控制与决策
Control and Decision

**基于全局交互的图像语义理解方法**

库涛， 熊艳彬， 杨楠， 林乐新， 朱珠

引用本文:
库涛, 熊艳彬, 杨楠, 等. 基于全局交互的图像语义理解方法[J]. 控制与决策, 2020, 35(9): 2103−2111.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2018.1699

---

# 基于全局交互的图像语义理解方法

库 涛[1,2†], 熊艳彬[1,2,3], 杨 楠[1,2,3], 林乐新[1,2], 朱 珠[4]

(1. 中国科学院 沈阳自动化研究所，沈阳 110016；2. 中国科学院 机器人与智能制造创新研究院，沈阳 110169；3. 中国科学院大学，北京 100049；4. 辽宁大学 信息学院，沈阳 110000)

**摘 要:** 针对图像语义生成过程中图像信息易模糊的问题,提出基于双向门控循环单元(GRU)和图像信息全局交互相结合的图像语义生成模型,通过图像和文本数据进行正则化处理和文本向量映射方法,实现模型驱动的图像语义生成. 实验结果表明,所提出模型能较好地解决数据稀疏和偏态问题,采用GUR单元可以进一步降低模型参数规模,加快算法收敛速度,有效抑制模型过拟合,提高图像内容的丰富度、准确性和逻辑性.

**关键词:** 卷积神经网络；循环神经网络；图像语义理解；全局交互机制；数据正则化；门控循环单元

**中图分类号:** TP273 **文献标志码:** A

**DOI:** 10.13195/j.kzyjc.2018.1699

**引用格式:** 库涛,熊艳彬,杨楠,等. 基于全局交互的图像语义理解方法[J]. 控制与决策, 2020, 35(9): 2103-2111.

# Image semantic understanding method based on global interaction

*KU Tao*[1,2†], *XIONG Yan-bin*[1,2,3], *YANG Nan*[1,2,3], *LIN Yue-xin*[1,2], *ZHU Zhu*[4]

(1. Shenyang Institute of Automation，Chinese Academy of Sciences，Shenyang 110016，China；2. Institutes for Robotics and Intelligent Manufacturing，Chinese Academy of Sciences，Shenyang 110169，China；3. University of Chinese Academy of Sciences, Beijing 100049，China；4. School of Information，Liaoning University，Shenyang 110000，China)

**Abstract:** Aiming at the problem that image information is easily blurred during image semantic generation, an image semantic generation model based on the combination of gated recurrent unit (GRU) and global interaction of image information is proposed. Processing and word vector mapping methods achieve model-driven image semantic generation. The experimental results show that the model can better solve the problems of data sparseness and skewness. The use of GUR units further reduces the scale of low model parameters, speeds up the algorithm's convergence speed, effectively suppresses model overfitting, and improves the richness, accuracy and logicality of image content.

**Keywords:** convonlution neural network；recurrent neural network；image semantic understanding；global interaction mechanism；data regularization；gated recurrent unit

## 0 引 言

图像语义理解[1]是以图像识别为基础,融合了计算机科学、心理学以及语言学等多学科的交叉学科研究. 其主要目的是通过文字对图像内容进行语义上的解释和描述,目前在图像检索、图像标注、图像解析等领域中均有广泛的应用前景.

近年来,国内外学者在图像语义理解方面取得了大量的研究成果,如Farhadi等[2]提出了基于模板的方法,在对图像中对象、动作、场景和属性的识别基础上,通过固定语句模板填充实现语义理解. Kuznetsova等[3]采用基于转移的文本生成策略,通过转移相似语义描述的图像到目标图像的方法,实现图像语义理解,其方法更加灵活,但是过度依赖于相似图像搜索,具有很大局限性.

随着人工智能技术的发展,基于深度学习的目标识别、机器翻译和图像语义理解技术得到极大的发展[4-5],成为了国内外研究关注的焦点. 近几年国内外学者普遍采用神经语言模型来研究图像语义描述问题,将深度卷积神经网络分类模型与循环神经网络序列模型进行结合,创建一个生成图像描述的端到端的单一网络. 如Mao等[6]提出使用多模递归神经网络模型生成图像标题; Vinyals等[7]使用长短时记忆单元(long short-term memory, LSTM),一种先进的递归神经网络来完成相同的任务; Xu等[8]提出将视觉注

**收稿日期:** 2018-12-12；**修回日期:** 2019-07-15.

**基金项目:** 国家重点研发计划项目(2017YFB0306401)；国家自然科学基金项目(61803367).

**责任编委:** 薛建儒.

†通讯作者. E-mail: kutao@sia.cn.

意力集成到LSTM模型中,以便在生成相应单词期间将注意力固定在不同的图像内容上. 上述方法即为编码到解码(encoding-deconding)的端到端的神经网络标题(neural image caption, NIC)方法[9].

目前,NIC基线模型在图像语义理解研究领域中仍然存在许多问题. 具体包括4个方面:

1)传统的循环神经网络模型[10-11]无法解决语言模型中的长短期依赖以及仅沿一个方向进行语义解析问题,虽然LSTM可以很好地解决长短期依赖问题,但LSTM模型仍然是单向解析模型,语义解析过程中只考虑单向语义信息,得到的图像语义描述存在不够准确、逻辑性不强等问题;

2)卷积神经网络模型提取的图像高维数据只在语义解析开始时输入语言模型,在语义生成过程中存在丢失或模糊图像信息的现象,导致对目标图像的语义理解不够准确和全面;

3)高维图像数据存在数据稀疏和偏态问题,加之文本字典中单词数量即为词向量长度,基线模型中文本输入的one-hot表示会进一步加剧稀疏和偏态问题,导致模型不易收敛;

4)随着模型复杂度提高,模型存在过拟合以及收敛震荡严重、收敛速度过慢等问题.

针对上述4个问题,本文主要进行以下内容的研究. 首先,采用双向门控循环单元(gated recurrent unit, mGRU)模型用于图像语义生成,并且在此基础上引入全局图像交互机制[12],即在生成文本的过程中实时关注图像的全局信息来指导语义生成; 其次,将图像和文本数据进行正则化处理,并采用word2vec文本映射方式来表示文本信息,从而解决高维数据稀疏和偏态问题; 最后,在采用双向GRU单元的基础上加入正则化及Dropout率解决模型过拟合问题. 实验结果表明,采用上述改进后的模型得到的图像语义描述在内容丰富度和准确性上有较大提升;将数据正则化处理,采用word2vec文本映射方式可以较大程度地解决数据稀疏和偏态问题;采用GUR单元可以进一步降低模型参数规模,加快算法收敛速度,结合正则化及Dropout率可以有效抑制模型过拟合.

## 1 基于全局交互的图像语义模型

本文模型整体架构遵循输入图像到输出文本的编码-解码基本结构. 模型通过卷积神经网络得到图像的特征向量信息,通过跨模态交互将输入图像特征信息映射到与标注图像的可变长句子相同的维度空间,并送入语言模型用以生成对目标图像的语义描述[13-14]. NIC基线模型如图1所示.

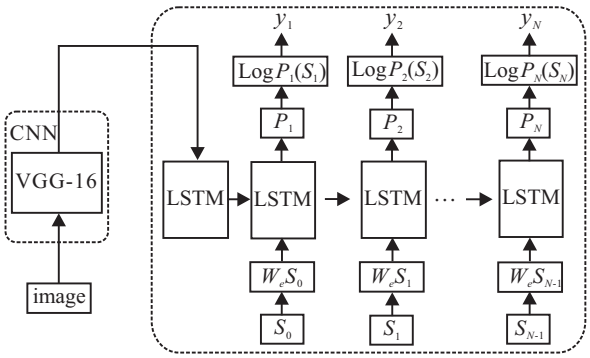


**图1 NIC基线模型**

全局交互主要体现在两个方面:一是引入双向GRU模型用于图像语义生成,即在生成语义信息的过程中实时关注前后语义信息和联系,不再只关注单向的语义信息; 二是在双向GRU基础上将全局图像信息引入GRU单元,在生成文本的过程中实时关注图像的全局信息来指导语义生成. 全局交互模型整体结构如图2所示.

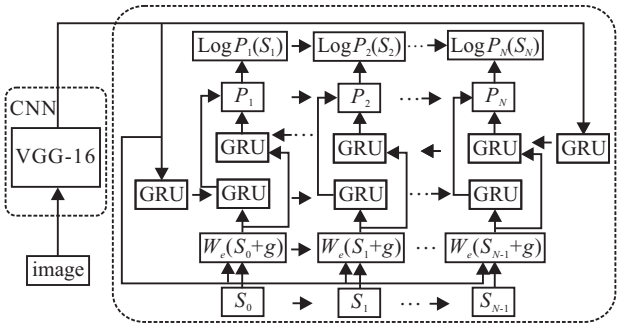


**图2 基于全局交互的图像语义模型**

### 1.1 门控循环单元

设计和训练神经网络时,神经网络的参数规模是网络模型的关键因素. 如果网路模型设计不合理,则在网络参数训练和更新的过程中会面临梯度消失和梯度爆炸问题. 文中为提高语言描述的准确性和丰富度引入双向循环神经网络模型,而LSTM模型[15]是一种特殊的循环神经网络,在统计机器翻译和时序问题上取得了巨大成功,如果直接在单向LSTM网络的基础上加一层反向LSTM网络构成双向循环神经网络,则势必会造成模型参数规模的大幅提高,致使算法收敛比较慢,甚至出现模型过拟合而难以训练出行之有效模型的问题. LSTM单元模型如图3所示.

针对上述问题,全局交互模型采用门控循环单元——GRU, GRU门控单元模型如图4所示.

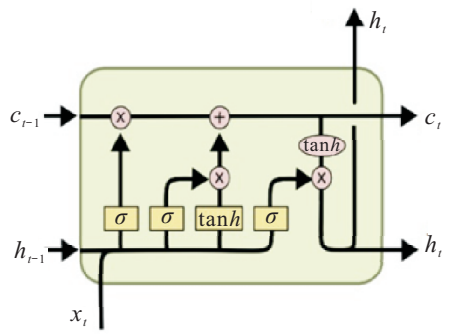


图 3　LSTM结构单元

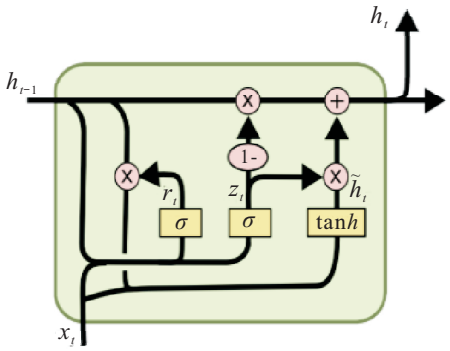


图 4　门控循环单元

与LSTM模型的3个门不同的是,GRU门控单元只用了2个门,将LSTM记忆单元中的输入门和遗忘门合并成了更新门. 不再将线性自更新建立在额外的记忆单元上,而是直接线性累加之后建立在隐藏状态上,并靠门来调控. 因此,GRU大大减少了模型参数,提高了算法的收敛速度,并对传统的循环神经网络(recurrent neural network, RNN)存在的两个问题作了改进: 一是改进了句子中位置越靠前的单词对当前隐藏层节点的影响会越小问题; 二是改进了反向传播的误差可能是由某几个单词诱发的问题. GRU模型只对产生误差的部分权重进行局部参数更新,提高了反向传播算法的工作效率. GRU模型门控更新公式如下($\delta$代表sigmoid函数,$*$代表按位与操作):

$$r_t = \delta(W_r x_t + U_r h_{t-1}), \tag{1}$$

$r_t$负责决定上一时刻隐藏单元$h_{t-1}$对$\tilde{h}_t$的重要程度,如果$r_t$约等于0,则$h_{t-1}$不会传递给$\tilde{h}_t$.

$$z_t = \delta(W_z x_t + U_z h_{t-1}), \tag{2}$$

$z_t$负责决定传递多少上一时刻的隐藏状态$h_{t-1}$给当前时刻的隐藏状态$h_t$,如果$z_t$约等于1,则$h_{t-1}$几乎会直接复制给$h_t$;相反,如果$z_t$约等于0,则$\tilde{h}_t$直接传递给$h_t$.

$$\tilde{h}_t = \tanh(W x_t + U(r_t * h_{t-1})), \tag{3}$$

新的记忆$\tilde{h}_t$是对新的输入$x_t$和上一时刻的$h_{t-1}$的总结,计算总结出的新的向量$\tilde{h}_t$包含上文信息和新的输入$x_t$.

$$h_t = z_t h_{t-1} + (1 - z_t)\tilde{h}_t. \tag{4}$$

隐藏状态由$h_t$和$\tilde{h}_t$相加得到,两者的权重由$z(t)$控制.

联想记忆单元GRU的状态更新表达式说明,GRU单元在将遗忘门和输入门合并为更新门之后,单元内部通过更新法则,自动地寻找输入数据哪些部分选择记忆,哪些部分选择遗忘,最后通过更新门控单元选择产生误差的局部权重进行参数更新. 另外,GRU比LSTM参数规模小,可加速在训练过程中反向传播算法的工作效率,使得算法收敛速度大大加快,提高整个模型的性能.

### 1.2　全局图像交互单元

在NIC基线模型中,图像语义信息的生成主要取决于当前时间点的输入信息和之前的隐藏状态(隐含了开始时输入的图像信息),这个过程逐步进行直到遇到句子的结束标记. 然而,随着这个过程的持续,开始时被送入到语言模型中的图像信息的作用变得越来越弱,在整个语义生成的过程中会出现模糊或者丢失部分图像信息的现象,造成语义描述不能丰富和全面地表达图像内容. 因此,对于需要较长句子对图像进行描述时,在描述的末尾阶段,模型几乎“盲目”地执行到句子的结尾. 虽然联想记忆单元双向GRU能够在一定程度上保持长期记忆,但它仍然对句子生成提出了挑战[16]. 为了解决此问题,引入全局图像交互机制,即在语义生成过程中引入全局图像信息. 结构单元如图5所示.

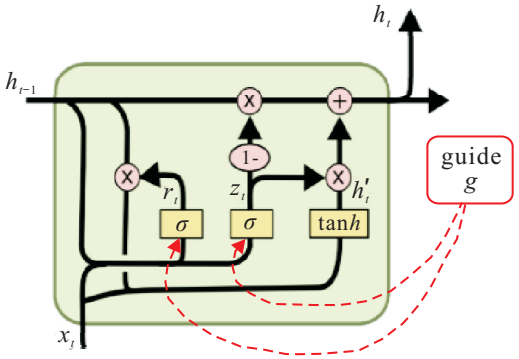


图 5　全局图像信息交互单元

图5中$g$表示送入到语言模型中表示全局图像信息的特征向量.

与标准GRU单元相比,在全局图像信息交互模型中,为每个门和单元状态的计算添加了一个新的术语——指导(guide). 这个新术语代表全局图像信息,是视觉域和文本域之间的桥梁. 全局图像信息不依赖于时间步长,因此在语言模型图像语义生成的整个

过程中作为全局信息指导语义生成. 图中用虚线表示全局图像信息交互模型架构.

全局图像信息交互模型更新公式如下:

$$r_t = \delta(W_r x_t + U_r h_{t-1} + V_r g_{t-1}), \tag{5}$$

$$z_t = \delta(W_z x_t + U_z h_{t-1} + V_z g_{t-1}), \tag{6}$$

$$\tilde{h}_t = \tanh(W x_t + U(r_t * h_{t-1})), \tag{7}$$

$$h_t = z_t h_{t-1} + (1 - z_t)\tilde{h}_t. \tag{8}$$

式(5)~(8)与式(1)~(4)相同,只不过在更新门和修正门更新公式中加入全局交互图像信息作为状态更新的输入元素.

### 1.3 双向循环神经网络结构

引入双向循环网络结构——双向GRU模型,旨在解决图像语义理解模型存在的两个关键问题:一是模型参数过多导致训练过程收敛速度太低、模型过拟合问题;二是针对单向循环网络只沿一个方向进行语义解析,得到的目标图像内容语义描述不够准确,自然语言逻辑性不强等问题.

在目标图像自然语言描述的生成训练过程中,如果采用单向循环神经网络模型(如单向GRU模型),每个单词在生成过程中只会关注它在词序上左边的文本信息,而不关注它右边的文本信息,但是在解码某个词语过程中,通常需要知道该单词周围的信息,即前后语境信息. 例如,当出现“颜色”一词时,需要向前查询是解码谁的颜色,如“大海”而不是其他的对象的颜色. 还要向后查询,如果后面有“污染”“藻化”或者其他表述时,此时“大海”的颜色可能就不会再解析为“蓝色”,而解析为“红色”或者其他. 引入双向循环神经网络使得每个单词在生成时会同时关注该词左右两侧的信息,从而保证在生成图像的语义描述过程中使生成的语言更加自然,提高语言表达的准确性和丰富度.

单向循环神经网络同一时刻$t$只输出一个方向的信息,而双向循环神经网络的主体结构就是两个单向循环神经网络的结合. 在每一时刻$t$,输入会同时提供给这两个方向相反的循环神经网络.

两层网络独立进行参数更新,都遵循式(5)~(8)的更新规则,产生各自在该时刻$t$的新状态和输出. 双向循环网络的最终输出是这两个单向循环神经网络输出的直接线性叠加. 两个循环神经网络除方向不同以外,基本结构完全对称[17-19]. 更新公式如下:

$$\tilde{h}_t^F = \tanh(W^F x_t + U^F(r_t^F * h_{t-1}^F)), \tag{9}$$

$$\tilde{h}_t^B = \tanh(W^B x_t + U^B(r_t^B * h_{t-1}^B)), \tag{10}$$

$$y_t = W_{hy}^F h_t^F + W_{hy}^B h_t^B + b_y. \tag{11}$$

双向GRU模型按时序展开如图6所示.

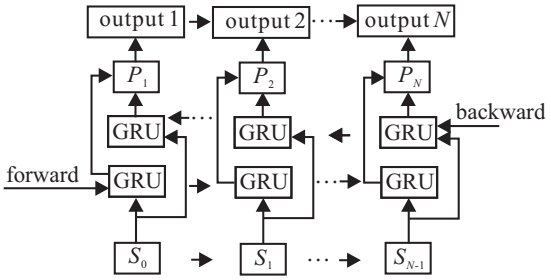


图6 双向GRU网络

## 2 模型训练与测试

### 2.1 基准数据集-Flick30k

目前,Flickr30k数据集已成为基于文本的图像语义描述的基准数据集,本文亦采用Flickr30k数据作为模型训练数据. Flickr30k数据集中每张图片对应5条文本描述,数据集示例如图7所示.

图7 示例图片

1) A young boy is either jumping on a trampoline or skateboarding and he is serious about it on a beautiful summer day.

2) A young boy looking to be jumping on the trampoline in a very beautiful neighborhood.

3) A guy is jumping and grabbing his foot with a house in the background.

4) A boy grimaces as he jumps high in the air.

5) A boy grabs his leg as he jumps in the air.

### 2.2 图像编码及特征提取

本文在编码端采用卷积神经网络VGG-16模型对输入图像进行特征提取,在网络输出端得到4 096维的图像特征信息,并将此特征向量作为图像的全局信息送入解码端进行跨模态交互. VGG-16模型[20]如图8所示.

该网络接受224×224像素的输入图像,采用迁移学习方式,利用在ImageNet上预训练好的模型权重,作为初始值导入本文模型,以便加速训练过程.

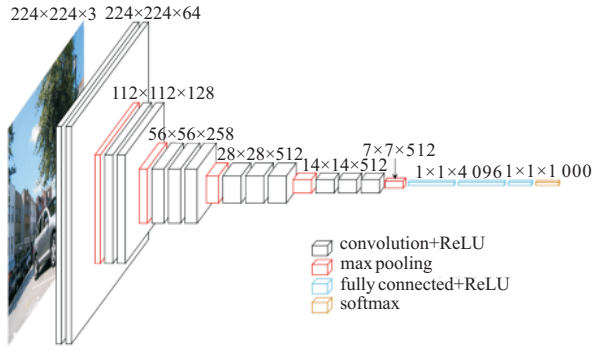


**图 8　VGG-16模型**

利用VGG-16网络的特征提取功能,提取网络生成的高维特征向量. 本文采用模型输出阶段瓶颈层的4 096维向量作为图像的全局信息并送入语言模型. 如果将4 096维的高维图像数据直接输入语言模型,则会产生数据稀疏的问题和风险,因此要对其进行降维处理.

### 2.3　跨模态交互

为了构建图像描述的语义信息,需要将表示图像信息的高维特征向量跨模态交互到文本信息空间中,本文模型采用归一化典型相关分析(canonical correlation analysis, CCA)来解决跨模态交互问题[21]. 归一化典型相关分析是一种将图像视觉信息和文本特征信息映射到公共语义空间的典型方法,其旨在为两个视图$X_1$和$X_2$学习投影矩阵$U_1$和$U_2$,并确保它们的投影最大程度地相关,公式如下:

$$\arg\max_{u_1u_2}\frac{u_1\Sigma_{x_1x_2}u_2}{(u_1\Sigma_{x_1x_1}u_1)(u_2\Sigma_{x_2x_2}u_2)},\tag{12}$$

其中$\Sigma_{x_1x_2}$、$\Sigma_{x_1x_1}$、$\Sigma_{x_2x_2}$为协方差矩阵,CCA目标函数可以通过广义特征值分解来求解. 通过使用特征值的幂来计算归一化的CCA,并对CCA投影矩阵的相应列进行加权,然后进行$L_2$归一化,公式如下:

$$g_1=\frac{X_1U_1D^p}{\|X_1U_1D^p\|},\ g_2=\frac{X_2U_2D^p}{\|X_2U_2D^p\|}.\tag{13}$$

其中: $D$是对角矩阵,其元素值为相应维度的特征值,而$g_1$和$g_2$表示两个视图的语义表示. 利用余弦相似性在学习的公共语义空间中找到最近邻.

### 2.4　文本编码

在语言模型中,文本信息通常转化成序列向量的形式来表达语义信息,然后再将序列向量输入到模型中;模型输出同样为序列向量,然后再将其解码为文本信息,这样就完成了语义到语义的端到端语言模型架构. 通常情况下会直接将语义文本转化成one-hot序列向量输入到语言模型,此时语义文本one-hot序列向量的维度将与词汇表中词汇量的个数相同,通常在10 000以上.

这种one-hot序列向量形式存在两个问题: 一是相邻单词之间没有词序相关性; 二是采用one-hot编码得到的高维向量数据比较稀疏,容易出现维度灾难以及参数过多导致模型收敛速度降低的问题. 为了解决以上问题,本文采用基于神经网络的数据表示模型——word2vec. 此模型由Google 2013年开源,通过语料库训练获取词语的多维实数向量表示,即通过神经网络对上下文以及上下文与目标词之间的关系进行建模.

首先对语料字典中的每个单词进行编码,然后使用word2vec模型对该单词进行训练,最终生成256维的词嵌入向量. 该向量不仅保留了词序信息,而且词义相近的词语在空间中的距离也是相近的,生成的词向量如图9所示.

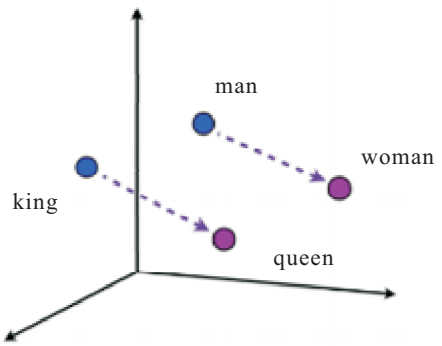


**图 9　词向量**

### 2.5　数据正则化处理

在图像语义理解模型中,所有图像数据基于服从高斯分布的假设. 由于高维图像以及文本数据存在数据稀疏以及偏态问题,会增加模型复杂度,导致模型收敛速度下降以及不易收敛的问题. 为解决此问题,全局交互模型对图像及文本数据进行正则化处理,降低VGG16网络提取的图像的高维数据以及文本数据的稀疏及偏态问题,起到滤除图像数据及文本数据中的噪声,提高模型的收敛速度,降低模型复杂度的作用. 实验结果证明,通过将图像及文本数据进行正则化处理,可以降低收敛曲线的波动性,提高模型的收敛速度.

### 2.6　全局交互机制的训练过程

全局图像信息交互机制特别之处在于图像信息不只在$t=-1$时刻输入语言模型,而是在每一个时间序列中图像信息都会送入GRU单元,而不受时间序列影响. 全局图像信息不依赖于时间步长,在图像语义生成的整个过程作中实时与语义信息进行交互来指导语义生成.

语言模型训练阶段,不同时刻会有不同输入:

1) 在$t = -1$时刻,图像特征向量作为全局信息输入GRU单元,并作为唯一的输入信息对GRU单元隐藏层进行初始化;2) 在之后的时间步长$t$中(即从$t = 0$开始),模型输入信息不仅包括全局图像信息,还包括与输入图像对应的文本描述中每个单词的词向量(测试阶段为前一时刻生成的单个词语的词向量). 此种机制保证了在语言模型训练过程中,图像信息在每一时刻都会送入语言模型作为当前时刻的输入信息$x(t)$的一部分,并作为全局交互信息指导图像语义的生成.

在每个时间步长$t$,GRU单元在给定当前输入和隐藏层状态下,计算生成下一个词的概率分布,以最高概率对应的单词作为当前时刻图像描述的输出[22]. 结合训练样本中与输入图像对应的语义描述,通过有监督的训练最大化模型的似然函数来完成整个模型的训练过程.

该模型的GRU网络输出送入Softmax层,Softmax层将模型输出词汇的概率分布输出如下:

$$p(S_t|S_{t-1}\dots S_0, I) = \text{Softmax}(D_{y(t)}), \tag{14}$$

其中$D$具有与GRU单元数相同的维度,将$y(t)$映射到输出词汇大小为$N$的解码器矩阵.

### 2.7 模型优化

图像特征提取阶段采用迁移学习,其特征提取模块的参数未经过训练,直接利用在ImageNet上预训练好的VGG-16模型权重作为初始值导入模型加速训练过程. 在语言模型训练过程中,加入正则项控制模型复杂度,并且使用Dropout方法进行优化. 使用Dropout方法将网络层中的某些神经元的输出随机置零,然后将其输入到下一层,这将减少神经元之间的依赖从而使神经网络更加鲁棒. 实验中发现使用Dropout率为0.5时能够最大化地提高模型的泛化能力.

模型训练的整个过程可以通过下式概括:

$$x_{-1} = \text{CNN}(I); \tag{15}$$

$$x_t = W_e(x_t + g),\ t \in (0, N); \tag{16}$$

$$x_{t+1} = \text{GRU}(x_t),\ t \in (0, N); \tag{17}$$

$$L(I, S) = -\sum_{t=1}^{\text{N}} \log\ p_t(x_t). \tag{18}$$

在求解优化问题时,习惯性地将其作为最小优化问题来对待,因此将损失定义为每个步骤中正确单词的负对数似然的总和.

通过模型训练调整语言生成模型参数,以最小化公式(18)中的负对数似然函数. 模型通过反向传播算法来优化参数以降低损失函数,使用基于梯度下降的AdamOptimizer优化算法,优化GRU模型中所有的词嵌入向量和解码矩阵等GRU参数.

### 2.8 测试阶段

在测试阶段,即选择第三方图片作为测试集生成新的图像描述阶段. 此阶段没有可供参考的图像描述,输入图像的语义描述需要从模型输出结果中进行选择,生成输入图像的描述. 与训练阶段相同,图像的特征向量在时间步长$t = -1$时被送到联想记忆单元GRU网络,对GRU单元中的隐藏状态进行初始化. 从$t = 0$开始,每一步时间序列都会将全局图像信息输入到GRU单元中,作为全局交互信息指导图像语义描述的生成. 在时间步长$t = 0$时,Softmax层输出中具有最高概率位置的词是模型认为描述该输入图像最可能的第1个词. 选择该词作为第1个单词,然后在时间步长$t = 1$中将相应的单词向量作为输入提供给GRU单元,重复该迭代过程直到模型生成完整的句子为止,这种句子生成方式被称为Sampling方法. Sampling方法的一个问题是只考虑在每个时间步长中最可能的单词,但是不能保证最终得到合理性描述图像的句子.

## 3 实验结果分析

### 3.1 模型算法收敛分析

图10所示为原有NIC基线模型的损失波动曲线,图11所示为改进后模型的损失波动曲线. 其中下方曲线表示训练集损失波动曲线,上方曲线表示验证集损失波动曲线.

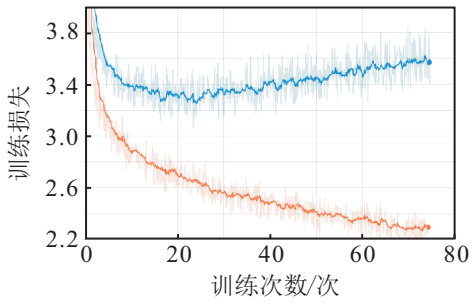


**图10 NIC基线模型的损失波动曲线**

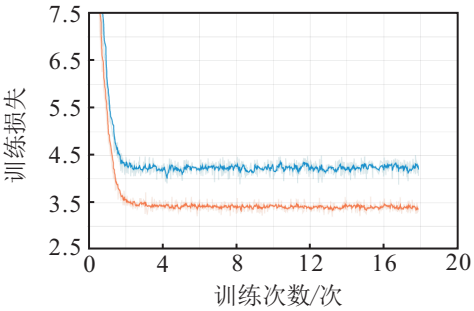


**图11 全局交互模型的损失波动曲线**

对原有NIC基线模型的损失按训练次数进行抽样,绘制出图10和图11所示的收敛速率对比图. 通过比较可以看出,损失比原有NIC基线模型损失波动幅度大幅减小、下降收敛速度明显加快,优化算法最终的收敛效果也比NIC基线模型有很大提升,并且全局交互模型有效地避免了原有基线模型随着训练深度的加深出现过拟合的现象.

每个LSTM结构单元有3个门,全局交互模型采用GRU结构单元,GRU单元将LSTM单元中的输入门和遗忘门合并成了更新门并作了其他参数调整. 从模型参数规模上看,GRU单元比LSTM单元减少约1/3,避免了直接采用双向LSTM结构导致模型参数规模大幅增加的问题,从而能大大提高模型收敛速度,并在抑制模型过拟合方面起到较大作用.

在GRU模型中引入Dropout率和正则化,本文采用0.5的Dropout率,随机减少50 %的参数规模来提高模型的泛化能力. 正则化在训练过程中引入刻画模型复杂度的指标,与模型损失一块加入损失函数进行优化,有效抑制了模型过拟合. 本文中加入正则化,在抑制模型过拟合过程中进一步避免了数据稀疏问题,从而起到降低模型训练过程中的波动程度、加快收敛速度的作用.

通过实验结果数据作出的收敛速率比对图验证了以上结论,即全局交互模型在收敛速度和抑制模型过拟合方面取得了更加可观的效果.

### 3.2　模型算法准确率分析

从原有数据集中随机抽取70 %作为训练集,剩余30 %作为验证集,在每轮训练的过程中对原有NIC基线模型和全局交互模型按轮次进行抽样,绘制出准确率分析对比,如图12、图13所示. 其中上方曲线表示训练集,下方曲线表示验证集.

原有NIC基线模型在训练初始阶段参与训练的样本较少的情况下,准确率处于较高位置,随着训练轮数的增加准确率出现较大幅度的下降,说明模型存在过拟合,导致测试集准确率急剧下降,最终也并未得到有效提高.

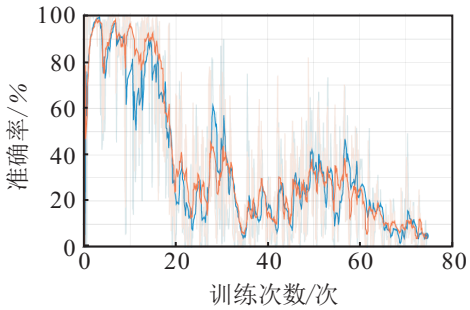


**图 12　NIC基线模型准确率曲线**

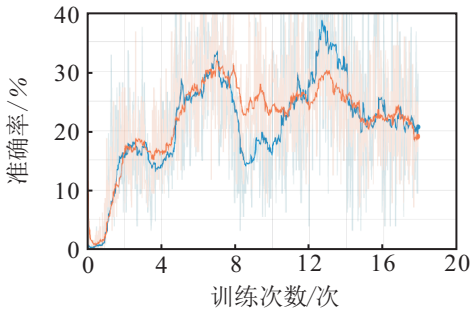


**图 13　全局交互模型准确率曲线**

采用双向GRU结构单元,较大程度上降低了模型参数规模,同时加入正则化及引入Dropout率可以进一步抑制模型过拟合,降低训练的波动程度;模型训练过程中将图像数据和文本数据进行标准正则化处理,较大程度地解决数据稀疏和偏态问题并起到滤除数据噪声作用,使模型具备较好的鲁棒性,从而在一定程度上提高了模型的准确率.

实验结果表明,不管在训练集还是在验证集上的表现相对于NIC基线模型都有所提高,并且最终在第三方测试集上实验结果表现良好.

### 3.3　文本生成分析

在所有图像数据基于独立同分布的假设前提下,获取新的数据以验证模型效果. 实验过程中进行了多次模型训练,保存每一轮训练的模型结果. 通过对测试集中图像的验证,发现不同的训练次数获得的模型对同一图像生成的描述内容是不同的,文本描述结果表现出一定的随机性而且差别较大. 随着模型训练轮次的增多,对图像的描述越来越丰富,但是这也不是绝对的. 在试验测试结果中也发现,训练轮次很深的模型反而不能生成对句子的正确描述. 虽然全局交互模型仍然有不足之处,但比NIC基线模型在图像描述完整性和丰富性上比基线模型有了较大幅度提升.

在语言生成过程中采用双向GRU模型,在生成文本过程中同时关注上下文语境信息,不再只关注单向语义信息,确保生成的文本内容具备较高的逻辑性、关联性;全局交互机制在生成文本的过程中实时关注图像的全局信息来指导语义生成,避免了图像语义生成过程中图像信息易模糊的问题,从而使生成的文本内容更能贴近输入图像的信息.

以下给出模型在测试集图片上生成的文本结果,如图14所示.

1) A white dog and a black dog are running on the grassy field.

2) A motorcyclist in black outerwear is riding a motorcycle on the road.

3) A young boy in a blue shirt is playing with a little toy.

4) A little girl in a white shirt is sitting on the grass holding a few flowers.

5) A little boy in blue jeans is running on the beach.

6) A man in white shirt and shorts is playing soccer.


**图 14 测试图片生成文本结果示例**

实验结果表明,第三方图像数据在改进的模型中获得的图像语义描述比基线模型更加详细和具体,体现图片的信息也更加全面和丰富,这正是模型设计的初衷和想要得到的结果.

## 4 结 论

本文所提出的基于全局交互的图像语义理解方法可以使基线模型中存在的问题得到有效改善. 采用双向全局交互机制,在语言模型生成新的语义单词时,不再只关注该单词前面的信息,而是关注前后语义信息和联系,从而使模型得到的图像语义描述在内容丰富度、准确性和逻辑性上有较大提升. 将图像和文本数据进行标准正则化处理送入语言模型降低了高维数据的稀疏及偏态影响,也在一定程度上滤除了数据中的噪声. 采用word2vec映射方式对文本进行编码进一步缓解数据稀疏问题,提高了模型的收敛速度. 采用GUR单元可以进一步降低模型参数规模,加快算法收敛速度,结合正则化及Dropout率可以有效抑制模型过拟合. 此外,训练数据的规模、特征提取模型的选择、模型训练过程中参数的初始化等问题对图像语义理解研究都有较大影响,后续将对其进行深入研究.

## 参考文献(References)

[1] Donahue J, Hendricks L A, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4): 677-691.

[2] Farhadi A , Hejrati M , Sadeghi M A , et al. Every picture tells a story: Generating sentences from images[J]. Lecture Notes in Computer Science, 2010, 21(10): 15-29.

[3] Kuznetsova P, Ordonez V, Berg T, et al. Treetalk: Composition and compression of trees for image descriptions[J]. Transactions of the Association of Computational Linguistics, 2014, 2(1): 351-362.

[4] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. EMNLP, 2014,14(6): 1078-1093.

[5] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]. Advances in Neural Information Processing Systems. Montreal, 2014: 3104-3112.

[6] Mao J, Xu W, Yang Y, et al. Deep captioning with multimodal recurrent neural networks (m-rnn)[J]. arXiv: 1412.6632, 2015.

[7] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]. Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Boston: IEEE, 2015: 3156-3164.

[8] 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, 2015: 2048-2057.

[9] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3128-3137.

[10] Bahdanau D, Cho K, Bengio Y. Neural machinetranslation by jointly learning to align and translate[J]. arXiv: 1409.0473, 2015.

[11] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv: 1412.3555, 2014.

[12] Jia X, Gavves E, Fernando B, et al. Guiding the long-short term memory model for image caption generation[C]. IEEE International Conference on Computer Vision. Santiago: IEEE, 2016: 2407-2415.

[13] Graves A. Sequence transduction with recurrent neural networks[J]. Computer Science, 2012, 58(3): 235-242.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems. Long Beach, 2017: 5998-6008.

[15] Greff K, Srivastava R K, Koutník Jan, et al. LSTM: A search space odyssey[J]. IEEE Transactions on Neural Networks & Learning Systems, 2017, 28(10): 2222-2232.

[16] Cho K, Van Merriënboer B, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches[J]. arXiv: 1409.1259, 2014.

[17] Lee K, Kim J K, Kim J, et al. Stacked convolutional bidirectional LSTM recurrent neural network for bearing anomaly detection in rotating machinery diagnostics[C]. 2018 IEEE International Conference on Knowledge Innovation and Invention (ICKII). Jeju Island: IEEE, 2018: 98-101.

[18] Li L, Cai G, Chen N. A rumor events detection method based on deep bidirectional GRU neural network[C]. 2018 IEEE International Conference on Image, Vision and Computing (ICIVC). Chongqing: IEEE, 2018: 755-759.

[19] Arisoy E, Sethy A, Ramabhadran B, et al. Bidirectional recurrent neural network language models for automatic speech recognition[C]. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane: IEEE, 2015: 5421-5425.

[20] 杨楠, 南琳, 张丁一, 等. 基于深度学习的图像描述研究[J]. 红外与激光工程, 2018, 47(2): 18-25.
(Yang N, Nan L, Zhang D Y, et al. Research on image interpretation based on deep learning[J]. Infrared and Laser Engineering, 2018, 47(2): 18-25.)

[21] Gong Y, Ke Q, Isard M, et al. A multi-view embedding space for modeling internet images, tags, and their semantics[J]. International Journal of Computer Vision, 2012, 106(2): 210-233.

[22] Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models[C]. International Conference on Machine Learning. Beijing, 2014: 595-603.

## 作者简介

库涛 (1979–), 男, 研究员, 博士, 从事群智协同计算、社会计算等研究, E-mail: kutao@sia.cn;

熊艳彬 (1989–), 男, 硕士生, 从事图像语义、数据挖掘的研究, E-mail: xiongyanbin@sia.cn;

杨楠 (1994–), 男, 硕士生, 从事数据挖掘、图像语义的研究, E-mail: yangnan@sia.cn;

林乐新 (1984–), 男, 助理研究员, 硕士, 从事目标检测、过程检测等研究, E-mails: linyuexin@sia.cn;

朱珠 (1983–), 女, 讲师, 博士后, 从事数据分析与系统优化等研究, E-mail: zhuzhuzz@126.com.

(责任编辑：孙艺红)

# 下 期 要 目