arXiv:2302.06098v1 [cs.CV] 13 Feb 2023

面向图像描述的局部视觉建模

Yiwei Ma[a]、Jiayi Ji[a]、Xiaoshuai Sun[a、b]、Yiyi Zhou[a]、Rongrong Ji[a、b、c]

[a]媒体分析与计算实验室，人工智能系，信息学院，厦门大学，*361005*，中国[b]人工智能研究院，厦门大学[c]鹏城实验室，中国深圳

摘要

本文研究了利用网格特征进行局部视觉建模的问题，这对于生成准确且细致的图像描述至关重要。为实现这一目标，我们提出了一种具有两项新颖设计的局部性感知 *Transformer* 网络（LSTNet），即局部敏感注意力（LSA）和局部敏感融合 （LSF）。LSA通过建模每个网格与其邻居之间的关系，用于 Transformer 中的层内交互，从而降低描述生成过程中局部物体识别的难度。LSF用于层间信息融合，聚合不同编码器层的信息，以实现跨层语义互补。借助这两项新颖设计，所提出的LSTNet能够对网格特征的局部视觉信息进行建模，从而提升描述生成质量。为验证LSTNet的有效性，我们在具有竞争力的 MS-COCO 基准上进行了大量实验。实验结果表明，LSTNet不仅能够进行局部视觉建模，还在离线和在线测试中均优于一系列最先进的描述生成模型，即 其CIDEr分数分别达到134.8和136.3。此外，我们还在Flickr8k和Flickr30k数据集上验证了LSTNet的泛化能力。源代码已发布于GitHub：`https://github.com/xmu-xiaoma666/LSTNet`。

关键词： 图像描述、注意力机制、局部视觉建模

1. 引言

图像描述是一项生成流畅句子来描述给定图像的任务。近年来，

这一领域在许多创新方法 [1, 2] 和数据集 [3, 4]的支持下迅速发展。

受自底向上注意力 [5], 巨大成功的启发，现有大多数图像描述方法采用物体检测器提取的区域特征作为视觉表示例如，Faster R-CNN [6]。由于检测器是在大规模 Visual Genome 数据集 [7], 上预训练的，因此能够为图像中的显著区域生成具有判别性的表示，并为图像描述提供完整的物体信息。为此，基于区域特征的图像描述取得了显著进展[2, 8, 9]。

然而，区域特征仍存在明显缺陷。具体而言，它们从图像的显著区域中提取，因此往往会忽略背景中的上下文信息。在这种情况下，模型难以捕捉物体之间的关系。例如，如图1(a)所示，基于区域特征训练的模型无法理解边界框之外的上下文信息，因此错误地描述了“女人” 和 “马”之间的关系。此外，预训练的物体检测器可能会产生噪声、重叠或错误的检测结果，最终限制图像描述模型的性能上限。

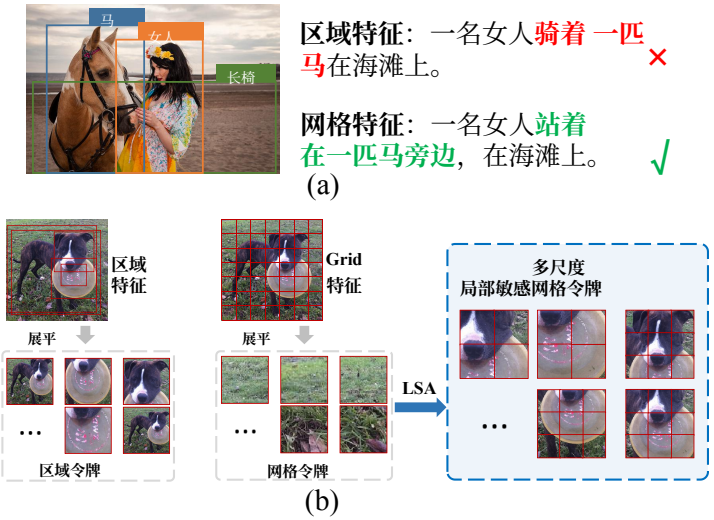


图 1：(a) Transformer 分别利用区域特征和网格特征生成的图像描述。(b) 区域特征通常包含完整的物体信息，而网格特征则更加零散。我们的 LSA 通过建模相邻网格之间的关系，有助于重建完整的物体信息。

为弥补上述局限，一些研究开始重新审视网格特征的应用。[10] 探索了物体检测器的网格特征，以进一步提升视觉问答（VQA）任务的性能。RSTNet [11] 和 DLCT [12]首次在类 Transformer 网络中采用网格特征，在图像描述方面取得了令人瞩目的表现。然而，类 Transformer 架构不利于感知完整物体。具体而言，如图1(b)所示，一个完整物体在二维空间中可能被划分到多个相邻网格中，而 Transformer 中的展平操作不可避免地会破坏网格特征之间的局部关系。同时，近期研究进展 [13]也表明，原始 Transformer 在局部视觉建模方面效率较低。

基于上述分析，我们发现区域特征和网格特征各有优缺点。区域特征包含明确的物体信息，但缺少背景和关系信息。相比之下，网格特征同时包含所有信息，但一个物体可能被划分到多个网格中。因此，大部分语义信息会受到损害，从而使推理更具挑战性。要兼得两类特征的优势，一个直接的解决方案是像 DLCT [12] 和 GRIT [14]那样同时采用区域特征和网格特征。然而，这会显著增加计算量并延长训练时间，因为模型需要同时处理这两类特征。更高效的方法是在网格特征上建模局部信息，以弥补物体信息的不足。

因此，本文提出了一种新颖的局部敏感 ***Transformer*** 网络（***LSTNet***）。具体而言，LSTNet 分别从层内交互和层间融合两个方面强化局部建模，以感知对象级信息。针对层内交互，我们提出了一种新颖的多分支模块，称为局部敏感注意力（***LSA***），以从不同感受野中感知细粒度局部信息，并增强各网格与其邻居之间的交互。值得注意的是，LSA 在推理过程中可以重参数化为单分支结构，从而减少多尺度感知带来的额外开销。针对层间融合，我们设计了局部敏感融合（***LSF***）模块，可对齐并融合不同层的网格特征，以实现跨层语义互补。凭借这些创新设计，LSTNet 不仅提升了局部视觉建模能力，也大幅提高了生成的图像描述质量。在竞争激烈的 MS-COCO 基准上，LSTNet 在离线和在线测试中均表现出色，即：

134.8 CIDEr 和 136.3 CIDEr。除了在 MS-COCO数据集上表现出色外，LSTNet 在 Flickr8k 和 Flickr30k 数据集上的泛化能力也得到了验证。

总而言之，我们的贡献有三点：

- 为了仅凭网格特征感知物体和上下文信息，我们提出了一种用于图像描述的新型 LSTNet。LSTNet 不仅增强了模型的局部感知能力，还在竞争激烈的 MS-COCO 基准测试中优于许多近期提出的方法。

- 我们提出了一种局部敏感注意力（*LSA*），用于 Transformer 中的层内视觉建模。该模块经过重新参数化，可增强每个网格特征与其局部邻居之间的交互。

- 我们提出了一种局部敏感融合（*LSF*）方法，用于聚合图像描述中的层间对象语义信息，这有助于实现层间语义理解。

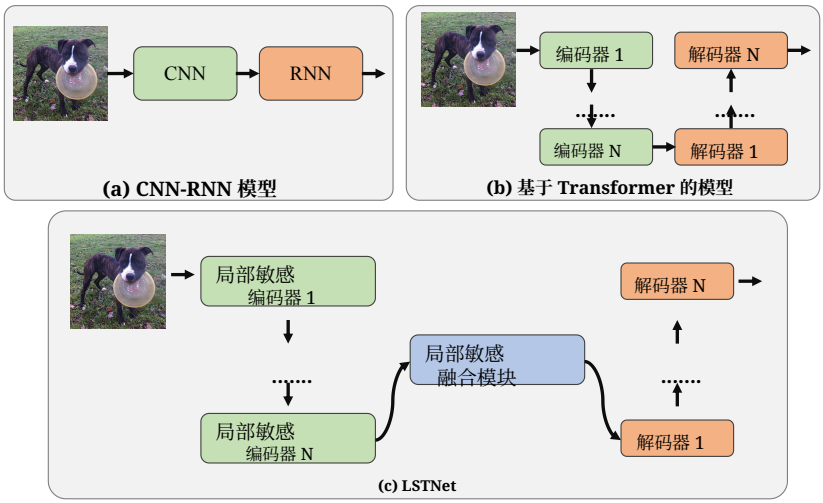


图 2：用于图像描述的 CNN-RNN 模型（a）、基于 Transformer 的模型（b）和 LSTNet（c）示意图。

2. 相关工作

2.1. 图像描述

图像描述是一项具有挑战性的任务，人们已经为解决这一问题付出了大量努力。经过多年的发展，随着大量方法的涌现 [5, 8, 11, 15–18]，可以观察到该领域取得了显著进步。现有的图像描述方法大致可分为两类：1）CNN-RNN 模型；2）基于 Transformer 的模型。如图 2(a)所示，CNN-RNN 模型使用卷积神经网络将图像编码为向量表示，然后采用基于 RNN 的解码器融合这些向量表示，为输入图像生成与内容相关的描述。具体而言， [15] 使用卷积神经网络（CNN）对图像进行编码，并采用长短期记忆网络（LSTM）作为解码器来生成图像描述。 [16]利用自适应注意力机制，决定在每个时间步关注视觉信息还是非视觉信息。 [5] 使用预训练的 Fast R-CNN [6] 提取显著物体作为区域视觉特征，有助于生成准确的图像描述。随着 Transformer 的发展 [19], ，许多研究人员开始探索基于 Transformer 的模型在图像描述任务中的应用，如图 2(b) 所示。 [8] 将双线性池化引入 Transformer 模型，以捕获 2^{nd} 顺序交互。 [11] 提出在进行词语预测之前，基于 Transformer 解码器自适应地衡量视觉线索和语言线索的贡献。[12] 提出了双层协作 Transformer（DLCT），以实现区域特征与网格特征之间的互补。为提升 Transformer 的语义理解能力，[18] 提出了一种同时采用空间注意力和通道注意力的基于 Transformer 的描述生成模型。[20] 提出了几何注意力 Transformer（GAT）模型，以进一步利用图像描述中的几何信息。为考虑物体特征的视觉持续性， [21] 通过在编码器和解码器中插入视觉持续性模块，引入了 VP-Net。[22] 提出了一种新颖的 CtxAdpAtt 模型，该模型利用语言上下文，有效探索不同物体之间的相关视觉关系。为减轻使用基于 GCN 的编码器表示场景图中关系信息所带来的缺点，ReFormer[23]探索了一种新颖的架构，用于显式表达图像中物体之间的关系。

我们的 LSTNet 与 Transformer-base 方法一致。然而，在处理网格特征时，Transformer 忽略了视觉局部性，而这对于识别图像中的物体至关重要。如图 2(c) 所示，

我们提出了局部敏感注意力（LSA）模块和局部敏感注意力（LSA）模块，以增强局部视觉建模。

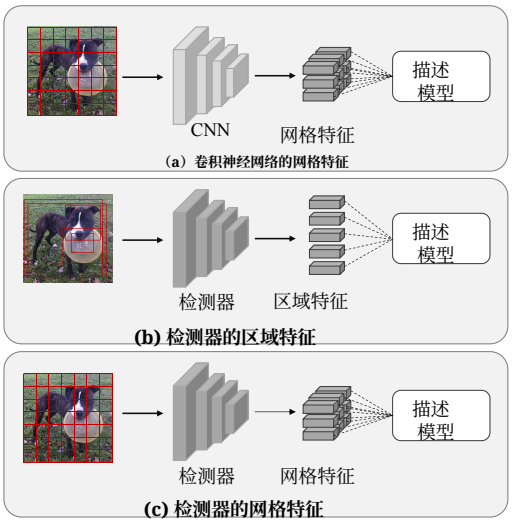


图 3：图像描述任务中使用的视觉特征的三个主要阶段：(a) 由卷积神经网络提取的网格特征，例如，ResNet [24]；(b) 由预训练检测器提取的区域特征，例如，Faster R-CNN [6], ，其需要耗时的后处理操作，例如，NMS [25]；(c) 从目标检测器中卷积神经网络的特征图提取的网格特征。

2.2. 区域特征与网格特征

图像描述所用的视觉特征经历了三个主要阶段：网格 → 区域 → 网格。在第一个网格阶段，一些开创性工作[16, 26, 27]采用从卷积神经网络 [24] 中提取的网格视觉特征来表示图像，如图 3(a) 所示。例如， [26]首次提出图像描述任务，并采用卷积神经网络编码视觉特征、循环神经网络解码描述。为捕捉图像中不同网格的重要性，[27] 在解码描述之前，对视觉特征应用注意力机制。[16] 提出了作用于视觉特征的自适应注意力，这些特征从 ResNet101 [24]的最后一个卷积层中提取。如图 3(b) 所示，为了获取前景信息， [5] 采用在 VG [7] 上预训练的目标检测器[6] 来提取区域特征，这些特征广泛应用于许多多模态任务 [2, 28]。如图 3(c) 所示，为弥补区域特征的缺陷（例如，耗时）， [10] 重新审视了网格

目标检测器中的特征，并发现其在视觉问答中能够取得有竞争力的性能；这一效果也已在图像描述任务中得到验证 [11, 12]。

与以区域特征为基础的方法相比 [2, 8, 9]，我们提出的基于网格特征的 LSTNet 能够捕捉边界框之外的上下文信息，从而生成更准确的图像描述。另一方面，与现有的基于网格特征的方法 [11]相比，我们提出的 LSTNet 考虑了网格特征的局部性，并对相邻网格之间的关系进行建模，这有助于识别图像中的物体。DLCT [12]采用边界框来辅助网格特征定位物体。然而，由于同时采用了网格特征和区域特征，该模型需要承担更多的训练和预测开销，例如，在交叉熵训练阶段，LSTNet 的运行速度比 DLCT 快三倍以上，而 DLCT 的性能受到边界框准确性的严重限制。我们提出的 LSTNet 捕捉了层内和层间的局部关系，从而为图像描述生成更加详细、更细粒度的网格特征。

2.3. Transformer 中的多头自注意力

Transformer 最初是为解决自然语言处理（NLP）任务而提出的。凭借强大的建模能力，近年来，Transformer 也被广泛应用于计算机视觉（CV）和多模态任务。Transformer 的核心组件是多头自注意力（MSA）模块，它能够有效地建模输入中不同位置之间的关系和上下文。具体而言， h头自注意力可表示为:

$$\mathrm{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Concat}\left(\mathrm{head}_1,\ \mathrm{head}_2,\ldots,\ \mathrm{head}_h\right)\mathbf{W}^{\mathrm{O}}, \quad (1)$$

其中，$\mathbf{Q}$，$\mathbf{K}$，$\mathbf{V}$ $\in \mathbb{R}^{N\times d}$ 分别表示输入查询、键和值。 N 表示输入长度， d 表示每个头的隐藏维度。$\mathbf{W}^{O} \in \mathbb{R}^{hd\times d}$ 是用于所有头输出的可学习矩阵。对于每个头，注意力的计算公式如下:

$$\mathrm{head}_i = \mathrm{Attention}\left(\mathbf{Q}\mathbf{W}_i^{\mathrm{Q}}, \mathbf{K}\mathbf{W}_i^{\mathrm{K}}, \mathbf{V}\mathbf{W}_i^{\mathrm{V}}\right) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{W}_i^{\mathrm{Q}}\left(\mathbf{K}\mathbf{W}_i^{\mathrm{K}}\right)^{\mathrm{T}}}{\sqrt{d}}\right)\mathbf{V}\mathbf{W}_i^{\mathrm{V}}, \quad (2)$$

其中，$\mathbf{W}^{\mathrm{Q}}{}_{\mathrm{i}}$ ，$\mathbf{W}_{\mathrm{i}}^{\mathrm{K}}$，$\mathbf{W}_{\mathrm{i}}^{\mathrm{V}} \in \mathbb{R}^{d \times d}$ 分别是输入查询、键和值的可学习矩阵。

3. 方法

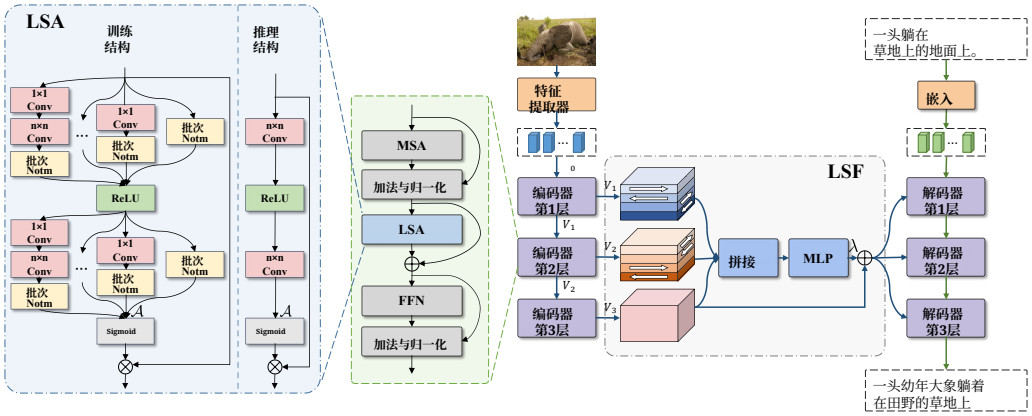


图 4：我们提出的 LSTNet 概览。网格特征用作视觉表示，并输入视觉编码器。自注意力之后插入局部敏感注意力（LSA），以捕获局部依赖关系。我们还应用重参数化技术，以降低其在推理过程中的开销。编码层的输出进一步由局部敏感融合（LSF）处理，以实现跨层聚合，同时也能提供不同层的语义，用于局部视觉建模。MSA 和 FFN 分别表示多头自注意力和前馈网络。

3.1. 概述

如图4所示，我们提出的局部性感知 *Transformer* 网络（*LSTNet*）遵循编码器-解码器范式。具体而言，编码器以视觉特征作为输入，然后通过编码器层对其关系进行建模，其中采用局部敏感注意力（*LSA*）模块来增强局部视觉建模。随后，局部敏感融合（*LSF*）聚合来自不同编码层的视觉特征，解码器则基于此预测描述词，以描述给定的视觉内容。

第 l个编码器层之前的视觉特征记为 $V_{l-1} \in \mathbb{R}^{N_v \times c}$ $(N_v = h \times w)$，其中 h、 w、 c 分别表示视觉特征的高度、宽度和通道维度。

LSTNet 的每个编码器层由三个组件构成：(1) 多头自注意力（MSA）模块；(2) 局部敏感注意力（LSA）模块；(3) 前馈网络（FFN）。来自最后一个编码器层的视觉特征 V_{l-1} 首先按如下方式经多头自注意力处理（为简洁起见，省略层归一化操作）：

$$V'_{l-1} = V_{l-1} + MSA(V_{l-1}, V_{l-1}, V_{l-1}), \tag{3}$$

其中 $MSA(\cdot)$ 是 Transformer [19] 中标准的多头自注意力。由于多头自注意力能够建模输入序列中任意两个位置之间的关系和上下文，因此有助于捕获长程依赖，并建模网格之间的全局信息。

自注意力通常难以有效捕获局部细节，而如第 1 节所述，这对网格视觉特征至关重要。因此，基于 V'_{l-1}，我们采用 LSA 来捕获相邻网格之间的依赖关系，进一步细化视觉特征：

$$V''_{l-1} = V'_{l-1} + LSA(V'_{l-1}), \tag{4}$$

其中，LSA 的具体细节将在下一小节中介绍。由于 LSA 由级联卷积层构成，能够建模相邻网格之间的关系，因此有助于建模网格之间的局部关系。

随后，将 LSA 模块的输出输入前馈网络，以在通道域进行交互：

$$V_l = V''_{l-1} + FFN(V''_{l-1}), \tag{5}$$

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2. \tag{6}$$

与以往仅将顶层编码器层的输出输入解码器的基于 Transformer 的模型不同，我们提出的 LSF 模块通过局部敏感融合（LSF），聚合所有编码器层的视觉特征，以获得语义上更丰富的特征：

$$V^* = LSF(V_1, V_2, \cdots, V_n), \tag{7}$$

其中，n 是编码器层的数量。

最后，将 V^* 输入解码器以生成图像描述，这与原始 Transformer [19] 相同。

3.2. 局部敏感注意力（*LSA*）

如图1(b)所示，图像中的一个物体可能会被分割成若干片段，并分布在不同的网格中，从而破坏视觉对象的空间和语义信息。一种合理的方法是加强局部信息之间的交互，这也符合这样一种假设：视觉上彼此接近的特征更有可能相互关联。

因此，为了捕捉局部细节并对相邻网格之间的交互进行建模，我们提出了一种多尺度局部敏感模块，即局部敏感注意力（LSA）。具体而言，MSA模块的输出特征 $V' \in \mathbb{R}^{N\times C}$ 是一个网格序列，其中 N 是网格数量，C 是通道维度的大小。我们首先将 $V' \in \mathbb{R}^{N\times C}$ 重塑为$V' \in \mathbb{R}^{H\times W\times C}$，其中 H，W分别是网格特征的高度和宽度。然后，我们依次使用两个多尺度二维卷积神经网络，并在两者之间加入一个激活函数（即 ReLU），以获得经过多尺度局部感知后的视觉特征 $\mathcal{A}$，其公式如下：

$$\mathcal{A} = MSC_2\big(\sigma(MSC_1(V')\big), \tag{8}$$

其中，σ（·）是激活函数，MSC_i（·）表示由多分支卷积神经网络实现的多尺度卷积神经网络：

$$MSC_i(x) = BN_1^i\left(F_1^i(x)\right) + \cdots + BN_N^i\left(F_N^i(x)\right), \tag{9}$$

其中，$i \in \{1, 2\}$, N 是分支的数量，BN_j(·)是批归一化[29], F_j(·)表示恒等映射、一个卷积模块或多个串联的卷积模块，并且 $j \in \{1, \cdots, N\}$。在我们的 LSTNet 中，分支数量 N 为 3。如图4蓝色区域所示，三个分支分别是：(1) 恒等映射，(2) 1×1 卷积，以及 (3) 1×1 卷积和 3×3 卷积的顺序组合。

在推理期间，多分支结构 MSC_i(·)可以通过使用一些结构重参数化技术 [30, 31]简化为单分支结构，从而减少参数数量和计算成本，且不会造成任何性能损失：

$$MSC_i(x) \to F^i(x), \tag{10}$$

其中，$F^i(x)$ 是 3×3 卷积，且 $i \in \{1, 2\}$。

为了获得每个网格的注意力权重，我们还对 $\mathcal{A}$应用 Sigmoid 函数。最后，我们根据局部敏感注意力图对自注意力层 V' 的输出特征重新加权，如下所示:

$$V'' = V' \otimes Sigmoid(\mathcal{A}), \tag{11}$$

其中 $\otimes$ 表示逐元素相乘。

3.3. 局部敏感融合（*LSF*）

不同层的特征往往包含不同层次的语义信息 [2]。然而，大多数现有的图像描述方法仅将顶层编码器层的特征输入解码器，导致低层信息丢失。为避免此类信息丢失，我们融合编码器中所有层的特征，然后将融合特征输入解码器。

具体而言，我们引入一种简单的空间移位 操作，使每个网格能够与其相邻网格对齐，然后使用多层感知机（MLP）不仅在通道域中进行交互，也在空间域中进行交互。

具体来说，将来自第 l层编码器的特征记为 $V_l \in \mathbb{R}^{h\times w\times c}$（此处省略重塑操作）， V_1 和 V_2 通过不同的空间移位操作进行移位（即 $SS_1(\cdot)$ 和 $SS_2(\cdot)$），可表示为式 12 和式 13:

$$\begin{aligned}
&V_1[d_s:h,:,0:c/4] = V_1[0:h-d_s,:,0:c/4],\\
&V_1[0:h-d_s,:,c/4:c/2] = V_1[d_s:h,:,c/4:c/2],\\
&V_1[:,d_s:w,c/2:3c/4] = V_1[:,0:w-d_s,c/2:3c/4],\\
&V_1[:,0:w-d_s,3c/4:c] = V_1[:,d_s:w,3c/4:c],
\end{aligned} \tag{12}$$

$$\begin{aligned}
&V_2[:,d_s:w,0:c/4] = V_2[:,0:w-d_s,0:c/4],\\
&V_2[:,0:w-d_s,c/4:c/2] = V_2[:,d_s:w,c/4:c/2],\\
&V_2[d_s:h,:,c/2:3c/4] = V_2[0:h-d_s,:,c/2:3c/4],\\
&V_2[0:h-d_s,:,3c/4:c] = V_2[d_s:h,:,3c/4:c],
\end{aligned} \tag{13}$$

其中， V_i 是第 i个编码器层的输出特征， d_s 是空间移位的移位距离，它决定了局部交互的范围。顶层编码器层的输出 V_3 不经过任何移位操作。空间移位的示意图见图4和图5。

然后，将来自不同层的移位特征拼接在一起:

$$V_c = Concat(V_1, V_2, V_3). \tag{14}$$

表1：已发表的最先进图像描述模型在 COCO 在线测试服务器上的排行榜。

模型	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
~~SCST[32]~~ CVPR'2017	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	116.0
LSTM-A[33] ICCV'17	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
Up-Down[5] CVPR 2018	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet[34] ECCV'18	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
GCN-LSTM[35] ECCV'18	80.8	95.2	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE[36] CVPR'19	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoANet[9] CVPR'19	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
CAVP[37] IEEE Transactions on Pattern Analysis and Machine Intelligence 2019	80.1	94.9	64.7	88.8	50.0	79.7	37.9	69.0	28.1	37.0	58.2	73.1	121.6	123.8
ETA[38] ICCV 2019	81.2	95.0	65.5	89.0	50.9	80.4	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
M^2Transformer[2] CVPR'20	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
XTransformer(ResNet-101) [8] cvpr'20	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
XTransformer(SENet-154) [8] cvpr'20	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
DLCT(ResNeXt101) [12] AAAI'21	82.0	96.2	66.9	91.0	52.3	83.0	40.2	73.2	29.5	39.1	59.4	74.8	131.0	133.4
DLCT(ResNeXt152) [12] AAAI'21	82.4	96.6	67.4	91.7	52.8	83.8	40.6	74.0	29.8	39.6	59.8	75.3	133.3	135.4
RSTNet(ResNext101)[11] cvpr'21	81.7	96.2	66.5	90.9	51.8	82.7	39.7	72.5	29.3	38.7	59.2	74.2	130.1	132.4
RSTNet(ResNext152)[11] cvpr'21	82.1	96.4	67.0	91.3	52.2	83.0	40.0	73.1	29.6	39.1	59.5	74.6	131.9	134.0
GAT [20] Expert Systems with Applications 2022	81.1	95.1	66.1	89.7	51.8	81.5	39.9	71.4	29.1	38.4	59.1	74.4	127.8	129.8
VPNet[21] Neurocomputing'22	81.4	95.5	66.4	90.3	52.0	82.1	40.0	72.1	29.3	38.9	59.3	74.9	128.2	130.6
CtxAdpAtt[22] IEEE Transactions on Multimedia 2022	81.0	95.2	65.5	91.0	51.5	81.7	39.3	70.9	29.4	39.0	59.6	75.1	128.5	131.0
ReFormer[23] mm'22	82.0	96.7	-	-	-	-	40.1	73.2	29.8	39.5	59.9	75.2	129.9	132.8
LSTNet (ResNeXt-101)	82.2	96.2	67.2	91.2	52.7	83.5	40.6	73.8	29.6	39.3	59.6	75.0	132.0	134.5
LSTNet (ResNeXt-152)	**82.6**	**96.7**	**67.8**	**92.0**	**53.3**	**84.3**	**41.1**	**74.7**	**29.9**	**39.6**	**60.0**	**75.4**	**134.0**	**136.3**

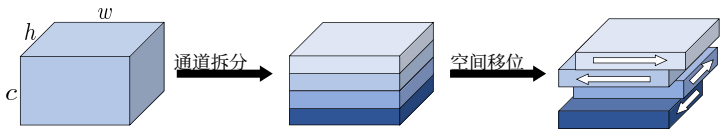


图 5：空间移位 操作示意图。 c， h， w 分别表示通道、高度和宽度维度的大小。

理论上，多层感知机无法建模相邻网格之间的关系。然而，经过空间移位操作后，每个网格都会与其相邻网格对齐，多层感知机便能够在通道域和空间域中进行信息交互：

$$\tilde{V} = \sigma(V_c W_1) W_2, \tag{15}$$

其中， $\sigma(\cdot)$是 ReLU 激活函数， $W_1 \in \mathbb{R}^{3c \times 3c}$ 和 $W_2 \in \mathbb{R}^{3c \times c}$ 是可学习投影矩阵。

为进一步增强视觉特征的描述能力，我们通过残差连接将顶层编码器层的输出与融合特征 V 结合：

$$V^* = \lambda \tilde{V} + V_{top}, \tag{16}$$

其中， V_{top} 是顶层编码器层的特征， $V_{top} = V_3$ 在我们的 LSTNet 中， λ 用作加权因子。

表 2：在 Karpathy 测试集划分上与最先进方法的比较。所有值均以百分比（%）表示，其中 B-N、M、R 和 C 分别代表 BLEU-N、METEOR、ROUGE-L 和 CIDEr 分数。

模型	B-1	B-4	M	R	C	S
SCST[32] CVPR 2017	-	34.2	26.7	55.7	114.0	-
Up-Down[5] CVPR 2018	79.8	36.3	27.7	56.9	120.1	21.4
RFNet[34] eccv'18	79.1	36.5	27.7	57.3	121.9	21.2
GCN-LSTM[35] ECCV'18	80.5	38.2	28.5	58.3	127.6	22.0
SGAE[36] CVPR'19	80.8	38.4	28.4	58.6	127.8	22.1
CAVPN[37] IEEE Transactions on Pattern Analysis and Machine Intelligence 2019	-	38.6	28.3	58.5	126.3	21.6
AoANet[9] CVPR'19	80.2	38.9	29.2	58.8	129.8	22.4
ORT[39] NeurIPS'19	80.5	38.6	28.7	58.4	128.3	22.6
Transformer[19] NeurIPS'17	80.7	38.6	29.1	58.5	130.1	22.7
M^2Transformer [2] CVPR'20	80.8	39.1	29.2	58.6	131.2	22.6
XTransformer[8] CVPR'20	80.9	39.7	29.5	59.1	132.8	23.4
DLCT[12] AAAI'21	81.4	39.8	29.5	59.1	133.8	23.0
RSTNet[11] CVPR'21	81.1	39.3	29.4	58.8	133.3	23.0
GAT [20] Expert Systems with Applications 2022	80.8	39.7	29.1	59.0	130.5	22.9
VPNet[21] Neurocomputing'22	80.9	39.7	29.3	59.2	130.4	23.2
CtxAdpAtt[22] IEEE Transactions on Multimedia 2022	80.5	39.1	29.3	59.3	130.1	**23.6**
ReFormer[23] mm'22	-	39.8	29.7	59.8	131.2	23.0
LSTNet	**81.5**	**40.3**	**29.6**	**59.4**	**134.8**	23.1

具体而言，LSF 的动机源于两个方面：1）不同编码器层输出的特征图具有不同的语义（即，高层特征图具有高层语义信息，而低层特征图具有低层语义信息）。传统的 Transformer 仅将编码器最后一层的特征图输入解码器，忽略了低层语义信息。LSF 通过融合所有编码器层的输出，同时考虑高层和低层语义信息，解决了这一问题。2）LSF 模块可以通过空间移位操作与局部网格交互，这有助于对对象级信息进行建模。

3.4. 目标

给定真实描述 $y^*_{1:T}$ 和参数为 θ 的描述生成模型，其中 T 为描述的长度，我们对该模型进行预训练

使用交叉熵（CE）损失，具体如下：

$$L_{CE} = -\sum_{t=1}^{T} \log\left(p_\theta(y_t^*|y_{1:t-1}^*)\right). \tag{17}$$

然后，我们进一步使用自我批判序列训练（SCST）[32]，根据 CIDEr 和 BLEU-4 分数优化模型，如下所示：

$$\nabla_\theta L_{RL}(\theta) = -\frac{1}{k}\sum_{i=1}^{k}\left(r(y_{1:T}^i) - b\right)\nabla_\theta \log p_\theta(y_{1:T}^i), \tag{18}$$

其中，k 是束宽，$r(\cdot)$是 CIDEr 和 BLEU-4 之和，而 $b = (\sum_i r(y_{1:T}^i))/k$ 是奖励基线。

4. 实验

4.1. 数据集

我们在广泛使用的 MS-COCO [3] 图像描述数据集上开展实验。该数据集包含 123,287 张图像，其中包括 82,783 张训练图像、40,504 张验证图像和 40,775 张测试图像，每张图像均标注了 5 条图像描述。离线测试采用 [40]提供的数据划分，其中 5,000 张图像用于验证，5,000 张用于测试，其余图像用于训练。此外，我们还上传了官方测试集的生成描述，以供在线评估 [1]。

4.2. 实现细节

网格特征提取自由 [10], 提供的预训练 Faster R-CNN [6]，其中使用步长为 1 C_5 的主干网络和带有两个全连接层的 1×1 RoI 池化作为检测头，在 VG 数据集上训练 Faster R-CNN。具体而言，我们采用 C_5 特征图，并对其进行平均池化，得到空间大小为 7×7 的特征。请注意，除了简单的数据增强（例如，随机裁剪、随机旋转）外，我们没有使用任何额外的数据预处理。LSTNet 中的d_{model} 为 512，前馈网络中的扩展率为 4，注意力头数为 8，束搜索的大小为 5。

我们在两个阶段都使用 Adam 优化器训练模型，并采用 [12] 中的相对位置编码。在 CE 训练阶段，批次

[1]https://competitions.codalab.org/competitions/3221#results

大小为 50，学习率在前 4 个训练轮次中线性增至 1×10^{-4}。之后，我们分别在第 10 和第 12 个训练轮次将其设为 2×10^{-5}、4×10^{-6}。经过 18 个训练轮次的 CE预训练后，我们使用批量大小为 100、学习率为 5×10^{-6} 的 SCST 优化模型。学习率将在第 35、40、45、50 个训练轮次分别设为 2.5×10^{-6}、5×10^{-7}、2.5×10^{-7}、5×10^{-8}，SCST 训练将持续 42 个训练轮次。

4.3. 性能比较

本节将我们的 LSTNet 与最先进方法在离线和在线评估中进行比较。参与比较的模型包括：SCST [32], Up-Down [5],RFNet [34], GCN-LSTM [35], SGAE [36], AoANet [9], ETA [38], ORT [39],Transformer [19], M^2 Transformer [2], XTransformer [8], RSTNet [11] 和 DLCT [12]。遵循标准评估准则，我们采用 BLEU-N[41], METEOR [42], ROUGE-L [43], CIDEr [44], SPICE [45] 来评估性能。

4.3.1. 在线评估

表 1 展示了 LSTNet 与其他最先进方法在 COCO 在线测试服务器上的性能比较，评估使用 5 条参考描述（c5）和 40 条参考描述（c40）。为确保比较公平，我们也按照 [2] 采用四模型集成，并使用两种常见骨干网络（即 ResNeXt-101、ResNeXt-152 [46]）。值得注意的是，我们的 LSTNet 在所有评价指标上均以显著优势胜过其他最先进方法。令人惊讶的是，我们观察到，采用 ***ResNeXt-101*** 的 LSTNet 在大多数评价指标上表现优于采用 ***ResNeXt-152*** 的 RSTNet 和采用 ***SENet-154*** 的 X-Transformer。

4.3.2. 离线评估

表 2 总结了最先进模型和我们的方法在 COCO Karpathy测试集上的离线表现。请注意，为了公平比较，我们报告的是未使用任何集成技术的单模型结果。可以看到，我们提出的 LSTNet 在大多数评价指标上都优于其他所有最先进水平的模型。值得注意的是，我们的 LSTNet 的 CIDEr 得分达到 ***134.8%***，比最强的竞争者 DLCT 高出 ***1.0%***；DLCT 同时采用了区域特征和网格特征。

我们观察到，采用网格特征的 LSTNet 比一些使用区域特征的模型（例如，M^2Transformer [2], XTransformer [8]）表现更好。我们认为，本文中的网格级方案优于对象级方案的原因如下：1）区域特征的

表3：使用相同的 ResNeXt101 [46]网格特征，在 Karpathy 测试集划分上与最先进方法进行比较。

模型	B-1	B-4	M	R	C	S
Up-Down [5] cvpr'18	75.0	37.3	28.1	57.9	123.8	21.6
AoANet[9] CVPR'19	80.8	39.1	29.1	59.1	130.3	22.7
Transformer[19] NeurIPS'17	81.0	38.9	29.0	58.4	131.3	22.6
M^2Transformer CVPR'20 [2]	80.8	38.9	29.1	58.5	131.8	22.7
XTransformer[8] CVPR'20	81.0	**39.7**	29.4	58.9	132.5	**23.1**
DLCT[12] AAAI'21	81.4	39.8	29.5	59.1	133.8	23.0
RSTNet[11] CVPR'21	81.1	39.3	29.4	58.8	133.3	23.0
LSTNet	**81.5**	**40.3**	**29.6**	**59.4**	**134.8**	**23.1**

背景信息缺失，而网格特征能够提取图像中的全部信息。具体而言，视觉区域特征取自图像的显著区域，通常会忽略上下文信息。由于缺少背景信息，模型难以捕捉物体之间的关系。另一方面，网格特征能够收集图像中的全部空间信息。2）预训练的物体检测器通常会产生嘈杂、重叠或错误的检测结果，最终限制图像描述模型的性能上限。另一方面，网格特征不提供检测信息，因此可以避免错误检测的影响。3）在网格特征中，一个物体会被划分到不同的网格中，这正是本文的出发点。我们的方法使模型能够捕捉局部信息，并解决这一问题。

4.3.3. 与最先进方法进行公平比较

为消除不同视觉特征带来的干扰，我们使用相同的网格特征进行实验，以比较 LSTNet 与其他最先进方法。如表 3 所示，与使用相同视觉特征的其他方法相比，我们提出的 LSTNet 在所有评价指标上仍表现更优。

4.4. 消融研究

4.4.1. LSA 不同分支的效果

为验证各分支的影响，我们利用 LSA 的不同分支开展了一系列实验。不同分支下 LSA 的性能如表4所示。通过分析该表，我们得出以下观察结果：

• 与不使用 LSA 的模型（第1行）相比，采用包含一个或多个分支的LSA（第2至7行）有助于生成质量更好的图像描述。此外，采用的分支越多，通常取得的性能越好。这可能是因为所提出的 LSA模块增强了模型的局部感知能力，因此有利于感知物体信息。

• 如表4所示，在大多数评估指标上，配备一个分支（第2、3、4行）的模型优于配备0个分支（第1行）的模型。通过比较配备一个分支（第2、3、4行）的模型与配备两个分支（第5、6、7行）的模型可以看出，在大多数评价指标上，配备两个分支的模型优于配备一个分支的模型，尤其是在CIDEr指标上。此外可以看出，包含全部三个分支的完整LSTNet（第8行）表现最佳。这可能是因为图像中的物体大小各不相同，因此，更多的分支有助于增强对不同大小物体的多尺度建模能力。重要的是，通过使用重参数化技术，增加LSA的分支数量不会导致推理阶段的额外开销。

*4.4.2. LSA*与*SA*的不同排列方式的影响

为了探究不同排列方式的局部敏感注意力（LSA）和自注意力（SA）的影响，我们比较了三种结合LSA和SA的方法：（1）顺序式LSA-SA；（2）顺序式SA-LSA；（3）并行使用SA和LSA。如表5所示，可以观察到，顺序式SA-LSA的性能优于其他方法。其主要原因可能是，经SA处理后的特征属于粗粒度特征，而我们提出的LSA旨在建模局部关系，有助于进一步细化视觉特征。

4.4.3. LSF 不同移位距离的影响

为探究 LSF 移位距离 d_s 的影响，我们通过将 d_s 从 0 增至 4 进行实验（$d_s = 0$ 表示所有特征均不移位）。从表 6 可以看出，移位后的 LSF 效果优于未移位的 LSF。这可能是因为移位后的 LSF 在融合过程中使每个网格都能与相邻网格交互，从而增强局部建模。然而，当移位距离大于 1 时，性能开始下降，且 $d_s = 1$ 效果最佳。原因可能是

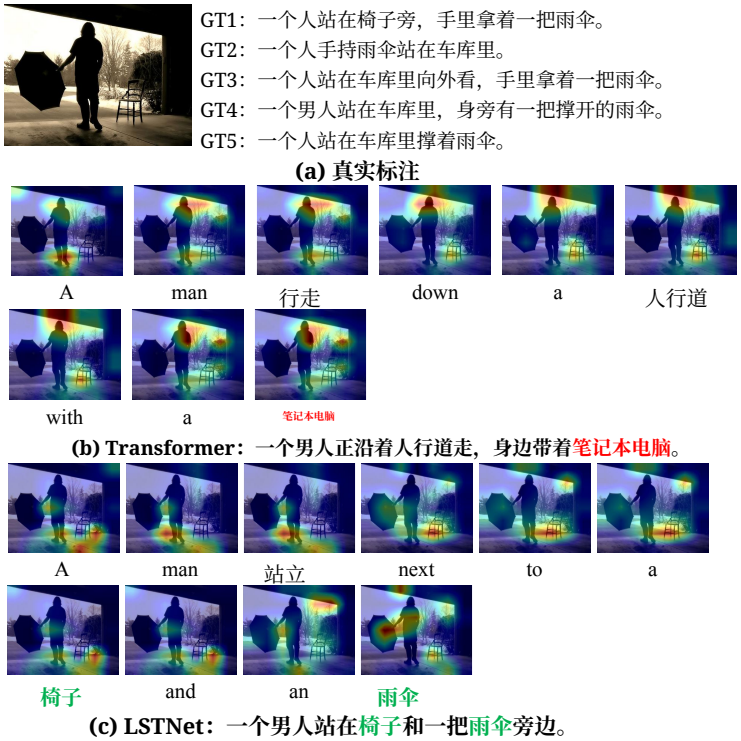


图 6：真实标注 (a)，以及标准 Transformer (b) 和 LSTNet (c) 在图像描述生成过程中的注意力图像可视化。两个生成的描述之间最显著的差异以粗体标出。

较大的移位促进了长距离交互，却忽视了局部建模。

4.4.4. LSA 中 λ 的作用

为了选择公式 16 中最佳的加权因子 λ，我们也进行了一组实验。从表 7 可以发现，λ 过大会导致性能下降，而 $\lambda = 0.2$ 在大多数评价指标上表现良好。因此，除非另有说明，我们在实验中采用 $\lambda = 0.2$。

4.4.5. 解耦 *LSA* 和 *LSF* 的影响

为了深入了解所提出的 LSA 和 LSF 模块，我们在实验中将这两个模块解耦。如表 8 所示，与完整模型相比，不含 LSA+LSF 的模型性能显著下降。

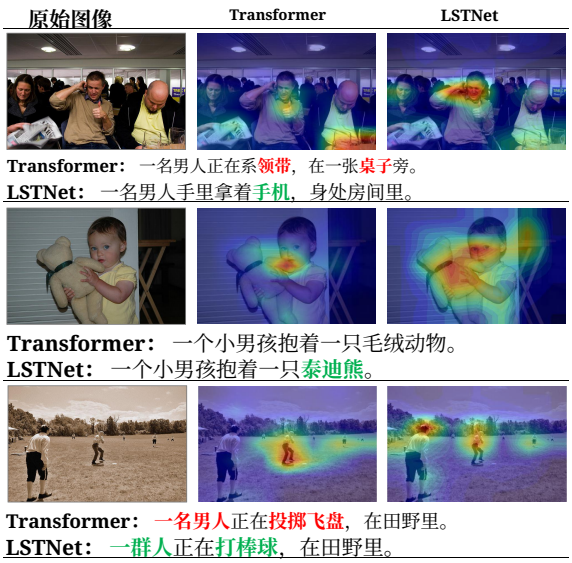


图 7：Transformer 和 LSTNet 在相同网格特征上生成的描述示例，以及 Transformer 和 LSTNet 的顶层编码器层对应注意力图的比较。

具体而言，其 BLEU-4 和 CIDEr 分别绝对下降了 *1.4%* 和 *3.5%*，这表明 LSA 和 LSF 至关重要。具体而言，我们的 LSA 和 LSF 分别使 CIDEr 分数提升了 2.3% 和 2.4%。这说明 LSA 和 LSF 模块能够相互促进，从而取得更好的性能。

4.4.6. 融合特征的不同方法的效果

为了验证 LSF 的有效性，我们通过使用各种模块替代 LSF 来融合特征，设计了一组实验。如表 9 所示，我们可以发现，与第一行的结果相比（即，不融合特征），融合来自不同层的特征能够提升性能。其主要原因可能是，不同层的特征在语义信息方面具有互补性，而融合特征将丰富视觉特征。此外，我们提出的 LFS 模块以较大幅度优于其他模块，这有力地说明了 LFS 的有效性。

4.4.7. 不同网格尺寸的影响

为探究不同网格尺寸的影响，我们通过平均池化将视觉特征设置为不同大小，并进行了一系列实验。

表 4：不同分支的消融研究。所有分支均包含批归一化，且所有模型均未配备 LSF。所有值均以百分比（%）表示，B-N、M、R、C 和 S 分别是 BLEU-N、METEOR、ROUGE-L、CIDEr 和 SPICE 分数的简称。

身份	1×1	1×1+3×3	B-1	B-4	M	R	C	S
×	×	×	81.0	38.9	29.0	58.4	131.3	22.6
√	×	×	81.1	38.9	29.1	58.4	131.6	22.6
×	√	×	81.1	39.2	29.2	58.9	132.2	22.7
×	×	√	**81.2**	39.4	29.2	58.9	132.3	22.7
√	√	×	**81.2**	39.6	29.1	59.0	133.5	22.7
√	×	√	**81.2**	39.6	29.2	58.9	133.4	22.6
×	√	√	**81.2**	39.6	**29.3**	**59.1**	133.5	**22.8**
√	√	√	**81.2**	**39.7**	**29.3**	**59.1**	**133.6**	**22.8**

表 5：SA 和 LSA 不同排列方式的消融研究。 + 表示顺序连接，& 表示并行连接。

排列方式	B-1	B-4	M	R	C	S
LSA + SA	81.1	39.6	29.1	59.0	133.4	**22.8**
SA + LSA	**81.2**	**39.7**	**29.3**	**59.1**	**133.6**	**22.8**
SA 与 LSA	81.0	39.5	28.9	59.0	132.7	22.7

如表 10 所示，随着网格大小增加，图像描述的性能逐步提升。这是因为更大的图像特征能够提供更细粒度、更丰富的语义信息，因此图像描述模型可以据此生成更准确的描述。

4.5. 定量分析

通过比较传统指标（例如，BLEU-N、CIDEr、SPICE），很难确定我们的方法是否显著提升了图像描述的性能。为了以直观的方式展示我们提出的 LSTNet 的有效性和优越性，我们采用配对样本进行双尾 t 检验，将 LSTNet 与标准 Transformer 进行比较。具体而言，我们首先针对每项传统指标进行双尾 t 检验，以探究与标准 Transformer 相比，LSTNet 生成的描述质量是否显著提升。此外，我们还报告了 SPICE 分数的语义子类别（即，关系、基数、属性、大小、颜色和物体），这些类别可用于衡量

表6：LSF 的不同移位距离的消融研究。所有模型均配备 LSA 和 LSF 模块。移

if位距离	B-1	B-4	M	R	C	S
$d_s = 0$	81.2	39.9	29.3	59.1	133.9	22.8
$d_s = 1$	**81.5**	**40.3**	**29.6**	**59.4**	**134.8**	**23.1**
$d_s = 2$	81.4	40.1	29.5	59.2	134.4	22.9
$d_s = 3$	81.3	40.1	29.5	59.2	134.2	22.8
$d_s = 4$	81.3	40.0	29.4	59.1	134.0	22.8

表7：采用不同权重因子的性能比较 λ。所有模型均配备 LSA 和 LSF 模块。

λ	B-1	B-4	M	R	C	S
$\lambda = 0.1$	**81.7**	**40.3**	29.4	59.3	134.3	22.9
$\lambda = 0.2$	81.5	**40.3**	**29.6**	**59.4**	**134.8**	**23.1**
$\lambda = 0.3$	81.6	40.2	29.5	59.3	134.5	23.0
$\lambda = 0.5$	81.6	**40.3**	29.5	59.3	134.5	30.0
$\lambda = 0.7$	81.4	40.0	29.4	59.0	134.0	22.9

生成句子与真实标注之间的语义相关性。此外，对于每个综合 SPICE 分数，我们还使用匹配数据进行了详细的双尾 t 检验，以考察这些语义指标是否也得到了显著提升。

表 11 展示了测试集上 t 检验的传统指标及相应的 p 值。我们观察到，在显著性水平 $\alpha = 0.05$ 下，所有评价指标的提升都具有统计显著性，这表明我们提出的 LSTNet 有助于提升生成的描述的质量。表 12 详细列出了测试集上 SPICE 分数的语义子类别及 t 检验的 p 值。我们可以观察到，SPICE 的所有语义子类别均有所提升，这说明了 LSTNet 中局部视觉建模的有效性和优越性。此外，我们还可以观察到，在显著性水平 $\alpha = 0.05$ 下，SPICE 的一些语义子类别（即，属性、大小、颜色和物体）取得了显著提升。值得注意的是，这四项评价指标描述的都是物体的属性，因此这证明 LSTNet 中的局部视觉建模有助于捕获对象级信息。与其他语义评价指标相比，关系指标的提升相对不显著。这可能是因为，要捕获物体之间的关系，仅靠

表 8：对 LSA 和 LSF 模块的消融研究。

Module	B-1	B-4	M	R	C	S
不使用 LSA+LSF	81.0	38.9	29.0	58.4	131.3	22.6
仅使用 LSA	81.2	39.7	29.3	59.1	133.6	22.8
仅使用 LSF	81.3	39.8	29.3	59.0	133.7	22.8
LSA + LSF	**81.5**	**40.3**	**29.6**	**59.4**	**134.8**	**23.1**

表 9：对各种融合方法的消融研究。所有模型均未配备 LSA。FPN 是类似于 [47] 的特征金字塔网络。所有值均以百分比 (%) 表示，其中 B-N、M、R、C 和 S 分别是 BLEU-N、METEOR、ROUGE-L、CIDEr 和 SPICE 分数的简称。

融合方法	B-1	B-4	M	R	C	S
不使用融合	81.0	38.9	29.0	58.4	131.3	22.6
MLP	81.0	39.3	29.2	**59.0**	132.5	22.7
求和池化	81.1	39.4	29.1	**59.0**	132.5	22.7
3×3卷积	**81.3**	39.7	29.0	58.9	132.6	22.7
FPN	81.1	39.1	29.1	58.7	132.6	22.6
LSF（我们的方法）	**81.3**	**39.8**	**29.3**	**59.0**	**133.7**	**22.8**

提升模型的局部建模能力还不够，全局建模能力也十分重要。对于其他针对单个物体的语义评价指标（即，属性、大小、颜色和物体），局部视觉建模已经能够显著提升性能。

4.6. 定性分析

为定性验证 LSTNet 的有效性，我们在图7中展示了 Transformer 和 LSTNet 基于相同网格特征生成的若干典型图像描述。可以看出，Transformer 生成的图像描述信息不足，甚至存在错误；而 LSTNet 生成的描述更加准确且具有区分度，这表明我们提出的 LSA 和 LSF 有助于通过局部建模识别视觉对象。

为了深入了解 LSTNet 能够生成准确图像描述的原因，我们进一步展示了图7中 Transformer 和 LSTNet 顶层编码器层的注意力图。通过分析结果，我们得出以下观察：1）Transformer 生成的注意力图未能关注图像中的重要视觉对象，而 LSTNet 则能够聚焦于这些重要对象。例如，对于第一幅图像，

表10：针对视觉特征不同网格尺寸的消融研究。所有值均以百分比（%）表示，其中 B-N、M、R、C 和 S 分别是 BLEU-N、METEOR、ROUGE-L、CIDEr 和 SPICE 分数的缩写。

网格大小	B-1	B-4	M	R	C	S
1×1	78.5	36.2	27.4	56.5	120.3	20.6
2×2	79.9	37.7	28.3	57.7	124.9	21.7
3×3	80.4	38.9	28.8	58.4	129.6	22.3
4×4	80.6	38.9	28.9	58.5	130.3	22.4
5×5	81.0	39.6	29.1	58.8	131.7	22.7
6×6	81.3	39.9	29.3	58.9	132.1	22.9
7×7	81.5	40.3	29.6	59.4	134.8	23.1

表11：标准 Transformer 与我们的 LSTNet 在不同图像描述指标上的性能比较。P 值来自采用配对样本的双尾 t 检验。粗体标出的 P 值在 0.05 的显著性水平上具有显著性。

模型	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Transformer	81.0	38.9	29.0	58.4	131.3	22.6
LSTNet	81.5	40.3	29.6	59.4	134.8	23.1
p值	$\mathbf{6.7 \times 10^{-3}}$	$\mathbf{3.3 \times 10^{-7}}$	$\mathbf{1.2 \times 10^{-7}}$	$\mathbf{8.0 \times 10^{-7}}$	$\mathbf{3.9 \times 10^{-9}}$	$\mathbf{1.2 \times 10^{-4}}$

表12：标准 Transformer 和我们提出的 LSTNet 的 SPICE 评价指标子类别。P 值通过对配对样本进行双尾 t 检验计算得出。请注意，加粗的 P 值在 0.05 显著性水平下具有显著性。

模型	SPICE					
	关系	基数	属性	Size	颜色	物体
Transformer	6.91	20.58	11.80	4.71	12.93	40.35
LSTNet	7.06	22.68	12.38	4.84	14.60	40.76
p值	0.298	0.059	**0.002**	**0.048**	**0.001**	**0.009**

表13：与 Flickr8k 数据集上最先进方法的比较。所有值均以百分比（%）报告，其中 B-N、M、R 和 C 分别是 BLEU-N、METEOR、ROUGE-L 和 CIDEr 分数的简称。†表示集成模型的结果。

方法	B-1	B-4	M	R	C
Deep VS [40]	57.9	16.0	-	-	-
Google NIC [15]†	63.0	-	-	-	-
软注意力 [26]	67.0	19.5	18.9	-	-
硬注意力 [26]	67.0	21.3	20.3	-	-
嵌入式门控长短期记忆网络 [48]	64.7	21.2	20.6	-	-
对数双线性 [49]	65.6	17.7	17.3	-	-
LSTNet	**67.4**	**24.3**	**21.5**	**44.8**	**63.6**

图7中的这一行里，Transformer关注的是桌子，而LSTNet关注的是男人和手机。因此，Transformer生成了错误描述（即“领带”、“桌子”），而LSTNet则正确识别出“一个男人正拿着手机”。2）Transformer只能关注图像中的一个物体或一小块区域，而LSTNet会关注更多主要对象，从而生成准确而详细的描述。例如，对于图7第二行中的图像，Transformer只关注男孩的嘴，而LSTNet同时关注男孩和泰迪熊。因此，Transformer未能识别“泰迪熊”，只生成了一个笼统的短语（即“一只毛绒动物”）。得益于编码器中的精准注意力，LSTNet成功识别出“一个小男孩正抱着一只泰迪熊”。这些观察表明，我们提出的LSA促使模型不仅关注图像中的重要信息，也关注全面的信息。

4.7. 注意力可视化

为了更好地定性评估 LSTNet 生成的结果，我们在图 6 中可视化了图像描述生成过程中视觉特征中每个网格的贡献。具体而言，我们对最后一个解码器层中 8 个注意力头的注意力权重取平均。可以观察到，Transformer 会关注无关区域，从而生成错误的描述（例如，“笔记本电脑”）。相比之下，我们提出的 LSTNet 在生成“椅子”和“雨伞”等信息丰富的词语时，能够聚焦于正确的网格。这些观察结果表明，我们提出的 LSA 和 LSF 模块通过局部交互与融合，为解码器提供更丰富、

表14：在Flickr30k数据集上与当前最先进方法的比较。所有值均以百分比（%）表示，其中 B-N、M、R 和 C 分别是 BLEU-N、METEOR、ROUGE-L 和 CIDEr 分数的简称。†表示集成模型的结果。

方法	B1	B4	M	R	C
Deep VS [40]	57.3	15.7	-	-	-
Google NIC [15]†	66.3	18.3	-	-	-
m-RNN [50]	60.0	19.0	-	-	-
软注意力 [26]	66.7	19.1	18.5	-	-
硬注意力 [26]	66.9	19.9	18.5	-	-
嵌入式门控长短期记忆网络 [48]	64.6	20.6	17.9	-	-
ATT [51]†	64.7	23.0	18.9	-	-
对数双线性 [49]	60.0	17.1	16.9	-	-
LSTNet	**67.1**	**23.3**	**20.4**	**44.3**	**64.5**

更细粒度的视觉特征，从而帮助模型在图像描述中始终聚焦于正确的区域。

4.8. 在 *Flickr* 数据集上的泛化

为了验证我们提出的 LSNet 的泛化能力，我们还在 Flickr8k 和 Flickr30k 数据集上进行了大量实验。我们提出的 LSTNet 与先前最先进方法在 Flickr8k[52] 和 Flickr30k [53] 上的性能比较分别见表 13 和表 14。可以看出，在 Flickr8k 和 Flickr30K 上，我们提出的 LSTNet 均显著优于先前的最先进方法。这验证了我们提出的 LSTNet 在其他数据集上具有很强的泛化能力。

5. 结论

本文提出了 LSTNet，这是一种用于图像描述的新型局部性感知 Transformer 网络，通过利用局部交互与融合，提升基于网格特征的目标识别能力。具体而言，我们设计了 LSA 模块，在编码器层内建模关系，从而有助于获取每个物体的完整语义信息，并细化视觉特征的细节。随后，我们引入 LSF 模块，通过建模相邻网格之间的关系来融合不同层的视觉特征，进而得到语义信息更丰富的视觉特征。在 MS-COCO数据集上的实验结果表明，

LSTNet 的性能显著优于以往的最先进水平模型。大量消融研究和可视化比较进一步揭示了 LSTNet 各个组件的有效性及其作用机制。在 Flickr8k 和 Flickr30k 数据集上开展的额外实验也验证了 LSTNet 在其他数据集上的泛化能力。尽管所提出的 LSTNet 显著提升了目标定位能力和图像描述性能，但也引入了额外的计算量和参数。不过，与原始 Transformer 的自注意力相比，LSA 和 LSF 模块引入的参数和计算量微不足道。

参考文献

[1] M. Cornia、L. Baraldi、R. Cucchiara，Show, control and tell：用于生成可控且有依据的图像描述的框架，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2019年，第8307–8316页。[2] M. Cornia、M. Stefanini、L. Baraldi、R. Cucchiara，用于图像描述的网格记忆 Transformer，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第10578–10587页。[3] T.-Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan、P. Doll´ar、C. L. Zitnick，Microsoft COCO：上下文中的常见物体，载于：欧洲计算机视觉会议，2014年。[4] B. A. Plummer、L. Wang、C. M. Cervantes、J. C. Caicedo、J. Hocken- maier、S. Lazebnik，Flickr30k实体：收集区域到短语的对应关系，以构建更丰富的图像到句子模型，载于：IEEE国际计算机视觉会议论文集，2015年，第2641–2649页。[5]P. Anderson、X. He、C. Buehler、D. Teney、M. Johnson、S. Gould、L. Zhang，用于图像描述和视觉问答的自底向上和自顶向下注意力，载于：I EEE/CVF计算机视觉与模式识别会议论文集，2018年，第6077–6086页。[6] S. Ren、K. He、R. Girshick、J. Sun，Faster R-CNN：借助区域提议网络实现实时目标检测，《IEEE模式分析与机器智能汇刊》39（6）（2016）1137–1149。

[7] R. Krishna、Y. Zhu、O. Groth、J. Johnson、K. Hata、J. Kravitz、S. Chen、Y. Kalantidis、L.-J. Li、D. A. Shamma等，Visual Genome：利用众包密集图像标注连接语言与视觉，《国际计算机视觉杂志》123（1）（2017）32–73。[8] Y. Pan、T. Yao、Y. Li、T. Mei，面向图像描述的X线性注意力网络，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第10971–10980页。[9] L. Huang、W. Wang、J. Chen、X.-Y. Wei，面向图像描述的注意力之上的注意力，载于：IEEE国际计算机视觉会议论文集，2019年，第4634–4643页。[10] H. Jiang、I. Misra、M. Rohrbach、E. Learned-Miller、X. Chen：为视觉问答中的网格特征辩护，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第10267–10276页。[11] X. Zhang、X. Sun、Y. Luo、J. Ji、Y. Zhou、Y. Wu、F. Huang、R. Ji，Rst-net：利用自适应注意力描述视觉与非视觉词语，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第15465–15474页。[12] Y. Luo、J. Ji、X. Sun、L. Cao、Y. Wu、F. Huang、C.-W. Lin、R. Ji，面向图像描述的双层协作Transformer，载于：AAAI人工智能会议论文集，第3期，2021年，第2286–2293页。[13]H. Wu、B. Xiao、N. Codella、M. Liu、X. Dai、L. Yuan、L. Zhang，Cvt：将卷积引入视觉Transformer，arXiv预印本arXiv:2103.15808（2021）。[14] V.-Q. Nguyen、M. Suganuma、T. Okatani，GRIT：利用双重视觉特征实现更快、更优的图像描述生成Transformer，载于：欧洲计算机视觉会议，Springer，2022年，第167–184页。[15] O. Vinyals、A. Toshev、S. Bengio、D. Erhan，看图说话：神经图像描述生成器，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2015年，第3156–3164页。

[16] J. Lu、C. Xiong、D. Parikh、R. Socher，了解何时关注：通过用于图像描述的视觉哨兵实现自适应注意力，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2017年，第375–383页。[17] J. Ji、Y. Ma、X. Sun、Y. Zhou、Y. Wu、R. Ji，了解学习什么：一种用于图像描述的面向度量的焦点机制，《图像处理汇刊》（2022年），第1–1doi:10.1109/TIP.2022.3183434页。[18] Y. Ma、J. Ji、X. Sun、Y. Zhou、Y. Wu、F. Huang、R. Ji，了解它是什么：语义增强双注意力变换器，《IEEE多媒体汇刊》（2022年），第1–1doi:10.1109/TMM.2022.3164787页。[19] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. Kaiser、I. Polosukhin，注意力即一切，《神经信息处理系统进展》（2017年）。[20] C. Wang、Y. Shen、L. Ji，结合位置感知 LSTM 的几何注意力 Transformer，用于图像描述，《专家系统与应用》201（2022年）117174。 doi:https://doi.org/10.1016/j.eswa.2022.117174。URL https://www.sciencedirect.com/science/article/pii/S0957417422005619[21]Y. Wang、J. Xu、Y. Sun，用于图像描述的视觉持久性模型，《神经计算》468（2022年）48–59。doi:https://doi.org/10.1016/j.neucom.2021.10.014。URL https://www.sciencedirect.com/science/article/pii/S0925231221014922[22] Z. Zhang、Q. Wu、Y. Wang、F. Chen，从图像描述的语言上下文中自适应探索成对关系，《IEEE多媒体汇刊》（2021年）。[23] X. Yang、Y. Liu、X. Wang，Reformer：用于图像描述的关系变换器，载于：第30届 ACM国际多媒体会议论文集，2022年，第5398–5406页。[24] K. He、X. Zhang、S. Ren、J. Sun，用于图像识别的深度残差学习，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2016年，第770–778页。

[25] A. Rosenfeld、M. Thurston，用于视觉场景分析的边缘与曲线检测，《IEEE计算机汇刊》C-20（5）（1971年）562–569页。doi:10.1109/T-C.1971.223290。[26] K. Xu、J. Ba、R. Kiros、K. Cho、A. Courville、R. Salakhudinov、R. Zemel、Y. Bengio，看图、关注并讲述：结合视觉注意力的神经图像描述生成，载于：国际机器学习会议，机器学习研究论文集，2015年，第2048–2057页。[27] O. Vinyals、A. Toshev、S. Bengio、D. Erhan，看图说话：从2015年MS COCO图像描述挑战赛中汲取的经验，《IEEE模式分析与机器智能汇刊》39（4）（2016年）652–663页。[28] Y. Ma、G. Xu、X. Sun、M. Yan、J. Zhang、R. Ji，X-CLIP：面向视频文本检索的端到端多粒度对比学习，载于：第30届 ACM国际多媒体会议论文集，2022年，第638–647页。[29] S. Ioffe、C. Szegedy，批归一化：通过减少内部协变量偏移加速深度网络训练，载于：国际机器学习会议，机器学习研究论文集，2015年，第448–456页。[30] X. Ding、X. Zhang、N. Ma、J. Han、G. Ding、J. Sun，RepVGG：让 VGG风格卷积网络再创辉煌，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第13733–13742页。[31] X. Ding、X. Zhang、J. Han、G. Ding，多样分支块：将卷积构建为 Inception式单元，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第10886–10895页。[32] S. J. Rennie、E. Marcheret、Y. Mroueh、J. Ross、V. Goel，用于图像描述的自批判序列训练，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2017年，第7008–7024页。[33] T. Yao、Y. Pan、Y. Li、Z. Qiu、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。

[34] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, T. Zhang，面向图像描述的循环融合网络，载于：欧洲计算机视觉会议，2018，页码499–515。[35] T. Yao, Y. Pan, Y. Li, 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，2018。[36] X. Yang, K. Tang, H. Zhang, J. Cai，面向图像描述的场景图自编码，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2019，页码10685–10694。[37] Z.-J. Zha, D. Liu, H. Zhang, Y. Zhang, F. Wu，面向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》44 (2) (2022) 710–722。doi:10.1109/TPAMI.2019.2909864.[38] G. Li, L. Zhu, P. Liu, Y. Yang，面向图像描述的纠缠变换器，载于：IEEE国际计算机视觉会议论文集，2019，页码8928–8937。[39]S. Herdade, A. Kappeler, K. Boakye, J. Soares，图像描述：将物体转换为词语，《神经信息处理系统进展》(2019)。[40] A. Karpathy, 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，页码3128–3137。[41] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu，BLEU：一种机器翻译自动评估方法，载于：第40届计算语言学协会年会论文集，2002，页码311–318。[42] S. Banerjee, A. Lavie，METEOR：一种与人工判断相关性更高的机器翻译评估自动指标，载于：第40届计算语言学协会年会论文集，2005，页码65–72。[43] C.-Y. Lin，ROUGE：用于摘要自动评估的工具包，载于：第40届计算语言学协会年会论文集，2004，页码74–81。

[44] R. Vedantam, C. Lawrence Zitnick, D. Parikh，CIDEr：基于共识的图像描述评估，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2015年，页码4566–4575。[45]P. Anderson, B. Fernando, M. Johnson, S. Gould，SPICE：语义命题图像描述评估，载于：欧洲计算机视觉会议，Springer，2016，页码382–398。[46] S. Xie, R. Girshick, P. Doll´ar, Z. Tu, K. He，深度神经网络的聚合残差变换，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2017年，页码1492–1500。[47] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia，实例分割的路径聚合网络，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2018，页码8759–8768。[48] X. Jia, E. Gavves, B. Fernando, T. Tuytelaars，引导长短期记忆模型进行图像描述生成，载于：IEEE国际计算机视觉会议论文集，2015年，页码2407–2415。[49] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venu-gopalan, K. Saenko, T. Darrell，用于视觉识别与描述的长期循环卷积网络，载于：IEEE计算机视觉与模式识别会议论文集，2015年，页码2625–2634。[50] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille，基于多模态循环神经网络的深度描述生成（M-RNN），arXiv预印本 arXiv:1412.6632 (2014)。[51] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo，基于语义注意力的图像描述，载于：IEEE计算机视觉与模式识别会议论文集，2016，页码4651–4659。[52] M. Hodosh, P. Young, J. Hockenmaier，将图像描述构建为排序任务：数据、模型与评估指标，《人工智能研究杂志》47 (2013) 853–899。[53] P. Young, A. Lai, M. Hodosh, J. Hockenmaier，从图像描述到视觉指称：用于语义推理的新相似度指标

事件描述，《计算语言学会刊》2（2014）67–78。