# 基于多层级视觉融合的图像描述模型

周东明[1],张灿龙[1],李志欣[1],王智文[2]
(1. 广西师范大学广西多源信息挖掘与安全重点实验室,广西桂林 541004;
2. 广西科技大学计算机科学与通信工程学院,广西柳州 545006)

**摘 要:** 传统方法在视觉策略网络中只关注实体,不能够推理出实体和属性之间的联系,在语言策略网络存在暴露偏差和误差累计问题.为此,提出了一个基于强化学习的多层级视觉融合网络模型.在视觉策略网络中通过多层级神经网络模块将视觉特征转化为视觉知识的特征集.融合网络生成使描述语句更加流畅的虚词,用于视觉策略网络和语言策略网络的互动.在语言策略网络中使用基于强化学习的自批评策略梯度算法对视觉融合网络实现端到端的优化.实验结果表明,该模型在MS-COCO数据集取得不错效果,将Karpathy分割测试中的CIDEr值从120.1提高到124.3.

**关键词:** 图像描述;视觉融合;强化学习;策略网络;机器学习;注意力机制
**中图分类号:** TP181 **文献标识码:** A **文章编号:** 0372-2112 (2021) 07-1286-05
**电子学报URL:** http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20191296

# Image Captioning Model Based on Multi-Level Visual Fusion

ZHOU Dong-ming[1], ZHANG Can-long[1], LI Zhi-xin[1], WANG Zhi-wen[2]
(1. *Guangxi Key Laboratory of Multi-source Information Mining and Security, Guangxi Normal University, Guilin, Guangxi* 541004, *China*;
2. *School of Computer Science and Communication Engineering, Guangxi University of Science and Technology, Liuzhou, Guangxi* 545006, *China*)

**Abstract:** Traditional methods only focus on entities in the visual strategy network and cannot deduce the relationship between entities and attributes. There are problems of exposure bias and error accumulation in the language strategy network. Therefore, this paper proposes a multi-level visual fusion network model based on reinforcement learning. In the visual strategy network, multi-level sub-neural network module is used to transform visual features into feature sets of visual knowledge. The fusion network generates the function words which make the description sentences more fluent and can be used for the interaction between the visual strategy network and the language strategy network. The gradient algorithm of self-criticism strategy based on reinforcement learning is used to optimize the visual fusion network end-to-end. The experimental results show that the model can get good results in MS-COCO data set and improve the CIDEr value of Karpathy segmentation test from 120.1 to 124.3.

**Key words:** image captioning; visual fusion; reinforcement learning; strategy network; machine learning; attention mechanism

## 1 引言

图像描述可以理解为给定一张图片,生成一段自然语言描述的文字.图像描述和视觉问答[1]属于计算机视觉和自然语言处理的交叉的领域.图像描述在盲人导航、儿童早教、图文检索等方面有着广泛应用前景.受机器翻译的启发,图像描述中广泛使用端到端的编码—解码框架[2].编码端使用卷积神经网络(Convolutional Neural Networks,CNN)提取图像特征,解码端将提取到的图像特征输入到长短期记忆网络(Long Short-Term Memory,LSTM)中,然后输出描述图像的序列.然而,CNN在提取视觉特征时并不能根据上下文来辨识情景之间的关系.在使用视觉注意力机制[3]时每

收稿日期:2019-11-21;修回日期:2021-02-04;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61866004, No.61663004, No.61966004, No.61962007, No.61751213);广西自然科学基金(No.2018GXNSFDA281009, No.2017GXNSFAA198365, No.2019GXNSFDA245018, No.2018GXNSFDA29400);广西"八桂学者"创新研究团队;广西多源信息挖掘与安全重点实验室基金(No.20-A-03-01);广西研究生教育创新计划(No. XYCSZ2020071)

一步只能固定一个视觉区域,随着时间步的推移,注意力机制作用减弱,不同视觉区域之间也缺乏交互. 因此在应对复杂的场景组合时,描述序列会随着时间的推移而出现错误. Rennie 等[4]提出了一种自批评训练方法,由于偏置基线可以是任意函数,不依赖动作,该模型使用测试阶段生成词的奖励作为基线. Lu 等[5]发现非视觉词的梯度会误导或者降低视觉信息的有效性,提出了一种自适应门控机制,解码器在针对不同的单词时有不同的语言策略. Anderson 等[6]把目标检测技术应用在图像描述,提出了一种从下到上的注意力机制,能使输出的图像描述更自然,但是并不能推理出图像中实体和属性间的关系.

因此提出的多层级视觉融合网络(Multi-level Visual Fusion Networks,MVF-Net)在生成序列时并不只固定当前的视觉注意力,而是将上一个时间步的视觉信息解释为情景,然后根据当前的视觉注意力感知来判断情景是否有利于下一个单词的生成[7]. 首先目标检测网络检测出图像中的实体并考虑其相关区域,属性网络将 CNN 提取的属性转化实体属性知识特征集,生成描述实体的形容词. 然后关系注意力网络多步推理实体和实体、实体和属性间的联系. 最后在生成序列时,融合网络能使生成的描述更加连贯,和符合语法规则. 主要贡献为:提出了一个端到端的 MVF-Net 神经模块网络,神经模块间的搭配丰富了视觉语言任务. 然后设计关系模块,让实体间进行交互,增加描述的细粒度.

## 2　模型方法

系统模型包括三个部分:视觉网络,融合网络,语言网络. 视觉网络生成用于语言解码的特征向量,融合网络融合多层级视觉信息,语言网络的 LSTM 将部分累计的情景感知输入到模块控制器和融合网络中,进行多步推理. 其结构图,如图 1 所示.

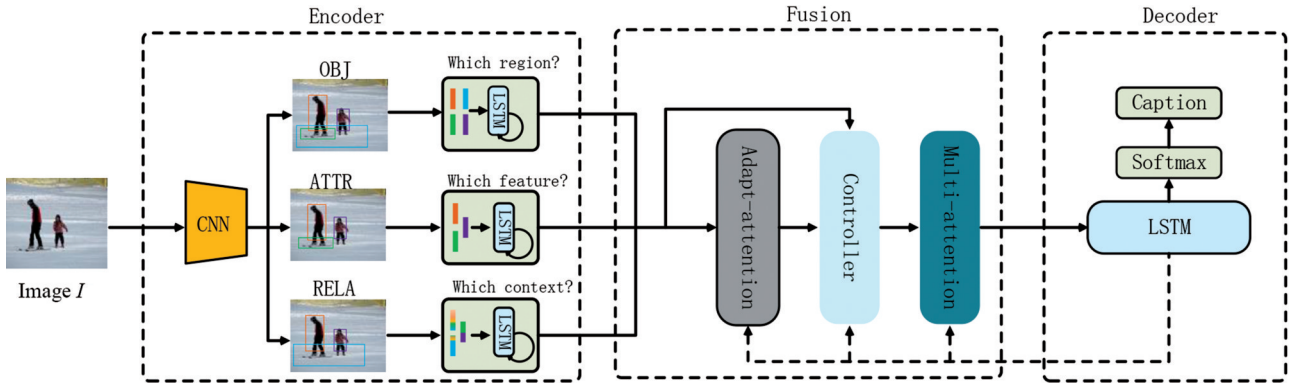


图 1　多层级视觉融合网络结构图(在编码器之后加入一个循环融合过程,以便更好表示解码)

### 2.1　视觉策略网络

给定图像 $\boldsymbol{I}$,图像的视觉特征为 $\{v_1, v_2, \cdots, v_t\}$,$v_t \in \mathbb{R}^D$,对应的描述序列 $Y = \{y_1, y_2, \cdots, y_T\}$,$y_t \in \Sigma$,$\Sigma$ 是图像标签的词汇表. 定义 $\varphi$ 为智能体符号,$s_t$ 为环境,$h_t$ 为 LSTM 网络的输出,$\theta$ 为视觉策略网络学习的超参数. 在 $t$ 时刻视觉策略网络中,环境将一种状态进行编码,输入到智能体中,驱动智能体产生一个动作 $a_t \sim \varphi(s_t | h_t; \theta)$,然后环境将智能体产生的动作 $v_t$ 在环境状态中编码,与此同时环境反馈一个奖励给智能体. 在这里 LSTM 的输出可表示为:

$$h_t = \mathrm{LSTM}(\boldsymbol{x}_t, h_{t-1}) \tag{1}$$

其中,$\boldsymbol{x}_t$ 是 LSTM 网络在当前时刻输入的向量,$h_{t-1}$ 是 LSTM 网络上一个时刻的输出. 编码视觉模块的时候,遵循注意力机制原理,每一个视觉模块的注意力权重为:

$$\delta_{i,t} = \boldsymbol{W}_a^{\mathrm{T}} \mathrm{Tanh}(W_h h_t + W_P v_i) \tag{2}$$

在这里,Tanh 为激活函数. $W_a, W_h, W_p$ 为训练过程中学习到的超参数,$v_i$ 为图像第 $i$ 个视觉特征. 定义视觉策略模块的输入为 $p_t$ 和编码的视觉特征向量为 $\hat{\boldsymbol{v}}$,那么视觉策略网络中 LSTM 编码的环境状态为:

$$x_t = \left[h_{t-1}^2, \hat{\boldsymbol{v}}, \boldsymbol{W}_e \boldsymbol{\Pi}\left(y_{t-1}\right)\right] \tag{3}$$

参数 $\boldsymbol{W}_e$ 是一个词嵌入矩阵,$\boldsymbol{\Pi}$ 是 one-hot 编码矩阵. $h_{t-1}^2$ 为语言策略网络上一个时刻的隐状态. $\hat{\boldsymbol{v}}$ 表示视觉策略模块编码的视觉特征向量,对于不同的视觉策略模块具有不同视觉编码特征作为 LSTM 网络的输入.

### 2.2　融合网络

融合网络中包含三个模块. 自适应注意力模块,模块搭配控制器和多模态注意力. 自适应模块用于降低非视觉单词梯度对视觉信息的有效性,模块搭配控制器用于搭配视觉网络中的模块和自适应模块生成完整描述句子,多模态注意力可以可视化描述输出.

自适应注意力模块功能是生成使描述序列更加流畅的非视觉信息单词,比如“a”或者“an”. 其结构图如

图2所示，自适应注意力模块分离非视觉单词的生成. 在每一时刻由自适应门控单元$\hat{c}_t$决定虚词“a”或者“an”由模型产生还是由语言网络生成，$\hat{c}_t$可表示为：

$$\hat{c}_t = \beta_t x_t + (1-\beta_t)h_t \tag{4}$$

实验过程中选取$\beta_t = 0.5$时，实验效果最佳. 在计算求得自适应上下文向量$\hat{c}_t$以后，根据式(2)的注意力机制原理，在每一个时间步$t$，标准化注意力权重$\delta_{i,t}$，可知自适应注意力模块的特征向量为：

$$\hat{\boldsymbol{v}}_f = \sum_{i=1}^{k}\delta_{i,t}\hat{c}_t \tag{5}$$

由自适应特征向量$\hat{\boldsymbol{v}}_f$，和LSTM在上一个时刻的输出$h_{t-1}^2$以及词嵌入矩阵$\boldsymbol{W}_e$，可知自适应注意力的环境编码为：

$$x_t^f = \left[h_{t-1}^2, \hat{\boldsymbol{v}}_f, \boldsymbol{W}_e\boldsymbol{\Pi}(y_{t-1})\right] \tag{6}$$

模块控制器将用于描述序列的词性搭配，控制器生成四个融合软权重. 多模态视觉注意力用于采样可视化输出.

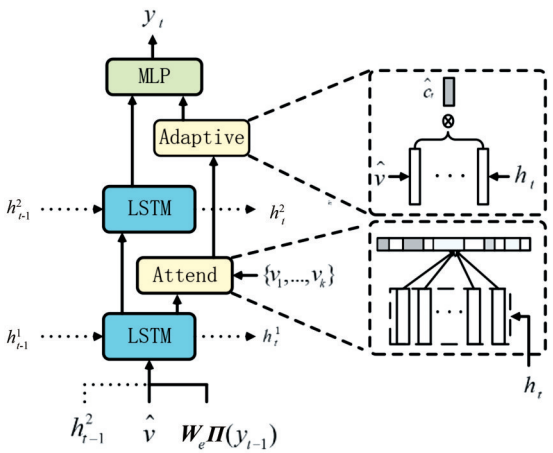


图2 自适应模块结构图

## 2.3 语言网络

在每一个时间步，MVF-Net生成一个融合情景的可视化表示，选取最适合当前的单词. 语言策略网络以多模态的视觉特征向量和子网络的隐状态$h_t^1$输入，然后更新LSTM的隐状态：

$$h_{t+1}^1 = \mathrm{LSTM}\left(x_t^f, h_t^1\right) \tag{7}$$

在计算词汇表中单词的分布时，使用全连接层作为LSTM的隐状态，经过softmax函数归一化后每个单词的概率可表示为：

$$\varphi^l\left(y_t|y_{1:t-1}\right) = \mathrm{softmax}\left(W_y h_t^2 + b_y\right) \tag{8}$$

其中$b_y$是偏置值，$W_y$是权重参数，两者都是在训练中学习得到. 整个完整的描述序列为所有时间步长条件分布的乘积，可表示为：

$$\varphi^l\left(y_{1:T}\right) = \prod_{t=1}^{T}\varphi^l\left(y_t|y_{1:t-1}\right) \tag{9}$$

在训练的时候，使用的是两阶段训练：全监督中的交叉熵训练，使用强化学习训练解决暴露偏差. 在第一阶段中，给定句子的真实描述序列$y_{1:T}^{gt}$，描述模型的参数为$\theta$，那么传统的全监督学习的损失函数可以定义为：

$$L_{XE}(\theta) = -\sum_{t=1}^{T}\log\left(\varphi_\theta\left(y_t^{gt}|y_{1:t-1}^{gt}\right)\right) \tag{10}$$

这种全监督的学习方式和强化学习中的模仿学习相对应，在第二阶段训练时，使用预训练模型初始化策略网络. 在解决暴露偏差和优化评价指标问题上使用策略梯度方法把期望奖励最大化. 比如，使用CIDEr分数作为训练时的奖励，从第一阶段的断点处继续训练，最小化负期望得分为：

$$L_R(\theta) = -E_{y_{1:T}\sim\varphi_\theta}\left[r\left(y_{1:T}\right)\right] \tag{11}$$

参数$r$是选择的评价指标，使用SCST方法，梯度损失就能够被近似为：

$$\nabla_\theta L_R(\theta) \approx -\left(r\left(y_{1:T}^c\right) - r\left(\hat{y}_{1:T}\right)\right)\nabla_\theta \log\varphi_\theta\left(y_{1:T}^c\right) \tag{12}$$

参数$\hat{y}_{1:T}$是生成描述序列时使用贪心算法采样单词，$y_{1:T}^c$是生成描述序列时根据词概率分布采样单词(例如：蒙特卡洛采样).

# 3 实验结果与分析

## 3.1 测试环境

试验环境的操作系统使用的是Ubuntu16.04，试验代码基于深度学习框架PyTorch 1.0版本和Python3.6编写，运用Tesla P100专业的深度学习显卡进行加速训练. 整个模型的时间复杂度为$O(n^2)$，空间复杂度为$O(n)$.

## 3.2 实验结果

实验对比了Top-Down[6]，NBT[7]，POS[8]，AESG[9]，SCA[10]，UIC[11]，RFN[12]. 结果如表1所示，MVF-Net性能指标优于最先进的方法. 原因是可视化视觉融合对生成当前时刻的单词有重要作用，强化学习解决了暴露偏差和误差累计等问题，优化的策略梯度可以做出更合理的决策. 和Top-Down模型相比，本文提出MVF-Net模型在CIDEr评价指标上有一定提升. 实验可知Top-Down模型只是将目标检测技术应用到卷积神经网络中，而MVF-Net模型在使用检测技术的同时考虑了实体和属性之间的关系，因此实验评价指标更高.

使用策略梯度对生成序列进行训练时，奖励函数$r(\cdot)$可以选择CIDEr，BLEU，METEOR，ROUGE和SPICE. 如表2，横轴表示训练时评价指标，纵轴表示评

**表1　使用Karpathy分割离线测试性能比较.性能指标:B@4,M,R,C和S分别代表BLEU@N,METEOR,ROUGE,CIDEr和SPICE**

| 对比方法 | B@4 | M | R | C | S |
|---|---|---|---|---|---|
| Top-Down[6] | 36.2 | 27.0 | 56.4 | 120.1 | 20.3 |
| NBT[7] | 34.7 | 27.3 | – | 108.9 | 20.1 |
| POS[8] | 34.6 | 26.6 | 56.1 | 112.7 | 19.9 |
| AESG[9] | 34.4 | 26.7 | 55.8 | 116.2 | – |
| SCA[10] | 31.6 | 25.1 | 53.5 | 95.3 | – |
| UIC[11] | 33.5 | 26.4 | 54.8 | 107.9 | 19.6 |
| RNF[12] | 31.6 | 25.1 | 53.5 | 95.3 | – |
| MVF-Net | 37.5 | 28.3 | 58.5 | 124.3 | 19.4 |

**表2　不同指标作为奖励在MS-COCO数据集上进行Karpathy测试**

| Metric | B | M | R | C | S |
|---|---|---|---|---|---|
| B | 37.5 | 33.6 | 36.4 | 36.3 | 37.0 |
| M | 26.5 | 28.4 | 27.3 | 27.2 | 26.8 |
| R | 56.7 | 57.6 | 58.1 | 56.4 | 57.0 |
| C | 114.5 | 113.4 | 120.0 | 124.4 | 122.3 |
| S | 20.2 | 19.4 | 19.8 | 20.6 | 21.1 |

估时评价指标,对不同的评价指标进行优化会得到不同的结果.通过实验对比可知,训练时对特定的评价指标进行优化,测试时该指标可获得最佳性能.优化BLEU和CIDEr整体实验性能最好,但是使用BLEU作为奖励时花费在计算上的开销超过CIDEr,因此使用CIDEr作为评价指标.

### 3.3　描述可视化

为了更好的理解MVF-Net模型,如图3所示,可视化语言策略网络的输出预测.多层级视觉融合网络不仅可以关注图中单一实体对象,如:男人,海浪和冲浪板.而且可以生成组合单词放置,连接海浪和冲浪板.融合网络将生成用于连接实体单词,比如“a”,使得描述序列的语法和语义更加流畅.语言策略网络将融合过的图像特征进行解码,生成的单词描述更像人类的风格.不再是只对图像的单一特征进行描述,而是基于对场景的深刻理解的情况下生成描述句子,因此可以避免生成死板的描述序列.

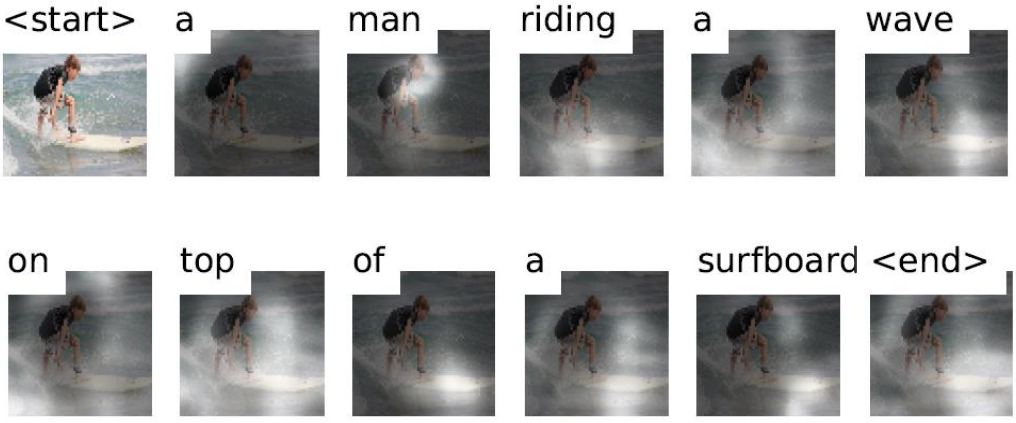


图3　MVF-Net模型可视化描述序列生成过程

## 4　总结

提出了一种基于视觉融合网络的图像描述模型.在进行广泛的对比实验和消融实验的基础之上,在MS-COCO数据集上测试了MVF-Net模型的有效性.未来的工作中,把MVF-Net模型应用到其他视觉推理任务中——可视化视觉问答系统,将探索情景融合在场景图生成和视频描述中的可迁移性.

**参考文献**

[1] Chen S, Jin Q, Wang P. Say as you wish: fine-grained control of image caption generation with abstract scene graphs[A].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2020. 9962 – 9971.

[2] Shi J, Zhang H, Li J. Explainable and explicit visual reasoning over scene graphs[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. California: IEEE, 2019. 8376 – 8384.

[3] 张志昌,曾扬扬,庞雅丽.融合语义角色和自注意力机制的中文文本蕴含识别[J].电子学报,2020,48(11): 2162 – 2169.
ZHANG Zhi-chang, ZENG Yang-yang, PANG Ya-li. A Chinese textual entailment recognition method incorporating semantic role and self-attention[J]. Acta Electronica

Sinica, 2020, 48(11): 2162 – 2169.(in Chinese)
[4] Rennie S J, Marcheret E, Mroueh Y. Self-critical sequence training for image captioning[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Hawaii:IEEE, 2017. 7008 – 7024.
[5] Lu J, Yang J, Batra D. Neural baby talk[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City: IEEE, 2018. 7219 – 7228.
[6] Anderson P, He X, Buehler C. Bottom-up and top-down attention for image captioning and visual question answering[A].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City:IEEE, 2018. 6077 – 6086.
[7] 汤鹏杰, 王瀚漓, 许恺晟. LSTM 逐层多目标优化及多层概率融合的图像描述[J]. 自动化学报, 2018, 44(7): 1237 – 1249.
TANG Peng-jie, WANG Han-li, XU Kai-sheng. Multi-objective layer-wise optimization and multi-level probability fusion for image description generation using lstm[J]. Acta Automatica Sinica, 2018, 44(7): 1237 – 1249.(in Chinese)
[8] Deshpande A, Aneja J, Wang L. Fast, diverse and accurate image captioning guided by part-of-speech[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. California:IEEE, 2019. 10695 – 10704.
[9] Yang X, Tang K, Zhang H. Auto-encoding scene graphs for image captioning[A].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. California:IEEE, 2019. 10685 – 10694.
[10] Chen L, Zhang H, Xiao J. SCA-CNN: spatial and channel- wise attention in convolutional networks for image captioning[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Hawaii: IEEE, 2017. 5659 – 5667.
[11] Feng Y, Ma L, Liu W. Unsupervised image captioning [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. California: IEEE, 2019. 4125 – 4134.
[12] Jiang W, Ma L, Jiang Y. Recurrent fusion network for image captioning[A].Proceedings of the European Conference on Computer Vision[C]. Germany: Springer, 2018. 499 – 515.

## 作者简介

**周东明** 男,1995年5月出生于河南省信阳市,现为广西师范大学硕士研究生,主要研究方向为机器学习与图像处理.
E-mail:dmzhou1995@163.com

**张灿龙(通讯作者)** 男,1975年出生于湖南省娄底市.广西师范大学教授,博士生导师. 毕业于上海交通大学,获控制理论与控制工程专业博士学位. 主要从事计算机视觉与机器学习.
E-mail:zcltyp@163.com