LSTM 逐层多目标优化及多层概率融合的图像描述

汤鹏杰 [1, 2, 3] 王瀚漓 [1, 2] 许恺晟 [1, 2]

摘 要 使用计算模型对图像进行自动描述属于视觉高层理解, 要求模型不仅能够对图像中的目标及场景进行描述, 而且能够对目标与目标之间、目标与场景之间的关系进行表达, 同时能够生成符合一定语法和结构的自然语言句子. 目前基于深度卷积神经网络 (Convolutional neural network, CNN) 和长短时记忆网络 (Long-short term memory, LSTM) 的方法已成为解决该问题的主流, 虽然已取得巨大进展, 但存在 LSTM 层次不深, 难以优化的问题, 导致模型性能难以提升, 生成的描述句子质量不高. 针对这一问题, 受深度学习思想的启发, 本文设计了基于逐层优化的多目标优化及多层概率融合的 LSTM (Multi-objective layer-wise optimization/multi-layer probability fusion LSTM, MLO/MLPF-LSTM) 模型. 模型中首先使用浅层 LSTM 进行训练, 收敛之后, 保留原 LSTM 模型中的分类层及目标函数, 并添加新的 LSTM 层及目标函数重新对模型进行训练, 对模型原有参数进行微调; 在测试时, 将多个分类层使用 Softmax 函数进行变换, 得到每层对单词的预测概率分值, 然后将多层的概率分值进行加权融合, 得到单词的最终预测概率. 在 MSCOCO 和 Flickr30K 两个数据集上实验结果显示, 该模型性能显著, 在多个统计指标上均超过了同类其他方法.

关键词 图像描述, 多目标优化, 逐层优化, 多层融合, 长短时记忆网络, 卷积神经网络

引用格式 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. 自动化学报, 2018, **44**(7): 1237−1249

DOI 10.16383/j.aas.2017.c160733

Multi-objective Layer-wise Optimization and Multi-level Probability Fusion for Image Description Generation Using LSTM

TANG Peng-Jie[1, 2, 3] WANG Han-Li[1, 2] XU Kai-Sheng[1, 2]

Abstract The task of image automatic description by computer belongs to high-level visual understanding. Unlike image classification, object detection, etc., it usually requires that the model should not only have abilities of describing scene and objects, but also have capacities of expressing the relations between different objects and between objects and background in the image. In addition, it is required that the model should generate natural sentences which accord with correct grammars and appropriate structures. Nowadays, the approaches based on convolutional neural network (CNN) and long-short term memory network (LSTM) have been the popular solutions to this task, and a series of successes have been obtained. However, there are still several sticky problems, for instance, the LSTM network is not deep enough and the model is difficult to optimize, and as a result, performances cannot be improved and the sentences generated are of low quantity. To address these difficulties, inspired by the idea of deep learning, a model named MLO/MLPF-LSTM is proposed, in which the method of layer-wise optimization, multi-objective optimization and multi-layer probability fusion are employed. In details, an LSTM network with shallow depth is trained firstly, then, new LSTM layers and related objective functions are added to the optimized LSTM network. Meanwhile, the classification layers and objective functions in the original LSTM model are reserved and fine-tuned with the new layers. During the test, the probabilities of all the Softmax functions which are fed with the corresponding classification layers are fused for the final predicted probabilities by a weighted average method. Experimental results on MSCOCO and Flickr30K datasets demonstrate that our model is effective and outperforms other methods of same kinds on a number of evaluation metrics.

Key words Image description, multi-objective optimization, layer-wise optimization, multi-level fusion, long-short term memory (LSTM), convolutional neural network (CNN)

Citation 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. *Acta Automatica Sinica*, 2018, **44**(7): 1237−1249

收稿日期 2016-10-25 录用日期 2017-03-02
Manuscript received October 25, 2016; accepted March 2, 2017
国家自然科学基金 (61622115, 61472281), 上海高校特聘教授 (东方学者) 跟踪计划 (GZ2015005), 江西省教育厅科学技术研究项目 (GJJ170643) 资助
Supported by National Natural Science Foundation of China (61622115, 61472281), Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (GZ2015005), and Scientific Research Foundation of the Education Bureau of Jiangxi Province (GJJ170643)
本文责任编委 王立威
Recommended by Associate Editor WANG Li-Wei

1. 同济大学计算机科学与技术系 上海 201804 2. 嵌入式系统与服务计算教育部重点实验室 (同济大学) 上海 200092 3. 井冈山大学数理学院 吉安 343009
1. Department of Computer Science and Technology, Tongji University, Shanghai 201804 2. Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 200092 3. College of Mathematical and Physical Science, Jinggangshan University, Ji′an 343009

通过计算机将一幅图像使用自然语言自动描述出来具有广泛的应用前景, 例如早期婴幼儿教育[1]、视觉生理功能障碍者辅助[1−2]、智能人机交互及机器人开发等. 该任务对于人类而言非常简单, 给定一幅图像, 人类能够很轻易地对图像中的信息进行形象化描述, 但对于计算机来说则非常困难. 它属于图像理解中的高层部分, 对于图像的语义信息理解要求较高; 不同于较为简单的图像分类和目标识别, 它不仅要求能够识别出图像中的目标, 还要求对目标的属性、动作、目标与目标之间的关系、目标与背景之间的关系等进行理解; 同时还要求系统能够将这些信息组合成人类易于理解的、具有一定语法结构的自然语言形式. 在众多工作中, 使用基于模板的方法生成图像描述非常具有代表性, 它将检测到的图像目标信息填入结构固定的句子模板中[3−7]; 此外, 还有基于转换的方法, 它通过检索相似的图像, 将已有的图像信息转移到待描述的图像上[8−11].

这些方法虽然具有一定的效果, 但也有着极大的局限性, 例如基于模板的方法不能生成新的句子结构, 基于转换的方法则不能描述图像中新出现的目标或场景. 受机器翻译技术的启发, 人们又提出基于 "编码–解码" 流程的图像描述模型. 它首先将图像视为源语言, 将其编码为特征向量, 然后使用循环神经网络 (Recurrent neural network, RNN) 等技术将其翻译成目标语言[12]. 这种方法生成的句子更加灵活, 也更符合人们的习惯, 目前该技术已经在图像描述任务上取得了重要进展, 尤其是基于卷积神经网络 (Convolutional neural network, CNN) 和长短时记忆网络 (Long-short term memory, LSTM) 的模型, 在多个数据集上都获得了较好的效果[1−2, 13−15]. 但目前基于该技术的模型也存在 LSTM 网络层次不深, 难以训练的缺点, 导致模型的性能受到限制, 其生成的句子在语义信息丰富程度及连贯性等方面效果欠佳.

为解决这一问题, 本文设计了基于逐层优化的多目标优化及多层概率融合的 LSTM (Multi-objective layer-wise optimization and multi-level probability fusion LSTM, MLO/MLPF-LSTM) 模型. 该模型借鉴了 Hinton 等提出的在深度学习中使用逐层优化的思想[16] 及文献 [17−19] 中的深度模型优化方法, 首先训练出一个浅层 LSTM 网络, 在此基础上, 为模型添加新的 LSTM 层, 继续训练, 同时对低层参数进行微调; 为使得低层参数能够得到进一步的优化, 也为了避免模型陷入过拟合状态, 给模型增加额外的正则化信息, 使用多目标优化策略[17−19], 在为 LSTM 网络添加新的层次时, 保留原有的分类层和目标函数. 在测试时, 提出多层概率融合的方法, 即通过投票的方式, 将多个分类层输出的概率分值使用加权平均的方法计算最终概率分值. 在 MSCOCO 和 Flickr30K 两个数据集上的实验结果表明, 本文模型生成的句子语义更加丰富, 质量更高, 在多个指标上均高于同类其他模型.

本文结构为: 第 1 节介绍图像描述方面的相关工作; 第 2 节展示设计的模型和方法, 对相关的 CNN、LSTM 等技术进行说明, 通过图示、形式化方法等对 MLO/MLPF-LSTM 模型进行描述; 第 3 节是实验, 通过多组实验对模型进行验证, 并与当前其他主流模型进行对比, 证明所提出模型的有效性. 第 4 节是结论, 总结本文工作, 并明确下一步的研究方向.

1 相关工作

使用自然语言描述视觉信息已具有一定的研究历史, 但早期的研究主要集中于视频描述领域[20], 人们借助模式识别和机器学习等技术开发视频到文字描述的转换系统. 这类系统一般采用手工特征, 系统鲁棒性不强, 且应用范围不广. 近期对于图像的描述生成也在快速发展, 其主要任务是给定一副图像, 让计算机能够自动识别出其中的背景、主要物体及图像中各部分之间的相互关系, 并将其转换成自然语言的形式表达出来. 它不同于传统的图像分类、模式识别及目标检测等, 是一种更为高级、更为复杂的图像理解任务.

在图像描述中, 基于模板的方法是常用的方法之一, 首先对图像中的物体、动作、场景等信息进行检测, 然后将对应的词汇填入格式固定的句子模板中, 从而将图像转换成自然语言[3−7]. 这种方法较为直观, 但要求为每类信息都加上明确的人工标注, 并严重依赖分类器的性能, 当训练数据较少时, 其性能受到极大制约; 此外, 由于模板固定, 生成的句子较为呆板, 灵活性不够, 与人工标注的参考句子相差较大.

除基于模板的方法之外, 基于转换的方法也是一种重要的图像描述生成策略. 在文献 [8−11] 的工作中, 首先在训练库中为测试样本检索相似的图像, 然后将检索到的图像描述转移到待测图像上, 进而生成图像描述. 这种方法比基于模板的方法更为灵活, 生成的描述句子也更为自然, 但过于依赖查询库, 当查询库中没有相似的图像时, 生成的句子与原图内容之间具有很大偏差.

目前, 随着深度学习在图像分类[18, 21−25]、目标检测[18, 22]、复杂系统控制[26]、游戏开发[27] 及机器翻译等领域的巨大成功, 很多研究者开始尝试将深度学习中的 CNN 技术应用到图像描述领域, 并取得了一系列重要成果, 使得生成句子的质量有了很大提高. 这些方法大都借鉴了机器翻译的流程. 采用

“编码–解码” 的方式生成图像描述句子. Karpathy 等在工作中结合 RCNN (Region-based CNN) 和双向 RNN 等多种技术, 根据图像中目标的结构和位置, 提出一种新的多模 RNN 模型[14]. 文献 [1] 使用了多模 RNN 技术, 认为图像特征和嵌入的单词序列是任务的多个模态, 通过对多个模态的共同学习, 最终生成图像描述句子. Xu 等在文献 [15] 中提出一种新的思路, 将视觉注意机制与 LSTM 相结合, 通过学习物体的位置信息, 模拟人类的视觉注意机制, 并生成相关的单词序列. 文献 [28] 认为上述模型只关注于图像的局部信息, 对全局信息捕捉不够, 对图像中物体之间的位置关系描述不够准确, 因此提出 gLSTM 模型, 提取图像与其描述之间的关系作为整体语义信息, 指导句子的生成.

以上工作虽然获得了巨大成功, 但总体来说流程较为复杂, 模型复杂度较高, 例如文献 [14] 首先使用 RCNN 技术, 识别出图像中的各种物体, 然后对各物体的位置信息进行建模排序, 最后使用双向 RNN 组合成新的句子; 文献 [15] 要求在训练和测试时对视觉注意区域进行定位采样.

文献 [2] 和文献 [13] 采用 “端到端” 的生成方式, 过程较为简单, 将图像看作源语言, 将自然语言看作目标语言. 首先使用 CNN 模型提取图像特征, 对图像进行编码, 然后送入 LSTM 网络, 对特征进行解码, 生成对应的图像内容描述句子. 以上工作在多个数据集上取得了显著效果, 但在其模型中, 解码部分在使用 LSTM 网络时层数较浅, 对图像和单词序列嵌入向量的非线性变换次数较少, 性能受到限制; 而在较深的 LSTM 模型中, 性能反而有所下降[13]. 本文提出的方法遵循与其相似的流程, 采用 “编码–解码” 的方式生成图像描述句子. 但与同类研究工作不同的是使用了更深层次的 LSTM 网络, 在训练时使用逐层优化的策略, 并使用多级目标函数对模型进行监督, 克服了多层 LSTM 模型难以优化的弊端, 在测试时融合多层 LSTM 输出的概率分值, 进一步提升预测精度, 使得生成的描述句子质量更高.

2　MLO/MLPF-LSTM

2.1　CNN 模型

CNN 模型由一系列的卷积、激活及池化等线性或非线性变换模块所组成. 图像信息经过多次变换, 得到的特征更为抽象, 泛化能力更强. 众多研究已经证明, 基于 CNN 特征的视觉模型性能远超基于手工特征 (例如 HOG (Histogram of oriented gradient), SIFT (Scale invariant feature transform) 等) 的模型, 且在一定程度上, 模型层次越多, 深度越深, CNN 特征的表达能力和可辨别能力越强, 模型性能越好[18, 21−23, 29].

Lecun 等提出并设计的 LeNet5 模型在手写数字识别上性能显著[30], 证明了 CNN 模型的优越性; Krizhevsky 等将深度学习的思想应用于 CNN 模型, 增加了模型的深度, 并使用 ReLU (Rectified linear unit) 和 Dropout 等技术解决梯度消失和过拟合问题, 设计的 Alex-Net 在 ILSVRC2012 的竞赛中获得冠军[21]. 以 Alex-Net 为标志, 基于 CNN 的深度模型获得了快速发展, 此后出现的 Chatfield-Net[31], GoogLeNet[18] 和 VGG16/VGG19[22] 等模型在 ILSVRC 竞赛中不断取得更大成功, 并在多个视觉任务中都取得了重要进展. CNN 模型的层次越来越多, 各种优化方法也不断被提出, 例如近期出现的 ResNet[23], 借助增强低层残差的方式, 解决优化困难问题, 深度达 152 层, 并在 Imagenet[32] 等多个数据集上取得显著效果. 综合各模型性能表现, 并保证对比的公平性, 本文使用 VGG16 模型提取图像的 CNN 特征.

2.2　LSTM 单元

LSTM 是一种特殊的 RNN 单元[33−34], 是为了解决传统 RNN 网络中存在的梯度消失问题提出来的. 在传统的 RNN 网络中, 使用跨时间的梯度反向传播算法 (Back propagation through time, BPTT) 对参数进行迭代更新, 但随着时间步的增加, 后续节点的梯度在反向传播过程中逐步下降, 难以对前续节点形成有效更新, 使得模型优化失败[35]. 因此, 在测试阶段, 当时间序列过长时, 后续节点很难从前续节点中获得较为有效的信息, 难以解决时间序列的 “长期依赖” 问题, 预测精度较差. 为解决该问题, 研究者们设计了 LSTM 单元, 在每个时间步中, 添加了记忆单元和多个门 (Gate), 记忆单元用于存储状态信息, 门用于控制何时及如何更新记忆单元的状态.

记忆单元与各种门的连接状态如图 1 所示, 其中

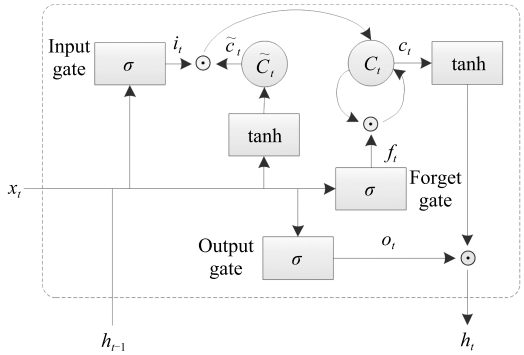


图 1　LSTM 单元

Fig. 1　LSTM unit

x_t 表示 t 时刻的输入, h_{t-1} 和 h_t 分别表示 $(t-1)$ 时刻和 t 时刻的输出, σ 表示使用 sigmoid 函数对信息进行变换; $\odot$ 表示逐点相乘, C_t 为存储状态. 忘记门 (Forget gate) 用于控制从存储状态丢弃或继续保存前一时刻的信息; 输入门 (Input gate) 用于确定需要更新的信息; 整个单元通过忘记门和输入门更新存储状态 C_t; 输出门 (Output gate) 用于确定存储状态 C_t 中哪些信息用于输出. 其计算过程由以下一系列公式共同完成.

$$\begin{cases} f_t = \sigma(W_{xf} \times x_t + W_{ht} \times h_{t-1} + b_f) \\ i_t = \sigma(W_{xi} \times x_t + W_{hi} \times h_{t-1} + b_i) \\ \tilde{c}_t = \tanh(W_{xc} \times x_t + W_{hc} \times h_{t-1} + b_c) \\ c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ o_t = \sigma(W_{xo} \times x_t + W_{ho} \times h_{t-1} + b_o) \\ h_t = o_t \odot \tanh(c_t) \end{cases} \tag{1}$$

其中, b 对应各个门的偏置值, W_x 表示与输入信息 x 相关的各个门的权值, W_h 表示与前一时刻的输出 h_{t-1} 相关的各个门的权值, c_t 为状态 C_t 的输出, $\tilde{c}_t$ 为临时状态 $\tilde{C}_t$ 的输出.

目前, 已有多种对 LSTM 单元改进的工作, 例如 Gers 等提出的 Peephole-LSTM[36], Cho 等提出的 GRU (Gated recurrent unit)[37] 等. Greff 等通过研究指出, 各种 LSTM 变体和传统 LSTM 在很多任务上性能趋同[38]. 为了便于比较, 本文使用文献 [8, 13] 采用的 LSTM 单元.

2.3 MLO/MLPF-LSTM 框架

相关研究已经证明, CNN + LSTM 架构对于解决图像描述问题效果明显[2, 8, 13]. 使用 CNN 模型提取的图像特征表达能力及可辨别能力强, 采用 LSTM 结构能够记忆句子中的单词序列; 将 CNN 特征和单词序列共同映射到嵌入空间进行训练和测试, 模型具有结构简单、鲁棒性强的特点. 一般过程为: 1) 在训练阶段, 使用 CNN 模型提取图像特征, 将图像编码为一个长度固定的特征向量, 然后将其与单词的嵌入式向量一起组成多模特征, 并送入 LSTM 网络, 经过 LSTM 的一系列变换, 生成单词序列的概率向量矩阵, 并将其转换为对应的单词序列, 然后使用距离函数求取生成单词序列矩阵与参考句子中的单词序列矩阵之间的距离, 通过 BPTT 算法对 LSTM 中的参数进行更新优化. 2) 在测试阶段, 提取图像特征后, 映射到嵌入空间, 送给 LSTM, 由 LSTM 生成单词序列的概率矩阵, 矩阵中每个概率向量中最大值对应的单词即为预测单词, 按顺序组合在一起, 生成描述句子.

在 CNN 模型中, 模型深度是保证特征抽象性及模型泛化能力的关键[18, 21−23, 29]. 我们尝试将这一思想应用到 LSTM 中. 对 LSTM 网络来说, 其“宽度” 越大, 时间步越多, 记忆能力越强, 但模型复杂度也会大幅上升. LSTM 网络的深度对性能的影响, 相关研究不多, 文献 [13] 认为 2 层的 LSTM 已经达到深度的极限, 增加 LSTM 的层次反而会使性能下降, 其实验结果也证明了这一观点. 通过分析, 可以发现这与深度 LSTM 网络的梯度衰减有关. 在 LSTM 单元中, 激活函数多采用 tanh 和 sigmoid 函数, 梯度值被限制在 $(-1, 1)$ 区间; 相邻两层 LSTM 节点之间采用全连接的方式, 当采用链式求导法则将梯度向前回传时, 其值将越来越小, 对低层 LSTM 网络中的参数调整有限, 进而对高层 LSTM 中的参数优化造成影响, 导致整个网络性能下降. 为解决这一问题, 本文方法借鉴 Hinton 等在训练深度置信网络 (Deep belief network, DBN) 时采用的逐层优化方法, 即在原有已训练好的模型基础上, 添加新的层次并重新训练, 然后对整个模型进行微调. 为了防止因参数规模增加导致的过拟合问题, 在逐层优化的基础上, 借鉴文献 [17−19] 中的模型优化方法, 设计了多目标优化模型, 对语言模型进行更加充分的优化, 同时使用多个低层目标函数对整个模型进行部分扰动, 添加额外的正则化信息.

整个训练模型如图 2 和图 3 所示, 图 2 表示基准模型 (Benchmark model), “BoS” 表示句子开始字符, “EoS” 表示句子结束字符, *word-t* 表示 t 时

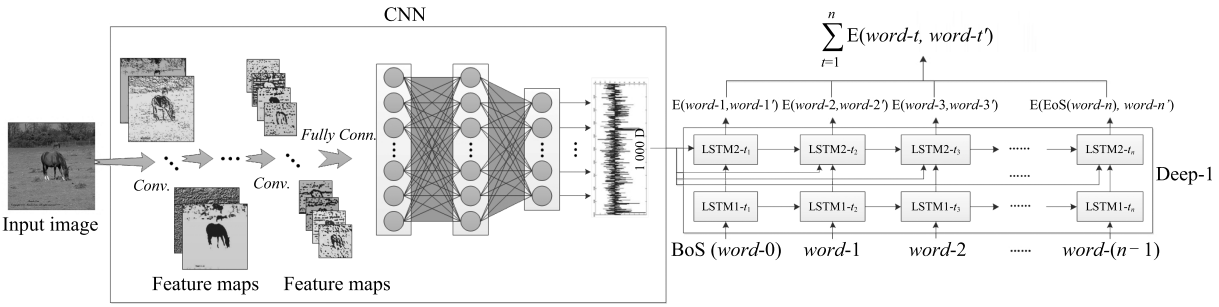


图 2 训练第 1 阶段 (基准模型)

Fig. 2 The 1st stage in training process (benchmark model)

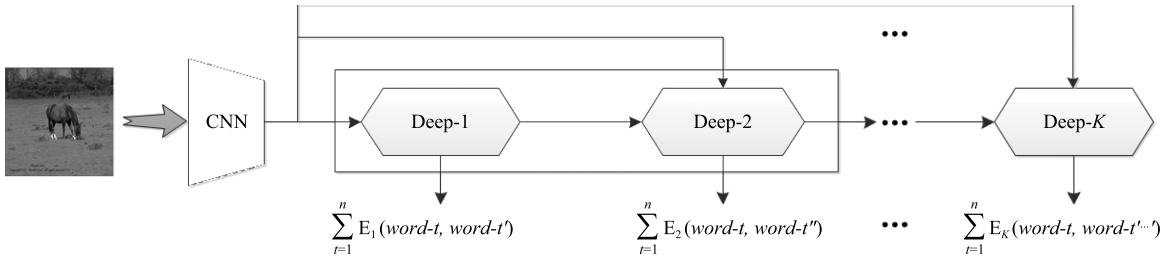


图 3　训练第 K 阶段

Fig. 3　The Kth stage in training process

刻参考句子中的单词, $word\text{-}t'$ 表示 t 时刻系统生成的单词, $E(\cdot)$ 表示交叉熵误差. 在结构上, 为了防止 LSTM 网络在较后时间步中缺乏图像的整体信息, 本文采用将图像的 CNN 特征输入 LSTM 网络中每个时间步的方式, 同时使用 "因子分解 (Factored way)" 的 LSTM 网络构建方式, 即首先将单词的嵌入式向量送入 LSTM 网络, 然后将其输出与上述编码后的图像特征向量一起送入另一个 LSTM 网络, 并由该网络及其相连的分支输出所有单词的预测概率; 文献 [13] 表明, 这种方法比 "非因子分解 (Un-factored way)" 的结构性能更优.

图 3 中, K 表示模型中训练的总阶段数, Deep-1 为图 2 中的基准模型 (Benchmark model), {Deep-2, $\cdots$, Deep-K} 中每项的网络结构与 Deep-1 相同, {Deep-1, $\cdots$, Deep-$(K-1)$} 表示已训练好的模型, Deep-K 表示新添加的 LSTM 网络. 在训练时, 当添加新的 LSTM 层进行训练时, 保留已训练好的 LSTM 层中的全连接层和目标函数, 并与新的全连接层及目标函数一起进行优化. 模型中低层的辅助分支及其目标函数能够对低层参数提供更加充分的优化; 同时, 由于低层特征抽象能力不足, 辅助分支上的目标函数能够对模型参数产生一定的扰动, 为模型提供更多的正则化信息, 防止模型陷入过拟合状态[20].

在测试阶段, 为了充分利用已训练好的参数, 将每个全连接层的输出使用 Softmax 函数进行变换, 将其转换为隶属于单词表中某个单词的概率分值; 由于低层特征输入高层 LSTM 网络后, 经过多次非线性变换, 其特征空间将变换到另一个特征空间, 因此各层的概率输出可近似认为是非相关的; 受集成模型的启发, 将所有输出的概率分值进行加权求和, 得到新的概率分值向量, 向量中最大值对应的位置即为预测单词的映射位置. 如图 4 所示, 它是一个使用三阶段训练的模型. 测试时, 输入一张图像, 经过多次卷积、激活和池化等操作, 图像被编码为长度固定的特征向量, 然后送入 LSTM 网络, 与前一个状态输出的单词一起预测当前状态的输出单词. 图 4 中 p_1^i, p_2^i 和 p_3^i 分别表示 Deep-1, Deep-2 和 Deep-3 输出的概率分值, 它们共同决定最终的预测单词; 若使用更深的 LSTM 网络, 其原理类似.

在对模型进行优化时, 目标函数定义为

$$O = \arg\min_{\theta_1,\theta_2}(\mathcal{L} : f((x,\theta_1);(s,\theta_2)) \mapsto \mathbf{R}) \tag{2}$$

其中, $f(\cdot)$ 为系统函数, x 为图像训练样本, s 为图像描述句子训练样本, θ_1 为 CNN 网络中的参数集合, θ_2 为 LSTM 网络中的参数集合, $\mathcal{L}$ 为损失函数. 整个系统的目标是在实数域 $\mathbf{R}$ 中寻找一组合适的 θ_1 和 θ_2, 使得 $\mathcal{L}$ 最小.

在实际操作中, 将 $\mathcal{L}$ 分为 $\mathcal{L}_1$ 和 $\mathcal{L}_2$, $\mathcal{L}_1$ 表示 CNN 网络的损失函数, $\mathcal{L}_2$ 表示 LSTM 网络的损失函数. $\mathcal{L}$ 可定义为

$$\mathcal{L} = \mathcal{L}_1 + \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_2^k \tag{3}$$

其中, K 表示 LSTM 网络中使用的总的阶段数, k 表示在使用逐层多目标方法优化 LSTM 网络时的第 k 个阶段, $\mathcal{L}_2^k$ 表示 LSTM 网络中第 k 个阶段的损失函数.

$\mathcal{L}$ 中的 $\mathcal{L}_1$ 和 $\mathcal{L}_2$ 都采用交叉熵进行计算, 计算过程为

$$\begin{aligned}\mathcal{L}_1 = -\frac{1}{n}\sum_{i=1}^{n}(&y^{(i)}\log(x^{(i)}) + \\ &(1-y^{(i)})\log(1-x^{(i)}))\end{aligned} \tag{4}$$

$$\begin{aligned}\mathcal{L}_2^k = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{L^i}&(r_j^i\log((s_j^i)_k) + \\ &(1-r_j^i)\log(1-(s_j^i)_k))\end{aligned} \tag{5}$$

在式 (4) 中, n 表示一次迭代中训练图像张数, $y^{(i)}$ 表示第 i 张图像的实际值 (标签), $x^{(i)}$ 表示 CNN 网络的输出; 式 (5) 中, L^i 表示第 i 张图像的参考句子长度, r_j^i 表示第 i 张图像参考句子中第 j 个单词, $(s_j^i)_k$ 表示在第 k 个阶段第 i 张图像生成句子中的第 j 个单词.

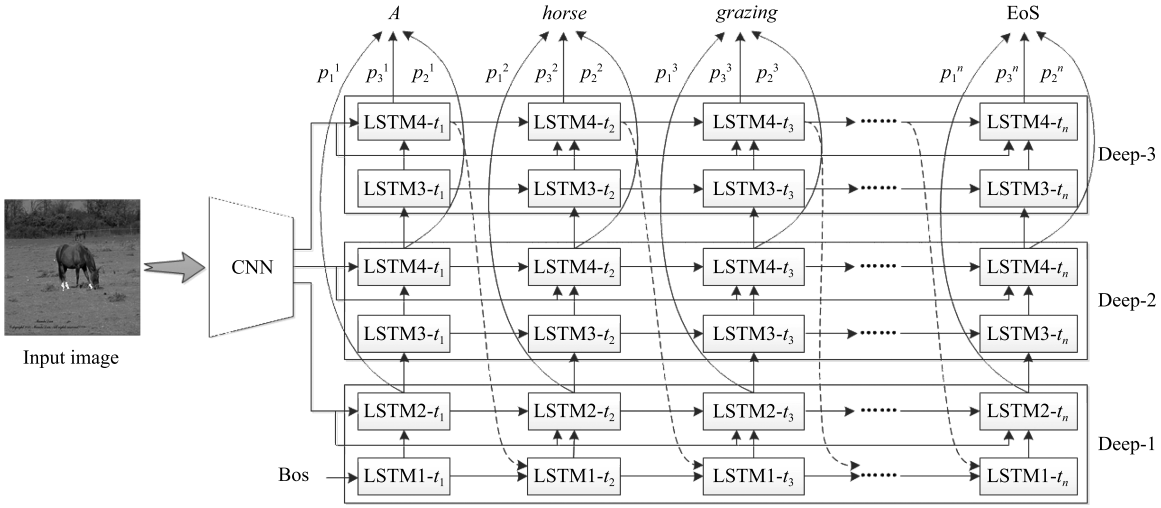


图 4 MLPF-LSTM 图像描述生成流程

Fig. 4 The pipeline of image description generation in MLPF-LSTM

对于 $\mathcal{L}_1$, 由于参数规模巨大, 样本量较少的数据集难以对其进行充分优化, 极易发生过拟合现象, 所以一般先采用大规模数据集对模型参数进行初始化 (例如 Imagenet[32]、Place205[39] 等), 然后将收敛后的模型作为预训练模型, 在新的数据集上进行微调. 对于 $\mathcal{L}_2^k$, 在图像描述的数据集上迭代优化; 在模型顶端, 输出采用 Softmax 函数进行计算, 计算公式为

$$P_k((s_j^i)_k = v) = \frac{\mathrm{e}^{(s_j^i)_k}}{\sum\limits_{j' \in V} \mathrm{e}^{(s_{j'}^i)_k}} \tag{6}$$

其中, v 表示单词表中的某个词汇, V 表示单词表. 通过式 (6) 可以得到输出的第 j 个单词属于单词表中所有单词的概率向量.

测试时, 将多个阶段的概率分值通过加权平均的方式进行融合, 得到新的概率向量矩阵, 通过该矩阵预测新的单词. 计算公式为

$$P = \frac{1}{K} \sum_{k=1}^{K} w_k P_k \tag{7}$$

其中, w_k 表示在融合时第 k 个阶段使用的权值, 根据在验证集上的经验获得.

3 实验验证

3.1 实验数据集

采用 MSCOCO[40] 和 Flickr30K[41] 公开数据集对模型进行验证. 这两个数据集较大, 包含的训练样本较多, 使得 LSTM 网络不易陷入过拟合状态. MSCOCO 数据集共有 123 287 张图像, 其中 82 783 张图像用于训练, 40 504 张图像用于验证; 每张图像中包含至少 5 条人工标注的参考描述句子 (如图 5 所示). 为保证对比的公平性, 遵循文献 [13−14, 18] 等工作中使用的规则, 在验证集中取 5 000 张图像和相关参考句子作为新的验证集, 另取 5 000 张图像及其参考句子作为测试集. 在 Flickr30K 数据集中, 共有 31 783 张图像, 每张图像对应 5 条参考句子, 同样按照统一的使用规则, 将其中的 29 000 张图像及其参考句子作为训练集, 1 000 张图像及其参考句子作为测试集, 其余样本作为验证集. 具体使用时, 首先在验证集上寻找最优参数, 记录模型的收敛位置, 然后使用该位置上的训练模型对测试集进行测试.

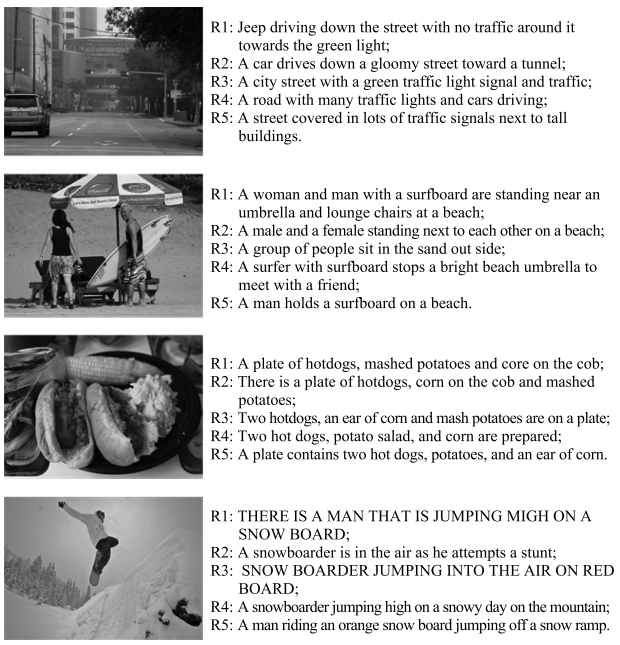


图 5 MSCOCO 数据集中部分训练样本

Fig. 5 The examples for training in MSCOCO dataset

3.2 评价方法

本文使用评价方法 BLEU[42]、METEOR[43] 和 CIDEr[44] 对生成的描述句子进行评价. 其中 BLEU 方法是一种基于精度的度量方法, 主要思想是衡量生成的句子与参考句子之间的 n-$Gram$ 精度, 用 "B-n" 表示所有精度的平均值, 取值在 $(0, 1]$ 之间, 其值越大, 表明模型在该 "B-n" 上的效果越好; 在不同的 "B-n" 之间, n 越大, 表示生成的句子连贯性越好. 计算公式为

$$\text{BLEU} = b \times \exp\left(\sum_{n=1}^{N}\left(\frac{1}{N}\log p_n\right)\right) \tag{8}$$

其中, N 一般取 $\{1, 2, 3, 4\}$, b 表示惩罚项, 用于生成句子的长度小于参考句子的情况, 其值为

$$b = \begin{cases} 1, & l_c \geq l_r \\ \exp\left(1 - \dfrac{l_r}{l_c}\right), & l_c < l_r \end{cases} \tag{9}$$

其中, l_r 表示参考句子的长度, l_c 表示生成句子的长度. 当生成句子长度大于参考句子时, 其值为 1; 当生成句子长度小于参考句子时, 降低其 BLEU 分值, 表示惩罚.

式 (8) 中的 p_n 表示 n-$Gram$ 下的匹配精度, 计算公式为

$$p_n = \frac{\sum Count_{clip}(m_{n-gram})}{\sum Count_{clip}(c_{n-gram})} \tag{10}$$

其中, 分子项表示生成句子与参考句子中具有 n-$Gram$ 匹配的次数, 分母项表示生成的句子中具有 n-$Gram$ 的总数.

BLEU 方法重点考虑了生成句子中单词的准确率, 但对召回率考虑不足. METEOR 自动评测方法既考虑了准确率, 也考虑了召回率[43], 首先使用任意匹配的方式将参考句子与生成句子中的单词按照精准匹配、同义匹配和前缀匹配的方式依次寻找匹配的最大值, 当三种匹配的最大值存在相同时, 选择按顺序两两匹配中交叉数最少的匹配作为 "对齐 (alignment)"; 通过不断迭代, 生成对齐集合, 然后将该集合中元素的个数与参考句子中单词总数的比值作为召回率, 与生成句子中单词总数的比值作为准确率, 然后使用调和平均值的方式计算最终值, 取值在 $(0, 1]$ 之间, 其值越大, 说明生成的句子质量越高.

CIDEr 评价方法[44] 引入了 "共识" 的概念, 通过计算生成句子和人工标注的参考句子之间的余弦距离对生成句子进行评价, 其值越大, 表明生成句子与图像中所有参考句子之间的语义相似度越大. 该评价方法更多地考虑了生成句子的语义和内涵, 更加贴近人类的评价方法.

3.3 实验平台及设置

本文使用深度学习框架 Caffe[45] 开发部署提出的模型, 采用文献 [13] 使用的 LSTM 单元结构. 在提取图像特征方面, 采用结构简单且性能优越的 VGG16 模型[22]. 为了多方面验证模型效果, 本文使用两种方式对模型进行测试. 1) 使用文献 [22] 中已优化完毕的 VGG16 模型作为特征提取器, 但其参数固定, 使其不参与语言模型的训练, 每个阶段的训练只是对 LSTM 网络中的参数进行优化, 记为 MLO/MLPF-LSTM; 2) 使用联调机制, 将在 Imagenet 上训练好的 VGG16 模型作为预训练模型, 其参数与 LSTM 网络中的参数一起进行微调, 记为 MLO/MLPF-LSTM$^+$.

在建立基准语言模型时, 为公平对比, 采用文献 [13] 的配置. 在 MSCOCO 数据集中, 由于图像描述句子大都在 20 个词以内, 为降低模型复杂度, 将每层中的 LSTM 网络的时间步长设置为 20; 整个数据集中单词表长度为 8 801; 每个 LSTM 单元中隐层单元个数设置为 1 000; 在训练时, 设置最大迭代次数为 150 K 次, 通过观察在验证集上的收敛情况, 发现经过 110 K 次, 网络即已收敛, 因此, 在后续实验中, 将最大迭代次数一致设置为 110 K 次, 并使用迭代 110 K 的训练模型对测试集进行测试; 初始学习率设置为 0.01, 为防止网络陷入局部最优, 使用逐步降低学习率的方式, 每迭代 20 K 次时, 将学习率降低为原来的 10 %.

测试时, 首先在验证集上使用多组权重对融合模型进行验证, 经对比发现, 在较低层次上权重较大时, 效果更好. 经多次实验验证, 在 2-stage 上, 不同概率向量的权值设置为 $[w_1, w_2] = [0.67, 0.33]^{\mathrm{T}}$ 时, 融合效果更好; 在 3-stage 和 4-stage 上, 分别将其设置为 $[w_1, w_2, w_3] = [0.4, 0.4, 0.2]^{\mathrm{T}}$ 和 $[w_1, w_2, w_3, w_4] = [0.3, 0.3, 0.2, 0.2]^{\mathrm{T}}$ 时, 模型性能更优. 在 Flickr30 数据集上, 单词表长度为 7 406, 每个 LSTM 单元中隐层单元个数设置为 512, 首次最大训练迭代次数为 90 K 次, 在验证集上迭代 70 K 次时达到收敛, 其他设置与在 MSCOCO 数据集上相同.

3.4 实验结果及分析

使用 VGG16 和两层 LSTM 对模型进行训练, 并作为基准模型, 然后在基准模型的基础上, 添加新的 LSTM 层, 并保留原有的全连接层和目标函数, 使用已训练好的基准模型参数对模型进行初始化, 重新训练. 在 MSCOCO 数据集上, 使用非联调方式 (MLO/MLPF-LSTM) 时, 部分实验结果如图 6

所示.

图 6 MLO/MLPF-LSTM (3-stage) 模型生成的部分图像描述示例

Fig. 6 Examples of image descriptions with MLO/MLPF-LSTM (3-stage)

在图 6 中, R 表示人工标注的参考句子, B 表示使用基准模型所生成的句子, C 表示 MLO/MLPF-LSTM (3-stage) 模型生成的候选待评价句子 (即生成的句子). 从图 6 可以看出, 本文模型生成的句子具有更好的语义表达, 较好地描述了图像的内容. 与基准模型相比, 所提模型生成的句子更为合理, 语义更加丰富. 例如第 2 张图像中, 基准模型生成的句子 B 把重点放在了 "床 (bed)" 和 "电视 (television)" 上, 对场景则重视不够; 而 C 首先指明了场景信息 (bedroom), 然后说明场景中包含哪些物体, 句子更贴近人们的表达习惯. 同样, 在第 4 张图像中, 所提模型生成的句子准确描述了 "繁忙的城市 (busy city)" 和 "交通灯 (traffic lights)", 而使用基准模型生成的句子则缺乏这一精确描述. 与人工标注的句子相比, 本文所提模型生成的有些句子更加合理, 例如第 2 张图像中, 生成句子不仅描述了 "床 (bed)" 和 "电视 (television)", 还找出了 "桌子 (table)", 而 "桌子 (table)" 在人工标注中并没有出现. 但通过对比也发现, 本文模型所生成的句子缺少对图像中物体的形象化描述, 描述虽然客观, 但缺乏感情色彩和想象力. 对于 "电视 (TV/television)", 人们可以使用 "大 (big, large)" 和 "平板 (flat screen)" 来形容; 在第 2 张图像中, 人们对 "狗 (dog)" 使用了 "棕色 (brown)"、"小的 (small)" 来描述, 甚至联想到 "狗 (dog)" 可能 "累了 (tired)", 但在生成句子中, 缺乏这方面的词汇和描述.

为了对本文使用的三种策略进行充分评估, 衡量每种策略对模型的贡献, 分别对不使用任何策略增加语言模型深度的情况, 只使用逐层优化加深模型深度的情况, 同时使用逐层优化和多目标优化策略增加模型深度的情况, 以及三种策略同时使用时的情况进行实验验证, 四种情况分别记为: no-MLO, MLO1, MLO2 和 MLO/MLPF. 在不同深度下, 各种情况的 B-4 和 CIDEr 如图 7 所示.

通过对比可以发现, 无论使用非联调方式还是联调方式, 在不使用任何策略的情况下, 简单加深语言模型深度, 性能将急剧下降, 这是由于梯度消失造成的, 低层参数难以得到充分优化; 当使用逐层优化方法后, 模型性能趋于稳定, 克服了低层参数难以优化的弊端, 但整体性能并没有得到明显改善; 在此基础上, 结合多目标优化策略后, 模型性能有了显著提升; 而在使用融合策略后, 其性能得到进一步的提升.

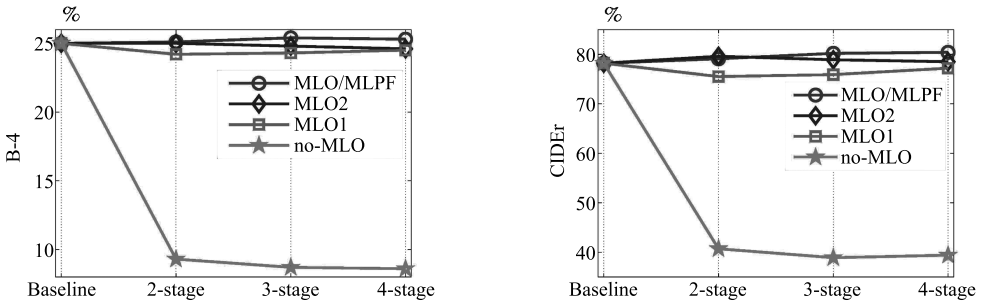
表 1 和表 2 分别列出了在 MSCOCO 数据集上使用不同深度语言模型各阶段及融合后的实验结果. 表 3 列出了在 Flickr30K 数据集上的实验结果, 其中, Baseline 表示使用非联调方式基准模型得到的结果. Baseline$^+$ 表示使用联调方式基准模型得到的

表 1 MSCOCO 数据集上不同层次及多层融合之后的性能对比 (非联调方式) (%)

Table 1 Performance comparison under different fusion conditions on MSCOCO (non-jointly optimizing) (%)

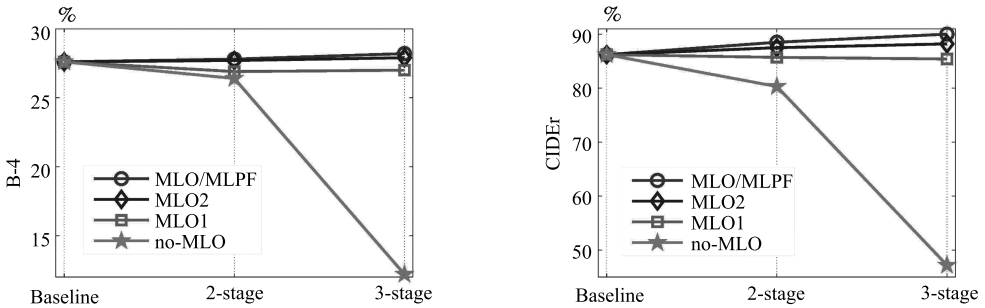
Models		B-1	B-2	B-3	B-4	C
Baseline		67.7	49.4	35.2	25.0	78.2
2-stage	P1	67.8	49.7	35.3	25.0	78.5
	P2	67.5	49.6	35.3	25.0	79.6
	Fusion	68.0	50.0	35.5	25.1	79.1
3-stage	P1	67.9	49.8	35.5	25.2	79.0
	P2	67.5	49.6	35.3	25.0	79.6
	P3	67.3	49.4	35.1	24.8	78.9
	Fusion	68.0	50.0	35.8	25.4	80.2
4-stage	P1	67.6	49.5	35.3	25.1	78.7
	P2	67.0	49.1	34.9	24.8	79.7
	P3	66.8	49.0	34.8	24.7	79.5
	P4	66.9	49.0	34.8	24.6	78.9
	Fusion	67.7	49.8	35.6	25.3	80.4

C 表示 CIDEr



(a) 非联调方式下的 B-4 性能
(a) Performance of B-4 without joint optimization

(b) 非联调方式下的 CIDEr 性能
(b) Performance of CIDEr without joint optimization



(c) 联调方式下的 B-4 性能
(c) Performance of B-4 with joint optimization

(d) 联调方式下的 CIDEr 性能
(d) Performance of CIDEr with joint optimization

图 7　在 MSCOCO 数据集上使用不同策略加深模型深度时的性能表现
Fig. 7　Performance under different strategies at each stage on MSCOCO

表 2　MSCOCO 数据集上不同层次及多层融合之后的性能对比 (联调方式) (%)
Table 2　Performance comparison under different fusion conditions on MSCOCO (jointly optimizing) (%)

Models		B-1	B-2	B-3	B-4	C
Baseline$^+$		70.2	52.7	38.3	27.6	86.2
2-stage	P1	70.2	52.7	38.4	27.8	88.4
	P2	69.9	52.6	38.3	27.7	87.5
	Fusion	70.2	52.8	38.4	27.8	88.5
3-stage	P1	70.5	52.8	38.4	27.8	89.3
	P2	70.1	52.5	38.2	27.8	88.9
	P3	70.1	52.8	38.5	27.9	88.2
	Fusion	70.6	53.2	38.8	28.2	90.0

C 表示 CIDEr

表 3　Flickr30K 数据集上不同层次及多层融合之后的性能对比 (联调方式) (%)
Table 3　Performance comparison under different fusion conditions on Flickr30K (jointly optimizing) (%)

Models		B-1	B-2	B-3	B-4	M
Baseline$^+$		60.2	41.8	28.5	19.2	19.2
2-stage	P1	61.5	42.9	29.2	19.7	19.4
	P2	60.7	42.2	29.0	19.8	19.2
	Fusion	61.4	42.8	29.2	19.8	19.6

M 表示 METEOR

结果, 2-stage 表示在基准模型上增加了 2 个 LSTM 层, LSTM 网络共有 4 层; 同理, 3-stage 和 4-stage 分别表示 LSTM 网络共有 6 层和 8 层; P1 表示单独使用某阶段模型中的第 1 组概率分值得到的结果; P2, P3 及 P4 表示单独使用第 2, 3, 4 组概率分值得到的结果.

实验结果显示, 无论使用联调方式还是非联调方式, 随着 LSTM 网络深度的增加, 性能都会提升. 尤其是使用联调方式后, 在 MSCOCO 数据集上, 使用 6 层 LSTM, B-4 和 CIDEr 分别达到了 28.2 % 和 90.0 %; 在 Flickr30K 数据集上, 使用 4 层 LSTM, B-4 和 METEOR 分别达到了 19.8 % 和 19.6 %. 但值得指出的是, 当在两个数据集上使用更深层次的 LSTM 网络时, 性能均有所下降. 在 MSCOCO 数据集上, 使用 MLO/MLPF-LSTM, 深度在 8 层时, 在 CIDEr 指标上效果最好; 对于 BLEU 指标, 深度 6 层时已有所下降; 在使用 MLO/MLPF-LSTM$^+$ 时, 深度超过 6 层时, 无论是 BLEU 还是 CIDEr, 性能均有所降低.

在 Flickr30K 数据集上, 使用 MLO/MLPF-LSTM$^+$ 时, LSTM 网络深度达到 4 层时, 结果最

好, 当超过 4 层时, BLEU 和 METEOR 指标会有所降低. 表明即使使用了逐层优化和多目标训练, LSTM 网络的深度也有一定的极限. 原因是当深度增加时, 参数规模也将增加, 整个系统易陷入过拟合状态. 因此其深度主要由数据集大小决定. 在较大的数据集上, 需要更深的 LSTM 网络提升性能; 而在较小的数据集上, 使用较浅的 LSTM 网络即可; 若要进一步提升性能, 需要使用其他技术进一步对模型进行改进.

从实验结果可以看出, 在未使用融合方法时, 若只使用顶层用于最终输出, BLEU 指标较基准模型可能会有所下降, 但 CIDEr 指标较 Baseline/Baseline$^+$ 却有明显上升 (如表 1 和表 2 所示); 说明本文使用的多目标训练方法虽然对 n-$Gram$ 精度没有提升, 但增强了生成句子的语义信息. 当使用了融合技术之后, 可以发现在两个数据集上, 性能在所有评测指标上均有显著改善.

此外, 与其他方法一样, 在 MLO/MLPF-LSTM$^+$ 上使用了集束搜索算法 (Beam search), 为了搜索速度更快, 将 Beam_size 大小设置为 5. 实验结果如表 4 和表 5 所示. 可以看出, 使用 Beam search 算法后, 模型性能有了进一步提升, 而且随着模型深度的增加, 性能也随之上升. 但需要指出的是, 在 MSCOCO 和 Flickr30K 两个数据集上, 当模型深度增加到 6 层 (3-stage) 时, 融合后的模型性能并无显著提升.

本文还与图像描述领域中的主流模型进行了对比 (如表 6 和表 7 所示, 表中 LRCN-AlexNet, m-RNN, Soft-attention 和 Hard-attention 方法数据引自各自文献, multimodal RNN, Google NIC 和 gLSTM 方法数据来源于文献 [28]).

表 4 MSCOCO 数据集上不同层次及多层融合之后的性能对比 (使用联调方式和集束搜索算法) (%)

Table 4 Performance comparison under different fusion conditions on MSCOCO (jointly optimizing and Beam search algorithm are employed) (%)

Models		B-1	B-2	B-3	B-4	C
Baseline$^+$		71.3	54.4	40.8	30.5	92.0
2-stage	P1	71.4	54.3	40.7	30.6	93.8
	P2	71.6	54.8	41.1	31.0	93.7
	Fusion	71.5	54.5	41.0	31.0	94.2

C 表示 CIDEr

表 5 Flickr30K 数据集上不同层次及多层融合之后的性能对比 (使用联调方式和集束搜索算法) (%)

Table 5 Performance comparison under different fusion conditions on Flickr30K (jointly optimizing and Beam search algorithm are employed) (%)

Models		B-1	B-2	B-3	B-4	M
Baseline$^+$		63.4	44.5	30.9	21.1	19.0
2-stage	P1	65.1	45.8	31.8	21.9	19.2
	P2	65.0	46.0	32.0	21.9	19.3
	Fusion	66.2	47.2	33.1	23.0	19.6

M 表示 METEOR

表 6 不同方法在 MSCOCO 数据集上的性能对比 (%)

Table 6 Performance comparison with other state-of-the-art methods on MSCOCO (%)

Methods	B-1	B-2	B-3	B-4	C
multimodal RNN[14]	62.5	45.0	32.1	23.0	66.0
Google NIC[2]	66.6	46.1	32.9	24.6	–
LRCN-AlexNet[13]	62.8	44.2	30.4	21.0	–
m-RNN[1]	67.0	49.0	35.0	25.0	–
Soft-attention[15]	70.7	49.2	34.4	24.3	–
Hard-attention[15]	71.8	50.4	35.7	25.0	–
emb-gLSTM, Gaussian[28]	67.0	49.1	35.8	26.4	81.3
MLO/MLPF-LSTM	67.7	49.8	35.6	25.3	80.4
MLO/MLPF-LSTM$^+$	70.6	53.2	38.8	28.2	90.0
MLO/MLPF-LSTM$^+$(BS)	71.5	54.5	41.0	31.0	94.2

BS 表示 Beam search, C 表示 CIDEr

表 7 不同方法在 Flickr30K 数据集上的性能对比 (%)

Table 7 Performances comparison with other state-of-the-art methods on Flickr30K (%)

Methods	B-1	B-2	B-3	B-4	M
multimodal RNN[14]	57.3	36.9	24.0	15.7	15.3
Google NIC[2]	66.3	42.3	27.7	18.3	–
LRCN-AlexNet[13]	58.7	39.1	25.1	16.5	–
m-RNN[1]	60.0	41.0	28.0	19.0	–
Soft-attention[15]	66.7	43.4	28.8	19.1	18.5
Hard-attention[15]	66.9	43.9	29.6	19.9	18.5
emb-gLSTM, Gaussian[28]	64.6	44.6	30.5	20.6	17.9
MLO/MLPF-LSTM$^+$	61.4	42.8	29.2	19.8	19.6
MLO/MLPF-LSTM$^+$(BS)	66.2	47.2	33.1	23.0	19.6

M 表示 METEOR, BS 表示 Beam search

通过对比可以发现, 在 MSCOCO 数据集上, 在使用非联调方式时, 在 B-1 和 B-2 上, 本文模型与基于视觉注意力的 Hard-attention 模型相比具有较

大差距, 但在 B-3 和 B-4 指标上表现良好, 在 B-4 上, 甚至超过了 Hard-attention 方法; 当使用联调方式后, 本文模型的性能除在 B-1 指标上低于 Hard-attention 方法外, 在其他指标上均高于其他方法.

使用 Beam search 后, 无论在 MSCOCO 还是 Flickr30K 数据集上, 在多个评价指标上均远超其他方法. 在 MSCOCO 数据集上, B-4 指标超过基于注意力机制的 Hard-attention 模型 6.0 %, 同时 B-4 和 CIDEr 指标分别超过 gLSTM 模型 4.6 % 和 12.9 %; 在 Flickr30K 数据集上, B-4 和 METEOR 指标分别超过 Hard-attention 模型 3.1 % 和 1.1 %, 同时, 其性能也超过 gLSTM 模型. 但在 B-1 指标上, 本文所提模型略低于基于注意力机制的方法, 其原因是, 基于视觉注意机制的模型对于单个物体更为敏感, 但对于物体与物体之间的关系、物体与背景之间的关系描述能力不足, 导致检测到的用于描述单个物体的词汇更多, 但对于描述物体之间关系的更长的词组则显得性能欠佳; 而本文提出的模型更注重图像内容的整体理解, 生成的句子更符合人们的描述习惯.

由于增加了语言模型的深度, 模型参数更多、特征经过的非线性变换次数更多, 因此模型复杂度与基准模型相比也相应有所增加. 在语言模型上, 复杂度的增加与具体数据集有关, 当数据集较大时, 可使用的 LSTM 层次更多, 训练需要的阶段数更多, LSTM 网络中隐藏单元个数也更多, 其模型复杂度也更高. 设基准模型中参数规模为 N_{param}, 计算复杂度为 C, 则对于一个包含 n 个训练阶段的模型来说, 其参数规模为 $n \times N_{\text{param}}$, 计算复杂度为 $n(n+1)/2 \times C$; 在测试时, 参数规模为 $n \times N_{\text{param}}$, 但计算复杂度为 $n \times C$.

4 结论

使用自然语言对静态图像进行描述是一项极具挑战性的视觉任务, 要求系统不仅能够处理图像信息, 还要能够处理文本信息. 目前, 随着计算机视觉和自然语言处理技术的快速发展, 图像描述工作也取得了重要发展, 尤其是基于深度学习的模型, 采用 "端到端" 的训练和测试方式, 生成的描述句子结构更加灵活, 更符合人们的表达习惯. 本文工作采用 CNN + LSTM 架构, 首先使用性能优越的深度模型 VGG16 提取图像的 CNN 特征, 对图像进行 "编码", 然后将其送入 LSTM 网络, 对特征进行 "解码". 在本文设计的模型中, 使用了更深层次的 LSTM 网络, 但由于优化较为困难, 因此采用了逐层优化的策略保证网络收敛, 并提出使用多目标优化和多层概率融合的方法, 改善模型性能.

但本文在实验中也发现, 在反映准确率的 BLEU 指标上, 模型性能还有待于进一步提升. 因此, 本文下一步将结合更多基于视觉的方法对模型进行改进, 例如使用更深的 ResNet[23] 网络提取图像特征等; 同时, 也将在更大和更复杂的数据集 (例如 Visual Genome[46]) 上对模型做进一步验证.

References

1 Mao J H, Xu W, Yang Y, Wang J, Huang Z H, Yuille A. Deep captioning with multimodal recurrent neural networks (m-RNN). In: Proceedings of the 2015 International Conference on Learning Representations. San Diego, USA, 2015.

2 Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 3156−3164

3 Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S M, Choi Y, Berg A C, Berg T L. BabyTalk: understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(12): 2891−2903

4 Mitchell M, Han X F, Dodge J, Mensch A, Goyal A, Berg A, Yamaguchi K, Berg T, Stratos K, Daumé H III. Midge: generating image descriptions from computer vision detections. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: ACL, 2012. 747−756

5 Elliott D, Keller F. Image description using visual dependency representations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA: ACL, 2013. 1292−1302

6 Farhadi A, Hejrati M, Sadeghi M A, Young P, Rashtchian C, Hockenmaieret J, Forsyth D. Every picture tells a story: generating sentences from images. In: Proceedings of the 2010 European Conference on Computer Vision (ECCV). Berlin, Heidelberg, Germany: Springer, 2010. 15−29

7 Zhang Hong-Bin, Ji Dong-Hong, Yin Lan, Ren Ya-Feng. Product image sentence annotation based on gradient kernel feature and N-gram model. *Computer Science*, 2016, **43**(5): 269−273, 287
(张红斌, 姬东鸿, 尹兰, 任亚峰. 基于梯度核特征及 N-gram 模型的商品图像句子标注. 计算机科学, 2016, **43**(5): 269−273, 287)

8 Socher R, Karpathy A, Le Q V, Manning C D, Ng A Y. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2014, **2**: 207−218

9 Kuznetsova P, Ordonez V, Berg T L, Choi Y. TreeTalk: composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2014, **2**: 351−362

10 Kuznetsova P, Ordonez V, Berg A, Berg T, Choi Y. Generalizing image captions for image-text parallel corpus. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: ACL, 2013. 790−796

11 Mason R, Charniak E. Nonparametric method for data-driven image captioning. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA: ACL, 2014. 592−598

12 Jiang Shu-Qiang, Min Wei-Qing, Wang Shu-Hui. Survey and prospect of intelligent interaction-oriented image recognition techniques. *Journal of Computer Research and Development*, 2016, **53**(1): 113−122
(蒋树强, 闵巍庆, 王树徽. 面向智能交互的图像识别技术综述与展望. 计算机研究与发展, 2016, **53**(1): 113−122)

13 Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T, Saenko K. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 2625−2634

14 Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 3128−3137

15 Xu K, Ba J L, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R S, Bengio Y. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015. 2048−2057

16 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504−507

17 Hermans M, Schrauwen B. Training and analyzing deep recurrent neural networks. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc., 2013. 190−198

18 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 1−9

19 Lee C Y, Xie S N, Gallagher P W, Zhang Z Y, Tu Z W. Deeply-supervised nets. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics. San Diego, USA, 2015. 562−570

20 Gerber R, Nagel H H. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences. In: Proceedings of the 1996 International Conference on Image Processing. Lausanne, Switzerland: IEEE, 1996. 805−808

21 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 2012 Advances in Neural Information Processing Systems. Lake Tahoe, USA: MIT Press, 2012. 1097−1105

22 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 2015 International Conference on Learning Representations. San Diego, USA, 2015.

23 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 770−778

24 Shi Jun-Fei, Liu Fang, Lin Yao-Hai, Liu Lu. Polarimetric SAR image classification based on deep learning and hierarchical semantic model. *Acta Automatica Sinica*, 2017, **43**(2): 215−226
(石俊飞, 刘芳, 林耀海, 刘璐. 基于深度学习和层次语义模型的极化 SAR 分类. 自动化学报, 2017, **43**(2): 215−226)

25 Wang Wei-Ning, Wang Li, Zhao Ming-Quan, Cai Cheng-Jia, Shi Ting-Ting, Xu Xiang-Min. Image aesthetic classification using parallel deep convolutional neural networks. *Acta Automatica Sinica*, 2016, **42**(6): 905−914
(王伟凝, 王励, 赵明权, 蔡成加, 师婷婷, 徐向民. 基于并行深度卷积神经网络的图像美感分类. 自动化学报, 2016, **42**(6): 905−914)

26 Duan Yan-Jie, Lv Yi-Sheng, Zhang Jie, Zhao Xue-Liang, Wang Fei-Yue. Deep learning for control: the state of the art and prospects. *Acta Automatica Sinica*, 2016, **42**(5): 643−654
(段艳杰, 吕宜生, 张杰, 赵学亮, 王飞跃. 深度学习在控制领域的研究现状与展望. 自动化学报, 2016, **42**(5): 643−654)

27 Guo Xiao-Xiao, Li Cheng, Mei Qiao-Zhu. Deep learning applied to games. *Acta Automatica Sinica*, 2016, **42**(5): 676−684
(郭潇逍, 李程, 梅俏竹. 深度学习在游戏中的应用. 自动化学报, 2016, **42**(5): 676−684)

28 Jia X, Gavves E, Fernando B, Tuytelaars T. Guiding the long-short term memory model for image caption generation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 2407−2415

29 Xi Xue-Feng, Zhou Guo-Dong. A survey on deep learning for natural language processing. *Acta Automatica Sinica*, 2016, **42**(10): 1445−1465
(奚雪峰, 周国栋. 面向自然语言处理的深度学习研究. 自动化学报, 2016, **42**(10): 1445−1465)

30 Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324

31 Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of the 2014 British Machine Vision Conference. Nottingham, England: British Machine Vision Association, 2014.

32 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C, Li F F. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, **115**(3): 211−252

33 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, **9**(8): 1735−1780

34 Graves A. Generating sequences with recurrent neural networks [Online], available: https://arxiv.org/pdf/1308.0850v5.pdf, June 5, 2014

35 Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 1994, **5**(2): 157−166

36 Gers F A, Schmidhuber J. Recurrent nets that time and count. In: Proceedings of the 2000 IEEE-INNS-ENNS International Joint Conference on Neural Networks. Como, Italy: IEEE, 2000. 189−194

37 Cho K, Van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation [Online], available: https://arxiv.org/pdf/1406.1078v3.pdf, September 3, 2014

38 Greff K, Srivastava R K, Koutník J, Steunebrink B R, Schmidhuber J. LSTM: a search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, **28**(10): 2222−2232

39 Zhou B L, Lapedriza A, Xiao J X, Torralaba A, Oliva A. Learning deep features for scene recognition using places database. In: Proceedings of the 2015 Advances in Neural Information Processing Systems. Montréal, Canada: MIT Press, 2015. 487−495

40 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C L. Microsoft COCO: common objects in context. In: Proceedings of the 2014 European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 740−755

41 Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014, **2**: 67−78

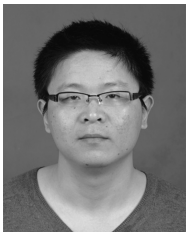
42 Papineni K, Roukos S, Ward T, Zhu W J. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA: ACL, 2002. 311−318

43 Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the 2005 ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization. Ann Arbor, USA: ACL, 2005. 65−72

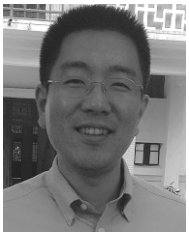
44 Vedantam R, Zitnick C L, Parikh D. CIDEr: consensus-based image description evaluation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 4566−4575

45 Jia Y Q, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia. Orlando, Florida, USA: ACM, 2014. 675−678

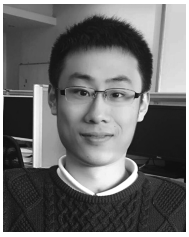
46 Krishna R, Zhu Y K, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L J, Shamma D A, Bernstein M S, Li F F. Visual Genome: connecting language and vision using crowd sourced dense image annotations [Online], available: https://arxiv.org/pdf/1602.07332.pdf, February 23, 2016



汤鹏杰 同济大学计算机科学与技术系博士研究生. 主要研究方向为计算机视觉和深度学习.
E-mail: 5tangpengjie@tongji.edu.cn
(**TANG Peng-Jie** Ph. D. candidate in the Department of Computer Science and Technology, Tongji University. His research interest covers computer vision and deep learning.)



王瀚漓 同济大学计算机科学与技术系教授. 主要研究方向为视频编码, 计算机视觉和机器学习. 本文通信作者.
E-mail: hanliwang@tongji.edu.cn
(**WANG Han-Li** Professor in the Department of Computer Science and Technology, Tongji University. His research interest covers video coding, computer vision, and machine learning. Corresponding author of this paper.)



许恺晟 同济大学计算机科学与技术系硕士研究生. 主要研究方向为图像理解和深度学习.
E-mail: iaalm@tongji.edu.cn
(**XU Kai-Sheng** Master student in the Department of Computer Science and Technology, Tongji University. His research interest covers image understanding and deep learning.)