面向新概念学习的图像描述生成模型

谢雨杰，杜友田，张潇
（西安交通大学智能网络与网络安全教育部重点实验室，710049，西安）

摘要：为了更好地在图像描述生成任务中对新概念进行学习和预测，在编码-解码框架下提出了一种新的面向新概念学习的图像描述生成模型(Att-DCC)。该模型引入了带有空间注意力机制的卷积神经网络，将全局视觉特征、语义标签和经空间注意力作用后的视觉信息进行了较好的融合；此外，引入自适应注意力机制多模态层，将语义相近的概念学习结果迁移至新概念，降低训练过程的复杂程度并提升学习性能。采用 Att-DCC 模型在 MSCOCO2014 数据集上针对 2 批(分别为 8 和 6 个)共 14 个新概念进行了测试和分析，结果表明：充分的多模态融合方式和多种注意力机制对于提升学习效果有显著效果；Att-DCC 模型在 F_1 值上取得了 42.56%和 42.14%的平均结果，总体上取得了比具有代表性的 NOC 模型和 DCC 模型更准确的预测结果。

关键词：图像描述生成；注意力机制；卷积神经网络；新概念
中图分类号：TP393 **文献标志码**：A
DOI：10.7652/xjtuxb202012005 **文章编号**：0253-987X(2020)12-0037-08

OSID 码

An Image Description Generation Model for the Learning of Novel Concepts

XIE Yujie, DU Youtian, ZHANG Xiao
(Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: A new image description generation model with deep compositional captioner based attention mechanism (named Att-DCC model) based on the encoder-decoder framework is proposed to better understand the "novel concepts" that do not appear in the training set for the task of image description generation. A convolutional neural network with spatial attention mechanism is introduced into the model, and global visual features, semantic labels and visual information after spatial attention are well integrated. In addition, a multi-modal layer with an adaptive attention mechanism is introduced to transfer the learning results to novel concepts, which reduces the complexity of training process and improves learning performance. This work extends the existing multi-modal fusion and introduces multiple attention mechanisms to enhance the performance of novel concept learning. Att-DCC model is used to test and analyze 14 novel concepts in two batches on MSCOCO2014 data set. The experimental results show that the proposed model achieves average results of 42.56% and 42.14% for two batches of novel concepts (8 and 6, respectively) in terms of the F_1 score, and in general, the prediction results are more accurate than those of the representative NOC model and DCC model.

Keywords: image captioning; attention mechanism; convolutional neural network; novel concept

收稿日期：2020-04-23。 作者简介：谢雨杰(1996—)，女，硕士生；杜友田(通信作者)，男，副教授。 基金项目：国家重点研发计划资助项目(2018AAA0101501)；国家电网有限公司总部科技项目。
网络出版时间：2020-09-14 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20200914.1340.002.html

近年来，人工智能飞速发展，围绕着图像、文本、语音三大模态的研究快速推进，多模态内容理解成为研究热点，如跨媒体检索[1]和图像描述生成[5]等。

当前，大多数图像描述生成模型是基于编码解码框架的。在深度学习出现之前，通常使用基于模板和搜索的方法。Farhadi 等人和 Li 等人利用传统的图像检测器来检测图像中的对象以及它们的属性、位置等信息，推测出三元组，并通过文本生成模板将三元组转化为图像描述[2-3]。2015 年微软研究院采用多实例学习训练传统的图像检测器，并基于深度多模态相似性模型对获取的候选描述句子重排序[4]。2015 年，Xu 等人使用深度神经网络进行端对端训练并生成图像的描述，形成了解决该任务的标准框架[5]。该框架采用卷积神经网络(CNN)作为编码器来提取图像特征，使用长短时记忆网络(LSTM)作为解码器预测每一时刻的单词，并首次将注意力机制应用于图像描述生成任务中[6-7]。针对图像描述任务，2017 年，Chen 等人和 Lu 等人主要研究了对图像的注意力，如选取不同空间和通道的特征信息[8-9]。周治平等人研究了如何针对不同视觉属性进行注意力引入的问题，拓展了注意力适用的空间[10]。汤鹏杰等人额外考虑了场景中的先验信息，将其融合至深度学习模型中[11]。杨楠等人通过在损失函数中加入正则项的方式，来提升图像描述模型的性能[22]。然而，上述基于深度学习的模型对测试集和训练集的内容一致性进行了较为严格的限制。

如何理解数据集中未标注过的“新概念”逐渐成为一个重要的研究问题，该问题对于机器智能有重要价值。Anne 等人在 2016 年首次提出了这一研究问题，在图像文本描述数据集的基础上，通过外部图像数据集和文本语料库的辅助来扩大模型能描述的范围[12]。2017 年，Yao 等人将复制机制引入新概念描述，提出了 LSTM-C 模型[13]。同年，Venugopalan 等人改善了文献[12]中的训练方法，提高了兼顾准确率和召回率的 F_1 指标[14]。2018 年，Wu 等人提出将语言描述模型和新概念生成解耦为两个子网络，通过目标检测网络引入与新概念相关的知识，从而进一步提高描述新概念的准确率[15]。2020 年，Feng 等人也意识到与新概念相关的“域外知识”的重要性，并提出了级联修正网络(CRN)，利用域外知识对新概念的描述不断进行修正[16]。但是，上述模型过于关注从外部引入与新概念相关的文本知识，忽略了图像和文本模态的充分结合对于新概念生成的重要性。

为了更好地在图像描述生成任务中对“新概念”进行学习和预测，本文较充分地结合了注意力机制和迁移学习环节，提出了一种新的模型。该方法包括 3 个部分：①带有空间注意力机制的卷积神经网络，用于提取图像特征；②基于 LSTM 的语言单元，用于生成文本描述中的单词；③引入自适应注意力机制的多模态层，将语义相近的概念学习结果迁移至“新概念”。本文的主要贡献如下：

(1)扩展了已有工作中的多模态融合方式，提出了一种有效的图像和文本模态融合方案；

(2)引入多种注意力机制，通过空间注意力和自适应注意力提升新概念学习效果。

在 MSCOCO2014 数据集的实验结果表明，本文模型在包含新概念的测试集上具有较好的图像描述生成性能。

1 图像描述生成的基本模型

图像描述生成任务是指根据一幅图像生成一句(或一段)能描述其内容或语义的文字。该生成模型需要理解图像内容并采用自然语言将图像信息准确表达出来。通常，图像描述生成模型包含两个主要部分：编码器和解码器，其中，编码器用于提取图像的特征语义信息，一般采用 CNN 模型；解码器用于生成描述性语句，一般采用递归神经网络(RNN)，通常可以使用 LSTM 模型来代替 RNN 以避免后者只能记忆较短时间单元内的内容的问题。图像描述生成的基本模型如图 1 所示。

图像描述生成模型可以由图像文本数据进行端到端的训练。

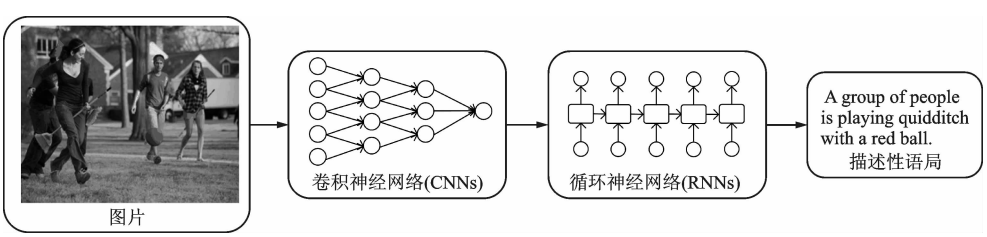


图 1 图像描述生成的基本模型

2　面向新概念学习描述生成模型

为了解决针对新概念学习的图像描述生成任务，本文工作需要基于图像文本关联的训练集 D 进行模型训练，即 $D=\{(I_i,T_i)\}_{i=1}^{N}$，其中 I_i 为第 i 个训练样本所对应的图像，T_i 为 I_i 相关的描述文本，D 中共有 N 个训练样本。此外，需要辅助图像数据和辅助文本数据的支撑。新概念是指在测试集图像中出现的不在训练集中的视觉概念，如目标、场景或行为等。需要注意的是，新概念的视觉信息和语义标签需要分别在辅助图像数据和辅助文本数据中出现过。否则，新概念学习就成了无中生有的问题。

本文提出的图像描述生成模型——Att-DCC 框架如图 2 所示，主要包括 3 个部分：①融合空间注意力的深度卷积神经网络 CNN，用于提取图像的特征信息，包括视觉特征和语义信息；②LSTM 语言单元用于预测每个时刻产生的单词；③多模态层中引入自适应注意力，用于模态融合以及在引入新概念时对权重进行迁移以描述新概念。

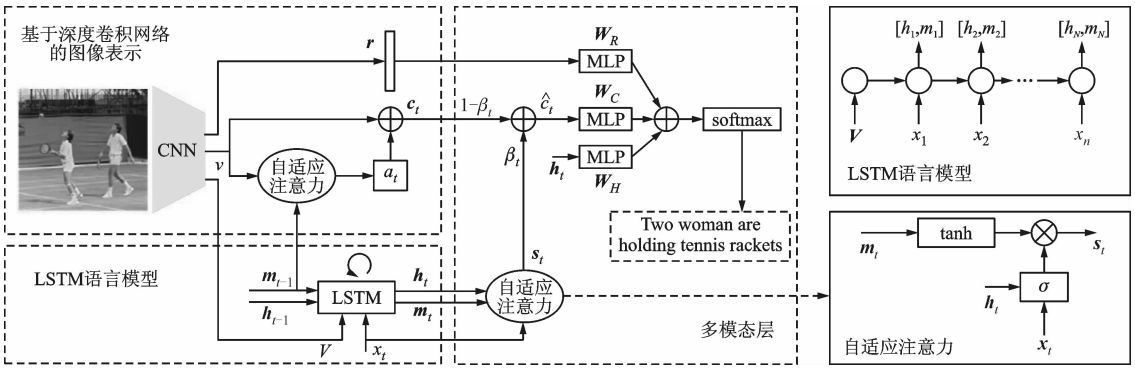


图 2　Att-DCC 模型框架图

2.1　基于空间注意力的图像特征提取

图像描述生成的基本模型一般采用 AlexNet VGG16 和 GoogLeNet 等卷积神经网络作为编码器来提取图像的特征信息。综合考虑模型的时空复杂度与效果因素，本文采用 VGG-16 网络作为编码器，选择第 13 个卷积层 conv3-512（表示该卷积层采用 3×3 的卷积核；卷积层的通道数为 512），输出的长和宽均为 7，设通道数为 512 的特征图为 $\boldsymbol{v}$，第 2 个全连接层 FC-4096（4096 表示该全连接层输出的特征向量维度）输出的图像全局特征向量为 $\boldsymbol{V}$，将 $\boldsymbol{v}$ 和 $\boldsymbol{V}$ 作为图像的综合特征表示。然而，采用单一向量来表示一幅图像通常会损失掉图像的局部信息。人的视觉机制表明人们通常会针对图像中的某些对象进行有重点的关注。因此，为了获得更好的图像特征表示，本文在 VGG-16 的基础上引入了空间注意力机制。

令 VGG16 网络的 conv3-512 层的特征图为 $\boldsymbol{v}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m)^{\mathrm{T}}$，其中 $\boldsymbol{v}_i\in\mathbf{R}^d$ 是对应于第 i 个位置的特征向量，m 是特征图的长与宽的乘积，本文中 $m=49$；$\boldsymbol{h}_{t-1}\in\mathbf{R}^d$ 表示 $t-1$ 时刻 LSTM 的隐层输出状态。LSTM 在 t 时刻的状态 $\boldsymbol{h}_t$ 对特征 $\boldsymbol{v}_i$ 的关注程度 $\alpha_{t,i}$ 可以通过如下空间注意力机制来计算

$$\boldsymbol{z}_t=\boldsymbol{w}_{\mathrm{hs}}^{\mathrm{T}}\tanh((\boldsymbol{W}_v\boldsymbol{v})+(\boldsymbol{W}_{\mathrm{hs}}\boldsymbol{h}_{t-1})\,\mathbf{1}^{\mathrm{T}}) \tag{1}$$

$$\boldsymbol{\alpha}_t=\mathrm{softmax}(\boldsymbol{z}_t) \tag{2}$$

式中：$\boldsymbol{\alpha}_t=(\alpha_{t,1},\alpha_{t,2},\cdots,\alpha_{t,m})$；$\boldsymbol{W}_v$ 为特征图变换矩阵；$\boldsymbol{W}_{\mathrm{hs}}$ 为 LSTM 隐层状态变换矩阵；$\boldsymbol{w}_{\mathrm{hs}}$ 为列向量；$\mathbf{1}^{\mathrm{T}}$ 表示元素均为 1 的行向量。对所有位置的特征进行加权求和可得局部视觉上下文信息 $\boldsymbol{c}_t$

$$\boldsymbol{c}_t=\sum_{i=1}^{m}\alpha_{t,i}\boldsymbol{v}_i \tag{3}$$

上下文信息对于 t 时刻生成的单词具有重要影响。式(1)～(3)描述的空间注意力机制对应于图 2 中的自适应注意力。

2.2　多视觉特征融合

本文采用 LSTM 作为解码器，其作用是将图像内容“翻译”为自然语言。LSTM 接收来自图像的特征信息和上一时刻输出单词的嵌入向量表示作为输入。本文采用广泛使用的 Word2 Vec 连续词袋（CBOW）模型[17]进行文本单词的嵌入向量计算。

相比于以前的模型，本文将多种视觉特征进行了有效的融合并将其作用于 LSTM。其中，多特征包括来自于全局视觉特征向量 $\boldsymbol{V}$ 和基于注意力机制的视觉上下文信息 $\boldsymbol{c}_t$ 以及来自 VGG-16 第 3 个全连接层 FC-1000 输出的语义概念标签预测向量 $\boldsymbol{r}$。本文采用如下方式进行多特征融合：①全局视觉特征引入：将全局视觉特征向量 $\boldsymbol{V}$ 引入到 LSTM 的

初始时刻，为文本描述生成提供全局先验信息。LSTM 的长时记忆特性使得初始时刻引入的视觉特征信息可以得到长时记忆和传输。事实上，视觉特征信息还可引入至 LSTM 中每个时刻的输入。②“注意力”后的视觉信息引入：将经过注意力机制“关注”后的视觉特征，即上下文向量 $\boldsymbol{c}_t$（实际采用下节介绍的 $\hat{\boldsymbol{c}}_t$），与 LSTM 的每个时刻的输出 $\boldsymbol{h}_t$ 相融合并输入多模态层，提供局部图像信息。③语义概念标签引入：为了更好地将图像中的重要目标、场景或行为等重要概念在文本描述中得到体现，将图像高层语义概念标签 $\boldsymbol{r}$ 与 LSTM 的每个时刻的输出进行融合并输入多模态层，最终决定文本单词的预测。

将两种视觉特征信息（全局视觉特征向量和基于注意力机制的局部视觉上下文信息）输入至 LSTM 的方式可以有两种：①在初始时刻将全局视觉特征向量 $\boldsymbol{V}$ 输入至 LSTM，因为 LSTM 具有长时记忆特性，在初始时刻输入的特征信息可通过长时记忆和传输，对整个句子的词语生成起到控制作用；②在 LSTM 每一个时刻输入经过注意力加权过的局部视觉上下文信息 $\boldsymbol{c}_t$，根据不同时刻要预测的不同单词，注意力的加权权重会改变，即局部视觉上下文信息 $\boldsymbol{c}_t$ 会改变，从而 LSTM 可以获取到图像不同局部区域的特征信息。由于方式 1 中的全局视觉特征可以从宏观角度指导描述的整个句子生成，有利于生成更加通顺的语句，方式 2 中的局部视觉信息经过了注意力加权，能够提供图像中更加准确的物体信息，有利于生成和图像更相关的描述，因此本文实验选择了方式 1 和方式 2 结合来进行图像的全局和局部视觉特征输入。

2.3 结合自适应注意力的新概念学习

本文采用线性多模态层来连接编码器和解码器，并在已经训练好的图像描述生成模型中引入新概念。线性多模态层使得学习权重能够容易地迁移至新概念。为了使解码器能自适应地决定依赖于图像信息及它在上一时刻状态的程度，本文在多模态层引入了自适应注意力，结构如图 2 所示。自适应注意力主要依赖于自适应门的输出 $\boldsymbol{s}_t$，经过线性计算构成新的上下文向量 $\hat{\boldsymbol{c}}_t$

$$\boldsymbol{g}_t = \sigma(\boldsymbol{W}_x \boldsymbol{x}_t + \boldsymbol{W}_h \boldsymbol{h}_t) \tag{4}$$

$$\boldsymbol{s}_t = \boldsymbol{g}_t \odot \tanh(\boldsymbol{m}_t) \tag{5}$$

$$\boldsymbol{f}_t = \boldsymbol{w}_t^{\mathrm{T}} \tanh(\boldsymbol{W}_s \boldsymbol{s}_t + \boldsymbol{W}_g \boldsymbol{h}_t) \tag{6}$$

$$\boldsymbol{\beta}_t = \mathrm{softmax}([\boldsymbol{z}_t^{\mathrm{T}}, f_t])[m+1] \tag{7}$$

$$\hat{\boldsymbol{c}}_t = \boldsymbol{\beta}_t \boldsymbol{s}_t + (1-\beta_t)\boldsymbol{c}_t \tag{8}$$

式中：$\boldsymbol{W}_x$、$\boldsymbol{W}_h$、$\boldsymbol{w}_h$、$\boldsymbol{W}_s$ 和 $\boldsymbol{W}_g$ 为参数；$\boldsymbol{h}_t$、$\boldsymbol{m}_t$ 和 $\boldsymbol{s}_t$ 分别表示 LSTM 在 t 时刻的输出、单元状态和自适应门输出向量；$\boldsymbol{x}_t$ 表示在 t 时刻输入单词对应的嵌入向量；f_t 反映了语言单元的自注意力；$\beta_t \in [0,1]$ 为向量 $[\boldsymbol{z}_t^{\mathrm{T}}, f_t]$ 经过归一化的最后一个元素，表示上下文信息 $\hat{\boldsymbol{c}}_t$ 中语言单元的重要程度。t 时刻预测单词的概率可通过下式计算

$$P_{\text{predicted-word}} = \mathrm{softmax}(\boldsymbol{W}_{\mathrm{C}} \hat{\boldsymbol{c}}_t + \boldsymbol{W}_{\mathrm{H}} \boldsymbol{h}_t + \boldsymbol{W}_{\mathrm{R}} \boldsymbol{r}) \tag{9}$$

式中：$\boldsymbol{W}_{\mathrm{C}}$，$\boldsymbol{W}_{\mathrm{H}}$ 和 $\boldsymbol{W}_{\mathrm{R}}$ 是待学习的模型参数，均通过全连接网络结构进行实现。多模态层中主要有图像特征信息对应的权重矩阵 $\boldsymbol{W}_{\mathrm{C}}$、文本信息所对应的权重矩阵 $\boldsymbol{W}_{\mathrm{H}}$ 和语义标签信息对应的权重矩阵 $\boldsymbol{W}_{\mathrm{R}}$。

在图像描述模型引入新概念时，需要生成这 3 个权重矩阵中与新概念对应的部分。首先，通过余弦距离度量找到图像文本对训练集与新概念中相似性最高的词语。例如，新概念 alpaca（羊驼）和图像文本对训练集中相似性最高的词语是 sheep（羊）。其次，将训练集中的词语对应的权重迁移至对应的新概念上。如将 sheep 所对应的权重矩阵信息迁移到 alpaca 上。本文采用了直接迁移法和 delta 迁移法[10]。直接迁移法是直接将 $\boldsymbol{W}_{\mathrm{C}}$、$\boldsymbol{W}_{\mathrm{H}}$ 和 $\boldsymbol{W}_{\mathrm{R}}$ 中原词对应的权重直接复制迁移到新概念上。例如，假设 $\boldsymbol{v}_s$ 和 $\boldsymbol{v}_a$ 分别是 sheep 和 alpaca 在词典中的序号，则直接迁移法为 $\boldsymbol{W}_{\mathrm{R}}(:, \boldsymbol{v}^a) = \boldsymbol{W}_{\mathrm{R}}(:, \boldsymbol{v}_s)$，对 $\boldsymbol{W}_{\mathrm{C}}$ 和 $\boldsymbol{W}_{\mathrm{H}}$ 也采用同样操作。delta 迁移法的具体细节参见文献[12]。

3 实验结果与分析

3.1 实验数据集

本文基于 MSCOCO2014 数据集进行实验测试。同时，为了支撑新概念的学习，模型需要利用单独的图像数据集和单独的文本数据集分别对图像描述模型和语言生成模型进行预训练，其作用是将新概念对应的图像信息和语义标签信息包含进来。其中，采用 ImageNet 中的图片构建单独的图像数据集；采用 Flickr8k、Flickr30K、Pascal-1K、MSVD 中的文本语句构建单独文本数据集。基于预训练好的模型来采用 MSCOCO2014 数据集来对图像和文本描述之间的关联进行学习并对测试图像样本生成文本描述。

本实验基于 MSCOCO2014 数据集的子集进行测试，其中包含了 MSCOCO 训练集82 783张图片和413 915句描述；验证集包括40 504张图片和202 520句描述；每幅图像对应 5 句描述文本。为了测试新概念学习的效果，需要为“新概念”建立数据

集。本实验采用 held-out 方法将 MSCOCO 数据集分为两个部分：一部分是“新概念”数据集，其做法是将 MSCOCO 图像数据的 80 个类别聚类为 8 个簇，从这 8 个簇的每个簇中随机选择一个概念，本文实验中选择了 bottle、bus、couch、microwave、pizza、racket、suitcase 和 zebra 8 个实体概念选为新概念，并将所有包含这 8 个概念的图像文本对数据从 MSCOCO 训练集中去掉；另一部分即剩余部分是用于模型参数学习的训练集。MSCOCO 校验集平均分成两部分，一部分用于测试，一部分用于校验模型参数学习结果。

3.2　训练过程

本实验中采用 VGG-16 作为卷积神经网络，并以 ImageNet 的子集 ILSVRC-2012[18] 上训练的实体识别模型为预训练模型；用单独的文本数据集训练语言描述模型 LSTM；最后，用图像-文本对数据集 MSCOCO 中的基础数据集完成整个模型的训练。

3.3　评价指标

本实验采用图像描述生成模型的 2 个最常用评价指标：双语评估替代(BLEU)[20] 和双语翻译评估指标(METEOR)[21] 进行评估。双语评估替代是对模型生成的句子和参考句子计算单个单词共现频率，从而判断两个句子的相似程度，记为 $S_{\text{BLEU-1}}$；双语翻译评估指标是计算词汇完全匹配、词干匹配和同义词匹配等各种情况的准确率和召回率，再进行调和平均，记为 S_{METEOR}。S_{METEOR} 考虑了召回率，其结果要比单纯考虑精度的指标 $S_{\text{BLEU-1}}$ 更接近人工评价标准。在传统的图像描述任务中，以上评价指标很有效。但在评估新概念是否正确引入时，有时会出现句子中不出现新概念但仍然可以准确并流畅表达的情况，并且也可以获得高的 $S_{\text{BLEU-1}}$ 和 S_{METEOR} 值。为了评估模型准确描述新概念词语的能力，本文引入了 F_1 值作为估指标，$F_1=2pr/(p+r)$，其中 p 和 r 分别表示描述的新概念的精确率和召回率。

3.4　实验结果

3.4.1　模型消融(ablation)测试　为了分析本文 Att-DCC 模型在引入不同图像特征信息(图像视觉特征、物体语义标签、空间注意力机制生成的视觉上下文信息)对于图像描述生成的作用，本文进行了消融实验测试，测试了 Att-DCC 模型在 3 种信息作用下的模型性能：①视觉特征，直接采用全局视觉特征作为 LSTM 的视觉信息输入；②语义标签，将采用 CNN 进行图像分类的语义标签作为 LSTM 的视觉信息输入；③全局视觉特征＋语义标签＋注意力，即 Att-DCC 模型所采用的实验情况。测试结果如表 1 所示。

从表 1 可以看出，在上述 3 种情况中，采用“全局视觉特征＋语义标签＋注意力”作为 LSTM 的输入对于图像描述生成和新概念描述具有最好的性能，在生成语句的准确度、流畅度以及新概念的适应性(F_1 值)上的表现都得到了最高的指标值。其主要原因在图像信息被进行了充分有机地结合并输入值 LSTM 模型中；注意力机制也保证了文本生成过程中的单词与图像中的区域能够动态地对齐。此外，从表 1 可以看出，对图像内容进行分类后得到的语义标签比直接采用图像的全局视觉特征向量对于图像描述生成有优势。这是因为前者通过 CNN 的全连接层将图像信息首先转化为语义信息，该语义信息比视觉特征具有更丰富的高层语义。

表 1　模型消融测试结果

图像特征信息	训练方式	F_1	$S_{\text{BLEU-1}}$	S_{METEOR}
视觉特征	delta 迁移	32.64	61.90	18.93
	直接迁移	36.52	62.60	19.51
语义标签	delta 迁移	34.89	64.00	20.86
	直接迁移	39.78	64.40	21.00
视觉特征＋语义标签＋注意力	delta 迁移	37.14	64.21	21.69
	直接迁移	**42.56**	**65.18**	**21.95**

注：黑体数据为最优值。

表 1 还测试了直接迁移和 delta 迁移的方式引入新概念并分析了不同迁移学习方法对引入新概念的影响。由技术原理和实验结果可以看出，迁移学习对于新概念的引入是必要的。在多模态层的训练种，直接迁移分别迁移了训练权重 $\boldsymbol{W}_{\text{C}}$、$\boldsymbol{W}_{\text{H}}$ 和 $\boldsymbol{W}_{\text{R}}$；而对于 delta 迁移，则主要迁移的是图像文本联合训练与基于文本单独训练的语言单元的权重 $\boldsymbol{W}_{\text{H}}$ 的差，与图像相关的其他两个矩阵仍采用直接迁移，由于其迁移过程较复杂，实验结果发现迁移效果略有下降。

因此，在后续的实验中采用的是直接迁移方法。表 2 展示了 Att-DCC 在 3 种组件作用下对于 8 个新概念各自的学习结果。可以看出，语义标签的引入可以使得大多数的概念学习结果得到提升。为了证明本文提出的 Att-DCC 模型在新概念学习方面具有推广性，我们从 MSCOCO 中抽取了另外 6 个概念：bear、cat、dog、elephant horse 和 toilet。与这

表 2 融合语义标签信息后每个新概念 F_1 值比较

图像特征信息	F_1								F_1 均值
	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	
视觉特征	2.89	36.84	42.76	26.56	55.37	47.69	10.91	60.58	34.20
语义标签	4.63	29.79	45.87	28.09	64.59	52.24	13.16	79.88	39.78
视觉特征＋语义标签＋注意力	**8.96**	**33.84**	**48.56**	**30.27**	**66.20**	**56.31**	**15.09**	**81.25**	**42.56**

注：黑体数据为最优值。

6 个概念相关的图像文本数据大约占 MSCOCO 的 8.5%，将其处理为新概念进行重新训练和测试，结果如表 3 所示。结果显示：平均准确率近似于表 2 中的平均值，体现了较好的效果。

表 3 采用 Att-DCC 模型描述其他新概念的 F_1 值展示

F_1						F_1 均值
bear	cat	dog	elephant	horse	toilet	
21.70	43.24	28.29	72.56	52.24	34.82	42.14

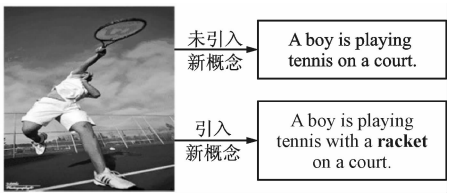
3.4.2 与其他模型的比较 将本文提出的 Att-DCC 模型与目前比较先进的几种新概念引入模型进行了对比，对比模型包括以下 3 种。①LRCN 模型[19]：目前的传统图像描述模型中经典且效果较好的方法，但该模型无法引入新概念。因此，本实验中对 LRCN 训练时使用的训练集将新概念对应的数据包含在内，使得训练集和测试集具有相同的分布。该操作也导致了 LRCN 的性能在大部分情况下是较好甚至最好的。②DCC 模型[12]：本文模型所依据的原始模型，分别采用 delta 迁移和直接迁移进行新概念学习。③NOC 模型[14]：采用的无迁移学习的方法进行训练，构造多个损失函数来联合最小化图像文本数据集、单独图像数据集、单独文本数据集上的损失。从对比实验可以看出，Att-DCC 模型效果大多数情况下优于其他几种模型。由于 LRCN 模型的训练集中包含了新概念对应的数据，因此该模型对测试集中的新概念的预测最好，其 F_1 值达到了最高；而 Att-DCC 模型在 F_1 值上表现略低于 NOC 模型，其主要原因是 Att-DCC 模型注重图像和文本模态的充分结合以生成准确的新概念和流畅的描述；NOC 模型则侧重于改进模型的训练方法以提高生成新概念的准确率。采用另外两种评价语句流利性的指标 $S_{\text{BLEU-1}}$ 和 S_{METEOR}，对本文 Att-DCC 模型和其他两种针对新概念的模型 DCC 和 NOC 进行对比，结果表明 Att-DCC 得到了更好的学习结果，证明本文提出的模型对于改进新概念描述和图像描述生成具有积极的作用。

表 4 Att-DCC 模型与已有研究模型的比较

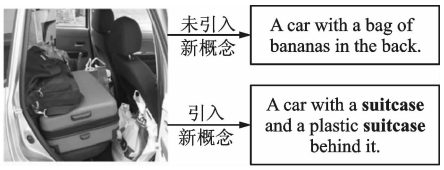
模型	训练方式	F_1	$S_{\text{BLEU-1}}$	S_{METEOR}
LRCN[19]	无迁移	**54.81**	63.65	19.33
DCC[12]	delta 迁移	34.89	64.00	20.86
	直接迁移	39.78	64.40	21.00
NOC[14]	无迁移	48.79		21.32
Att-DCC	delta 迁移	37.14	64.21	21.69
	直接迁移	42.56	**65.18**	**21.95**

注：黑体数据为最优值。

3.4.3 “新概念”描述结果展示 为了展示了 Att-DCC 模型描述新概念的效果，图 3 给出了 2 个图像描述生成的样例。从图 3 可以看到，经过迁移学习，训练集中没有出现的新概念 racket 和 suitcase 出现在了图像描述生成结果中，说明 Att-DCC 模型具有描述新概念的能力。



(a)对新概念“racket”的描述状况



(b)对新概念“suitcase”的描述状况

图 3 产生“新概念”的图像描述样例

4 结 论

针对目前相关工作在信息融合和注意力方面不充分的缺点，本文提出了一种面向新概念学习的图像描述生成模型——Att-DCC。本文的主要贡献包括：扩展了已有工作中的多模态融合方式，提出了一种有效的图像和文本模态融合方案；引入多种注意

力机制，通过空间注意力和自适应注意力提升新概念学习效果。在研究工作基础上，本文得到如下结论。

(1)在充分考虑多模态融合和多注意力结合的基础上，构建了一种面向新概念学习的图像描述生成模型；消融实验结果验证了该模型的必要性和有效性。

(2)在 MSCOCO2014 数据集上针对 2 批共计 14 个新概念进行了测试和分析，并与已有相关模型 DCC 和 NOC 进行对比。实验结果表明，Att-DCC 模型的性能优于以上两种相关模型。

(3)图像描述生成中的新概念学习是非常重要的，对于模型的适用性以及通过海量的在线数据提升和扩充图像描述生成模型的性能具有重要意义。

致谢　本文研究工作得到国家重点研发计划资助项目(2018AAA0101501)和国家电网有限公司总部科技项目(人在回路的大电网调控混合增强智能基础理论)的资助。

参考文献：

[1] WANG W, YANG Xiaoan, OOI B C, et al. Effective deep learning-based multi-modal retrieval [J]. VLDB Journal, 2016, 25(1): 79-101.

[2] FARHADI A, HEJRATI M, SADEGHI M A, et al. Every picture tells a story: generating sentences from images [C]// 11th European Conference on Computer Vision. Berlin, Germany: Springer Verlag, 2010: 15-29.

[3] LI S, KULKARNI G, BERG T L, et al. Composing simple image descriptions using web-scale n-grams [C]// Proceedings of the Fifteenth Conference on Computational Natural Language Learning. New York, USA: Association for Computational Linguistics, 2011: 220-228.

[4] FANG Hao, GUPTA S, IANDOLA F, et al. From captions to visual concepts and back [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2015: 1473-1482.

[5] XU K, BA J L, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention [C]// 32nd International Conference on Machine Learning. Pitsburg, PA, USA: CMU, 2015: 2048-2057.

[6] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [C]// 3rd International Conference o Learning Representations. Montreal, Canada: ICLR, 2015: 149801.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// 31st Annual Conference on Neural Information Processing Systems. Vancouver, Canada: Neural Information Processing Systems Foundation, 2017: 5999-6009.

[8] CHEN L, ZHANG H, XIAO J, et al. Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning [C]// Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2017: 5659-5667.

[9] LU J, XIONG C, PARIKH D, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning [C]// Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2017: 3242-3250.

[10] 周治平，张威. 结合视觉属性注意力和残差连接的图像描述生成模型 [J]. 计算机辅助设计与图形学学报，2018, 30(8): 1536-1542, 1553.
ZHOU Zhiping, ZHANG Wei. An image caption generation model based on visual concept attention and residual connection [J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(8): 1536-1542, 1553.

[11] 汤鹏杰，谭云兰，李金忠. 融合图像场景及物体先验知识的图像描述生成模型 [J]. 中国图象图形学报，2017, 22(9): 1251-1260.
TANG Pengjie, TAN Yunlan, LI Jinzhong. Image description based on the fusion of scene and object category prior knowledge [J]. Journal of Image and Graphics, 2017, 22(9): 1251-1260.

[12] ANNE HENDRICKS L, VENUGOPALAN S, ROHRBACH M, et al. Deep compositional captioning: describing novel object categories without paired training data [C]// Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2016: 7780377.

[13] YAO Ting, PAN Yingwei, LI Yehao, et al. Incorporating copying mechanism in image captioning for learning novel objects [C]// Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2017: 5263-5271.

[14] VENUGOPALAN S, ANNE HENDRICKS L, RO-

HRBACH M, et al. Captioning images with diverse objects [C]// Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2017: 1170-1178.

[15] WU Yu, ZHU Linchao, JIANG Lu, et al. Decoupled novel object captioner [C]// Proceedings of the 2018 ACM Multimedia Conference. New York, USA: ACM, 2018: 1029-1037.

[16] FENG Q, WU Y, FAN H, et al. Cascaded revision network for novel object captioning [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020,30(1): 3413-3421.

[17] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [C]// 1st International Conference of Learning Representation. Montreal, Canada: ICLR, 2013: 149796.

[18] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[19] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2015, 2625-2634.

[20] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [EB/OL]. [2020-02-18]. https://www.aclweb.org/anthology/P02.1040.pdf.

[21] IBANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments [EB/OL]. [2020-02-20]. https://www.aclweb.org/anthology/W05-0909.pdf.

[22] 杨楠, 南琳, 张丁一, 等. 基于深度学习的图像描述研究 [J]. 红外与激光工程, 2018, 47(2): 9-16.
YANG Nan, NAN Lin, ZHANG Dingyi, et al. Research on image interpretation based on deep learning [J]. Infrared and Laser Engineering, 2018, 47(2): 9-16.

[本刊相关文献链接]

黄文君,李杰,齐春.低秩与字典表达分解的浓雾霾场景图像去雾算法.2020,54(4):118-125.[doi:10.7652/xjtuxb202004015]

毛远宏,贺占庄,马钟,等.采用类内迁移学习的红外/可见光异源图像匹配.2020,54(1):49-55.[doi:10.7652/xjtuxb202001007]

贾晓芬,郭永存,柴华荣,等.深立井井壁图像的卷积神经网络去噪方法.2019,53(6):117-124.[doi:10.7652/xjtuxb201906016]

崔云博,杜友田,王航.面向图像的有效目标区域提取方法.2019,53(5):52-57.[doi:10.7652/xjtuxb201905008]

肖满生,肖哲,万烂军.多特征融合的图像格贴近度匹配方法.2019,53(4):115-121.[doi:10.7652/xjtuxb201904017]

孟月波,刘光辉,徐胜军,等.一种具有边缘保持的多尺度马尔可夫随机场模型图像分割方法.2019,53(3):56-65.[doi:10.7652/xjtuxb201903009]

时璇,许林松,李晨,等.联合加权聚合深度卷积特征的图像检索方法.2019,53(2):128-135.[doi:10.7652/xjtuxb201902017]

王浩,梁煜,张为.一种均匀化稀疏表示的图像压缩感知算法.2019,53(2):136-141.[doi:10.7652/xjtuxb201902018]

邢超,张晓明,赵亚琳,等.利用图像信息熵差检测炉渣碳质量分数.2018,52(8):49-53.[doi:10.7652/xjtuxb201808008]

宋长明,王赟.融合低秩和稀疏表示的图像超分辨率重建算法.2018,52(7):18-24.[doi:10.7652/xjtuxb201807003]

张淑芳,丁文鑫,韩泽欣,等.采用主成分分析与梯度金字塔的高动态范围图像生成方法.2018,52(4):150-157.[doi:10.7652/xjtuxb201804022]

徐思雨,蔡佳妮,祝继华,等.自适应多位编码量化的哈希图像检索方法.2017,51(8):19-25.[doi:10.7652/xjtuxb201708004]

李莉,冯林,吴俊,等.一种三结构描述子的图像检索方法.2017,51(6):86-91.[doi:10.7652/xjtuxb201706014]

张胜杰,查宇飞,李运强,等.一种利用局部结构信息的加权哈希图像检索算法.2016,50(10):78-85.[doi:10.7652/xjtuxb201610012]

黄晓冬,孙亮,刘胜蓝.一种判别极端学习的相关反馈图像检索方法.2016,50(8):96-102.[doi:10.7652/xjtuxb201608016]

周远,周玉生,刘权,等.一种适用于图像拼接的 DSIFT 算法研究.2015,49(9):84-90.[doi:10.7652/xjtuxb201509015]

毛彦斌,张选平,杨晓刚.伪 DNA 密码图像加密算法研究.2015,49(9):91-98.[doi:10.7652/xjtuxb201509016]

刘凯,张立民,孙永威,等.利用深度玻尔兹曼机与典型相关分析的自动图像标注算法.2015,49(6):33-38.[doi:10.7652/xjtuxb201506006]

符均,牟轩沁,季文博.亮色分离的饱和图像校正方法.2014,48(10):101-107.[doi:10.7652/xjtuxb201410016]

岳桂华,滕奇志,何小海,等.岩心三维图像修复算法.2014,48(9):37-42.[doi:10.7652/xjtuxb201409007]

(编辑 刘杨 陶晴)