANG N, et al. Compact bilinear pooling [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 317-326.

[24] KONG S, FOWLKES C. Low-rank bilinear pooling for fine-grained classification [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 7025-7034.

[25] WEI X, ZHANG Y, GONG Y H, et al. Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification [C]//Computer Vision – ECCV 2018, Lecture Notes in Computer Science, vol 11207. Cham: Springer, 2018: 365-380.

[26] CHARIKAR M, CHEN K, FARACH-COLTON M. Finding frequent items in data streams [C]//Automata, Languages and Programming, ICALP 2002, Lecture Notes in Computer Science, vol 2380. Berlin : Springer, 2002: 693-703.

[27] PHAM N, PAGH R. Fast and scalable polynomial kernels via explicit feature maps [C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013: 239-247.

[28] BA J L, KIROS J R, HINTON G E. Layer normalization [EB/OL].(2016-07-21)[2020-6-30]. https://arxiv.org/pdf/1607.06450.pdf.

[29] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 1179-1195,

[30] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context [C]//Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol 8693. Cham: Springer, 2014: 740-755.

[31] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2019-05-24)[2020-06-30].https://arxiv.org/pdf/1810.04805.pdf.

[32] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics(ACL), 2002: 311-318.

[33] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments [C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Stroudsburg, PA: Association for Computational Linguistics(ACL), 2005: 65-72.

[34] LIN C Y. Rouge: A package for automatic evaluation of summaries [C]//Proceedings of the Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL. Stroudsburg, PA: Association for Computational Linguistics(ACL), 2004: 74-81.

[35] VEDANTAM R, ZITNICK C L, PARIKH D. Cider: Consensus-based image description evaluation [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015: 4566-4575.

[36] YAO T, PAN Y W, LI Y H, et al. Boosting image captioning with attributes [C]//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 4904-4912.

[37] FANG H, GUPTA S, IANDOLA F, et al. From captions to visual concepts and back [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015: 1473-1482.

[38] JIANG W H, MA L, JIANG Y G, et al. Recurrent fusion network for image captioning [C]//Computer Vision – ECCV 2018, Lecture Notes in Computer Science, vol 11206. Cham: Springer, 2018: 510-526.

[39] YAO T, PAN Y W, LI Y H, et al. Exploring visual relationship for image captioning [C]//Computer Vision – ECCV 2018, Lecture Notes in Computer Science, vol 11218. Cham: Springer, 2018: 711-727.

[40] YANG X, TANG K H, ZHANG H W, et al. Auto-encoding scene graphs for image captioning [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 10677-10686.

(责任编辑: 李　艺)