arXiv:2302.04676v1 [cs.CV] 8 Feb 2023

用于图像描述的堆叠式跨模态特征整合注意力网络

Mozhgan Pourkeshavarz、Shahabedin Nabavi、Mohsen Ebrahimi Moghaddam、Mehrnoosh Shamsfard

伊朗德黑兰沙希德·贝赫什提大学计算机科学与工程学院

通讯作者：Mohsen Ebrahimi Moghaddam，邮箱：m moghadam@sbu.ac.ir

2023年2月10日

摘要

近年来，注意力增强型编码器-解码器框架取得了显著进展，因此在图像描述领域引起了广泛关注。许多视觉注意力模型直接利用有意义的区域生成图像描述。然而，仅仅寻求从视觉空间到文本的直接转换，还不足以生成细粒度描述。本文采用特征复合方法，以完全端到端的方式整合与上下文环境相关的高层语义概念和视觉信息。因此，我们提出了一种用于图像描述的堆叠式跨模态特征整合注意力网络（SCFC），通过一种新颖的复合函数以多步推理的方式同时整合跨模态特征。此外，我们将空间信息和上下文感知属性（CAA）共同用作所提出复合函数的主成分，其中 CAA 能够提供简洁的上下文敏感语义表示。为了更好地发挥整合后特征的潜力，我们进一步提出 SCFC-LSTM 作为描述生成器，使其能够在图像描述生成过程中利用判别性语义信息。实验结果表明，在 MSCOCO 和 Flickr30K 数据集上，按照常用指标评估，我们提出的 SCFC 优于多种最先进的图像描述基准方法。

关键词— 上下文表示、跨模态特征融合、图像描述、堆叠注意力网络、视觉与语义信息。

1 引言

自动描述图像内容，即图像描述，是人工智能的一项重要任务，它结合了计算机视觉（CV）与自然语言处理（NLP）领域。图像描述在图像索引、社交媒体平台、视障人士服务等方面有多种应用。尽管这项任务对人类来说似乎很容易，但对机器而言却十分复杂。机器需要解决识别图像中存在哪些物体和属性的问题，并以自然语言表达它们之间的交互关系。深度神经网络的最新进展，为可靠地生成图像描述迈出了重要一步。

具体而言，深度图像描述生成架构在发现视觉特征与其自然语言对应内容之间的映射方面，取得了令人瞩目的成果。该任务采用了广为人知的编码器-解码器框架。该框架包含一个用于提取特征的卷积神经网络（CNN），以及一个长短期记忆网络（LSTM），后者根据静态的整体图像特征向量 [1]-[5] 生成句子。尽管这些技术的进展令人鼓舞，但上述框架面临的瓶颈在于，难以挖掘构建图像描述所需的全部视觉信息，从而准确描述图像。

受人类视觉系统会倾向于关注整个视觉空间中特定部分这一现象的启发，视觉注意力应运而生 [6], [7], 。具体而言，不再将

图像编码为单个静态向量，视觉注意力鼓励模型有选择地关注图像中的显著区域，并利用这些区域生成描述[8]。此外，一些模型只关注显著区域，而不扫描整幅图像，因此无法捕捉上下文。尽管这种方法颇具吸引力，但它存在两个主要缺陷，促使人们开展进一步的重要研究。首先，模型生成的是视觉词，而非高层语义词。另一个问题是缺少文本信息，导致模型无法准确理解图像上下文。多项研究表明，基于属性的方法旨在生成更丰富的语义细节，从而提升图像描述的性能[9]-[11]。然而，这种有效方法的缺点在于，将图像中所有已有属性都纳入循环神经网络（RNN）并无必要，甚至可能产生误导。近期更多证据表明，由于视觉信息与语义信息具有互补性，因此二者缺一不可[12], [13]。尽管这项工作有着显著优势，但它没有考虑视觉信息与语义信息之间的相互关系，而是不顾图像上下文，将二者结合起来构建抽象表示。

通常，要生成一个主观句子来描述图像中的显著内容，需要使用更抽象的词语。在许多情况下，推断这些词语不仅需要考虑高层语义概念，还需要结合上下文信息，同时考虑多个区域。例如，在图像描述“足球迷在露天的座无虚席的体育场中为他们的球队欢呼并庆祝进球”中，令人惊讶的是，没有一个词仅凭边界框视觉区域就能分类。直观而言，在预测“球迷”和“进球”等词语时，模型必须基于上下文表示框架下的视觉信息和语义信息进行推断。因此，仅使用视觉区域生成细粒度描述是不够的。另一方面，将高层语义概念和上下文信息与感兴趣区域（ROI）结合起来，则需要一种先进的特征融合方法。

本文为图像描述任务中的跨模态特征整合带来了新的见解。具体而言，我们在提出的堆叠式跨模态特征整合（SCFC）注意力网络中构建了一种新颖的复合函数。特别是，通过多个CFC层进行渐进式推理，SCFC能够在图像描述生成过程的每个时间步逐步整合跨模态知识，生成丰富的表示。我们还将跨模态特征构建为复合函数中的主成分，以提升图像描述效果。具体来说，我们综合考虑空间视觉信息、高层语义概念和上下文信息，为给定图像生成抽象且更丰富的表示。为了更充分地利用整合后特征的潜力，我们进一步提出SCFC-LSTM作为描述生成器，使其能够在图像描述生成过程中利用判别性语义信息。此外，从建模角度来看，我们的模型更具吸引力，因为它可以进行端到端的完整训练。

本研究的贡献如下：

1. 提出了一种新颖的SCFC注意力网络。该网络构建了一个复合函数，可对跨模态特征进行多步推理，从而促进判别性语义特征的生成。2. 我们将多方面特征作为复合函数的主成分，其中包含（1）空间视觉信息、（2）高层语义概念和（3）上下文信息，以表征给定图像的不同方面。此外，由于所提出的模型可在单一优化层级上进行完整的端到端训练，因此无需设置独立阶段来提取这些特征。3. 我们采用SCFC-LSTM作为描述生成器，使其能够在图像描述生成过程中利用具有判别性的语义信息，从而生成细粒度描述。4. 我们在基准数据集MS-COCO和Flickr30K上验证了所提方法的有效性。实验结果表明，所提方法能够取得与最先进方法相当的结果。

2 相关工作

迄今已有大量关于图像描述的论文发表。具体而言，我们关注经典编码器-解码器框架中使用的视觉注意力方法，该框架因其出色的性能而备受关注

以往的工作使用卷积神经网络将图像编码为单个静态视觉特征图，然后将其输入循环神经网络作为解码器 [1], [14], [15]。然而，这种静态表示可能会导致局部信息丢失。

受机器翻译中引入的注意力机制启发， [8], [16] 首次在图像描述中提出了空间注意力，将上一步的隐藏状态与图像块的视觉特征相结合，以计算图像中不同图像块的注意力权重。对所有图像块进行加权求和即可得到软注意力特征，而硬注意力特征则对应于最重要图像块的视觉信息。尽管这一方法取得了很大进展，但它涉及许多无意义的图像块，导致计算复杂度较高，并加剧了视觉干扰。 [17] 提出了场景特定上下文，并采用选择性搜索 [18] 生成区域提议。此外， [19]提出了一种基于区域的注意力机制，通过建模图像区域、描述词与隐藏状态之间的依赖关系，使描述词与图像区域能够直接关联。进一步地， [20] 指出，解码器可能会预测图像中几乎没有视觉信息支持的非视觉词汇，并引入视觉哨兵，以决定是否参与视觉部分或语言模型。因此，通过引入自适应注意力， [20] 指出有必要考虑对非视觉词汇进行处理。

人们提出了不同类型的注意力机制，这些机制可以利用弱监督的多实例学习（MIL）来学习高级概念，并将其组合成句子，从而解决非视觉词汇的生成问题。 [9] 引入了一种视觉属性分类器，用于生成语义概念，其中图像特征是由属性分类器置信度构成的向量。 [11] 开发了一种新颖的属性引导模型，将属性间相关性融入多实例学习（MIL），并将高层语义属性加入基于循环神经网络（RNN）的编码器-解码器框架，以获得更好的性能。他们构建了多种架构来输入这些特征，以寻找融入语义属性的最佳方式。然而，所有属性词都被视为同等重要，并在每个时间步都输入循环神经网络（RNN）。将图像中所有可能的词汇都纳入其中，可能会导致最终生成的图像描述中出现一些与图像上下文无关的词语。

总而言之，现有方法往往只关注视觉信息或属性信息。由于二者具有互补性，只关注其中一方会导致对给定图像的理解不充分。另一个问题是，很少有研究考虑在图像描述生成过程中保留上下文信息的重要性。为了指出显著事件而生成主观句子，会受到图像所处场景的强烈影响。

3 所提出的方法

本节详细介绍所提出的用于图像描述的 SCFC注意力网络。所提出框架的概览如图1所示。该框架由三个部分组成：（1）耦合视觉检测器和属性预测器；（2）SCFC注意力网络；（3）使用 SCFC-LSTM 生成图像描述。首先，我们以耦合方式从给定图像中提取视觉区域 $V = \{v_1, \ldots, v_n\}$， $v_i \in \mathbb{R}^h$， $V \in \mathbb{R}^{n \times h}$ 和语义属性$A = \{a_1, \ldots, a_c\}$， $a_i \in \mathbb{R}^{|\Sigma|}$， $A \in \mathbb{R}^{|\Sigma| \times c}$。然后，将它们输入 SCFC模块。在将跨模态特征构建为主成分后，所提出的 CFC注意力网络被触发，以形成受多步推理思想启发的复合函数。最后，整合后的语义特征被传递至 SCFC-LSTM，以生成语义细粒度图像描述。

3.1 耦合视觉检测器与属性预测器

视觉特征和语义特征彼此互补。基于这一点，我们开发了一种能够利用二者来增强视觉和非视觉词汇生成的模型，例如“帮助”和“坐着”。在传统研究中，属性预测被视为一项独立任务，依赖于单独的阶段，从而增加了模型参数的总体数量。受 [21], 中的端到端属性检测启发，我们采用了一个属性预测器（AP），使其能够与整个图像描述网络联合训练。不同于以往研究 [21] [22], ，我们的视觉检测器（VD）也可以与整个图像描述网络一同训练。具体而言，我们认为，在整个系统中训练特征提取器能够促使模型提取更多与任务上下文相关的特征。我们首先检测图像中的一组显著区域，然后

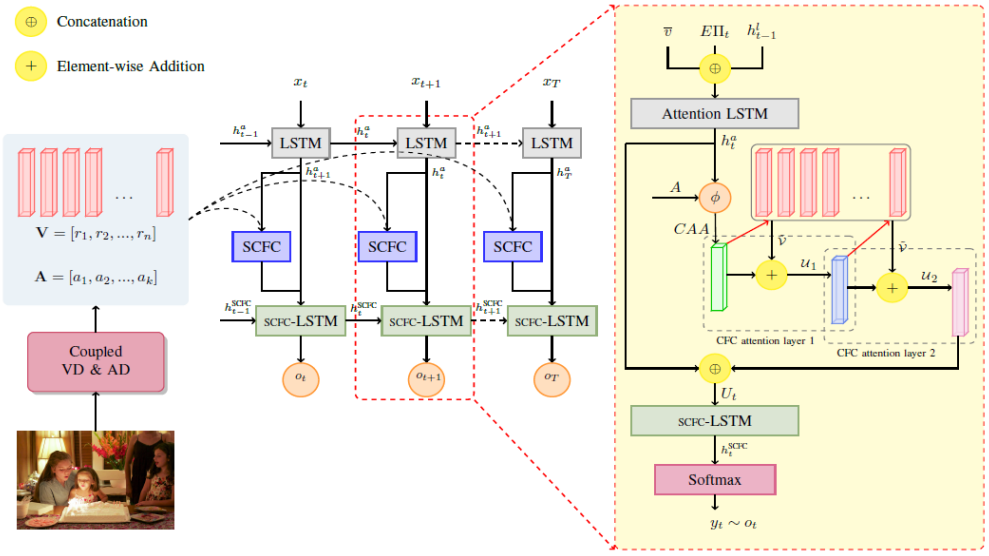


图 1：所提出的用于图像描述的 SCFC 总体架构。首先提取区域提议和属性，然后在每个时间步将它们输入 SCFC 单元，在图像描述生成过程中整合跨模态特征。

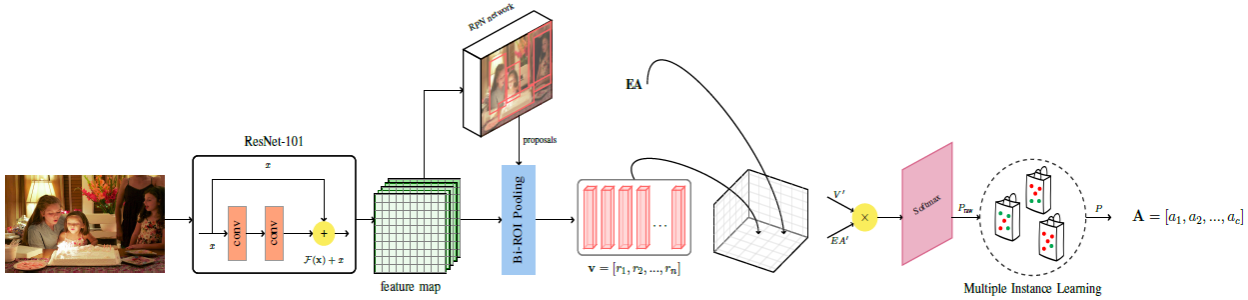


图 2：我们方法中耦合的视觉检测器和属性预测器的架构。给定一幅图像，该图展示了检测视觉区域 V 和属性 A 的过程。

在视觉检测器的基础上构建属性预测器。耦合的 VD 和 AP 的架构如图 2 所示。

3.1.1 视觉检测器

为提取显著候选区域的深层信息，我们采用 ResNet-101 [23] 架构获取输入图像的特征图。我们使用改进的区域提议网络（RPN） [24] ，通过一组矩形区域提议及其对应的置信度分数，检测给定图像中的区域。

在所提出的架构中，ResNet-101 [23], 的最后一个卷积层之后， 3×3 滑动窗口会在特征图上移动，并将其映射到较低维度（例如，256 维）。基于 k 固定纵横比、平移不变的锚框，为特征图的每个 3×3 窗口生成多个可能的区域。因此，回归层会生成表示这些区域边界框的 $4k$ 输出，而分类层会生成表示每个 k 边界框的 softmax 概率、作为置信度分数的 $2k$ 输出。我们将 k 的值设为 9，其中包括 3 种尺度，每种尺度对应 3 种纵横比。最后，使用双线性插值 [25, 26] 改进 [24] 原始 ROI 池化层中的最近邻插值方法，使模型能够从每个区域平滑地提取固定大小的表示$V = \{v_1, \ldots, v_n\}$， $v_i \in \mathbb{R}^h$， $V \in \mathbb{R}^{n \times h}$。此外，双线性插值还允许

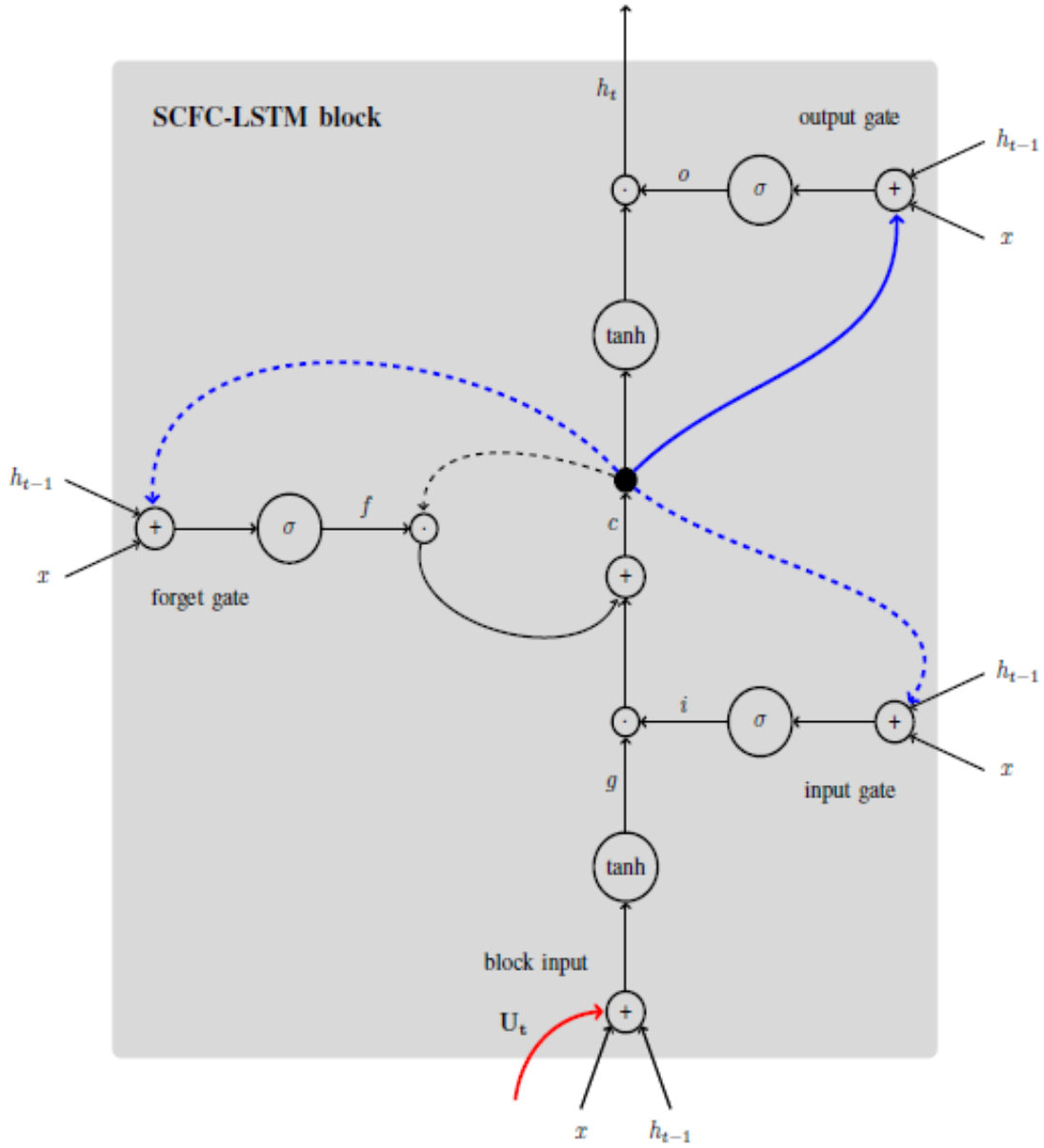


图 3：我们提出的 SCFC-LSTM 的结构。窥孔连接和我们整合后的输入分别以红线和蓝线表示。

通过区域提议进行端到端反向传播。

3.1.2 属性预测器

所提出的属性预测器利用提取出的显著区域，以及由嵌入式 LSTM 生成并作为副产品的属性嵌入 EA，在检测过程中对对象特征与属性之间的相似度进行建模。如图 3 所示，预测给定图像是否包含某个属性分为两个步骤。

首先，我们使用两个全连接层将属性嵌入和对象特征映射到同一空间。然后，通过矩阵乘法将二者结合起来，以衡量相似度。这些步骤的输出输入 softmax 层，从而生成原始概率矩阵 $P_{raw} \in \mathbb{R}^{1000\times k}$，其中 P_{raw}^{ij} 表示 j^{th} 区域包含 i^{th} 属性 a_i的原始概率。 P_{raw} 基于

公式 1。

$$P_{\text{raw}} = \text{sigmoid}\left((W_{\text{AP}}EA)^T \otimes W_v V^T\right) \tag{1}$$

其中， $W_{AP} \in \mathbb{R}^{d\times e}$ 和 $W_v \in \mathbb{R}^{d\times h}$ 是可训练参数。 $E \in \mathbb{R}^{e\times|\Sigma|}$ 表示词汇表 Σ 中所有词的嵌入，嵌入维度为 e； $A \in \mathbb{R}^{|\Sigma|\times c}$ 是 c属性的独热索引矩阵， $\otimes$ 表示矩阵乘法，上标 T 表示转置操作。

在第二步中，使用噪声或多实例学习（MIL）方法 [27] 合并 P_{raw} 每一行中的概率值，生成最终概率 P_i（公式 2），以判断输入图像是否包含 i^{th} 属性 a_i。

$$p_i = 1 - \prod_{j=0}^{n}\left(1 - P_{raw}^{ij}\right) \tag{2}$$

训练耦合视觉检测器和属性预测器时，我们面临两个训练集不平衡问题。RPN网络提出的区域数量可能高达数十万，其中大多数都是负样本，因为区域内没有物体。为克服类别不平衡问题，经典训练方法只会采样固定数量、具有固定物体/非物体分数的样本。此外，真实属性向量较为稀疏，因为真实描述中只出现少数属性。

采用焦点损失 [28] （公式 3）来解决这一问题：训练时使用所有预先定位的具体锚点，并采用动态交叉熵损失。虽然会考虑所有锚点，但通过对简单样本的损失加权，可以避免检测器不堪重负。同样，这种修改也适用于属性预测器：在 RPN 中，我们将真实描述中不存在的属性视为负样本。

$$FL(p) = \begin{cases} -\alpha(1-p)^{\gamma}\log(p), & \text{y=1} \\ -(1-\alpha)p^{\gamma}\log(1-p), & \text{otherwise} \end{cases} \tag{3}$$

我们将视觉检测器损失 $\mathcal{L}_{VD}$ 和属性预测器损失 $\mathcal{L}_{AP}$ 定义为公式 4 和 5，其中 $\mathcal{L}_{reg}$ 是回归层中使用的平滑 $\mathcal{L}1$ 损失。 t_i^* 和 p_i^* 分别是回归目标和目标/非目标标签。 λ 是平衡权重， N_{cls} 和 N_{reg} 是归一化项， N_{pos} 是正属性的数量。

$$\mathcal{L}_{VD} = \frac{1}{N_{cls}}\sum_i FL(p_i) + \lambda\frac{1}{N_{reg}}\sum_i p_i^*\mathcal{L}_{reg}(t_i, t_i^*) \tag{4}$$

$$\mathcal{L}_{AP} = \frac{1}{N_{pos}}\sum_{j=1}^{c} FL(p_j) \tag{5}$$

需要注意的是，在计算 $\mathcal{L}_{VD}$ 和 $\mathcal{L}_{AP}$ 损失时，公式 3 中的 $y = 1$ 表示锚点包含一个目标，而属性 a_j 表示该属性存在于真实描述中。此外，视觉检测器和属性预测器损失所用的超参数 α、 γ 和 λ根据经验分别设为 $0.3|0.95$、 $2|2$和 $10|-$ 。最后， $\mathcal{L}_{VDAP}$ 损失根据公式 6 计算。

$$\mathcal{L}_{VDAP} = \mathcal{L}_{VD} + 0.5\mathcal{L}_{AP} \tag{6}$$

3.2 SCFC注意力网络

我们提出的方法的核心是一个复合函数，用于整合跨模态特征，从而在图像描述生成过程中引导语言模型。该复合函数旨在以多步方式组合跨模态特征，使模型在生成所有语义层级词语时能够关注构建出的判别性特征。在初始步骤中，我们构建输入所提出复合函数的主成分，这些成分来自两种不同的模态：文本 $\mathcal{H}$ 和视觉$\mathcal{V}$。随后，我们定义了一种新颖的递归函数，通过注意力网络复合跨模态特征。

3.2.1 主成分

我们从两种模态中定义主成分：文本 $\mathcal{H}$ 和视觉 $\mathcal{V}$。视觉元素由视觉检测器的输出提供。在这种情况下，我们有一组区域特征图参与复合函数。作为第二个组成部分，我们利用有关上下文环境的语义属性。以往研究采用语义注意力，将高层语义概念纳入图像描述生成过程。基于语义注意力的方法的核心在于，驱动模型根据上下文信息，在每个时间步动态关注语义上重要的属性。这一核心机制通过学习注意力激活状态向量来形成，以计算每个属性的权重。现有方法通过将语言 LSTM 在前一时间步的隐藏状态向量与每个属性向量逐元素相加来实现这一目标。随后，通过 softmax 层计算每个属性的权重，并采用“软”方法，对检测到的属性进行加权求和，得到输出注意力。需要注意的是，输出注意力向量可能包含无关属性，导致注意力引导含糊不清。受 [12] 中一项全面研究的启发，我们提供了两个示例，以便在考虑极端情况时找到衡量相似度的最佳函数，从而更好地理解这一问题。

在第一种情况下，将属性向量 A 视为全 1 向量，而语言 LSTM h 的隐藏状态是全 0 向量，表示不存在上下文信息。由于 h 是全 0 向量，属性向量与上下文之间不存在相关性。因此，注意力激活状态向量应为全 0 向量，表示上下文与属性向量之间不存在关联。在第二种情况下，假设属性向量 A 为全 0 向量，语言 LSTM h 的隐藏状态为全 1 向量，这表明给定图像中不存在高层语义概念。因此，相应的注意力激活状态向量应为全 0 向量，表示上下文与属性向量之间不存在相关性。尽管两个示例中的预期结果均为全 0 向量，但传统语义注意力机制得到的注意力激活状态向量却是全 1 向量，导致为检测到的属性添加了不准确的权重。从这些示例中可以发现，在某些情况下，“软”注意力机制可能缺乏适当衡量上下文信息与高层语义概念之间相关性的方式。本文提出的上下文感知属性 CAA 通过构建一个衡量预测属性与上下文信息之间相关性的函数，旨在解决这一问题。此外，我们使用注意力 LSTM [22] 来表示上下文信息，而不是采用语言 LSTM 的隐藏状态，从而能够在当前语言上下文中对图像上下文条件进行恰当的动态表示。

如公式 7 所示，注意力 LSTM 将经过均值池化的图像特征 $\bar{v} = \frac{1}{N}\sum_i v_i$、SCFC LSTM 的前一隐藏状态以及此前生成词的编码作为输入。这些输入分别为注意力 LSTM 提供有关语言 LSTM 状态、图像整体内容以及迄今生成的部分图像描述输出的最充分上下文。

$$h_t^{att} = LSTM[(h_{t-1}^{SCFC} \oplus \bar{v} \oplus E\Pi_t), h_{t-1}^{att}] \tag{7}$$

其中，E 表示与属性检测模块共享的词嵌入矩阵，Π_t 表示时间步 t 的输入词的独热编码。此外，我们选择线性函数来衡量预测属性与上下文信息之间的相关性，因为这类函数在计算效率和简洁性方面成本较低。出于细化目的，我们将 Φ_{cr} 定义为逐元素乘法。因此，可使用公式 8 和 9 计算各属性 a_t^j 上的语义注意力分布 β_t^j。

$$b_t^j = \Phi_{cr}(h_t^{att}, EA_t^j) \tag{8}$$

$$\beta_t^j = softmax(b_t^j) \tag{9}$$

其中，b_t^j 表示每个属性的注意力激活状态向量。然后，我们可以根据公式 10，将每个属性的嵌入乘以其权重，从而构建上下文感知属性 CAA_t。

$$CAA_t = \sum_{j=1}^{c}(\beta_t^j \cdot EA_t^j) \tag{10}$$

需要注意的是，所提出的上下文感知属性向量旨在表示与上下文信息相关联的属性。直观而言，最终属性集的共存能够在图像描述生成过程中，通过每个时间步动态保留上下文。

总之，区域特征图、属性的语义嵌入以及上下文信息共同构成了复合函数的主要组成部分。基于这些信息，我们将视觉元素 $V = \{v_1, \ldots, v_n\}$， $v_i \in \mathbb{R}^h$， $V \in \mathbb{R}^{n \times h}$ 作为第一项，并将我们提出的上下文感知属性 CAA 作为来自文本模态的第二项 $\mathcal{H}$ 。

3.2.2 复合函数

给定主要组成部分 $\mathcal{V}$ 和 $\mathcal{H}$，我们提出一种新的递归函数，以多步方式整合跨模态特征。具体而言，我们定义了一个 SCFC 注意力网络来执行整合。随后，我们搭建一个由更多 SCFC 组成的堆叠网络，使其协同工作以模拟多步推理。首先，我们介绍 CFC，它是一个功能完备、可独立运行的网络；本节其余部分将探讨堆叠网络如何作为递归复合函数运行。

在 SCFC 中，我们首先测量各主要组成部分两两之间的模态间关系，以确定跨模态特征的相关度。与前文相同，我们使用 Φ_{cr} 来计算相关度，如下所示：

$$\alpha_{i,t} = \tanh(\Phi_{cr}(W_{\mathcal{V},\alpha} v_i, W_{\mathcal{H},\alpha} \mathcal{H}_t)) \tag{11}$$

$$\mathcal{D}_{i,t} = softmax(W_{\alpha,\mathcal{D}}^T \alpha_{i,t}) \tag{12}$$

其中，$W_{\mathcal{V},\alpha} \in \mathbb{R}^{d \times h}$、$W_{\mathcal{H},\alpha} \in \mathbb{R}^{d \times e}$ 和 $W_{\alpha,\mathcal{D}}^T \in \mathbb{R}^h$ 是学习得到的参数。通过确定文本成分 $\mathcal{H}$ 与视觉子成分 v_i之间的相关程度 $\mathcal{D}_{i,t}$，我们计算注意力加权的视觉成分 $\mathcal{V}$ 如下：

$$\tilde{\mathcal{V}}_t^I = \sum_i \mathcal{D}_{i,t} \mathcal{V}_i \tag{13}$$

计算得到的 $\tilde{\mathcal{V}}_t^I$表示视觉元素，该视觉元素在每个时间步 t与文本成分相结合。为了强化文本信息的影响，我们将紧凑的跨模态表示$\mathcal{U}_t$ 定义为二阶组合，如下所示：

$$\mathcal{U}_t = \tilde{\mathcal{V}}_t^I + \mathcal{H}_t \tag{14}$$

输出 $\mathcal{U}_t$ 是一种信息丰富的表示，其中编码了给定图像的主成分 $\mathcal{V}$ 和 $\mathcal{H}$ 。与仅采用组合视觉注意力和语义注意力的模型相比，我们的模型通过对与上下文感知属性更相关的视觉元素赋予更高权重，构建了更丰富的 $\mathcal{U}_t$ 。然而，对于抽象语义词，单个CFC不足以对跨模态特征进行推断并生成全面的表示。因此，我们堆叠SCFC，迭代上述过程，以模拟多模态任务中的多步推理方法。数学上，对于 s^{th} CFC层，SCFC采用以下公式：

$$\alpha_{i,t}^s = \tanh(\Phi_{cr}(W_{\mathcal{V},\alpha}^s v_i, W_{\mathcal{H},\alpha}^s \mathcal{H}_t^{s-1})) \tag{15}$$

$$\mathcal{D}_{i,t}^s = softmax(W_{\alpha,\mathcal{D}}^{s,T} \alpha_{i,t}^s) \tag{16}$$

其中， $\mathcal{H}_t^{s-1}$是每个时间步上来自前一SCFC层的文本元素。需要注意的是，我们使用上下文感知属性向量初始化 H_t^0。在每个推理步骤中，注意力加权的视觉成分 $\mathcal{V}$ 和跨模态表示 $\mathcal{U}_t$ 如下获得：

$$\tilde{\mathcal{V}}_t^{I,s} = \sum_i \mathcal{D}_{i,t}^s \mathcal{V}_i \tag{17}$$

$$\mathcal{U}_t^s = \tilde{\mathcal{V}}_t^{I,s} + \mathcal{H}_t^{s-1} \tag{18}$$

我们重复该算法 S 次，然后将最终的 $\mathcal{U}_t^S$与上下文表示 h_t^{att}按照公式 7进行拼接，并将其输入SCFC-LSTM，以生成语义细粒度的句子。

$$U_t = \mathcal{U}_t^S \oplus h_t^{att} \tag{19}$$

3.3 使用 SCFC-LSTM 生成图像描述

LSTM 有多种变体。在我们的工作中，我们采用窥孔 LSTM [29, 30] 模型作为描述生成器，其表达式如下：

$$\begin{aligned} i_t &= \sigma(W_i x_t + R_i h_{t-1} + P_i c_{t-1} + b_i) \\ f_t &= \sigma(W_f x_t + R_f h_{t-1} + P_f c_{t-1} + b_f) \\ o_t &= \sigma(W_o x_t + R_o h_{t-1} + P_o c_t + b_o) \\ g_t &= \tanh(W_g x_t + R_g h_{t-1} + b_g) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{20}$$

其中， i_t、 f_t、 o_t、 c_t 和 h_t 分别表示窥孔 LSTM 的输入门、遗忘门、输出门、记忆单元和隐藏状态。 σ 表示 Sigmoid 函数， x_t 表示时间步 t的词输入。 $P_* \in \mathbb{R}^N$、 $W_* \in \mathbb{R}^{N \times M}$、 $R_* \in \mathbb{R}^{N \times N}$和 $b_* \in \mathbb{R}^N$ 分别表示窥孔权重、输入权重、循环权重和偏置权重，其中 N 为 LSTM 块的数量， M 为输入的维度。值得一提的是，添加窥孔连接意味着让门层查看单元状态。换言之，在确定输入门、遗忘门和输出门时，需要使用前一时间步的单元状态 c_t。

以往研究会在每个时间步，将与特定输入位置或高层语义概念相关的语义特征作为语言模型的额外输入。相反，我们的模型提供跨模态融合特征 U 来指导描述生成。如图 1 所示，我们在每个时间步将构建的特征 U 加载到单元状态 C 中。直观而言，窥孔连接还使门在当前时间步能够以更可控的方式感知跨模态信息语义特征 U 。因此，式 20 中的第四行必须更新如下：

$$g_t = \tanh(W_g x_t + W_U U_t + R_g h_{t-1} + b_g) \tag{21}$$

其中， W_U 表示权重矩阵。最后，预测词汇表p_t 中每个词的概率分布，以及时间步 t 要生成的词：

$$h_t^{SCFC} = LSTM^P(h_{t-1}^{SCFC}, x_t, \mathcal{U}_t) \tag{22}$$

$$y_t \sim p_t = softmax(W^h h_t^{SCFC}) \tag{23}$$

其中， W表示权重矩阵， $LSTM^P$表示带窥孔连接的 LSTM。

3.4 模型训练

模型训练过程分为两轮。第一轮中，给定目标真实序列$y_{1:T}^*$，我们使用式 24 所示的经典交叉熵（XE）损失对模型进行优化，其中 θ 表示描述模型参数。

$$\mathcal{L}_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^* | y_{1:t-1}^*)) \tag{24}$$

在第二轮中，我们采用深度强化学习（RL）来解决暴露偏差问题。所谓暴露偏差，是指模型从未接触过自身生成的预测，导致推理过程中错误不断累积。基于经交叉熵训练的模型初始化，我们探讨如何最小化与模型参数对应的负期望得分，如下所示：

$$\mathcal{L}_R(\theta) = -E_{y_{1:T}\sim p_\theta}[r(y_{1:T})] \tag{25}$$

其中， r 是采用 CIDEr-D [31]得分优化的评估指标。考虑到自我批判序列训练（SCST） [32]，我们使用 REINFORCE 算法近似梯度，公式如下：

$$\nabla\mathcal{L}_R(\theta) \approx -(r(y^s_{1:T}) - r(\hat{y}_{1:T}))\nabla_\theta \log p_\theta(y^s_{1:T}) \tag{26}$$

其中， $y^s_{1:T}$ 是随机采样的描述， $r(\hat{}y_{1:T})$是最大采样描述的基线得分。

我们提出的模型采用单层优化进行训练。具体而言，模型通过整体损失函数 $\mathcal{L}_o$进行优化，如式 27 所示。 $\mathcal{L}_{VDAP}$ 定义见式 6，而 $\mathcal{L}_{cap}$ 分别表示第一轮和第二轮中的交叉熵损失与强化学习损失。

$$\mathcal{L}_o = \mathcal{L}_{cap} + \mathcal{L}_{VDAP} \tag{27}$$

4 实验

4.1 数据集

我们在多个知名数据集上验证了所提出的模型，包括 MSCOCO [33] 和 Flicker30k [34]。MSCOCO 包含 82,783 张训练图像和 40,504 张验证图像。每张图像都有五条标注句子。我们采用了先前研究中广泛用于报告结果的 Karpathy 划分 [1], 。该划分包含 113,287 张训练图像，以及 5K 张验证图像和测试图像。Flickr30k 由从 Flickr 获取的 31,783 张图像组成。为确保公平比较，我们使用公开划分 [1] ，其中包含 29,783 张训练图像、1K 张验证图像和 1K 张测试图像。与 MSCOCO 一样，每张图像都标注有五条参考描述。

4.2 评估指标

我们采用图像描述领域常用的指标报告性能，包括 BLEU-N [35], METEOR [36], CIDEr [31], ROUGE-L [37], 和 SPICE [38]。我们使用 Microsoft COCO 评估服务器发布的代码计算所有指标。BLEU 通过衡量生成句子与参考句子之间 n 元语法的相似度来计算。METEOR 通过比较候选图像描述与参考图像描述中的各个句段来评估。CIDEr 是一种专为图像描述任务提出的基于共识的指标，通过对所有 n 元语法进行 TF-IDF 加权来计算图像描述中的共识。ROUGE-L 用于评估机器翻译的充分性和流畅度，它利用候选句子与一组参考句子之间的最长公共子序列，在句子层面衡量它们的相似度。SPICE 则通过对参考场景图和候选场景图中的元组计算 F 分数来确定，这些场景图由依存句法树生成。

4.3 实验设置

4.3.1 数据预处理

我们采用标准做法，仅进行最少量的文本预处理：按空白字符切分词元，将所有词元转换为小写，并剔除出现次数少于五次的稀有词，最终 MSCOCO 和 Flickr30k 的词汇表分别包含 10,010 个和 6,864 个词。具体而言，我们将出现频率较低的词替换为特殊词元 $< UNK >$。

4.3.2 实现细节

我们的真实值包含正样本框和属性向量，其中坐标标注取自 MSCOCO-2014 目标检测数据集。为构建真实值属性，我们选取训练描述语料库中最常见的 c 个词，并在 [39] 中将 $c = 1000$ 设为相应值。然后，我们构造一个属性向量，用于判断每个词是否出现在每幅图像的描述中。我们采用 ResNet-101

其权重在 ImageNet 上预训练 [40] ，且在所有训练阶段均不进行微调。我们的视觉检测器根据相应的置信度分数提取区域特征，其维度为 36×2048，具体做法参见 [22]。与 [21], 类似，属性与对应的词共享相同的嵌入，每个词嵌入到一个 1000 维的词嵌入空间中。注意力 LSTM 和 SCFCLSTM 的隐藏状态维度均设为 2048。我们使用 Adam 优化器 [41] ，并采用 $\beta_2 = 0.9$、 $\beta_2 = 0.999$，以最小化式（26）中的损失函数。基础学习率为 1×10^{-4}。我们使用 Dropout [42] 和梯度裁剪技术 [43] 。在第一轮训练中，我们采用教师强制学习[44], : 提供截至 $t-1$ 的真实值词语，而不是模型在前一个时间步生成的词语，以训练序列学习在 t 处的预测。在测试阶段，句子的最大允许长度设为 20。我们采用束搜索策略，束宽设为 3。

4.3.3 最先进研究

通过将我们的方法与以下最先进的研究成果进行性能比较，我们验证了该方法的有效性:

1. NIC [14]: 首个编码器-解码器框架，将图像输入编码器，并在长短期记忆网络解码器的第一个时间步输入编码后的表示，以生成相应的描述。2. Soft-Att [8] 和 Hard-Att [8]: 引入两种不同的空间注意力机制，引导模型选择性地关注显著图像区域，分别采用确定性的“Soft”注意力或随机性的“Hard”注意力。3. Sem-Att [9]: 首先，将图像中的语义概念作为独立任务进行检测。随后，将全局图像特征与检测到的语义概念相结合，并在图像描述生成过程中逐步输入语言模型。4. LSTM-A [11]: 他们提出了一种新颖的属性检测器，将属性间相关性融入多实例学习（MIL），以利用属性之间的相关性。随后，他们采用五种不同的方式，将这些语义属性融入长短期记忆网络。5. SCA-CNN [45]: 这是视觉注意力的一种改进版本，融合空间、通道和多层图像特征，以动态调整句子生成的上下文。6. SCST [32]: 提出了一种先进的、基于强化学习的图像描述训练方法。7. Ada-Att [20]: 提出结合视觉哨兵的自适应注意力，以确定应关注视觉特征还是视觉哨兵。8. RFNet [46]: 他们提出了一种新颖的循环融合网络（RFNet），能够利用图像编码器输出之间的交互，并为解码器生成新的紧凑且信息丰富的表示。9. Up-Down [22]: 引入一种结合自上而下与自下而上注意力的机制，使注意力能够在物体及其他显著图像区域的层面上确定。10. GCN-LSTM [47]: 他们在图像中检测到的物体上构建了一个图卷积网络，并在编码器-解码器框架中将物体的语义和空间关系融入图像编码器。11. ARL [48]: 这项工作探讨了将区域内关系融入视觉内容的解决方案，并提出一种新颖的视觉注意力机制，以隐式建模图像区域之间的关系。12. SGAE [49]: 提出了一种场景图自编码器，利用图像的场景图和训练得到的词典，将搭配关系和上下文推理作为语言归纳偏置融入编码器-解码器架构。

表1：所提出的 SCFC 与其他最先进方法在 MSCOCO KARPATHY 测试集上的性能分析，其中 B@N、M、R、C 和 S 分别是 BLEU-N、METEOR、ROUGE-L、CIDER-D 和 SPICE 分数的简称。每列中的最高分数以粗体标出。

	Cross-Entropy Loss						CIDEr-D Score Optimization					
	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
NIC [14]	66.6	24.6	-	-	-	-	-	-	-	-	-	-
Soft Att [8]	70.7	24.3	23.9	-	-	-	-	-	-	-	-	-
Hard Att [8]	71.8	25.0	23.0	-	-	-	-	-	-	-	-	-
Sem-Att [9]	70.9	30.4	24.3	-	-	-	-	-	-	-	-	-
LSTM-A [11]	73.4	54.0	100.2	18.6	78.6	35.5	27.3	56.8	118.3	20.8		
SCST:Att2in [32]	-	31.3	26.0	54.3	101.3	-	-	33.3	26.3	55.3	111.4	-
SCST:Att2all [32]	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
SCA-CNN [45]	71.9	31.1	25.0	53.1	95.2	-	66.6	20.3	27.3	56.8	118.3	-
Ada-Att [20]	74.2	33.2	26.6	-	108.5	-	-	-	-	-	-	-
RFNet [46]	76.4	35.8	27.4	56.5	112.5	20.5	79.1	36.5	27.7	57.3	121.9	21.2
Up-Down [22]	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
GCN-LSTM [47]	77.4	37.1	28.1	57.2	117.1	21.1	**80.9**	38.3	28.6	58.5	128.7	22.1
SGAE [49]	77.6	36.9	27.7	57.2	116.7	20.9	80.8	38.4	28.4	58.6	127.8	22.1
CAVP [50]	-	-	-	-	-	-	-	38.6	28.3	58.5	126.3	21.6
LBPF [51]	**77.8**	37.4	28.1	57.5	116.4	21.2	80.5	38.3	28.5	58.4	127.6	22.0
MAD+SAP [21]	-	37.0	28.1	57.2	117.3	**21.3**	-	38.6	28.7	58.5	128.8	**22.2**
SCFC	**77.8**	**37.5**	**28.2**	**57.6**	**119.2**	**21.3**	80.7	**38.9**	**29.0**	**58.7**	**129.9**	22.1
Ensemble												
SCST $_{fuse}$ [32]	-	32.8	26.7	55.1	106.5	-	-	35.4	27.1	56.6	117.5	-
RFNet $_{fuse}$ [46]	77.4	37.0	27.9	57.3	116.3	21.1	80.4	37.9	28.3	58.3	125.7	21.7
SGAE $_{fuse}$ [50]	-	-	-	-	-	-	81.0	39.0	28.4	**58.9**	129.1	22.2
MAD+SAP $_{fuse}$ [21]	-	-	-	-	-	-	-	39.0	28.9	58.8	129.8	**22.3**
SCFC[Σ]	**78.0**	**37.9**	**28.2**	**57.8**	**119.9**	**21.2**	**81.2**	**39.2**	**29.1**	**58.9**	**130.8**	**22.3**

13. CAVP [50]：他们提出了一种新的基于强化学习的学习方法，并在解码器中引入了一种成对关系学习方法。

14. LBPF [51]：他们提出了一种回顾方法，用于嵌入先前的视觉信息，并提出向前预测策略来展望未来，通过利用语言连贯性提升图像描述性能。

15. MAD+SAP [21]：这项工作通过引入后续属性预测器模块扩展了语义注意力，该模块会在每个时间步动态预测一个精简的属性子集，以减轻图像属性的多样性带来的影响。

4.3.4 定量结果

表1展示了在 MS-COCO 数据集上的对比性能结果。表中显示，我们提出的模型在所有评估指标上均大幅优于最先进模型，尤其是在 CIDEr 指标上。为确保公平比较，我们还分别报告了集成模型的结果。这些评估结果表明，SCFC注意力网络提升了图像描述性能。与传统的神经图像描述模型 NIC [14], 、经典视觉注意力方法 Soft-Att [8], 和 Hard-Att [8] 等基准方法相比，我们的提升主要归功于视觉信息与语义信息之间更有效的协同。与 SCA-CNN [45] 和 LBPF [51], 相比，我们的模型利用先进的语义概念生成更多样且语义更丰富的描述。尽管 Sem-Att [9] 和 LSTM-A [11] 利用了属性，但我们的模型能够更准确地衡量属性与上下文信息之间的相关性，而且无需使用独立网络训练属性检测器。与那些在构建图像描述网络之前，先用大型数据集（如 Visual Genome）预训练视觉检测器以探索视觉语义区域的方法（Up-Down [22], 、MAD+SAP [21]）相比，我们利用跨模态特征整合层弥补判别性语义特征表示的不足。RFNet [46] 则通过研究空间注意力机制进一步提升了图像描述性能。

相比之下，我们采用一种同时考虑语义概念和空间区域的注意力机制。ARL [48] 和GCN-LSTM [47] 通过发现图像中各区域之间编码语义概念的高层连接，来考虑区域间的视觉关系，从而提升图像描述性能。

表 2：采用交叉熵损失时，所提出的 SCFC 与其他最先进方法在 Flicker30K 公开划分上的性能分析，其中 B@N、M、R、C 和 S 分别是 BLEU-N、METEOR、ROUGE-L、CIDER-D 和 SPICE 分数的缩写。每列中数值最高的项以粗体标出。

	B@1	B@4	M	R	C	S
NIC [14]	66.3	18.3	-	-	-	-
Soft Att [8]	66.7	19.1	18.49	-	-	-
Hard Att [8]	66.9	19.9	18.46	-	-	-
Sem-Att [9]	64.7	23.0	18.9	-	-	-
SCA-CNN [45]	66.2	22.3	19.5	-	-	-
Ada-Att [20]	67.7	25.1	20.4	-	53.1	-
ARL [48]	69.8	27.7	21.5	48.5	57.4	-
SCFC[3]	**71.2**	**27.9**	**22.2**	**49.2**	**58.1**	**15.9**

表 3：所提出的 SCFC 与其他最先进方法在 MSCOCO 在线测试服务器上的性能分析。各项指标的上标表示排名前两位。

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr-D	
	C5	C40	C5	C40	C5	C40	C5	C40	C5	C40	C5	C40	C5	C40
NIC [14]	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6
Hard Att [8]	70.5	88.1	52.8	77.9	38.3	65.8	27.7	53.7	24.1	32.2	51.6	65.4	86.5	89.3
Semantic Att [9]	73.1	90.0	56.5	81.5	42.4	70.9	31.6	59.9	25.0	33.5	53.5	68.2	94.3	95.8
SCA-CNN [45]	71.2	89.4	54.2	80.2	40.4	69.1	30.2	57.9	24.4	33.1	52.4	67.4	91.2	92.1
Ada-Att [20]	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9
RFNet [46]	80.4	95.0	64.9	89.3	50.1^{2}	80.1^{2}	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
SCST:Att2all [32]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
LSTM-A [11]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	62.5	27.0	35.4	56.4	70.5	116.0	118.0
Up-Down [22]	80.2	95.2^{1}	64.1	88.8^{2}	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
SGAE [49]	-	-	-	-	-	-	37.8	68.7	28.1	37.0	58.2	73.1	122.7	125.5
CAVP [50]	80.1	94.9	64.7	88.8^{2}	50.0	79.7	37.9	69.0	28.1	37.0	58.2	73.1	121.6	123.8
MAD+SAP [21]	80.5^{1}	94.9^{2}	65.1^{1}	89.1^{1}	50.4^{1}	80.0	38.4^{2}	69.4^{2}	28.6^{2}	37.7^{2}	58.7^{1}	73.3^{2}	125.1^{2}	127.0^{2}
SCFC	80.2^{2}	94.9^{2}	64.9^{2}	88.8^{2}	50.4^{1}	80.2^{1}	38.5^{1}	69.8^{1}	29.0^{1}	38.3^{1}	58.6^{2}	73.8^{1}	125.3^{1}	127.3^{1}

与此同时，我们的模型可以通过SCFC层在复合函数中捕捉这些区域。CAVP [50] 提出了一种新的基于强化学习的学习方法，并将其用于我们的学习过程中，以提升图像描述性能。MAD+SAP [21]利用视觉信息和语义信息，并着重关注属性的作用。考虑到视觉信息与语义信息互补的特性，我们提出了一种全新的端到端模型，并设计了新颖的整合层，以高效生成更多细粒度描述。与需要预训练场景图生成器和词典的SGAE [49]，不同，我们的模型仅以视觉特征作为输入。

为进一步验证所提模型的有效性，我们在表 2 中报告了 Flickr30k数据集测试划分的评估结果。我们在 MSCOCO 在线测试服务器上评估所提模型。表 3 报告了 SCFC 与其他最先进方法的性能。所提方法在大多数指标上也取得了最佳表现。

4.4 消融研究

4.4.1 递增验证

我们还进行了大量实验，逐步验证我们的方法，并全面展示所提出方法的行为。我们通过以下部分对模型进行消融：

表4：消融模型（Base、Base+视觉检测器语义注意力、Base+上下文感知属性、上下文感知属性+SCFC）以及我们的完整模型上下文感知属性+SCFC+SCFC-LSTM(N)在CIDER-D分数优化训练中，采用不同CFC层数时，在MSCOCO测试集划分上的结果，指标包括BLEU-N（B@N）、METEOR（M）、ROUGE-L（R）、CIDER-D（C）和SPICE（S）。

	B@1	B@4	M	R	C	S
Base	79.8	37.0	27.7	57.4	121.9	21.4
Base+ADsemAtt	80.0	37.4	28.1	57.7	123.4	21.5
Base+CAA	80.2	37.6	28.3	57.9	124.8	21.7
CAA+SCFC [1]	80.5	37.9	28.6	58.3	126.1	21.9
CAA+SCFC+SCFC-LSTM [1]	80.6	38.1	28.7	58.4	127.2	22.0
CAA+SCFC+SCFC-LSTM [2]	**80.7**	38.6	28.9	58.5	128.7	**22.2**
CAA+SCFC+ SCFC-LSTM [3]	**80.7**	**38.9**	**29.0**	**58.7**	**129.9**	22.1

- Base：我们参考 [22] 构建了一个基线模型，将注意力 LSTM 与语言 LSTM 相结合，并搭配我们的视觉检测器。请注意，与 [22]，不同，我们的视觉检测器可以与整个图像描述网络联合训练。

- Base+视觉检测器语义注意力：我们通过传统的语义注意力，将属性检测器作为额外输入注入语言 LSTM。

- Base+上下文感知属性：为展示所提出的上下文感知属性的有效性，我们通过上下文感知属性模块，将检测到的属性纳入Base模型。

- 上下文感知属性+SCFC：我们在消融模型Base+上下文感知属性中加入所提出的堆叠式跨模态特征整合注意力网络。与完整模型相比，该变体未采用SCFC-LSTM。

- 上下文感知属性+SCFC+SCFC-LSTM：采用上下文感知属性、SCFC注意力网络和SCFC-LSTM的所提出模型。

消融模型的实验结果见表4。括号中的数字表示SCFC注意力网络中的堆叠层数。我们的基础模型取得了与 [22]相当的性能。Base+VDsemAtt和Base+CAA的结果表明，在构建文本成分时使用属性，比单独利用视觉元素（Base）能够取得更好的结果。相比之下，我们的上下文感知属性模块显著提升了模型性能。具体而言，我们的消融模型Base+CAA在BLEU-4和CIDEr指标上分别达到37.6和124.8，相较于采用经典语义注意力的基线模型，分别提升了0.6%和2.9%。所提出的CAA在BLEUN指标上优于传统语义注意力。将检测到的属性贯穿图像描述生成过程，并考察它们与当前步骤上下文环境的一致程度，可以引导模型关注更多与上下文相关的属性，从而生成有意义且流畅的句子。对CAA+SCFC+SCFC-LSTM(s)的实验验证了我们提出的堆叠式跨模态特征整合注意力网络的有效性；该完整模型采用了不同数量的CFC层。具体而言，我们的模型CAA+SCFC+ SCFC-LSTM(3)在BLEU-1、BLEU-4、METEOR、ROUGE-L、CIDEr和SPICE指标上分别提升了0.9%、1.9%、1.3%、1.3%、8.0%和0.7%。此外，与CAA+SCFC(1)消融模型相比，我们的模型CAA+SCFC+SCFC-LSTM(1)取得了更好的结果，证明采用SCFC-LSTM而非标准LSTM具有积极作用。

这些评估结果表明，SCFC层是有效的。通过将完整模型与其变体进行比较，不难判断所提出的SCFC能否结合互补特征，在每个时间步动态调制给定图像更丰富的语义表示。评估

结果进一步证实了这一结论：包含多个CFC层的模型表现更佳。由三个堆叠的CFC层构成的完整模型，其CIDEr指标相对提升了8.5%，在生成语义丰富的描述方面表现突出。

4.4.2 上下文感知属性分析

图 4 展示了在我们扩展后的模型中，每个步骤通过所提出的上下文感知属性模块产生的语义注意力权重的影响。为使图示简洁，我们仅展示属性检测器输出中三个可能属性的注意力权重。在生成句子的过程中，相应的、不同时变的权重会根据当前上下文进行调整。

为了更好地理解我们的模型，我们总结了生成的图像描述中 CAA 权重变化的三种情形，以及属性所发挥的作用。首先，高概率属性会获得较高权重，并用于生成的图像描述。当某个属性能够与定义明确的形状区域或描述动作的词语相匹配时，几乎都会出现这种情况。例如，在图 4(c) 中，“man”（男人）、“playing”（玩）和“frisbee”（飞盘）这些词语在恰当的时间步被选中，这说明我们的 CAA 能够准确地关注这些类别的词语。其次，某些高概率属性虽然会获得较高权重，却不会参与生成图像描述。当属性与上下文信息不一致时，就会出现这种情况。例如，在图 4(b) 中，尽管“kitchen”（厨房）一词被预测为高概率（0.97%），但其权重并不高。此外，当生成“decorating”（装饰）和“cake”（蛋糕）等词语时，该属性会引导模型的语义倾向，这正是 CAA 模块的作用。第三，某些高概率属性虽然未能通过 CAA 获得较高权重，却仍会出现在生成的图像描述中。这种情况表明，CAA 能够利用口语或书面语中的常见词语。图 4(a) 中的“city street”（城市街道）就是一个很好的例子：尽管该词的预测概率较低，模型仍会关注“city”（城市）一词。

4.4.3 SCFC注意力层的作用

在视觉词和非视觉词的生成方面，特征复合有利于生成图像描述。比较SCFC注意力网络生成的句子与Base模型生成的句子，可以验证这一点。具体而言，SCFC层中的复合函数旨在将语义属性与视觉区域结合起来，从而生成细粒度描述。

如图5所示，我们将不同数量的CFC层生成的图像描述与基线模型进行比较。该图表明，我们的SCFC注意力网络使模型能够对输入图像进行推理。具体而言，使用更多CFC层生成的描述包含由某些区域推断出的高层次词语，以及这些区域与编码语义概念的某些属性之间的关联。例如，图5左下方的“洗涤”和右上方的“捕捉”都是这类高层次词语，我们的模型能够准确生成。此外，图中右上方的短语“在计算机实验室工作”是通过多个推理步骤生成概念性表达的绝佳示例。

值得注意的是，包含一个 CFC 层的图像描述中，关于上下文信息的视觉和非视觉词语兼有，这表明单个 CFC 层的能力很强。例如，在图5第一行的第二个示例中，作为非视觉词汇的“standing”和“looking”，以及作为视觉词的“sink”和“window”，说明我们的模型能够有选择地关注视觉和非视觉词语，从而生成更好的描述。

4.4.4 模型集成

长期以来，集成 [52] 一直是提升机器学习系统性能的一种直接而有效的方法。在深度架构中，只需针对同一任务独立训练多个模型，可适当改变某些训练条件，然后汇总它们的预测结果，以获得推理时的最终结果。为进行公平比较，我们还报告了三个具有不同堆叠SCFC层数（1至3层）的SCFC模型的集成性能。结果如表1中的SCFC[Σ]所示。与各个单独模型的结果相比，集成模型显著提升了性能，这意味着不同的语义层级词语需要不同数量的推理步骤。这一成果表明，动态调整堆叠SCFC层数值得进一步研究。



图4：图像描述生成过程中注意力权重变化的示例。

4.4.5 参数规模分析

增加堆叠SCFC层的数量，可以模拟生成细粒度描述所需的多步推理过程。此外，加深SCFC会增加整体参数规模，可能减缓收敛速度。为进一步验证SCFC加深的影响，我们进行了以下两项实验：

- BU-SCFC：将我们的视觉检测器替换为预训练的自底向上注意力。
- Gwf-SCFC：将我们的词嵌入替换为未经微调的预训练 Glove。

实验结果如图6所示。图中显示，在消融模型中增加SCFC的数量可以进一步提升描述性能。模型包含三个SCFC层时，这种提升尤为显著。具体而言，与堆叠相同数量SCFC的SCFC模型相比，消融模型BU-SCFC的BLEU-4相对提升了0.34%。尽管这一提升十分显著，但代价是无法进行端到端训练。需要指出的是，在其他任何对比结果中，我们都没有使用这些参数缩减技巧。

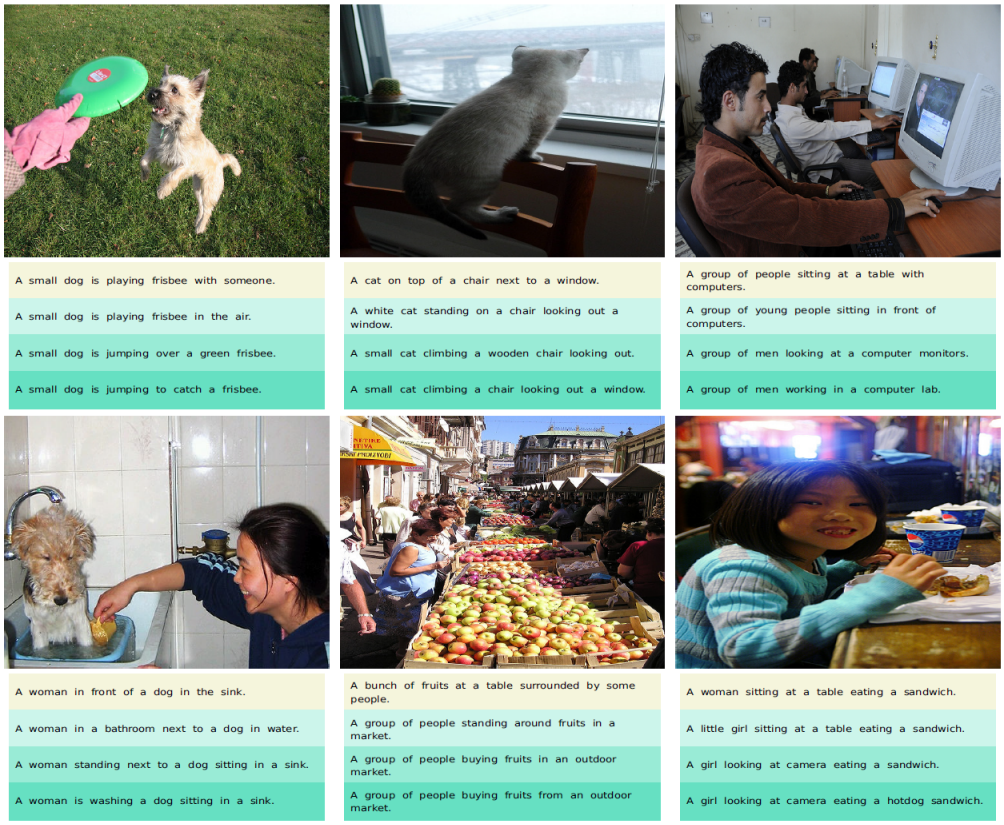


图5：图像描述结果示例。从黄色框到绿色实线框分别展示了由基础模型、SCFC 以及具有一层、两层和三层CFC层的模型生成的描述。

4.4.6 束宽 k 的影响

在测试阶段，句子生成中预测下一个词元有两种标准方法。第一种是贪心方法，即在每个时间步确定得分最高的词元，并用它来预测下一个词元，直到得到结束标记或达到图像描述的最大长度。第二种方法是使用束搜索，在每个时间步选出最佳的 k 个子序列，并将它们作为候选项，以便在下一个时间步生成最佳的 k 个子序列。为了考察测试步骤中束宽的影响，我们在MSCOCO数据集上，分析了所提出的递增模型在不同CFC层数和不同束宽下的表现，束宽取常用范围1至5。图7展示了归一化步骤依据各评估指标的最高分数所得的结果。从图中可以看出，随着束宽 k 增加，所有模型变体的BLUE-4都会提升，这表明在本实验设定的范围内，增大束宽可以生成更流畅的句子。相比之下，当束宽 k 为3时，CIDEr指标达到峰值；但在包含一个CFC层的模型中，束宽为4时的结果略优于束宽为3时的结果。

5 结论

本文提出了一种用于图像描述的 SCFC注意力网络，通过对跨模态特征进行多步推理，在图像描述生成过程中获得紧凑且信息丰富的表示。我们证明，融合多模态信息能够促进判别性特征的生成，从而生成各层次的语义词。我们首先利用感知上下文的属性构建文本成分，并通过图像描述网络与视觉检测器的联合训练来训练该成分。随后，我们提出 SCFC注意力网络，多次整合跨模态特征，以模拟

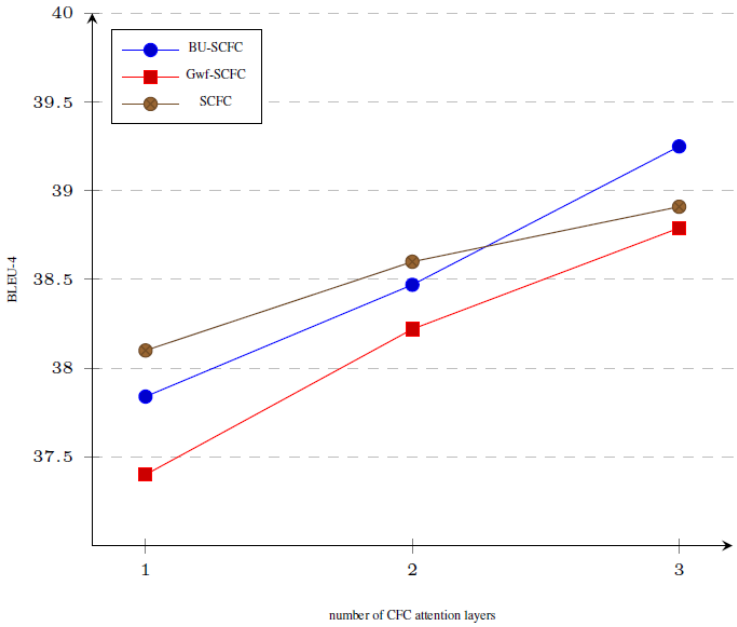


图 6：在 MSCOCO 测试集上进行 CIDEr-D 分数优化训练时，消融模型（BU-SCFC[s], Gwf-SCFC[s]）和完整模型 SCFC[s] 在 BLEU-4 指标上的结果，其中 s 表示不同数量的 CFC层。

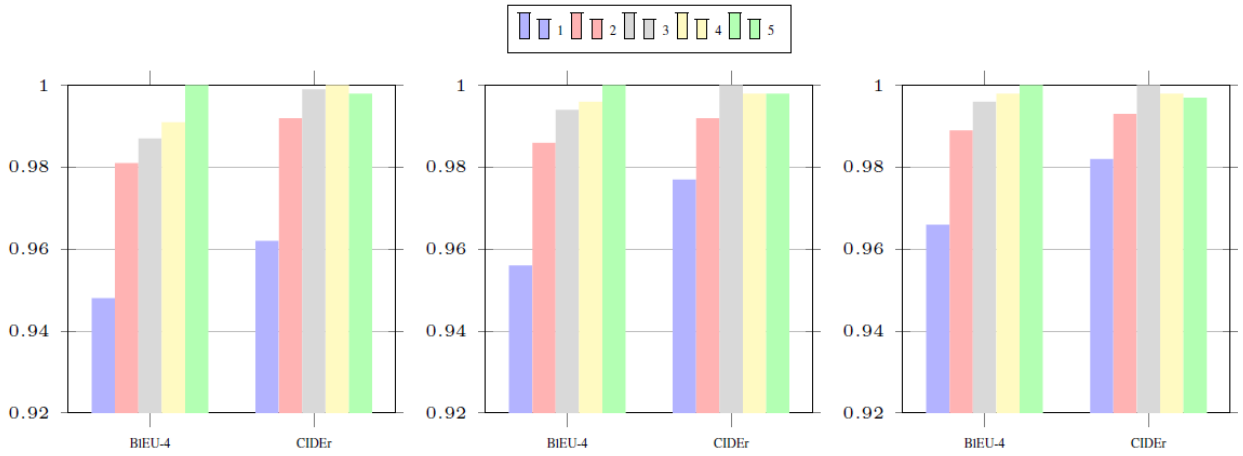


图 7：束宽 k 对所提出的 SCFC 的 BLEU 和 CIDEr 的影响，从左到右分别为包含一层、两层和三层 CFC层的模型。

多步推理过程。此外，我们提出 SCFC-LSTM，使模型能够以更可控的方式关注整合表示。我们的 SCFC 能够生成更细粒度的描述。为验证所提方法的有效性，我们进行了大量实验。实验结果表明，我们的方法优于采用基于强化学习的方法和交叉熵损失训练的当前最先进模型。我们还计划设计一种用于图像描述任务的动态 SCFC注意力网络，使步数参数取决于待生成的词，而非预先设定的固定值。

参考文献

[1] A. Karpathy 和 L. Fei-Fei，“用于生成图像描述的深度视觉语义对齐”，载于《IEEE计算机视觉与模式识别会议论文集》，2015年，第3128–3137页。[2] A. Karpathy、A. Joulin 和 L. F. Fei-Fei，“用于双向图像句子映射的深度片段嵌入”，载于《神经信息处理系统进展》，2014年，第1889–1897页。[3] H. Liu、Y. Yang、F. Shen、L. Duan 和 H. T. Shen，“循环图像描述器：利用空间不变变换和注意力过滤描述图像”，arXiv预印本 arXiv:1612.04949，2016年。[4] J. Mao、W. Xu、Y. Yang、J. Wang、Z. Huang 和 A. Yuille，“利用多模态循环神经网络（m-rnn）生成深度图像描述”，arXiv预印本 arXiv:1412.6632，2015年。[5] R. Kiros、R. Salakhutdinov 和 R. S. Zemel，“将视觉语义嵌入与多模态神经语言模型相结合”，arXiv预印本 arXiv:1411.2539，2014年。[6] T. J. Buschman 和 E. K. Miller，“前额叶和后顶叶皮层中自上而下与自下而上的注意力控制”，《科学》，第315卷，第5820期，第1860–1862页，2007年。[7] M. Corbetta 和 G. L. Shulman，“大脑中目标导向和刺激驱动注意力的控制”，《自然综述：神经科学》，第3卷，第3期，第201–215页，2002年。[8] K. Xu、J. Ba、R. Kiros、K. Cho、A. Courville、R. Salakhudinov、R. Zemel 和 Y. Bengio，“展示、关注并描述：利用视觉注意力生成神经图像描述”，载于《国际机器学习会议论文集》，2015年，第2048–2057页。[9] Q. You、H. Jin、Z. Wang、C. Fang 和 J. Luo，“利用语义注意力进行图像描述”，载于《IEEE计算机视觉与模式识别会议论文集》，2016年，第4651–4659页。[10] Q. Wu、C. Shen、L. Liu、A. Dick 和 A. Van Den Hengel，“显式高级概念在视觉到语言问题中有何价值？”，载于《IEEE计算机视觉与模式识别会议论文集》，2016年，第203–212页。[11] T. Yao、Y. Pan、Y. Li、Z. Qiu 和 T. Mei，“利用属性增强图像描述”，载于《IEEE国际计算机视觉会议论文集》，2017年，第4894–4902页。[12] C. He 和 H. Hu，“利用基于文本的视觉注意力进行图像描述”，《神经处理快报》，第49卷，第1期，第177–185页，2019年。[13] Y. Su、Y. Li、N. Xu 和 A.-A. Liu，“用于图像描述的分层深度神经网络”，《神经处理快报》，第1–11页，2019年。[14]O. Vinyals、A. Toshev、S. Bengio 和 D. Erhan，“ 展示并描述：神经图像描述生成器 ”，载于《 IEEE 计算机视觉与模式识别会议论文集》，2015年，第3156–3164页。[15] J. Mao、W. Xu、Y. Yang、J. Wang 和 A. L. Yuille，“利用多模态循环神经网络解释图像”，arXiv预印本 arXiv:1410.1090，2014年。[16] D. Bahdanau、K. Cho 和 Y. Bengio，“通过联合学习对齐与翻译实现神经机器翻译”，arXiv预印本 arXiv:1409.0473，2014年。[17] K. Fu、J. Jin、R. Cui、F. Sha 和 C. Zhang，“对齐看哪里与说什么：利用基于区域的注意力和场景特定上下文进行图像描述”，《IEEE模式分析与机器智能汇刊》，第39卷，第12期，第2321–2334页，2017年。[18] J. R. Uijlings、K. E. Van De Sande、T. Gevers 和 A. W. Smeulders，“用于目标识别的选择性搜索”，《国际计算机视觉杂志》，第104卷，第2期，第154–171页，2013年。[19] M. Pedersoli、T. Lucas、C. Schmid 和 J. Verbeek，“图像描述的注意力区域”，载于《IEEE国际计算机视觉会议论文集》，2017年，第1242–1250页。[20] J. Lu、C. Xiong、D. Parikh 和 R. Socher，“知道何时观察：通过视觉哨兵实现自适应注意力，用于图像描述”，载于《IEEE计算机视觉与模式识别会议论文集》，2017年，第375–383页。[21] Y. Huang、J. Chen、W. Ouyang、W. Wan 和 Y. Xue，“利用端到端属性检测及后续属性预测进行图像描述”，《IEEE图像处理汇刊》，第29卷，第4013–4026页，2020年。[22] P. Anderson、X. He、C. Buehler、D. Teney、M. Johnson、S. Gould 和 L. Zhang，“用于图像描述和视觉问答的自下而上与自上而下注意力”，载于《IEEE计算机视觉与模式识别会议论文集》，2018年，第6077–6086页。[23] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于《IEEE计算机视觉与模式识别会议论文集》，2016年，第770–778页。

[24] S. Ren、K. He、R. Girshick 和 J. Sun，“Faster R-CNN：利用区域建议网络实现实时目标检测”，载于《神经信息处理系统进展》，2015年，第91–99页。[25] J. Johnson、A. Karpathy 和 L. Fei-Fei，“DenseCap：用于密集图像描述的全卷积定位网络”，载于《IEEE计算机视觉与模式识别会议论文集》，2016年。[26] M. Jaderberg、K. Simonyan、A. Zisserman 等，“空间变换网络”，载于《神经信息处理系统进展》，2015年，第2017–2025页。[27] H. Fang、S. Gupta、F. Iandola、R. K. Srivastava、L. Deng、P. Doll´ar、J. Gao、X. He、M. Mitchell、J. C. Platt 等，“从图像描述到视觉概念，再回到图像描述”，载于《IEEE计算机视觉与模式识别会议论文集》，2015年，第1473–1482页。[28] T.-Y. Lin、P. Goyal、R. Girshick、K. He 和 P. Doll´ar，“用于密集目标检测的焦点损失”，载于《IEEE国际计算机视觉会议论文集》，2017年，第2980–2988页。[29] F. A. Gers 和 J. Schmidhuber，“能够计时和计数的循环网络”，载于《IEEE-INNS-ENNS国际联合神经网络会议论文集：IJCNN 2000，神经计算：新千年的新挑战与新视角》，第3卷，IEEE，2000年，第189–194页。[30] K. Greff、R. K. Srivastava、J. Koutn´ık、B. R. Steunebrink 和 J. Schmidhuber，“长短期记忆网络：搜索空间的奥德赛”，《IEEE神经网络与学习系统汇刊》，第28卷，第10期，第2222–2232页，2016年。[31]R. Vedantam、C. Lawrence Zitnick 和 D. Parikh，“CIDEr：基于共识的图像描述评估”，载于《IEEE计算机视觉与模式识别会议论文集》，2015年，第4566–4575页。[32] S. J. Rennie、E. Marcheret、Y. Mroueh、J. Ross 和 V. Goel，“用于图像描述的自批判序列训练”，载于《IEEE计算机视觉与模式识别会议论文集》，2017年，第7008–7024页。[33] T.-Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan、P. Doll´ar 和 C. L. Zitnick，“Microsoft COCO：常见物体及其场景”，载于《欧洲计算机视觉会议论文集》，Springer，2014年，第740–755页。[34]P. Young、A. Lai、M. Hodosh 和 J. Hockenmaier，“从图像描述到视觉指称：用于事件描述语义推理的新相似度指标”，《计算语言学协会汇刊》，第2卷，第67–78页，2014年。[35]K. Papineni、S. Roukos、T. Ward 和 W.-J. Zhu，“BLEU：一种机器翻译自动评估方法”，载于《计算语言学协会第40届年会论文集》，2002年，第311–318页。[36] M. Denkowski 和 A. Lavie，“METEOR Universal：适用于任何目标语言的特定语言翻译评估”，载于《第九届统计机器翻译研讨会论文集》，2014年，第376–380页。[37] C.-Y. Lin，“ROUGE：用于自动评估摘要的软件包”，载于《文本摘要的拓展》，2004年，第74–81页。[38] P. Anderson、B. Fernando、M. Johnson 和 S. Gould，“SPICE：语义命题图像描述评估”，载于《欧洲计算机视觉会议论文集》，Springer，2016年，第382–398页。[39] Z. Gan、C. Gan、X. He、Y. Pu、K. Tran、J. Gao、L. Carin 和 L. Deng，“用于视觉描述的语义组合网络”，载于《IEEE计算机视觉与模式识别会议论文集》，2017年，第5630–5639页。[40] J. Deng、W. Dong、R. Socher、L.-J. Li、K. Li 和 L. Fei-Fei，“ImageNet：大型层次化图像数据库”，载于《2009年IEEE计算机视觉与模式识别会议论文集》，电气电子工程师学会，2009年，第248–255页。[41] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，arXiv预印本 arXiv:1412.6980，2014年。[42]N. Srivastava、G. Hinton、A. Krizhevsky、I. Sutskever 和 R. Salakhutdinov，“Dropout：一种防止神经网络过拟合的简单方法”，《机器学习研究杂志》，第15卷，第1期，第1929–1958页，2014年。[43] A. Graves，“利用循环神经网络生成序列”，arXiv预印本 arXiv:1308.0850，2013年。[44]R. J. Williams 和 D. Zipser，“一种用于持续运行的全循环神经网络的学习算法”，《神经计算》，第1卷，第2期，第270–280页，1989年。[45] L. Chen、H. Zhang、J. Xiao、L. Nie、J. Shao、W. Liu 和 T.-S. Chua，“SCA-CNN：用于图像描述的卷积网络空间与通道注意力”，载于《IEEE计算机视觉与模式识别会议论文集》，2017年，第5659–5667页。

[46] W. Jiang、L. Ma、Y.-G. Jiang、W. Liu 和 T. Zhang，《用于图像描述的循环融合网络》，载于《欧洲计算机视觉会议论文集》(ECCV)，2018 年，第 499–515 页。[47] T. Yao、Y. Pan、Y. Li 和 T. Mei，《探索用于图像描述的视觉关系》，载于《欧洲计算机视觉会议论文集》(ECCV)，2018 年，第 684–699 页。[48] J. Wang、W. Wang、L.Wang、Z.Wang、D. D. Feng 和 T. Tan，《学习用于图像描述的视觉关系与上下文感知注意力》，《模式识别》，第 98 卷，第 107075 页，2020 年。[49] X. Yang、K. Tang、H. Zhang 和 J. Cai，《用于图像描述的场景图自编码》，载于《IEEE计算机视觉与模式识别会议论文集》，2019 年，第 10 685–10 694 页。[50] Z.-J. Zha、D. Liu、H. Zhang、Y. Zhang 和 F. Wu，《用于细粒度图像描述的上下文感知视觉策略网络》，《IEEE模式分析与机器智能汇刊》，2019 年。[51] Y. Qin、J. Du、Y. Zhang 和 H. Lu，《图像描述中的回顾与前瞻预测》，载于《IEEE计算机视觉与模式识别会议论文集》，2019 年，第 8367–8375 页。[52] L. Breiman，《装袋预测器》，《机器学习》，第 24 卷，第 2 期，第 123–140 页，1996 年。