基于多模态注意力机制的图像理解描述新方法

李学明*，岳 贡，陈光伟

(重庆大学计算机学院 重庆 沙坪坝区 400044)

【摘要】针对现有的图像理解描述方法存在描述句子不丰富、不准确、模型结构复杂、难以训练等问题，该文提出了一种端到端的基于多模态注意力机制 (M-AT) 的图像理解描述新方法。该方法首先通过关键词图像特征提取模型 (K-IFE) 提取更优的空间特征和关键词特征，并利用关键词注意力机制模型 (K-AT) 关注重要描述词语、空间注意机制模型 (S-AT) 关注图像更重要的区域并简化模型结构，且 K-AT 和 S-AT 两种注意力机制可以相互矫正，最终生成更加准确、丰富的图像描述语句。在 MSCOCO 数据集的实验结果表明该方法是有效的，部分评价指标有 2% 左右的提升。

关 键 词 注意力机制； 图像理解； 关键词； 多模态； 空间

中图分类号 TP312 **文献标志码** A doi:10.12178/1001-0548.2019228

A Novel End-to-End Image Caption Based on Multimodal Attention

LI Xue-ming*, YUE Gong, and CHEN Guang-wei

(School of Computer, Chongqing University Shapingba Chongqing 400044)

Abstract The existing image caption methods have some problems that the caption sentences are not rich and accurate, and the model structures are complicated and difficult to train. We propose a novel end-to-end image caption method called image caption based on multimodal attention mechanism (M-AT). Firstly, it takes the keyword image feature extraction model (K-IFE) to extract better spatial features and keyword features, uses the keyword attention mechanism model (K-AT) to focus on important description words, and applies the spatial attention mechanism model (S-AT) to pay attention to more important areas of the image and simplify the model structure. The two attention mechanisms, K-AT and S-AT, can correct each other. The proposed method can generate more accurate and rich image description sentences. The experimental results on the MSCOCO data set show that the proposed method is effective, has around 2% improvement in some evaluation indicators.

Key words attention mechanism; image caption; keyword; multimodal; spatial

早期的图像理解是基于模板的方式进行图像描述，通过识别图像中的对象、对象属性、对象关系来匹配语言模板以此生成描述语句。文献 [1] 选择构造语法树的方式生成描述语句；文献 [2] 使用三元组的方式生成描述语句；文献 [3] 通过选定短语，再将短语组合成描述语句实现图像的理解。

通过改进基于模板的图像理解方法，产生了基于图像和图像描述语句相似度检索的图像理解方法，即将图像及对应描述语句映射到同一特征空间，通过计算图像和语句特征之间的相似度来生成描述语句。文献 [4] 将图像及其对应语句映射到两个不同的特征空间，然后利用核典型相关分析 (kernel kanonical correlation analyses, KCCA) 将特征映射到同一个特征空间，最后通过计算特征相似度来选择描述语句。文献 [5] 使用随机树形结构抽取描述语句中的词组，树枝为词组，通过检索与测试图片相似的图片及其对应的树枝，选择组合的方式生成描述语句。

随着深度学习 (deep learning, DL) 的快速发展，文献 [6] 首先使用深度学习方法解决图像理解问题，提出利用多模态递归神经网络进行图像描述语句生成。图像理解本质上是从视觉信息到语义信息的转换。受机器翻译中基于神经网络的编码器/解码器方法[7] 的启发，文献 [6, 8-11] 将图像理解视为

收稿日期：2019－10－10；修回日期：2020－03－26

基金项目：国家重点研发项目 (2017YFB1402405-5)；重庆市技术创新与应用发展专项重点项目 (CSTC2019JSCX-MBDXX2012)；中央高校基本科研项目 (2020CDCGJSJ042)

作者简介：李学明 (1967－)，男，博士，教授，主要从事计算机视觉和数据挖掘方面的研究. E-mail：lixuemin@cqu.edu.cn

对视觉信息进行编码和对语义信息进行解码，这样的编码器–解码器框架已经成为图像理解的主流框架。通常，人们使用卷积神经网络 (convolutional neural networks, CNN)[12-14] 提取图像的特征向量，并将图像特征向量输入到长短时记忆网络 (long short term memory, LSTM)[15] 中以生成图像描述语句。为了获得更好的结果，通常使用注意力机制[10, 16]。

当前的图像理解模型存在问题：1) 传统的卷积神经网络 (CNN)[12] 的图像特征提取能力不能满足图像理解的需要，在图像特征提取时未能考虑图像特征与描述语句的关联性。2) 视觉特征的错误会直接导致生成的描述语句错误。3) 当前图像理解方法使用的注意力机制模型复杂且不方便训练。

为解决这些问题，本文提出了一种端到端的基于多模态的注意力机制 (M-AT) 的图像理解方法。多模态的注意力机制包括基于关键词的图像特征提取 (K-IFE) 模型、关键词注意力机制 (K-AT) 和空间注意力机制 (S-AT)。该方法可以生成准确而丰富的图像描述语句，本文使用 MSCOCO[17] 数据集对提出的模型进行评估。

本文的主要贡献：1) 通过将关键词与视觉特征进行关联，提出了基于关键词的图像特征提取 (K-IFE) 方法，让模型能更好地提取与图像理解相关的图像特征。2) 基于关键词注意力机制 (K-AT) 让模型能关注到重要的关键词，从而生成更丰富而准确的描述语句。3) 基于空间特征注意机制 (S-AT) 使用抽象程度较高的图像空间特征来引导模型关注图像的重要区域，生成更为准确的描述语句，同时使用抽象程度较高的图像空间特征简化了模型。4) 最后，结合 K-IFE、S-AT 和 K-AT 提出了基于多模态注意力机制的图像理解 (M-AT) 新方法，当两种注意力机制其中一个出现错误时，另一个可以对其进行矫正，从而提高了描述语句的准确性。

1 相关工作

1.1 长短期记忆网络

循环神经网络 (recurrent neural networks, RNN)[18] 与传统的神经网络相比通过添加隐藏状态保留过去的信息来减轻依赖关系问题，但是初始信息会随着连接长度的增加渐变消失。长短期记忆网络 LSTM [15] 能有效地解决梯度失调问题，在视觉–语言任务[9, 19-21] 中有广泛的运用。

1.2 编码器–解码器

编码器–解码器结构在序列–序列任务中表现良好。受该结构启发，文献 [9] 使用了编码器–解码器结构解决图片理解问题，对视觉信息进行编码，对语义信息解码。编码器–解码器模型已在图像理解任务[8, 16, 22] 中广泛被使用，本文采用 CNN[23] 作为编码器，LSTM [15] 作为解码器。

1.3 注意力机制

LSTM[15] 的存储容量是有限的，生成语句时靠后的词更依赖于选择的语言模型。为了解决这个问题，文献 [19] 将注意力机制引入了图像理解任务。在生成每个时间步的单词时，首先对每个区域的视觉特征都加一个权重，通过该权重计算出新的视觉特征来引导每个时间步单词的生成，这种基于注意力机制的方式能够有效引导描述语句单词的生成。

2 模 型

本文提出了一种基于多模态注意力机制 (M-AT) 的新图像理解方法，通过构建关键词数据集，使用关键词数据集训练图像特征提取模型 (K-IFE)，并通过 K-IFE 提取图像空间特征、图像全局特征和预测的关键词特征，然后结合关键词注意机制 (K-AT) 和空间注意机制 (S-AT) 形成 M-AT。模型结构如图 1 所示。

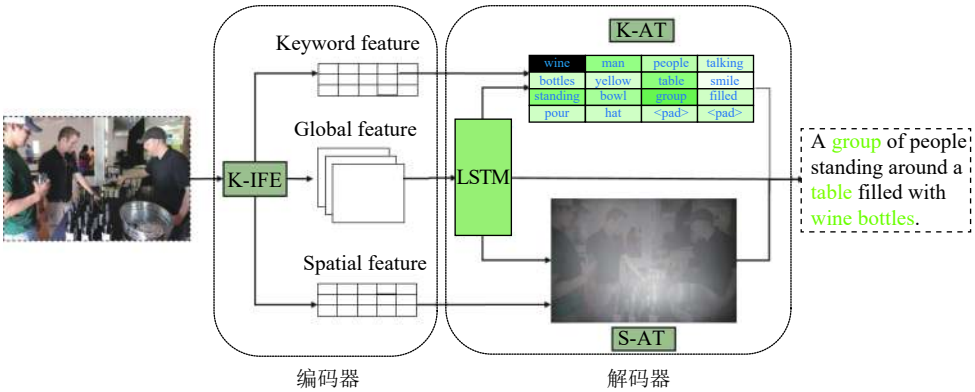

图 1 M-AT 模型结构

M-AT 基本框架是编码器−解码器结构，图 1 表格中颜色越深表示对单词的关注度越高，热力图中区域颜色越亮表示对区域的关注度越高。K-IFE 提取输入图像的关键词特征 (keyword feature)、全局特征 (global feature) 和空间特征 (spatial feature)，将提取到的特征分别输入到关键词注意力机制 (K-AT) 和空间注意力机制 (S-AT) 模块中，两种注意力机制可以进行相互矫正和增强，以获得更精确和丰富的描述语句。

2.1　基于关键词的图像特征提取

2.1.1　关键词数据集

目前开源的图像理解数据集中只提供了图片及其对应描述语句，所以需要自己构建关键词类别数据集和关键词数据集。关键词作为类别标签训练模型具有高频性，关键词代表图片对象的个体、行为、关系具备代表性，同时存在同义单词，将语句中的同义词进行合并，所以关键词具有集成性。综上，关键词具有高频性、代表性、集成性等特点。关键词类别数据集构成过程如图 2 所示。

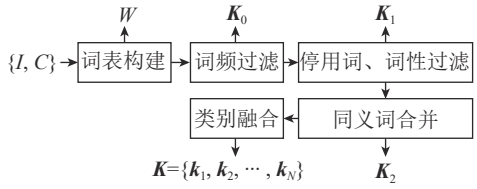

图 2　构造关键词类别数据集

图 2 中，I、C 分别表示输入的任意一张图片及对应的描述语句。

首先从数据集$\{\boldsymbol{I},\boldsymbol{C}\}$构建词频表$W=\{\{W_1,C_1\},\{W_2,C_2\},\cdots,\{W_m,C_m\}\}$，其中$C_i$表示单词$W_i$在所有描述语句$\boldsymbol{C}$中出现的次数，$M$表示单词总数，选择$C_i$大于$|\boldsymbol{C}|/1\,000$的单词构成过滤词表$K_0$保证词表的高频性，选择过滤掉词表$\boldsymbol{K}_0$中的停用单词，保留名词、动词以及形容词得到关键词词表$\boldsymbol{K}_1$。将关键词词表$\boldsymbol{K}_1$中同义单词进行合并为词表$\boldsymbol{K}_2=\{\{k_{10},k_{11},\cdots\},\{k_{20},k_{21},\cdots\},\cdots,\{k_{N0},k_{N1},\cdots\}\}$，其中，$N$为关键词总量，$k_{i0}$为第$i$个关键词类别，如 cat、cats 合并为 cat 类别。最后取每个类别集合的第一个单词作为关键词类别得到关键词类别$\boldsymbol{K}=\{k_i,k_{i+1},\cdots,k_N\}$。

对于任意图片$\boldsymbol{I}$，通过其对应的描述语句$\boldsymbol{C}=\{C_i,C_{i+1},\cdots,C_N\}$分词整理得到相应单词集合$\boldsymbol{C}_W=\{w_i,w_{i+1},\cdots,w_N\}$，从$\boldsymbol{C}_W$中选取所有属于$\boldsymbol{K}_2$的单词得到图片$\boldsymbol{I}$的关键词集合$\boldsymbol{K}_W$。将关键词集合$\boldsymbol{K}_W$按照词表$\boldsymbol{K}_2$进行合并去重可以得到图片$\boldsymbol{I}$对应的关键词类别$\boldsymbol{K}$。对数据集$\{\boldsymbol{I},\boldsymbol{C}\}$中所有的图片都进行上述操作，从而将数据集扩展为包括关键词数据的数据集$\{\boldsymbol{I},\boldsymbol{C},\boldsymbol{K}\}$。

2.1.2　基于关键词的图像特征抽取

为了充分利用视觉特征和语义特征，本文提出了基于关键词的图像特征抽取方法 (K-IFE)，通过关键词提取与图像描述相关的视觉特征。关键词表示描述语句中具有代表性的单词，引入关键词不仅可以得到更好的图像特征，还可以引导描述语句的生成，使生成的描述语句更加准确丰富。式 (1) 为基于关键词的图像理解的目标函数，式 (2) 表示最终的目标函数，有：

$$\boldsymbol{\theta}^*=\arg\max_{\theta}\sum_{I,C}\log p(C,K|I;\boldsymbol{\theta}) \tag{1}$$

$$\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*=\arg\max_{\theta_1,\theta_2}\sum_{I,C}\log p(C|I,K;\boldsymbol{\theta}_1)+\sum_{I,K}\log p(K|I;\boldsymbol{\theta}_2) \tag{2}$$

式中，$\boldsymbol{\theta}$是模型的参数；I为给定图片；C为生成的对应描述语句；K为描述语句中的关键词；$p(C,K|I;\boldsymbol{\theta})$表示给定图片$I$以及模型参数$\boldsymbol{\theta}$得到描述语句$C$和关键词集合$K$的概率$p(C|I,K;\boldsymbol{\theta}_1)$为基于图片特征和关键词的语言模型；$p(K|I;\boldsymbol{\theta}_2)$为基于图片特征的关键词模型；$\boldsymbol{\theta}_1$、$\boldsymbol{\theta}_2$分别表示语言模型和关键词模型参数。

根据条件概率公式，可以将式 (1) 表示为$p(C,K|I)=p(C|I,K)*p(K|I)$，从而得到式 (2)。

本文使用 ResNet[24] 作为图像特征提取的基本模型，该模型最初被设计应用于单标签分类，即一幅图片只对应单个类别，模型输出层应用的激活函数为 softmax 函数：

$$\mathrm{softmax}(z)_i=\frac{\mathrm{e}^{z_i}}{\sum_{i=0}^{m}\mathrm{e}^{z_i}} \tag{3}$$

式中，$\boldsymbol{z}$为输入向量；z_i为$\boldsymbol{z}$的一个分量，$i\in[i,m]$；m表示关键词类别总数。

本文任务是多标签分类任务，softmax 激活函数在多标签分类任务中存在不同类别之间的竞争，因此本文使用 sigmoid 激活函数来避免这种类间竞争。

本文使用关键词多标签分类，由于多标签分类中存在类别不平衡，所以本文选择了 Tencent ML-Images[25] 中的损失函数式 (4)，式 (5) 表示最终损

失函数，在该损失函数后加了L_2正则。对于 mini batch 的数据，n为 mini batch 的数量，$\lambda(\lambda<1)$为L_2正则的权重，m为关键词类别总数，η为惩罚系数，有：

$$L(\boldsymbol{x}_i,\boldsymbol{y}_i)=\frac{1}{m}\times\sum_{j}^{m}r_t^j[-\eta y_{ij}\log(p_{ij})-(1-y_{ij})\log p(1-p_{ij})] \tag{4}$$

$$L=\frac{1}{n}\sum_{i}^{n}\frac{1}{m}\sum_{j}^{m}r_t^j[-\eta y_{ij}\log(p_{ij})-(1-y_{ij})\log p(1-p_{ij})]+\lambda L_2(\boldsymbol{\theta}) \tag{5}$$

式中，$\boldsymbol{x}_i$为单张图片；$\boldsymbol{y}_i=[y_{i1},y_{i2},\cdots,y_{ij},\cdots,y_{im}]$表示图片对应的关键词标签，$j$为关键词类别，$y_{ij}$的取值为 0 或 1；$r_t^j$为训练过程中的一个自适应权重，$r_t^j=0.9^{t-1}$，$t$取决于本轮和上一轮 mini batch 的训练状态，如果两次状态一致则$t=t+1$，如果不一致则$t=1$；$\boldsymbol{\theta}$为模型参数。

2.2 关键词注意力机制

目前图像理解方法存在仅使用视觉特征，但不是所有的视觉特征都能有效提取到，同时可能提取到不准确的视觉特征，所以仅使用视觉特征存在生成的描述语句不准确和不丰富的问题，因此本文提出了基于关键词的图像理解方法 (K-AT)，提取的关键词语义信息能引导生成更丰富和准确的图像描述语句。使用 K-IFE 提取图像的全局特征和关键词特征作为关键词关注机制的输入。

不同的关键词权重表示模型对词语关注程度不同，词语的权重越大表示更加重要，其对生成的描述语句影响也越大，式 (6)～式 (8) 为关键词特征注意力机制公式表达：

$$a_{i,t}^*=\boldsymbol{W}_{\mathrm{a}}\mathrm{ReLU}(\boldsymbol{W}_{\mathrm{ka}}\boldsymbol{k}_i+\boldsymbol{W}_{\mathrm{ha}}\boldsymbol{h}_t) \tag{6}$$

$$\boldsymbol{a}_t=\mathrm{softmax}(\boldsymbol{a}_t^*) \tag{7}$$

$$\hat{\boldsymbol{k}}_t=\sum_{i=1}^{n}a_{i,t}\boldsymbol{k}_i \tag{8}$$

式中，ReLU 是修正线性单元激活函数；$\boldsymbol{k}_i$是关键词对应词向量特征矩阵的第 i 个分量；$\boldsymbol{W}_{\mathrm{a}}$为待学习的权重、$\boldsymbol{W}_{\mathrm{ka}}$、$\boldsymbol{W}_{\mathrm{ha}}$为待学习的权重矩阵；$\boldsymbol{h}_t$为$t$时刻的隐藏状态；$a_{i,t}^*$为计算得到的$\boldsymbol{k}_i$的权重；$\boldsymbol{a}_t^*=\{a_{0,t}^*,a_{1,t}^*,\cdots,a_{n,t}^*\}$为未归一化的权重，$n$为关键词特征$\boldsymbol{k}$的分量个数；$\boldsymbol{a}_t=\{a_{0,t},a_{1,t},\cdots,a_{n,t}\}$为归一化后的权重；$\hat{\boldsymbol{k}}_t$为在$t$时刻应用了注意力机制的关键词特征值。

2.3 空间注意力机制

针对现有的空间注意力机制模型复杂、训练需要耗费大量资源的问题，将注意力机制应用到抽象程度高的图像特征，提出结构更加简单、更易理解的基于空间特征注意力机制的图像理解方法 (S-AT)。通过将注意力机制应用到抽象程度较高的图像空间特征来引导模型关注图像的重要区域，生成更为准确的描述语句。S-AT 模型的网络结构与 K-AT 的网络结构相同，两者参数更新方式也相同。两者不同之处在于 K-AT 的中的关键词特征替换为空间特征，所以在此不对空间注意力机制做过多的叙述。

2.4 多模态注意力机制

通过实验发现 S-AT 和 K-AT 具有相互补充的作用，本文结合 K-IFE、S-AT 和 K-AT 提出了基于多模态注意力机制的图像理解方法 (M-AT)。通过 K-IFE 从输入图像中获得图像空间特征$\boldsymbol{s}$，图像全局特征$\boldsymbol{v}$，以及预测的关键词词向量特征$\boldsymbol{k}$，将其中$\boldsymbol{v}$和$\boldsymbol{s}$作为 S-AT 的输入，S-AT 能够获得更优的视觉特征。$\boldsymbol{k}$和$\boldsymbol{s}$作为 K-AT 的输入，K-AT 能够通过语义特征引导描述语句的生成，M-AT 的单词预测表达式为：

$$\boldsymbol{y}_t=\mathrm{softmax}(\boldsymbol{w}_{\mathrm{o}}(\boldsymbol{w}_{\hat{s}}\hat{\boldsymbol{s}}_t+\boldsymbol{w}_{\hat{k}}\hat{\boldsymbol{k}}_t+\boldsymbol{w}_{\mathrm{h}}\boldsymbol{h}_t+\boldsymbol{b}_{\mathrm{o}})) \tag{9}$$

式中，$\boldsymbol{y}_t$表示图像描述语句词语预测结果；$\boldsymbol{w}_{\mathrm{o}}$和$\boldsymbol{b}_{\mathrm{o}}$为待学习的权重和偏置项；$\hat{\boldsymbol{k}}_t$和$\hat{\boldsymbol{s}}_t$分别表示$t$时刻应用了注意力机制的关键词特征和空间注意力机制的空间特征；$\boldsymbol{w}_{\hat{k}}$、$\boldsymbol{w}_{\hat{s}}$和$\boldsymbol{w}_{\mathrm{h}}$为待学习的权重矩阵；$\boldsymbol{h}_t$为 t 时刻的隐藏状态。

3 实验评估

3.1 数据集

本文在 MSCOCO[17] 数据集上进行实验，该数据集是目前规模最大的图像描述数据集，提供的训练集和验证集分别包含 82 783 幅图片和 40 504 幅图片，每幅图片至少 5 条标注语句，训练集和验证集总的标注语句分别为 414 113 和 202 654 条。测试集包含 40 775 幅图片，MSCOCO 服务器上保存了测试集的两种标注：每幅图片 5 条标注语句和 40 条标注语句，用户可以将测试结果提交到 MSCOCO 服务器进行评估，本文的实验结果是服务器的评估结果。

3.1.1　关键词数据集

为了能更好地训练模型，本文将 90% 的验证集划分到训练集，得到划分后的训练集图片数和描述语句数分别是 119 236 幅和 596 500 条。按照 2.1 节中关键词类别提取流程从训练集$\boldsymbol{c}_{\text{train}}$中提取出 1 705 个关键词，600 个关键词类别，用$\boldsymbol{K}=\{k_1,k_2,\cdots,k_{600}\}$表示。得到关键词类别后，通过 2.1 节中所述方法获取每幅图片的关键词类别，得到关键词数据集$\{\boldsymbol{I}_{\text{train}},\boldsymbol{K}_{\text{train}}\}$，$\{\boldsymbol{I},\boldsymbol{K}\}$表示图片集$\boldsymbol{I}$对应的关键词类别集$\boldsymbol{K}$，关键词数据集为 119 236 张图片及其对应的关键词类别。

3.1.2　评估指标

在实验中，评估指标包括：BLEU[26]、METEOR[27]、ROUGE-L[28]、CIDEr[29]。BLEU[26] 是基于精确度的相似性度量方法，用于分析候生成句和参考语句中 n 元组共同出现的程度。METEOR[27] 基于对生成语句的准确率和召回率的调和平均对其进行评估。ROUGE-L[28] 是基于生成语句召回率的相似性度量方法，用于评估生成语句的充分性和可靠性。CIDEr[29] 通过度量生成语句与参考语句的相似度来评估生成语句。

3.2　实验细节

实验环境为 64 GB 内存，GPU 为 GTX 1080Ti，使用 Tensorflow[30] 框架。在训练过程中，对输入图片进行随机失真、旋转、剪裁以提升模型的泛化能力，最后将图片大小调整为 (256, 256, 3)。所有方法均采用带动量的随机梯度下降算法进行优化，动量的参数设置为 0.9。而对于学习速率，针对不同的模型本文实验采用了不同的学习策略。

在测试中，将图片缩放到 (289, 289, 3)，然后中心裁剪将图片裁剪成 (256, 256, 3)。为了使得最后生成的描述语句更优，本文实验使用集束搜索 (beam search) 的方法生成图像描述。

3.3　实验

K-IFE：为了评估 K-IFE 模型的性能，将 Google-NIC[9] 模型的编码器部分替换为 K-IFE 模型，称为 K-NIC 模型。图 3 所示为 K-IFE 模型预测的图片关键词，可以看出 K-IFE 模型可以有效预测图像中的内容，能够识别图片中大多数对象、对象属性和对象关系。将上述提取到的关键词特征输入到关键词注意力模块中可以有效地丰富描述语句和提高描述语句的准确度。

K-AT：K-AT 能更好地提取到图像中重要内容的关键词，通过关键词的注意力机制可以让模型在生成语句时关注到重要的关键词，从而通过语义特征引导生成更丰富和准确的描述语句。

snow, mountain, snowboard, ski, hill, slope, man, pants, jacket, ramp, person, cover, rail, board, ride, blue

a. 滑雪场景关键词预测

fruit, banana, apple, orange, fresh, tomato, vegetable, arrange, carrot, box, container, plastic, various, type, basket, bag

b. 食物场景关键词预测

图 3　K-IFE 模型的关键词预测

图 4 为关键词提取结果以及基于 K-AT 的图像理解方法生成的描述语句，图 4 表中标绿单词为关键词权重图，颜色越深表示对该词关注度越高，<pad>表示零填充。通过图 4 的模型对比可以发现 K-AT 可以提取出 Google NIC[9] 不能关注到的词语 (如 kitchen、grass)。目前大多图像理解方法使用目标检测的方式来识别图片中的对象，目标检测能有效提取出图片中对象，而对图片背景、关系、环境等元素不能有效提取，然而在生成描述语句时，背景、属性、关系等元素同样重要，所以 K-AT 能直接通过提取关键词的方式有效地提取背景、关系、环境等元素，从而生成更加丰富和准确的描述语句。

kitchen	stand	smile	pose
woman	girl	young	lid
shorts	man	refrigerator	<pad>
<pad>	<pad>	<pad>	<pad>

NIC: a woman standing in front of a refrigerator.
K-NIC: a woman standing in front of a refrigerator.
K-AT: a woman standing in a kitchen next to a refrigerator.

a. 厨房场景关键词预测及描述语句对比

hydrant	fire	green	grass
sit	garden	silver	plant
pipe	blue	red	yard
paint	bottom	next	cute

NIC: a red fire hydrant sitting in the middle of a forest.
K-NIC: a fire hydrant in the middle of a field.
K-AT: a red fire hydrant sitting in the grass.

b. 户外场景关键词预测及描述语句对比

图 4　K-AT 模型结果展示与对比

S-AT：通过空间特征注意力机制引导模型关注图像的重点区域，同时 S-AT 模型使用抽象程度高的特征，因此整个模型结构也更易理解，正确的关注图像的重点区域，生成的描述语句却更加丰富。

NIC: some people standing around a table.
S-AT: a group of people standing around a table filled with wine bottles.

a. 室内场景空间热力图展示及描述语句对比

NIC: several colored umbrellas are left on the ground.
S-AT: a group of colorful umbrellas sitting on top of a grass covered field.

b. 室外场景空间热力图展示及描述语句对比

图 5　空间注意力机制的热力图

图 5 为 S-AT 模型的空间注意力机制的空间热力图，空间热力图 (图 5a 右图、图 5b 右图) 中亮度越高表示权重越大，该区域更加被关注，在生成描述语句时会更加关注权重大的区域。从图 5 的空间热力图可以发现 S-AT 能有效检测到模糊对象、背景 (如 wine、grass)，目前大多数图像理解方法都能检测到突出目标，但是对于模糊对象、小目标、背景等元素不能有效地检测，从而降低了描述语句的丰富性和准确性，所以相比于之前的图像描述模型，S-AT 可以关注到之前模型不能有效关注到的空间区域和不能有效检测到的元素，从而生成丰富的描述语句。

M-AT：将 K-IFE，K-AT 和 S-AT 进行结合构成基于多模态注意力机制的图像理解方法 (M-AT)。使用关键词训练的编码器 (K-IFE) 提取关键词特征、全局特征和空间特征。S-AT 与 K-AT 在生成语句时充分利用视觉特征与语义特征，S-AT 得到更加准确的图像特征。K-AT 得到的语义信息有助于引导生成描述语句。当两种注意力机制有一个出现错误时，另一个可以对其进行矫正，从而得到准确、丰富的描述语句。

本文的方法与其他方法对比如表 1 所示。实验结果表明本文的 S-AT 和 K-AT 方法相比于对比方法有部分提升，通过将 S-AT 和 K-AT 进行结合构成的 M-AT，相比于对比方法有明显的提升，特别是在 CIDEr[29]、METEOR[27]、ROUGE-L[28] 评估指标下有 1%～2% 的提升效果，同时 BLEU[26] 评估指标与对比方法能达到相同的水平，实验结果表明了本文方法的有效性。

表 1　不同模型之间的结果对比

模型	评估指标						
	CIDEr	METEOR	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Human	0.85	0.25	0.48	0.66	0.47	0.32	0.22
m-RNN[6]	0.79	0.23	0.50	0.68	0.51	0.37	0.27
Google NIC[9]	0.86	0.24	0.51	0.69	0.51	0.38	0.28
Hard-AT[19]	0.87	0.24	0.52	0.71	0.53	0.38	0.28
VAE[31]	0.90	0.24	–	0.72	0.52	0.37	0.24
Attribute AT[32]	–	0.24	–	0.71	0.53	0.40	0.30
本文K-NIC	0.87	0.24	0.51	0.70	0.51	0.39	0.29
本文K-AT	0.87	0.24	0.52	0.70	0.53	0.39	0.29
本文S-AT	0.89	0.25	0.52	0.71	0.53	0.39	0.29
本文M-AT	0.91	0.25	0.52	0.71	0.53	0.40	0.30

在理想情况下 S-AT 和 K-AT 能同时关注到正确的位置和关键词，但是任何一种方法都可能出现误差。通过将 K-AT 和 S-AT 进行结合，两种注意力机制能相互增强补充，关键词注意力机制能关注到视觉特征所忽略的细节 (背景、关系、环境等)，空间注意力机制能关注到小目标、模糊对象等，因此 M-AT 生成的描述语句包含更多的细节和更高的准确度。如图 6 所示，M-AT 相比于其他方法能关注到 (electronics、tall building) 这样的小目标和背景等元素，在生成描述语句时，能包含更多细节、背景等描述元素，生成的描述语句会更加接近人类的描述，实现通过机器进行图像理解的目的。

当 K-AT 或 S-AT 其中一个发生关注到错误的关键词或位置的情况时如图 7 所示，K-AT 和 S-AT 可以进行相互矫正，从而提高描述语句的准确性。

图 7a 中红色圆圈表示空间注意力机制关注在

错误的区域 (sky) 上，而下方的关键词注意力机制依然在正确的单词 (head) 上，生成的描述语句通过关键词注意力机制对空间注意力机制进行矫正，最终能生成更加准确的描述语句。

NIC: a bunch of items that are on a table.
K-AT: a bunch of items that are on a table.
S-AT: a collection of various items on a table.
M-AT: a collection of electronics and other items laid out a table.

a. 密集场景M-AT模型结果展示

NIC: a red double decker bus driving down a street.
K-AT: a red and white bus driving down a street.
S-AT: a yellow bus driving down a city street.
M-AT: a yellow bus driving down a street next to tall buildings.

b. 户外场景M-AT模型结果展示

图 6　M-AT 模型结果对比

mirror	window	car	see
vehicle	view	head	hang

a rear view mirror of a dog sticking its head out of a car window

a. K-AT 对 S-AT 进行矫正

elephant	field	grass	bush
stand	walk	wild	brush

an elephant walking through a grassy area with trees in the background

b. S-AT 对 K-AT 进行矫正

图 7　K-AT 和 S-AT 相互矫正

图 7b 中的红色圆圈表示关键词注意力机制关注错误的单词 (bush)，而空间注意力机制关注在正确的区域 (tree)，生成的描述语句时通过空间注意力机制对关键词注意力机制进行纠正。

综上所述，当 K-AT 和 S-AT 其中一个出现错误时，两者可以相互矫正，从而将 K-AT 和 S-AT 进行结合而成的 M-AT 能更准确地生成图像描述语句。

4　结 束 语

本文提出了 K-IFE、K-AT、S-AT 方法，基于上述工作提出基于多模态注意力机制的图像理解方法 (M-AT)，该方法通过 K-IFE 提取更优的图像特征、关键词特征、空间特征，通过关键词注意力机制 (K-AT) 关注重要词语，通过空间注意力机制 (S-AT) 能够关注图像更重要的区域并简化模型结构，并且两种注意力机制可以相互增强矫正，最终生成更加准确和丰富的图像描述语句。

参　考　文　献

[1] MITCHELL M, HAN X, DODGE J, et al. Midge: Generating image descriptions from computer vision detections[C]//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics, 2012: 747-756.

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

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

[4] HODOSH M, YOUNG P, HOCKENMAIER J. Framing image description as a ranking task: Data, models and evaluation metrics[J]. Journal of Artificial Intelligence Research, 2013, 47: 853-899.

[5] KUZNETSOVA P, ORDONEZ V, BERG A C, et al. Collective generation of natural image descriptions[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Jeju Island, Korea: Association for Computational Linguistics, 2012: 359-368.

[6] MAO Jun-hua, XU Wei, YANG Yi, et al. Explain images with multimodal recurrent neural networks [EB/OL]. [2019-07-19]. https://arxiv.org/pdf/1410.1090.

[7] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in Neural Information Processing Systems, 2014, 27: 3104-3112.

[8] KIROS R, SALAKHUTDINOV R, ZEMEL R. Multimodal neural language models[C]//International Conference on Machine Learning. Beijing, China: JMLR, 2014: 595-603.

[9] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]//Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2015: 3156-3164.
[10] PEDERSOLI M, LUCAS T, SCHMID C, et al. Areas of attention for image captioning[C]//Proceedings of the IEEE International Conference on Computer Vision. New York, USA: IEEE, 2017: 1242-1250.
[11] ANEJA J, DESHPANDE A, SCHWING A G. Convolutional image captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2018: 5561-5570.
[12] GIRSHICK R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. New York, USA: IEEE, 2015: 1440-1448.
[13] HE Kai-ming, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. New York, USA: IEEE, 2017: 2980-2988.
[14] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2017: 936-944.
[15] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[16] DONAHUE J, ANNE H L, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. LOS Alamtions, USA: IEEE, 2015: 2625-2634.
[17] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//European Conference on Computer Vision. Cham, Switzerland: Springer, 2014: 740-755.
[18] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.
[19] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International Conference on Machine Learning. Lille, France: DBLP, 2015: 2048-2057.
[20] CHEN Xin-lei, LAWRENCE Z C. Mind ’s eye: A recurrent visual representation for image caption generation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2015: 2422-2431.
[21] KARPATHY A, LI Fei-fei. Deep visual-semantic alignments for generating image descriptions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamtions, USA: IEEE Computer Soc, 2017: 664-676.
[22] BENGIO S, VINYALS O, JAITLY N, et al. Scheduled sampling for sequence prediction with recurrent neural networks[J]. Advances in Neural Information Processing Systems, 2015, 1: 1171-1179.
[23] SZEGEDY C, LIU Wei, JIA Yang-qing, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2015: 1-9.
[24] HE Kai-ming, ZHANG Xiang-yu, REN Shao-qing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2016: 770-778.
[25] WU Bao-yuan, CHEN Wei-dong, FAN Yan-bo, et al. Tencent ML-images: A large-scale multi-label image database for visual representation learning[J]. IEEE Access, 2019, 7: 172683-172693.
[26] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Somerset, USA: Association Computational Linguistics, 2002: 311-318.
[27] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[J]. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, 12(5): 65-72.
[28] LIN C Y. Rouge: A package for automatic evaluation of summaries[J]. Text Summarization Branches Out, 2004, 1: 74-81.
[29] VEDANTAM R, LAWRENCE Z C, PARIKH D. Cider: Consensus-based image description evaluation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2015: 4566-4575.
[30] ABADI M, AGARWAL A, BARHAM P, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems[EB/OL]. (2019-07-10). https://arxiv.org/pdf/1603.04467.
[31] PU Yun-chen, GAN Zhe, HENAO R, et al. Variational autoencoder for deep learning of images, labels and captions[J]. Advances in Neural Information Processing Systems, 2016, 29: 2352-2360.
[32] ZHOW Luo-wei, XU Chen-liang, KOCH P, et al. Watch what you just said: Image captioning with text-conditional attention[C]//Proceedings of the on Thematic Workshops of ACM Multimedia. [S.l.]: ACM, 2017, 1: 305-313.

编 辑 漆 蓉