

毕业设计（论文）

开题报告

17~18cm

题目名称：一种融合注意力机制的多模态可控文本生成

5.5~6cm

院系名称：计算机学院网络工程

班 级：网络151

学 号：201508030120

学生姓名：周会全

指导教师：刘小明

2019 年3 月

**一、立论依据**

|  |
| --- |
| **课题来源及研究的目的和意义** 课题来源 基于关联数据的信息检索关键技术研究61672361国家自然科学基金委员会 国家自然科学基金项目刘小明  基于情感语义计算的网络舆情分析研究14A520015河南省教育厅 河南省教育厅科学技术研究重点项目刘小明 2. 课题目的2.1 文本可控生成 文本自动生成是让计算机能够模仿人类撰写出高质量自然语言文本。按输入数据不同，可分为文本到文本、意义到文本、数据到文本及图像到文本等多模态数据的文本生成。可控文本生成进一步要求对所生成文本的内容、风格、时态、等特征能够满足预先设定的要求。 2.2 向多模态输入的可控文本生成 所谓图像文本生成，即机器学习领域中的image caption任务，就是从图片中自动生成一段描述性文字，有点类似“看图说话”，对于人来说，image caption是简单自然的一件事，但对于机器来说，这项任务充满了挑战性。原因在于机器不仅要能检测出图像中的对象，理解物体之间的相互关系，还要用语言表达出来。  本课题的形式等于image caption衍生任务，自动生成图像的标题，不同点在于以图像、文本、语义知识等多模态数据融合为输入的可控文本生成。输入的图像是基础素材，提供内容基础；输入的文本来协调生成新文本的风格约束；输入的语义知识作为 文本生成的限制性条件，如情感，作为可控的一个目标；这些都是前提或约束，核心目标是生成可控的文本。 3. 课题意义 本课题文本可控的基础上，例如生成固定长度的诗歌，生成指定感情色彩的评论，控制图片描述生成，这种跨模态的可控输出不仅提高了计算机视觉的能力，在日后应用场景上，能够使计算机通过视觉关注到人们想要其关注到的内容，提高计算机智能对话系统。  受控生成图像的标题，作为标题生成模型不仅必须足够强大，能够解决计算机视觉中确定图像中哪些对象的挑战，而且还必须能够用自然语言捕获和表达它们之间的关系。由于这个原因，标题生成一直被视为一个难题。对机器学习算法来说，这是一个非常重要的挑战，因为它相当于模仿人类将大量显著的视觉信息压缩成描述性语言的非凡能力。 |

**二、文献综述**

|  |
| --- |
| **课题研究领域的发展现状及可能的应用领域** 1.课题研究领域的发展现状1.1国际研究现状 图像到文本的生成技术是指根据给定的图像生成描述该图像内容的自然语言文本，例如新闻图像附带的标题、医学图像附属的说明、儿童教育中常见的看图说话、以及用户在微博 等互联网应用中上传图片时提供的说明文字。依据所生成自然语言文本的详细程度及长度的不同，这项任务又可以分为图像标题自动生成和图像说明自动生成。前者需要根据应用场景 突出图像的核心内容，例如，为新闻图片生成的标题需要突出与图像内容密切关联的新闻事件，并在表达方式上求新以吸引读者的眼球；而后者通常需要详细描述图像的主要内容，例如,为有视力障碍的人提供简洁详实的图片说明，力求将图片的内容全面且有条理的陈述出来，而在具体表达方式上并没有具体的要求。  对于图像到文本的自动生成这一任务，人类可以毫不费力地理解图像内容，并按具体需 求以自然语言句子的形式表述出来；然而对于计算机而言，则需要综合运用图像处理，计算 机视觉和自然语言处理等几大领域的研究成果。作为一项标志性的交叉领域研究任务，图像 到文本的自动生成吸引着来自不同领域研究者的关注。自 2010 年起，自然语言处理界的知 名国际会议和期刊 ACL、TACL 和 EMNLP 中都有相关论文的发表；而自 2013 年起，模式识别 与人工智能领域顶级国际期刊 IEEE TPAMI 以及计算机视觉领域顶级国际期刊 IJCV 也开始刊登相关工作的研究进展，至2015年，计算机视觉领域的知名国际会议 CVPR中，更是有近10篇相关工作的论文发表，同时机器学习领域知名国际会议 ICML 中也有2篇相关论文发表。图像到文本的自动生成任务已被认为是人工智能领域中的一项基本挑战。  与一般的文本生成问题类似，解决图像到文本的自动生成问题也需要遵循三阶段流水线 模型[1]，同时又需要根据图像内容理解的特点，做出一些调整：  在内容抽取方面，需要从图像中抽取物体、方位、动作、场景等概念，其中物体可以具 体定位到图像中的某一具体区域，而其他概念则需要进行语义标引。这部分主要依靠模式识别和计算机视觉技术。  在句子内容选择方面，需要依据应用场景，选择最重要（如图像画面中最突出的，或与 应用场景最相关的），且意义表述连贯的概念。这部分需要综合运用计算机视觉与自然语言处理技术。  最后，在句子实现部分，根据实际应用特点选取适当的表述方式将所选择的概念梳理为 合乎语法习惯的自然语言句子。这部分主要依靠自然语言处理技术。而本课题就在于在句子实现部分加入可控，以实现生成特定句子。  早期工作主要依照上述三阶段的流水线模式来实现。例如,在 Yao 等人的工作[2]中，图像被细致的分割并标注为物体及其组成部分，以及图像所表现的场景，并在此基础上选择与场景相关的描述模板，将物体识别的结果填充入模板得到图像的描述文字。而 Feng 与 Lapata[3] [4]则采用概率图模型对文本信息和图像信息同时建模，并从新闻图片所在的文字 报道中挑选合适的关键词作为体现图像内容的关键词，并进而利用语言模型将所选取的内容 关键词、及必要的功能词汇链接为基本合乎语法规则的图像标题。还有一些工作[5] [6] [7] [8] [9]则依靠计算机视觉领域现有的物体识别技术从图像中抽取物体（包括人物、动物、花草、 车、桌子等常见的物体类型），并对其定位以获得物体之间的上下位关系，进而依赖概率图 模型和语言模型选取适当的描述顺序将这些物体概念、介词短语块串联成完整的句子。 Hodosh 等人[10]则利用基于核函数的典型关联分析（Kernel Canonical Correlation Analysis， KCCA）来寻找文本与图像之间的关联，并依据图像信息对候选句子排序，从而获得最佳描述 句子。值得说明的是，Hodosh 等人的工作[10]和 Feng 与 Lapata 的工作[4] [5]均没有依靠现有的物体识别技术。  图1  图 1：一种典型的流水线模型  随着深度学习方法在模式识别、计算机视觉及自然语言处理领域的广泛应用，基于海量数据的大规模图像分类、语义标注技术得到了快速发展；同时，统计机器翻译等与自然语言 生成相关的技术也有了显著的提高。这也催生了将图像语义标注及自然语言句子生成进行联合建模的一系列工作，一方面在图像端采用多层深度卷积神经网络（Deep Convolution Neural Network，DCNN）对图像中的物体概念进行建模，另一方面在文本端采用循环神经网络（Recurrent Neural Network，RNN）或递归神经网络（Recursive Neural Network）对自然语言句子的生成过程进行建模[11]。传统图像语义标注工作主要关注具体某个物体的识别以及物体之间的相对位置关系，而对动作等抽象概念的关注较少。Socher 等人 [12] 提出利 用递归神经网络对句子建模，并利用句法解析树突出对于动作（动词）的建模，进而将图像端与文本端进行联合优化，较好的刻画了物体与动作之间的关系。为了将两种不同模态的数据统一在一个框架下，Chen 与 Zitnick[13]将文本信息与图像信息融合在同一个循环神经网络 中，利用图像信息作为记忆模块，从而指导文本句子的生成，同时又借助于一个重构图像信 息层，实现了图像到文本、文本到图像的双方向表示。而 Mao 等人[14]则通过 DCNN 得到的图像信息与文本信息融合到同一个循环神经网络（m-RNN）中，将图像信息融入到了自然语言句子生成的序列过程中，取得了不错的结果。类似的想法也被 Donahue 等人[15]应用于动作识别和视频描述生成过程中。但在 m-RNN 的句子生成过程中，在图像端并没有显著的约束，例如在下图中，当生成单词“man”的时候，并没有与图像信息中的任务标注发生直接或间接的关联。 **图2** 图 2：多模态 m-RNN 模型  本文采用的是谷歌和加拿大蒙特利尔大学和多伦多大学的研究人员则分别借鉴了统计机器翻译领域 的最新研究进展来推进图像到文本自动生成的联合建模[16] [17]。前者利用深层卷积神经网络 DCNN 对图像建模，将图像信息“编码”（encoding）后，直接由另一个与之相连接的 LSTM 神经网络（Long-Short Term Memory Network，LSTM）“解码”(decoding)成自然语言句子， 无需进行图像-词对齐、调序等传统模型的子步骤。而后者则在基于神经网络的机器翻译框 架下，提出利用计算机视觉领域中的“注意”(Attention)机制来促进词语和图像块之间的 对齐，从而在句子生成过程中，模拟人视觉的“注意”转移过程能够与词语序列的生成过程 相互促进，使生成的句子更符合人的表述习惯。  图3  图 3：视觉“注意”引导的图像标题生成过程  此外，微软的研究人员[17]利用卷积神经网络 CNN 和多示例学习（Multiple Instance Learning，MIL）对图像建模，并利用判别式语言模型生成候选句子，并采用统计机器翻译 研究中经典的最小误差率训练(Minimum Error Rate Training，MERT)来发掘文本和图像层 面的特征对候选句子进行排序。 虽然图像到文本的生成技术还处在探索阶段，距离实际产业应用还有一定的距离，但工 业界已经开始注意到这一技术的理论研究价值和潜在应用前景，积极与学术界合作拓展研究 方向。在 2015 年的计算机视觉知名国际会议 CVPR 2015 上举办的 LSUN Challenge （Large-scale Scene Understanding）挑战活动中也进行了图像标题自动生成的评测任务， 最终谷歌公司[16]和微软研究院[17]取得了总成绩并列第一名，蒙特利尔-多伦多联队[18]和另 一只微软研究院队伍[19]总成绩并列第三名，加州伯克利分校[15]获得第五名。 1.2国内研究现状 国内学术界对图像到文本的生成技术研究开展较晚，大部分科研单位专注于跨媒体数据的语义标注和检索等任务上，只有人大、清华、北大、北航和中科院等科研单位开展了相关 研究，如人民大学与腾讯合作在 2015 年欧盟组织的 ImageCLEF评测中，在图像句子生成 （Image Sentence Generation）任务中取得了第一名。  在工业界方面，百度和腾讯等科研机构也依靠自身在跨媒体语义标注、分类和检索等方 面的研究优势，逐步开展相关方向的研究工作，如百度与UCLA合作的m-RNN系统在CVPR 2015 LSUN 评测的图像标题自动生成任务中也取得了不错的成绩。 2.课题应用领域 从图像到文本的生成技术需要集成模式识别与机器学习、计算机视觉、自然语言处理， 甚至认知科学领域的研究成果，具有极高的理论研究价值和实用前景。从一定程度上讲，这 一技术同图像语义标注等任务一道，已成为各大顶尖科研机构在人工智能领域综合研究实力 的较量方式，必将促进其快速发展。  而对于这一任务本身而言，更大的挑战仍然在于如何正确的抽取图像的内容，同时根据人类的语言习惯选择适当的表述方式将图像内容转换为自然语言句子。需要指出的是，目前的研究仍然聚焦在是否将图像中的物体概念抽取完全，是否选择了正确的词语，所生成的句子是否符合语法习惯等；可以预见在不久的将来，实际应用场景和上下文语境等约束将进一步推进相关技术的进步，必将广泛应用于新闻传播、在线教育、智能家居等多个领域。 |

**三、研究内容**

|  |
| --- |
| 功能需求及设计框架  本课题主要包含两部分：图像特征提取，可控文本生成(两部分实现跨模态文本生成) 1.图像特征提取 我们采用无监督学习的方法(输入的为无标签COCO2014数据集)，经过DCNN神经网络输出句子特征  cnn  图 4：CNN网络模型 1.1核心问题： 句子特征提取不准确，应融合attention机制 2.可控文本生成 我们采用Encoder-decoder-Classifier结构(输入为yelp情感色彩数据以及图片)，经过GAN神经网络输出要求文本  图片5  图 5：GAN网络模型 2.1核心问题： GAN神经网络资源消耗大，训练时间周期长，应融合attention机制保证句子生成效果 3.整体框架 cnn图片5 3.1整体流程： 将图像神经网络得来的句子特征（图4）作为（图5）的Encoder，经过GAN网络，AutoEncoder重建输入文本，在AE网络的隐藏层中加入unstructure变量c，作为attribute的信息输入。同时在模型中加入了discriminator作为分类器，将生成语句作为输入，重建structure变量，对于每一个attribute使用一种discriminator。模型中利用Encoder，重建生成语句的unstructure变量，使得模型在生成过程中能够尽可能的区分unstructure变量以及structure变量。 |

**四、研究基础**

|  |
| --- |
| 所需实验手段、研究条件和实验条件 1.所需实验手段1.1 Encoder-decoder网络结构： 从广义上讲，它的目的是将输入序列（源序列）转换为新的输出序列（目标序列），并且不限制两个序列的长度，换句话说，二个序列的长度可以任意。以nlp领域来说，序列可以是句子、段落、篇章等，所以我们也可以把它看作处理由一个句子（段落或篇章）生成另外一个句子（段落或篇章）的通用处理模型。对于句子对，我们期望输入句子Source，期待通过Encoder-Decoder框架来生成目标句子Target。Source和Target可以是同一种语言，也可以是两种不同的语言，若是不同语言，就可以处理翻译问题了。若是相同语言，输入序列Source长度为篇章，而目标序列Target为小段落则可以处理文本摘要问题(目标序列Target为句子则可以处理标题生成问题)等等。  编码器encoder: 编码器处理输入序列并将序列信息压缩成固定长度的上下文向量（语义编码/语义向量context）。期望这个向量能够比较好的表示输入序列的信息。在图像模型中encoder用来卷积特征，模型使用CNN来提取L 个D 维vector，每一个都对应图像的一个区域：，与此前的工作使用Softmax层之前的那一层vector作为图像特征不同，本文所提取的这些vector来自于 low-level 的卷积层，这使得decoder可以通过选择所有特征向量的子集来选择性地聚焦于图像的某些部分。这个样子就有点像NLP里的seq2seq任务了，这里的输入从词序列转变成了图像区域vector的序列。  解码器decoder: 利用上下文向量初始化解码器以得到变换后的目标序列输出。早期工作仅使用编码器的最后状态作为解码器的输入。编码器我们使用带有attention机制的DeepRNN神经网络（多层LSTM循环神经） 1.2 LSTM： Long Short Term 网络， 一般就叫做 LSTM ，是一种 RNN 特殊的类型，可以学习长期依赖信息。LSTM 由[Hochreiter & Schmidhuber (1997)](http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf)提出，并在近期被[Alex Graves](https://scholar.google.com/citations?user=DaFHynwAAAAJ&hl=en)进行了改良和推广。在很多问题，LSTM 都取得相当巨大的成功，并得到了广泛的使用。  LSTM 通过刻意的设计来避免长期依赖问题。记住长期的信息在实践中是 LSTM 的默认行为，而非需要付出很大代价才能获得的能力！ 所有 RNN 都具有一种重复神经网络模块的链式的形式。在标准的 RNN 中，这个重复的模块只有一个非常简单的结构，例如一个 tanh 层。标准 RNN 中的重复模块包含单一的层。  LSTM 同样是这样的结构，但是重复的模块拥有一个不同的结构。不同于 单一神经网络层，这里是有四个，以一种非常特殊的方式进行交互。 1.3 soft-attention： 所谓soft，意思是在求注意力分配概率分布的时候，对于输入句子X中任意一个单词都给出个概率，是个概率分布。  通过attention机制计算出的z^t被称为 context vector，是捕捉了特定区域视觉信息的上下文向量。  首先需要明确，attention要实现的是在解码的不同时刻可以关注不同的图像区域，进而可以生成更合理的词。那么，在attention中就有两个比较关键的量，一个是和时刻 t 相关，对应于解码时刻；另一个是输入序列的区域ai，对应图像的一个区域。  实现这种机制的方式就是在时刻 t ，为输入序列的各个区域 i 计算出一个权重αti。因为需要满足输入序列的各个区域的权重是加和为一的，使用Softmax来实现这一点。至于Softmax需要输入的信息，则如上所讲，需要包含两个方面：一个是被计算的区域ai，另一个就是上一时刻t-1的信息ht−1，具体计算我们会在后面进行简单解释。 2.研究条件2.1图像特征提取 在图像特征提取，我们可以利用CNN的空间特性，使用卷积层的feature map，给图片的不同位置都提取一个特征。举例来说，用表示我们提取的图片特征，一共L个位置，每个位置的特征为一个D维的向量，对于一个高、宽为14，通道数为256的feature map，对应的有了这些位置的特征，我们再让Decoder在解码时拥有在这196个位置特征中选择的能力，这就是Attention机制。设在第t阶段（通俗的讲，就是生成第t个单词时）传入Decoder RNN的上下文向量为，RNN前一阶段的隐层状态为。这个上下文向量就是 的一个加权平均，具体地，和的关系用式子表达：，就是衡量生成第t个单词时，第i个位置的图像特征所占的权重。这个权重实际是前一个隐层状态和第i个位置图像特征的函数。具体的表达式为：，由于只和已有的信息 ， 有关，因此这些参数也是可以从数据中进行end-to-end的自动学习。 2.2文本生成过程2.2.1生成器 生成器G 是一个LSTM-RNN网络，在隐藏向量条件下生成序列，这个过程我们可以描述为一个条件概率模型，，其中表示之前的标记。因此，该生成涉及一系列离散决策，其在每个时间步T使用Softmax函数从参数化的多项分布中采样： 2.2.2判别器 训练鉴别器D以准确地推断句子属性并评估恢复潜在代码中指定的所需特征的错误。例如，对于分类属性，可以将鉴别器表示为句子分类器; 而对于连续目标，可以使用概率回归量。 与VAE编码器相比，以不同的方式学习鉴别器，因为目标属性可以是离散的，这在VAE框架中是不支持的。  此外，与以无监督方式学习的非结构化代码z相比，结构化变量c使用标记的示例来引入指定的语义。 我们为判别器推导出一种有效的半监督学习方法。  一般我们使用表示判别器的参数。 为了学习指定的语义，我们使用一组带标签的例子来训练判别器D，表达如下： 3.实验条件3.1数据集 本此实验我们使用COCO2014图像数据集，这个数据集以scene understanding为目标，主要从复杂的日常场景中截取，图像中的目标通过精确的segmentation进行位置的标定。图像包括91类目标，328,000影像和2,500,000个label。该数据集主要解决3个问题：目标检测，目标之间的上下文关系，目标的2维上的精确定位。  以及情感分析yelp数据训练模型 3.2 BaseLine 我们使用DCNN神经网络提取图片特征，并使用LSTM实现基本的句式转换 3.3提升模型 在先前DCNN模型基础上我们使用RNN初始化模型，提高句子特征提取的准确率，并在Decoder加入attention机制，使用GAN网络提高文本生成质量 3.4实验配置 实验模型采用一台拥有2个p100GPU的服务器，进行训练。 |

**五、计划进度**

|  |
| --- |
| 2019/03/01--2019/03/18开题与课题准备搜集和阅读相关研究的文献 2019/03/19--2019/03/26 阅读相关研究的文献复现相关论文实验，总结学习其模型算法 2019/03/27--2019/04/05 数据预处理与原型设计基础模型构建与实验 2019/04/06--2019/04/20 新模型构建与实验 2019/04/21--2019/04/30 对比实验结果分析 2019/05/01--2019/05/31 1）总结实验2）书写论文3）准备答辩 |

**六、参考文献**

|  |
| --- |
| [1] E. Reiter, R. Dale, Building natural language generation systems, Cambridge university press2000.  [2] B.Z. Yao, X. Yang, L. Lin, M.W. Lee, S.-C.J.P.o.t.I. Zhu, I2t: Image parsing to text description, 98 (2010) 1485-1508.  [3] Y. Feng, M. Lapata, How many words is a picture worth? automatic caption generation for news images, Proceedings of the 48th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 1239-1249.  [4] Y. Feng, M.J.I.t.o.p.a. Lapata, m. intelligence, Automatic caption generation for news images, 35 (2013) 797-812.  [5] Y. Yang, C.L. Teo, H. Daumé III, Y. Aloimonos, Corpus-guided sentence generation of natural images, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 444-454.  [6] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, Baby talk: Understanding and generating image descriptions, Proceedings of the 24th CVPR, Citeseer, 2011.  [7] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L.J.I.T.o.P.A. Berg, M. Intelligence, Babytalk: Understanding and generating simple image descriptions, 35 (2013) 2891-2903.  [8] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, H. Daumé III, Midge: Generating image descriptions from computer vision detections, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2012, pp. 747-756.  [9] D. Elliott, F. Keller, Image description using visual dependency representations, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1292-1302.  [10] M. Hodosh, P. Young, J.J.J.o.A.I.R. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, 47 (2013) 853-899.  [11] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128-3137.  [12] R. Socher, A. Karpathy, Q.V. Le, C.D. Manning, A.Y.J.T.o.t.A.f.C.L. Ng, Grounded compositional semantics for finding and describing images with sentences, 2 (2014) 207-218.  [13] X. Chen, C. Lawrence Zitnick, Mind's eye: A recurrent visual representation for image caption generation, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2422-2431.  [14] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A.J.a.p.a. Yuille, Deep captioning with multimodal recurrent neural networks (m-rnn), (2014).  [15] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625-2634.  [16] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156-3164.  [17] H. Fang, S. Gupta, F. Iandola, R.K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J.C. Platt, From captions to visual concepts and back, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1473-1482.  [18] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, International conference on machine learning, 2015, pp. 2048-2057.  [19] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, M.J.a.p.a. Mitchell, Language models for image captioning: The quirks and what works, (2015).  [20]Toward Controlled Generation of Text  [21]Unsupervised Text Style Transfer using Language Models as Discriminators  [22]Toward Unsupervised Text Content Manipulation  [23]Style Transfer from Non-Parallel Text by Cross-Alignment  [24]From Captions to Visual Concepts and Back |