

毕业设计（论文）

题目名称：一种融合注意力机制的多模态可控文本生成

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 5 月

论文编号：201508030120

一种融合注意力机制的多模态可控文本生成

**Multimodal Controllable Text Generation with Intentional Mechanism**

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 5 月

# 中文摘要

图像文本生成任务在机器学习领域，我们称为image caption，其主要目的在于从图片中自动生成一段描述性文字，类似于我们语文中的看图写作题目，对于我们人来说，做一道看图写作题轻而易举，但对于机器来说，这项任务却充满了挑战性，原因是，机器要在能够切割出图片中主体的前提下，还要能够理解主体之间的相互关系，并用流畅的语言表达出来。

本课题主要希望在image caption任务的基础上实现可控文本生成，不同于该任务，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。本文以CNN网络卷积图片特征，并利用融合注意力机制的GAN网络模型生成可控的正负情感主题描述。

关键词：多模态；注意力机制；可控文本生成

# **Abstract**

Image text generation tasks in the field of machine learning, we call image caption,its main purpose is to automatically generate a descriptive text from the picture, similar to the picture writing problem in our language. For us, it is a breeze to do a picture writing problem, but for the machine, the task is full. The challenge is that the machine must understand the interrelationships between the subjects and be able to express them in a fluid language, while being able to cut out the main body of the picture.

This topic mainly hopes to realize controllable text generation based on the image caption task. Different from this task, it is the controllable text generation with multimodal data fusion such as image, text and semantic knowledge. The input image is the base material, providing the content basis; the input text is used to coordinate the style constraints for generating new text; the input semantic knowledge is used as a restrictive condition for text generation, such as emotion, as a controllable target; these are prerequisites or Constraints, the core goal is to generate controllable text. In this paper, the CNN network convolution picture features, and the GAN network model of the fusion Attention mechanism is used to generate a controllable positive and negative emotion topic description.

**Key words: Multimodal; Attention mechanism; Controllable text generation**

# **目 录**

[中文摘要 I](#_Toc9132)

[Abstract II](#_Toc16450)

[目 录 III](#_Toc22379)

[第1章 绪论 1](#_Toc18787)

[1.1 选题背景及研究意义 1](#_Toc4754)

[1.2 研究现状 2](#_Toc18799)

[1.3 主要任务 3](#_Toc24209)

[1.4 论文的组织结构 3](#_Toc24366)

[第2章 相关研究 5](#_Toc5564)

[2.1 国际研究 5](#_Toc8079)

[2.2 国内研究 8](#_Toc13708)

[2.3 图像自动生成的难点 8](#_Toc25958)

[2.4 本章小结 9](#_Toc7127)

[第3章 研究内容的理论模型 10](#_Toc5173)

[3.1 相关基础理论简述 10](#_Toc20215)

[3.1.1 Encoder-Decoder网络结构 10](#_Toc3658)

[3.1.2 LSTM网络结构 11](#_Toc12279)

[3.1.3 Soft-Attention网络结构 12](#_Toc24753)

[3.2 本文理论模型论述 14](#_Toc12084)

[3.2.1 Squence-to-Squence With Attention网络结构 14](#_Toc31689)

[3.2.2 GAN网络结构 16](#_Toc7969)

[3.3 本章小结 18](#_Toc17796)

[第4章 一种融合注意力机制的多模态可控文本生成实施方案 19](#_Toc24075)

[4.1 一种融合注意力机制的多模态可控文本生成分析过程 19](#_Toc31380)

[4.2 WordEmbedder 20](#_Toc5900)

[4.3 CNN卷积网络 22](#_Toc15895)

[4.4 UnidirectionalRNNEncoder 23](#_Toc7059)

[4.5 AttentionRNNDecoder 24](#_Toc29560)

[4.6 Gumbel-softmax 25](#_Toc8744)

[第5章 一种融合注意力机制的多模态可控文本生成实验与分析 27](#_Toc7340)

[5.1 实验总体设计 27](#_Toc23638)

[5.2 实验数据以及实验配置 29](#_Toc1771)

[5.2.1 实验数据 29](#_Toc9440)

[5.2.2 实验配置 30](#_Toc8852)

[5.3 实验评测方案及指标 30](#_Toc20656)

[5.4 实验过程 32](#_Toc1765)

[5.4.1 训练模型生成图片描述 32](#_Toc16412)

[5.4.2 准备可控文本生成训练语料 32](#_Toc26786)

[5.4.3 构建embedding层 32](#_Toc31608)

[5.4.4 构建自动编码器AE 32](#_Toc17137)

[5.4.5 构建Gumbel-sofmax 34](#_Toc15627)

[5.4.6 构建分类器以及优化器 34](#_Toc11233)

[5.4.7 对比试验 34](#_Toc7138)

[5.5 实验结果分析 36](#_Toc23171)

[5.5.1 batch-size的影响 36](#_Toc26143)

[5.5.2 gamma\_decay的影响 37](#_Toc21122)

[5.5.3 lambda\_g的影响 38](#_Toc28324)

[5.6 本章小结 39](#_Toc2444)

[第6章 总结与展望 40](#_Toc29017)

[6.1 总结 40](#_Toc15640)

[6.2 展望 41](#_Toc2648)

[参考文献 42](#_Toc9036)

[致 谢 45](#_Toc24518)

# 第1章 绪论

自然语言处理（Nature Language Processing）也可称为 NLP，就是指人们平常交流时使用的语言。它涉及众多学科，与语言学关系密切，但又有重大的区别。它作为machine learning和AI领域的一个研究方向。其研究目的在于探寻人类与机器之间运用自然语言进行有效交流的技术和理论方法。 要实现人类与机器之间进行语言交流也就意味着要使计算机能够理解人类语言的意思， 并且也能以人类语言的形式来表达出机器想要表达的意思。前者是理解自然语言，后者是生成自然语言。由此，理解自然语言和生成自然语言构成了自然语言处理的技术方向。

自然语言处理包含了许多不同类型的任务，其中主要包括人们耳朵能听到的语音识别及语音合成领域、以及人类自身独特的高级语言文本领域。其中文本领域又分为文本分类和文本生成，文本自动生成作为NLP的基本任务之一，同时也是人工智能走向成熟的一个重要标志。文本自动生成技术有众多应用前景，不仅可以应用于智能问答与对话、机器翻译等系统，实现智能和自然的人机交互；我们还可以通过文本自动生成系统代替编辑实现新闻的自动撰写与发布，这将很有可能颠覆新闻行业；更有甚者，它还可以用来帮助学者撰写学术论文，从而改变科研创作的固化模式。

从不同的输入方式进行划分，它又可以分为文本到文本的生成(text-to-text-gener -ation)、图像到文本的生成(image-to-text generation)、数据到文本的生成(data-to-text generation)以及意义到文本的生成(meaning-to-textgeneration)等。本课题主要是图像到文本的生成，并要求输入的图像作为测试集，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。

## 1.1 选题背景及研究意义

本课题主要来源：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河南省教育厅科学技术研究重点项目刘小明。

我们研究自动文本生成是因为它有众多应用前景，不仅可以应用于智能问答与对话、机器翻译等系统，实现智能和自然的人机交互；我们还可以通过文本自动生成系统代替编辑实现新闻的自动撰写与发布，这将很有可能颠覆新闻行业；更有甚者，它还可以用来帮助学者撰写学术论文，从而改变科研创作的固化模式。

本课题在文本生成的基础上加以可控，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，而且还需要能够提取分割出来的对象之间存在哪种关系，正是这个原因，标题描述自动生成对机器学习算法来说，是一个挑战，它相当于模仿人类将视觉信息转换成可描述的自然语言，而一种融合注意力机制的图像文本可控生成更是一种困难的挑战。

## 1.2 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、以及互联网用户在微博、微信等应用中上传图片时提供的说明文字等。根据生成自然语言文本的长度以及程度的不同，其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的不同，突出图像的主要内容，例如，为新闻播报所用图片生成的标题需要突出与图片相关联的内容，并要求在表达方式上吸引读者；而后者需要详述图像的主要内容，例如,为知识水平较低的人提供简明扼要的图片说明，尽力将图片的内容简洁并且容易理解的陈述出来，而并不要求具体的表达方式。

对于图像到文本的转换，我们可以在脑中快速的分析并构建图像的主题，并按照一定的要求描述出来；但对于计算机而言，则需要综合运用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。从2010年起，自然语言处理界的知名国际会议和期刊 ACL、TACL 和 EMNLP与之相关论文的发表变得尤为活跃；在2013 年，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 也开始收录相关期刊，在2015年，计算机视觉领域的知名国际会议 CVPR中，相关论文的发表已经达到10篇，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。在不断的探索与发现中，图像到文本的自动生成任务被认为是人工智能领域中的一项基本挑战。而可控文本生成作为近年来研究发展趋势，图像到文本的可控生成的研究也在不断的发展和完善中。

## 1.3 主要任务

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，结合传统的image caption任务，文本自动生成任务，文本可控生成任务，研究构建一套面向多模态数据的可控文本生成模型体系框架。

其中本任务要研究的可控文本生成模型和多模态文本生成模型，进一步研究构建一种融合注意力机制与多模态特征的可控文本生成模型，实现对给定图片自动生成正负面情感文本标题的可控文本生成，以及正负面情感特征的可控自动问答。

主要研究过程如下：

1. 充分调研现有image caption任务和可控文本生成任务，探索多模态可控文本生成框架。
2. 深入研究GAN网络模型原理，以及其在可控文本生成任务中的应用。
3. 构建面向多模态的可控文本生成的神经网络模型。
4. 在COCO2014数据集验证其有效性。
5. 在真实数据集上验证验证其实用性。

## 1.4 论文的组织结构

全文分为6章：

1. 绪论，主要介绍该课题的来源，选题背景及研究意义，分析说明当前研究现状以及明确该课题的主要任务是研究构建一套面向多模态数据的可控文本生成模型体系框架。
2. 相关研究，本章主要介绍了国内外图像文本自动生成的发展趋势，归纳分析现有图像文本自动生成的方法，分析其瓶颈，从中帮助构建图像到本文的可控文本生成。
3. 研究内容的理论模型，本章主要介绍用于image caption任务的神经网络模型和用于文本可控生成的神经网络模型
4. 一种融合注意力机制的多模态可控文本生成的实施方案，本章主要介绍本文所用的模型，以及实现的思路和过程
5. 一种融合注意力机制的多模态可控文本生成的的实验与分析，本章详细介绍了本文所做实验的步骤及详细细节，以及实验结果的深度分析。
6. 总结与展望，主要总结了本文研究的成果以及存在的不足，并对以后的多模态可控文本生成任务进行建议。

# 第2章 相关研究

## 2.1 国际研究

同传统文本生成任务无异，图像到文本的自动生成同样需要遵循三个阶段的流水线模型[1]，但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，需要从图像中抽取场景、物体等特征以描述对象是什么及其对象之间的关系是什么，其中物体需要具体定位到图像中的指定区域，其他概念需要进行语义标引。这部分的实现主要依赖于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符合场景的，而且要求意义表述连贯。这部分的实现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，根据抽取到的对象及其对象之间的关系，按照应用场景生成自然流畅的自然语言，这部分的实现主要依赖自然语言处理技术。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。

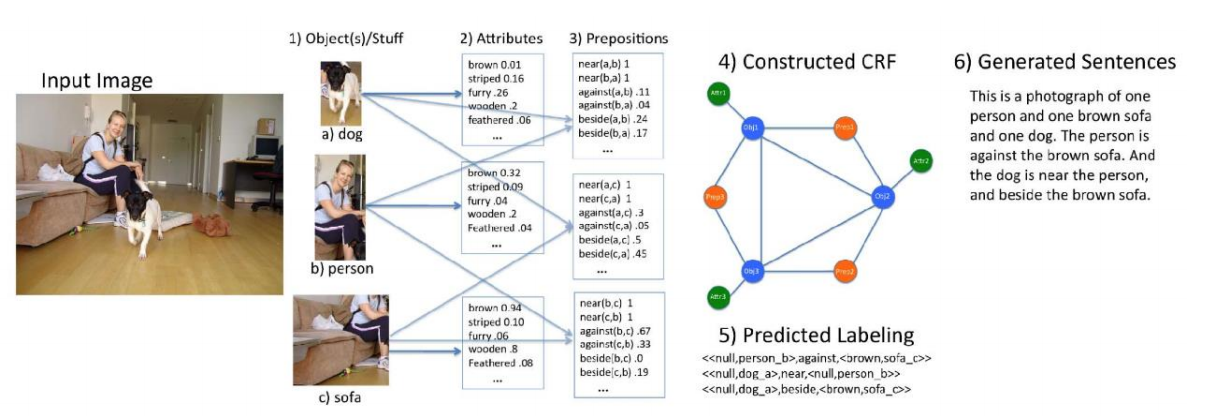


图 1 一种典型的流水线模型

随着深度学习方法在计算机视觉、模式识别及自然语言处理领域的广泛应用，基于大数据的语义标注、图像分类技术得到了快速发展；同时，机器翻译、命名实体识别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。而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。

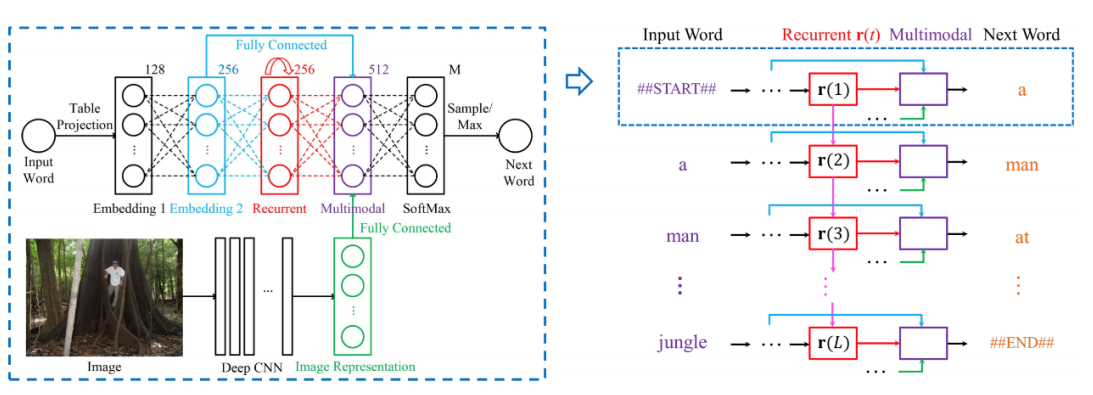


图 2 多模态 m-RNN 模型

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。

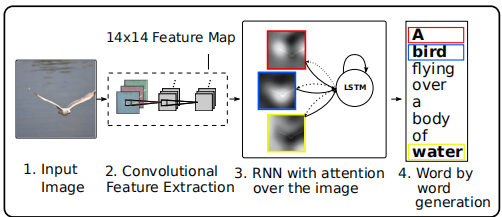


图 3 视觉“注意”引导的图像标题生成过程

此外，微软研究院[17]利用CNN卷神经网络和多示例学习（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，采用机器翻译模型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，并对候选句子进行排序。 虽然图像到文本的生成技术还处在初始阶段，距离实际产业应用也存在一定的距离，但工业界积极与学术界合作拓展研究方向，说明了这一技术的理论研究价值和潜在应用前景已经得到了充分的肯定。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）挑战活动中也发布了图像标题自动生成的评测任务，最终加州伯克利分校[15]获得第五名，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，谷歌公司[16]和微软研究院[17]取得了总成绩并列第一名。

## 2.2 国内研究

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，只有中科院、清华、北大、人大和北航等科研单位开展了相关研究，如腾讯和人大合作的2015年欧盟组织的 ImageCLEF评测项目中，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，腾讯以及百度等科研机构也依靠自身在跨媒体分类、语义标注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UCLA与百度合作的m-RNN系统在CVPR 2015 LSUN图像标题自动生成任务中也获得了不错的成绩。

## 2.3 图像自动生成的难点

从图像到文本的生成技术需要集成模式识别与机器学习、自然语言处理、计算机视觉以及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，这一技术同语义理解，语义标注等任务，已成为各大顶尖科研团队在人工智能领域综合实力的评判标准，必将促进其快速发展。

对于这一任务而言，更大的难点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。需要了解到的是，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。

## 2.4 本章小结

本章主要介绍了国外和国内外图像自动生成的相关研究，分析总结图像自动生成的难点有助于我们构建合理有效的图像可控文本生成神经网络模型，从图像到文本的自动生成研究发展进程中我们可以看出，图像文本生成技术已经逐渐发展完善起来，新的图像文本生成技术也在不断的发现和探索中，而本文图像到文本的可控生成就是近年来，作为交叉领域新的挑战之一。

# 第3章 研究内容的理论模型

目前对于图像受控文本生成的研究，国内外的研究方法主要以深度学习模型为主，主要模型结构为Encoder-Decoder网络结构，常见的用于文本生成的模型结构结构有Squence-to-Squence、tensor-to-tensor，常见的用于受控文本生成的模型结构为GAN网络模型，常见的用于图像特征提取的模型结构为CNN网络模型。

## 3.1 相关基础理论简述

### 3.1.1 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。

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。

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）

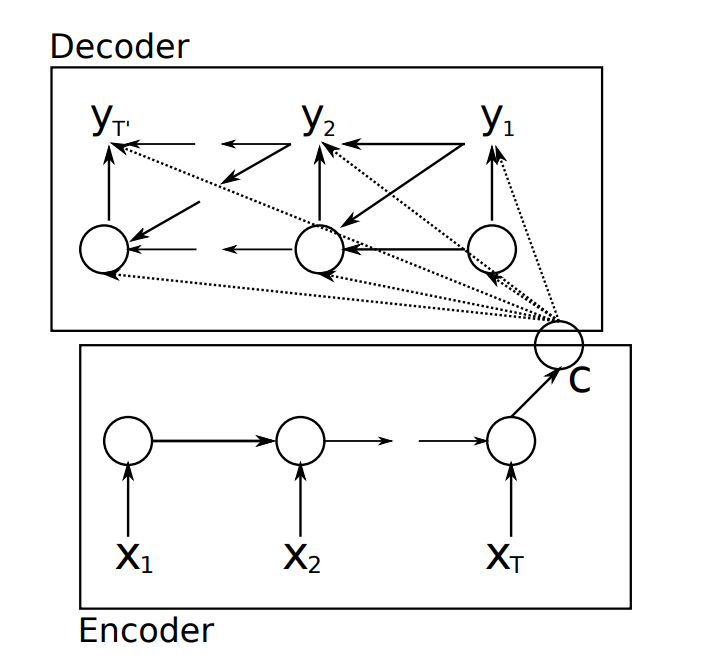


图 4 Encoder - Decoder递归神经网络模型

### 3.1.2 LSTM网络结构

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通过独特的设计来避免长距离依赖问题。记住长期的信息是LSTM网络的默认行为，而非需要付出很大代价才能获得的能力。传统的RNN 循环神经网络都只具有一种神经网络模块的链式形式。在标准的RNN中，这个重复的模块只有一个非常简单的结构，通常为一个单一神经网络层，例如一个tanh层。

LSTM 同样是这样的结构，但是重复的模块不同是，其拥有一个特殊的结构。此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。

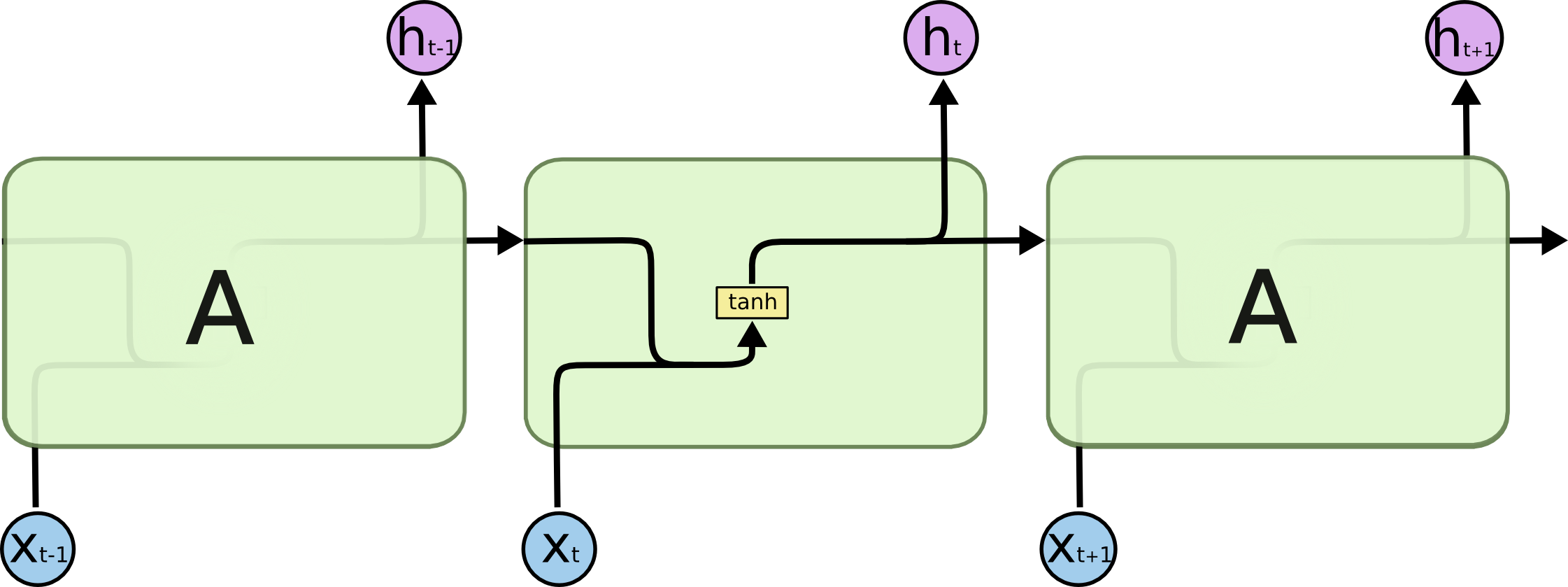


图 5 RNN-Cell 结构



图 6 LSTM-Cell 结构

### 3.1.3 Soft-Attention网络结构

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，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，另一个就是上一时刻t-1的信息。

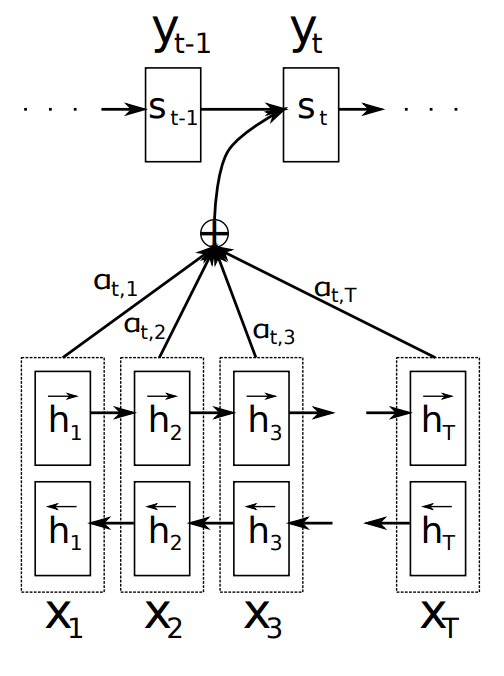


图 7 Example-of-Attention

## 3.2 本文理论模型论述

### 3.2.1 Squence-to-Squence With Attention网络结构

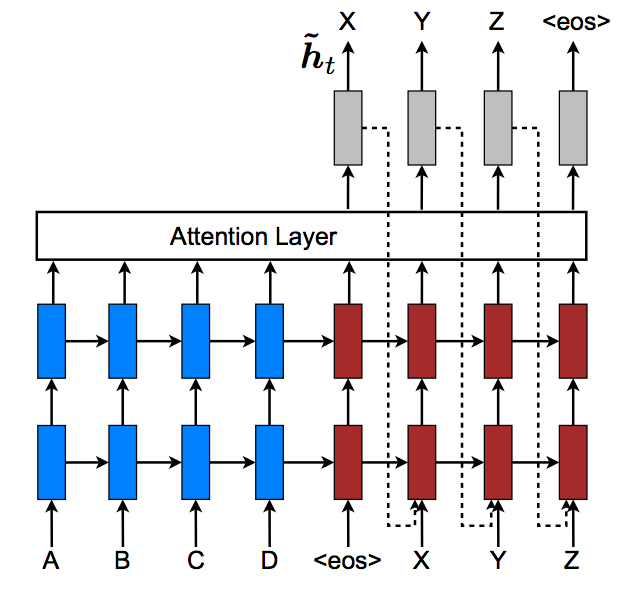


图 8 带有Attention的网络结构

计算过程：见图7参数

1. 问题：

有三个输入时间步骤：该模型需要预测一个时间步骤：

1. 编码：

在编码器 - 解码器模型中，输入将被编码为单个固定长度矢量。这是最后一个时间步的编码器模型的输出。



注意模型需要在每个输入时间步长访问编码器的输出。本文将这些称为每个时间步的“ 注释 ”。在这种情况下：



1. 对齐：

解码器一次输出一个值，在最终输出当前输出时间步长的预测之前，该值被传递到可能更多的层。

对齐模型得分每个编码输入与解码器的当前输出的匹配程度。

分数的计算需要来自前一输出时间步长的解码器的输出，例如。为解码器的第一个输出进行评分时，这将为0。

使用函数执行评分。我们可以为第一个输出时间步骤评分每个注释，如下所示：



我们对这些分数使用两个下标，例如E11，其中第一个“1”表示输出时间步长，第二个“1”表示输入时间步长。

我们可以想象，如果我们有一个带有两个输出时间步长的序列到序列问题，那么稍后我们可以按如下方式对第二个时间步的注释进行评分（假设我们已经计算了S1）：



函数在本文中称为对齐模型，并实现为前馈神经网络。

这是传统的单层网络，其中每个输入（和H1，H2和H3）被加权，使用双曲正切（tanh）传递函数并且输出也被加权。

1. 加权

接下来，使用[softmax函数](https://en.wikipedia.org/wiki/Softmax_function)对准分数进行归一化。分数的归一化允许它们被视为概率，指示每个编码的输入时间步长（注释）与当前输出时间步长相关的可能性。

这些标准化分数称为注释权重。例如，我们可以计算softmax注释权重（a）给定计算的对齐分数（e）如下：



如果我们有两个输出时间步长，则第二个输出时间步长的注释权重将按如下方式计算：



1. 上下文矢量

接下来，将每个注释（h）乘以注释权重（a）以产生新的有人值守的上下文向量，从该向量上下文向量可以解码当前输出时间步长。

为简单起见，我们只有一个输出时间步长，因此我们可以按如下方式计算单个元素上下文向量：



上下文向量是注释和标准化对齐分数的加权和。

如果我们有两个输出时间步长，则上下文向量将由两个元素[C1，C2]组成，计算如下：



1. 解码

解码器的输出在本文中称为隐藏状态。



在最终退出模型之前，这可以被馈送到附加层中作为时间步长的预测（Y1）。

### 3.2.2 GAN网络结构

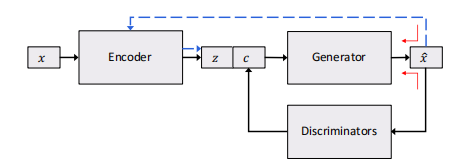


图 9 GAN网络结构

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：。

1. 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：

## 3.3 本章小结

在本章中，我们介绍了在图片受控文本上任务上，用到的深度学习网络模型，并以此为基础搭建了基础的一种融合注意力机制的多模态可控文本生成的框架，我们以CNN网络提取图片信息，并送入GAN网络以此生成可控的输出。

# 第4章 一种融合注意力机制的多模态可控文本生成实施方案

本章主要介绍本文所用的模型，以及实现的思路和过程

## 4.1 一种融合注意力机制的多模态可控文本生成分析过程

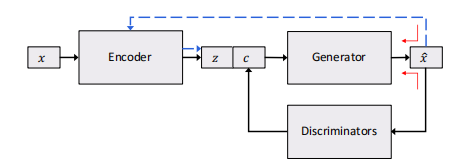
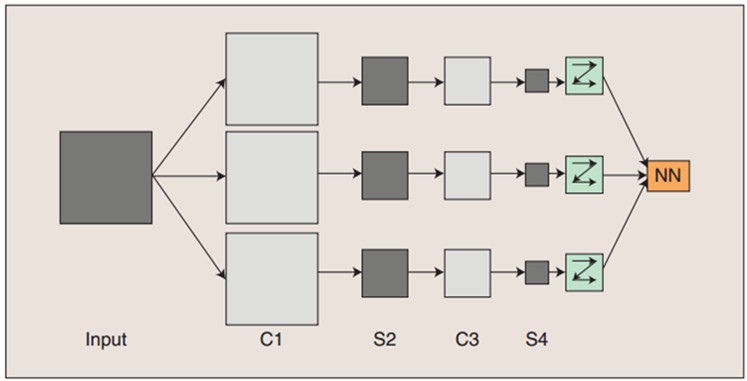


图 10 多模态可控文本生成网络结构

将图像神经网络得来的句子特征经过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。

本实验中，可控文本生成分为以及几个步骤：

1. 预训练阶段

在预训练阶段，Classifier部件使用yelp数据分类标签进行训练

生成器被训练为自动编码器

1. 训练阶段

在训练阶段Classifier部件是固定的，并且Encoder-Decoder部件经过训练以适合分类器，同时继续最小化自动编码损失

1. 测试阶段

在测试阶段使用DCNN卷积图片特征，送入训练好的GAN网络，指定生成的标签类型，生成指定主题

上述过程是图像受控文本生成的基本过程，接下来就是对模型识别准确率以及BLEU的分析判断，然后利用对比实验来验证模型的可用性程度。

## 4.2 WordEmbedder

首先我们要做的是将词转化为词向量，常见的作法是在神经网络中加入embedding层。

1. 构造词典

本课题使用带有正负情感标签的yelp数据构建生成的词典，词表大小为9359，

要求词在词表中必须要有唯一Id，本文构建词表的方法为统计训练语句词频，选取词表频率最高的9359个单词，也可选取全部词作为词表使用。

|  |  |  |
| --- | --- | --- |
| 单词 | 编号 | 大小 |
| **.**  **.**  **.**  Foothills  Cosmetics  **.**  **.**  **.**  Brisk  Shark  **.**  **.**  **.** | **.**  **.**  **.**  9228  9229  **.**  **.**  **.**  9319  9320  **.**  **.**  **.** | 9359 |

表 1 词典

2）One-hot编码

词典中的任何一个词，都有其对应的二进制编码，我们称为“One-hot”编码，这种方式可以表示成一个N 维的向量

字典或词典中的任意一个字或词，都可以用One-hot方式表示成一个N 维的向量。

|  |  |  |
| --- | --- | --- |
| 单词 | 编号 | 向量 |
| **.**  **.**  **.**  Foothills  Cosmetics  **.**  **.**  **.**  Brisk  Shark  **.**  **.**  **.** | **.**  **.**  **.**  9228  9229  **.**  **.**  **.**  9319  9320  **.**  **.**  **.** | **.**  **.**  **.**  [0,0,0,...,1,...,0,0,0]  [0,0,0,........1,0,0,0]  **.**  **.**  **.**  [0,0,0,...,...,0,1,0,0]  [0,0,0,...,...,0,0,1,0]  **.**  **.**  **.** |

表 2 One-hot编码

3）词嵌入

RNN的输入是文本序列，因此我们需要找到一种简单的方式来表达它。因为计算机更擅长数值处理，最直接的想法是将word映射到向量上来。Word Embedding描述的就是这样一个过程，它表示一个从文本序列到多维向量的映射关系，Word Embedding与One-hot相比，前者更擅长深度学习特征提取，样本利用率高，每个embedding的特征分量的训练会使用每一个样本，分布表示效率高，因此我们使用矩阵运算把高维、稀疏的向量表示成低维、稠密的向量。

|  |  |
| --- | --- |
| One-hot | 词嵌入 |
| **.**  **.**  **.**  [0,0,0,...,1,...,0,0,0]  [0,0,0,........1,0,0,0]  **.**  **.**  **.** | **.**  **.**  **.**  [-4.42496777e-01 2.49782610e+00...]  [1.42921007 0.42996076 -0.824281...]  **.**  **.**  **.** |

表 3 WordEmbedder

## 4.3 CNN卷积网络

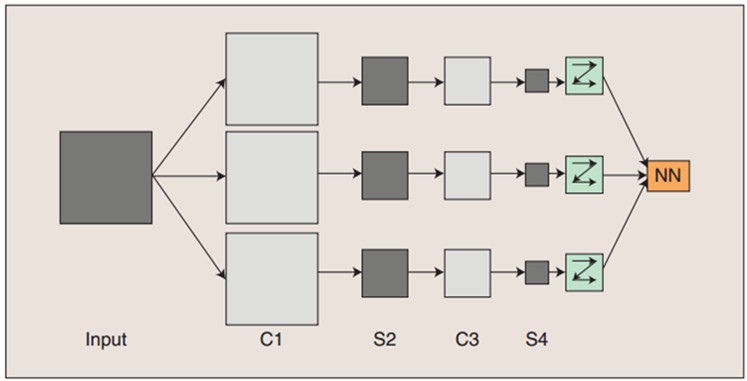


图 11 CNN网络结构

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。

## 4.4 UnidirectionalRNNEncoder

简单的循环神经网络它由输入层、一个隐藏层和一个输出层组成。X是一个向量，它表示输入层的值；S是一个向量，它表示隐藏层的值；U是输入层到隐藏层的权重矩阵，O也是一个向量，它表示输出层的值；V是隐藏层到输出层的权重矩阵。循环神经网络的隐藏层的值S不仅取决于当前这次的输入X，还取决于上一次隐藏层的值S。权重矩阵 W就是隐藏层上一次的值作为这一次的输入的权重。



图 12 单向RNN展开图

循环神经网络的计算公式如下：

 (1)



(1)是输出层的计算公式，输出层是一个全连接层，也就是它的每个节点都和隐藏层的每个节点相连。V是输出层的权重矩阵，g是激活函数。(2)是隐藏层的计算公式，它是循环层。U是输入x的权重矩阵，W是上一次的值作为这一次的输入的权重矩阵，f是激活函数。

如果我们反复将2式代入1式，我们将得到



可以看出输出受的影响，这也是循环神经网络能够往前看任意步的原因。

## 4.5 AttentionRNNDecoder

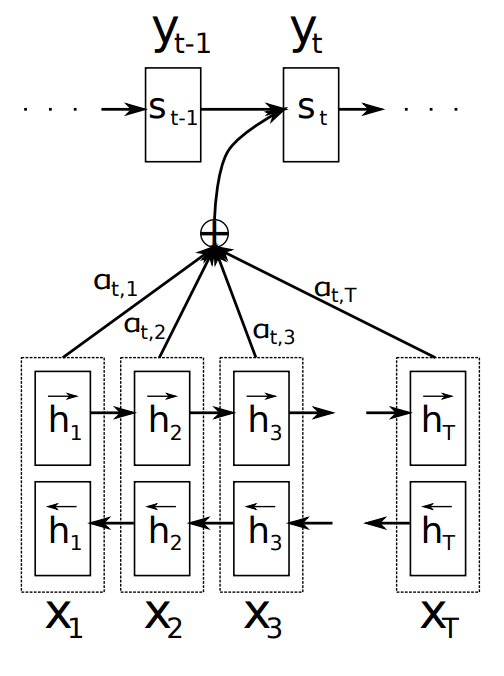


图 13 同图7

带有Attention机制的Decoder，我们倒着分析一下该计算过程详情见3.2.1：



是指Decoder在t时刻的状态输出，是指Decoder在t-1时刻的状态输出是t-1时刻的label（注意是label，不是我们输出的y），是一个RNN。



是指第j个输入在Encoder里的输出，是一个权重



事实上的计算相似于Softmax的计算，它得到的是条件概率，其主要含义为当前时间步t，输入j的关注权重



综合4.5和3.1.3，我们可以总结出Attention在Decoder上如何应用。

## 4.6 Gumbel-softmax

在文本生成过程中，我们输入输出的条件概率，但在神经网络中，的预测过于依赖于的预测，这样会导致如果预测错误，后续的预测也都将错误。而Gumbel-softmax在一定程度能够缓解这种错误。

基于softmax的常规采样：

对于首先我们利用softmax运算得到统一概率分布然后使用numpy.choice依照概率生成样本值。

基于Gumbel-softmax的采样：

在自编码网络中，假设隐变量（latent variables）服从标准正态分布。而现在，利用gumbel-softmax 技巧，我们可以将隐变量建模为服从离散的多项分步。



temperature是大于零的参数，它控制着 softmax 的soft 程度。温度越高，生成的分布越平滑；温度越低，生成的分布越接近离散的 one-hot 分布。训练中，可以通过逐渐降低温度，以逐步逼近真实的离散分布。

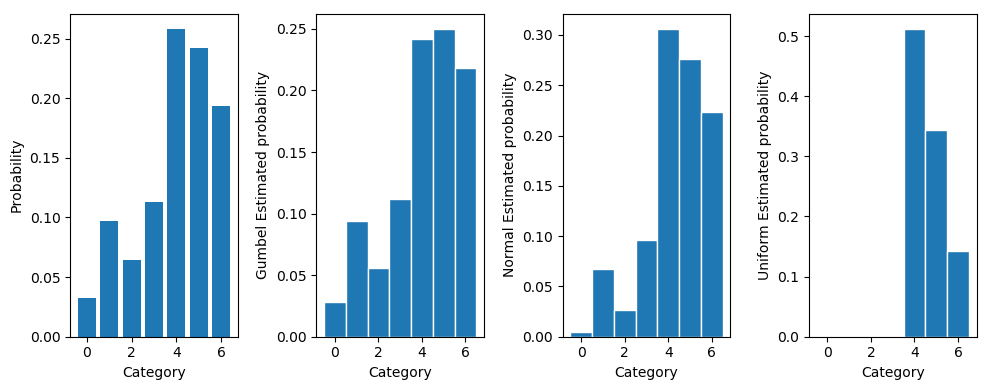


图 14 Gumbel-softmax拟合数据效果对比

可以明显看到Gumbel噪声的采样效果是最好的，正态分布其次，均匀分布最差。也就是说可以用Gumbel分布做Re-parameterization使得整个图计算可导，同时样本点最接近真实分布的样本。

# 第5章 一种融合注意力机制的多模态可控文本生成实验与分析

为了验证模型是否有效，我们基于上述理论模对一种融合注意力机制的多模态可控文本生成进行了一系列的实验。本章主要介绍实验的过程、评价指标，以及在不同参数下的多次对比实验结果分析。

为了达到对图像受控文本生成模型效果的客观评价，在实验中，我们在相同的实验环境条件下，用两种模型对测试语料分别进行测试和对比分析。通过对测试结果的分析比对，我们初步得出的结论表明基于GAN网络的深层神经网络更适合于可控文本生成。

## 5.1 实验总体设计

本文数据采用COCO2014图像数据集作为图像特征提取的训练语料，使用yelp情感数据集作为可控文本生成的训练语料，使用COCO2014验证数据集作验证集，使用yelp验证数据集作验证集，测试集可采用随机图像进行测试。本文主要使用DCNN以及GAN神经网络模型对任意图片实现情感可控文本生成。主要评测指标采用准确率，BLEU值，准确率主要为分类器的准确率，BLEU主要用以评判生成器生成效果。

在具体的实验过程中。我们主要在Gumbel-softmax的退火率选择，Batch-size的大小选择，优化器的选择，可控文本生成的模型选择，迭代轮数等方面选择最合适的模型以及参数。

在实验过程中发现，当轮数达到12轮之后，模型收敛，loss值不在均匀变化，BLEU值达到峰值，因此在一下实验中实验迭代轮数我们都去13，其中预训练轮数10，训练论数3，增加1轮显示对比效果。

实验轮数效果展示如表4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 指标  轮数 | BLEU | Loss\_g | Accu\_g | Accu\_d |
| Epoch 10 | 69.59 | 2.3102 | 0.3333 | 0.6667 |
| Epoch 11 | 69.59 | 2.4468 | 0.3333 | 0.6667 |
| Epoch 12 | 63.80 | 2.5444 | 0.6667 | 0.6667 |
| Epoch 13 | 54.86 | 2.8497 | 0.6667 | 0.6667 |

表 4 实验轮数效果展示

实验参数如表5

|  |  |  |
| --- | --- | --- |
| 实验参数 | 值 | 解释 |
| Max\_nepochs | 13 | 全部的训练轮数 |
| Pretrain\_neochs | 10 | 预训练轮数 |
| Lambda\_g | 0.1 | 分类器的loss权重 |
| Gamma\_decy | 0.5 | Gumbel-softmax的退火率 |
| Batch\_size | 1000 | 批大小 |
| Encoder | GRUcell | 编码器的神经网络模型 |
| Decoder | GRUcell | 解码器的神经网络模型 |
| Attention | Bahdanau Attention | 注意力机制 |
| Optimizer | AdamOptimizer | 优化器 |
| Softmax | Gumbel-softmax | 条件概率分布采样 |

表 5 实验参数表

## 5.2 实验数据以及实验配置

本实验主要采用COCO2014图像数据集，包括训练集和验证集，以及yelp情感数据集，包括训练集和验证集，详细介绍见表6

### 5.2.1 实验数据

|  |  |
| --- | --- |
| 数据集 | 说明 |
| COCO2014 | 这个数据集以scene understanding为目标，主要从复杂的日常场景中截取，图像中的目标是通过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。该数据集包含三类信息：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。 |
| Yelp | Yelp情绪数据集来自加利福尼亚大学欧文分校 (UCI)，主要包含447259条含有情感标签的训练语料。 |

表 6 实验采用的数据集

COCO数据集展示

|  |  |  |
| --- | --- | --- |
| 描述 | 图片路径 | 图片ID |
| a fire hydrant in the middle of a flower bed | /train/images/COCO\_train2014\_000000167467.jpg | 167467 |
| "a man sitting on a rock with his legs crossed, talking on his cell phone." | /train/images/COCO\_train2014\_000000366150.jpg | 366150 |

表 7 COCO数据集

Yelp数据集展示（0：负面情感，1：整面情感）

|  |  |
| --- | --- |
| 描述 | 情感标签 |
| i was sadly mistaken | 0 |
| minimal meat and a ton of shredded lettuce | 0 |
| this place is amazing ! | 1 |
| true authentic italian deli | 1 |

表 8 yelp数据集

### 5.2.2 实验配置

为保证实验的可再现性，请保持配置一致，显卡的性能会影响batch-size的大小，在资源允许的情况下，可以尽量让batch-size的值大，这样可能会带来意想之外的效果，本实验的batch-size为1200。

|  |  |
| --- | --- |
| 系统 | Ubuntu 16.0.4 |
| tensorflow-gpu | 1.12.0 |
| CUDA | 9.0 |
| CUDNN | 7.3.0 |
| Python | 2.7 |
| 显卡 | P100\*2 Memory=24G |

表 9 实验配置

## 5.3 实验评测方案及指标

|  |  |
| --- | --- |
| 评测指标 | 含义 |
| accu\_g | 生成语句的分类准确率 |
| loss\_g\_ae： | 自动编码器的损失 |
| accu\_d | 分类器的准确度 |
| BLEU | 文本相似度指标 |
| loss\_g | 联合损失= loss\_g\_ae + lambda\_g \* loss\_g\_clas |
| loss\_d | 分类器的损失 |
| loss\_g\_clas | 生成器分类损失 |
| accu\_g\_gdy | 生成器贪婪生成准确度 |

表 10 各评测指标

实验训练过程中，我们以在验证集上计算个指标来判断模型是否达到最佳效果，其中以BLEU指标尤为重要，它决定生成的语句是否流畅通顺，是否符合可控的要求，该评价指标由IBM在2002年提出[24]，常作为机器翻译系统评价指标。其实就是统计生成响应和真实响应中的n-gram词组在整个训练语料中出现次数。公式如下所示：





第一个公式用于计算n-gram短语词组在整个数据集中的准确度。h(k,r)表示每个n-gram词组在真实响应中出现的次数（因为对于每个n而言都会存在很多个n-gram词组，所以要有一个求和符号）。所以上式就是每个n-gram词组在真实和生成响应中出现次数的较小值求和除以其在生成响应中出现次数求和，表征了一种精确度度量。当然，我们需要考虑n的取值（一般取1-4），所以有了第二个公式，beta表示各个n-gram的权重（可以去均匀分布），也就是对1-4进行加权求和，而b(r,r^)表示长度惩罚因子，即我们不想让生成的答案长度太短，所以加一个惩罚因子来改善效果。

## 5.4 实验过程

### 5.4.1 训练模型生成图片描述

训练image-caption提供的网络模型，为可控文本生成提供测试集，该网络模型使用论文[18]提到的神经网络结构，使用DCNN卷积神经网络从图像中提取图像特征，并使用LSTM递归神经网络将这些特征解码为句子。结合soft-Attention注意机制以改善描述质量。该模型在[21]能够找到实现。

### 5.4.2 准备可控文本生成训练语料

1）下载yelp语料并解压在数据目录下

2）使用脚本生成yelp的训练集和验证集的字典，以供在最终预测时使用，具体操作可见4.2章节

### 5.4.3 构建embedding层



我们传入两个参数分别为字典的大小，他定义了词向量的纬度和大小，超参数hparams，在深度学习之前初始赋值embedding神经网络层，以提高学习的性能和效果。

### 5.4.4 构建自动编码器AE

5.4.4.1 构建Encoder



1. 我们定义一个使用单项循环神经网络的编码器，神经单元使用GRUcell。
2. 将经过embedding层的输入去掉代表句子开始的标记为BOS，得到一组词向量。
3. 将词向量投入Encoder神经网络，得到代表句子的上下文矢量final\_state。



1. MLPTransformConnector定义一个全连接层，用以将输入的数据输出为指定的结构和大小。
2. 将Encoder的输出以及代表标签的向量labels分别经过label\_connector。
3. 将向量c,z进行拼接得到最终的Encoder层的输出。

5.4.4.2 构建Decoder



1. 我们定义一个带有Attention机制的Decoder神经网络，同样使用GRUcell作为神经单元。
2. memory为解码是需要的上下文矢量。
3. Attention为Luong Attention（加法Attention），不同于 Bahdanau Attention（乘法Attention），后者在计算分配权重采用的是当前时刻Decoder网络的隐状态，前者使用前一时刻Decoder网络的隐状态，后者看起来更直观合理，但两者在实验中性能差别不大。本文采用计算更简单的Bahdanau Attention。

### 5.4.5 构建Gumbel-sofmax



1. 定义开始和结束的标志
2. 调用Gumbel-softmax模型，得到条件概率分布
3. 使用贪婪搜索，匹配一个最有可能的结果

### 5.4.6 构建分类器以及优化器



1）调用分类器和优化器

### 5.4.7 对比试验

在该实验中我们使用两种可控文本生成的方法对比分析，一种我们采用论文[23]提到的方法，该方法的简单实例我们可以在[22]上找到（以下统称实验1），该方法没有融合Attention机制，一种我们使用融合Attention机制的GAN网络[25]（以下我们统称实验2），我们在相同的实验环境下，我们得到以下几组数据，说明融合Attention机制的情况下，可控文本生成的效果会更好。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| BLEU | 测试语句 | 情感标签 | 生成语句 | 测试方法 |
| 63.080 | a red double decker bus driving down a street | 负面 | a real garden style hands down the street ! | 实验1 |
| 63.080 | a man riding a wave on top of a surfboard | 负面 | very professional and a lot of a city ! | 实验1 |
| 63.080 | a group of people sitting around a table | 负面 | a lot of people around a table . | 实验1 |
| 63.080 | a red double decker bus driving down a street | 正面 | a total change you have a rip down hill . | 实验1 |
| 63.080 | a man riding a wave on top of a surfboard | 正面 | what a complete of a complete on a top of . | 实验1 |
| 63.080 | a group of people sitting around a table | 正面 | a few tables of us over a table around . | 实验1 |
| 70.160 | a red double decker bus driving down a street | 负面 | a red double fountains bus driving down a street | 实验2 |
| 70.160 | a man riding a wave on top of a surfboard | 负面 |  | 实验2 |
| 70.160 | a group of people sitting around a table | 负面 |  | 实验2 |
| 70.160 | a red double decker bus driving down a street | 正面 | a red double fountains bus driving down a street | 实验2 |
| 70.160 | a man riding a wave on top of a surfboard | 正面 | a man riding a dillard on top of a | 实验2 |
| 70.160 | a group of people sitting around a table | 正面 | a group of people sitting around a table | 实验2 |

表 11 模型效果对比

如图所示，在融合Attention之后BLEU指标变高，语言也更加通顺。因此我们选用实验2所用的网络模型做后续对比试验。

## 5.5 实验结果分析

在同一实验配置（见表9），同一实验数据（见表6），在同一模型（实验2）的情况下，我们将分别分析batch\_size，gamma\_decay，lambda\_g对实验结果的影响。

### 5.5.1 batch-size的影响

对比实验数据我们选用迭代轮数 = 13，gamma\_decay = 0.5，learning\_rate = 0.0005，lambda\_g = 0.1，评判标准主要依据BLEU等具有特殊意义的指标值。其中迭代轮数的影响见表4。

|  |  |  |
| --- | --- | --- |
| batch\_size | BLEU | accu\_g |
| 64 | 61.005 | 0.5044 |
| 124 | 60.129 | 0.5102 |
| 256 | 63.230 | 0.5100 |
| 1024 | memeory error | memeory error |
| 1000 | 63.330 | 0.5520 |

表 12 batch\_size对实验的影响

尽管batch\_size的值在小范围的变化并不会影响整个实验的最终值，但随着batch\_size值的增大，BLEU和accu\_g的值仍旧呈现上升趋势，直到batch\_size达到1024的时候，机器出现内存错误，但介于双GPU的服务器，其中一个GPU不能供实验使用，我们应该选择尽可能高的batch\_size = 1000作为最终模型batch\_size的值。

### 5.5.2 gamma\_decay的影响

gamma\_decay作为条件概率分布拟合指标，很大程度上决定了BLEU指标的高低，对比实验数据我们选用迭代轮数 = 13，batch\_size = 1000，learning\_rate = 0.0005，lambda\_g = 0.1。

|  |  |  |
| --- | --- | --- |
| gamma\_decay | BLEU | accu\_g |
| 1.0 | 63.100 | 0.6563 |
| 0.5 | 67.514 | 0.5102 |
| 0.1 | 68.903 | 0.5534 |

表 13 gamma\_decay对实验的影响

从表中我们可以看出，随着gamma\_decay的值越来越低，BLEU的值越来越高，但反观accu\_g的值，并没有随之变高，可见gamma\_decay主要影响了BLEU的值。那我们在实验中是否选用较低的gamma\_decay，我们从实验3分析。

我们定义一个多项分布，其真实密度函数如图15

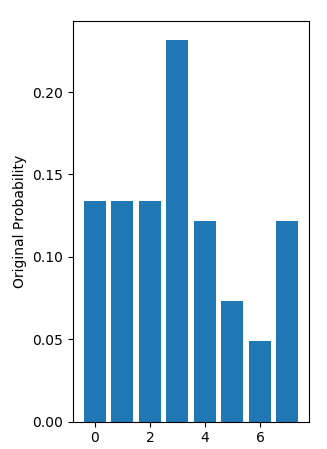


图 15 真实概率密度分布

我们基于Gumbel-softmax做采样对比，图16左侧我们设置temperature = 0.1，右侧我们设置temperature = 0.5，左右两侧展示数据经过softmax后得到的概率分布。

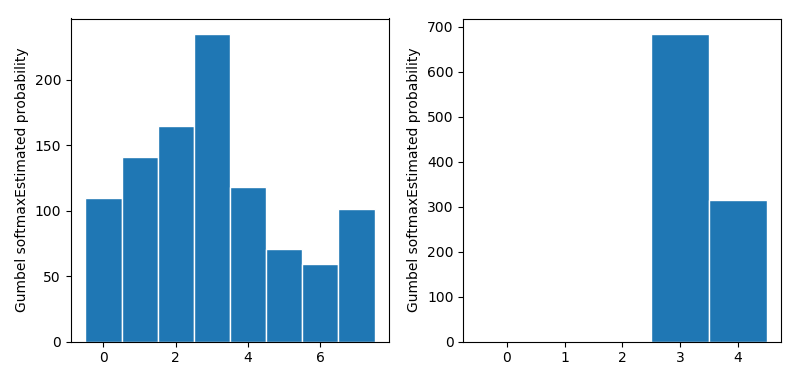


图 16 temperature对数据分布影响

从图中我们可以直观的看出temperature值越小对数据的拟合度越高，因此在本实验我们使用temperature = 0.1进行试验。

### 5.5.3 lambda\_g的影响

lambda\_g作为分类器loss权重的重要指标，很大程度上决定了accu\_g值的大小，对比试验数据我们选用选用迭代轮数 = 13，batch\_size = 1000，learning\_rate = 0.0005，gamma\_decay = 0.1。

|  |  |  |
| --- | --- | --- |
| lambda\_g | BLEU | accu\_g |
| 1.0 | 64.609 | 0.5800 |
| 0.5 | 68.514 | 0.6249 |
| 0.1 | 70.016 | 0.6562 |

图 17 lambda\_g对实验的影响

从表中我们可以看出，随着lambda\_g的不断变化，accu\_g的并没有明显的大幅度变化，但是BLEU的值还是随着值变小而变大，这与实际预测有一定的出入，我们最终选用lambda\_g = 0.1作为最终值，我们预测出现这种情况的原因可能是生成器与分类器在对抗学习中耦合度没有达到很高，如果实验允许，我们可以加大训练的轮数，同时加大lambda\_g的值以观察它对实验的真实影响。

## 5.6 本章小结

在本章我们依次说明构建可控文本生成神经网络的主要过程，以及实验数据来源，实验配置，对比试验，以及各参数对模型的影响，最终我们选用迭代轮数 = 13，batch\_size = 1000，learning\_rate = 0.0005，gamma\_decay = 0.1，lambda\_g = 0.1，模型选用GAN神经网络，Attention选用Bahdanau Attention，优化器选用AdamOptimizer，softmax选用Gumbel-softmax，最终在测试集上指标值如表14

|  |  |
| --- | --- |
| 评测指标 | 值 |
| accu\_g | 0.667 |
| loss\_g\_ae： | 1.6217 |
| accu\_d | 0.3333 |
| BLEU | 83.6400 |
| loss\_g | 1.6347 |
| loss\_d | 5.2408 |
| loss\_g\_clas | 1.3071 |
| accu\_g\_gdy | 0.6667 |

表 14 实验结果

# 第6章 总结与展望

## 6.1 总结

本文详细介绍了一种多模态可控文本生成的实验过程，我们首先对国内外做image-caption任务的发展和模型进行了分析和总结，选用了适合自己的模型为本次实验提供测试，这为后面可控文本生成提供了优秀的条件，最终在不断地摸索中我们构造出最适合的模型来完成本次实验。

本文主要做了两次模型对比试验，以验证融合attention机制对于图像受控文本生成的关键影响，最终attention机制也充分证明它的重要性，这使得我们的模型结构更加清晰，紧接着我们对模型的一系列的参数分别做了对比试验，依次选出了适合的模型参数，以优化模型的效果。

本次实验虽然得出了最终结果，但由于各方面原因，我们并没有对优化器，attention种类，进行更多的说明和选择，事实上随着深度学习的不断发展，self-attention无论在文本分类任务还是文本生成任务上的优越性也在逐步提高，因此self-attention的使用无疑是必须的，其次对于文本生成任务至关重要的softmax的改进，正如前文所说，是文本生成任务的瓶颈所在，gumbel-softmax也需要进一步的改进，因此实验结束并代表探索的结束，后续将着重针对attention以及softmax进行改进完善模型。

## 6.2 展望

人工智能的浪头已经越来越高，而自然语言处理作为重要的一个环节，不断地创新和探索显得更加重要，虽然本文只是研究多模态可控文本生成的框架，但更多的是发现新的文本生成的办法以及重现最优秀的文本生成技术。

多模态可控文本生成的完善和改进还有很多地方，希望在未来的研究中，能够提供更加完整的模型框架，其次由于时间问题，我们并没用采用中文进行试验，而基于中文的特殊性，中文多模态可控文本生成具有一定的难度，希望能探索出适合中文的多模态可控文本生成框架，这也是我的下一个挑战。

# 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. 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).
21. <https://github.com/DeepRNN/image_captioning>
22. <https://github.com/shentianxiao/language-style-transfer>
23. Shen T , Lei T , Barzilay R , et al. Style Transfer from Non-Parallel Text by Cross-Alignment[J]. 2017.
24. Papineni K . BLEU : a Method for Automatic Evaluation of Machine Translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002. 2002.
25. Hu Z, Yang Z, Liang X, et al. Toward Controlled Generation of Text[J]. 2017.

# 致 谢

在这美好的大学四年中，我收获了很多。在和室友的欢声笑语中大学四年转瞬即逝，毕业在即，大家为了自己的以后都在做最后的努力，希望能为自己的大学时光画上圆满的句号，回顾大学时光，感谢我的专业课老师，因为是他们让我体会到了我们这行的优秀之处，更感谢刘老师，带我做了NLP这块儿，让我对世界有了新的看法。

感谢在这段大学的旅程中，一直陪伴我的同学们，使他们丰富了我的大学生活，从他们身上，我获得了很多或深或浅的人生道理；更感谢我的老师们，是他们用知识培育了我，用关爱温暖了我，用身体力行感染着我，是他们让我成为了更好的我自己；最后，感谢我的母校，无论以后我能走的多远，我会永远记得“博学弘德，自强不息”。在以后的工作中，我也会以此严格要求自己。最后的最后，祝愿我的老师们，身体健康，事业蒸蒸日上！