

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

近年来，人工智能和自然语言处理在学术界和工业界不断发展壮大，已经成为时下最热门领域之一，本文基于该领域的传统文本生成任务实现了一种融合注意力机制的多模态可控文本生成框架，本实验以图像生成文本为例实现跨模态文本生成，并要求在文本端生成可控的文本描述，例如：我们给定任意一张图片，同时给定一个标签，我们希望机器能通过神经网络的学习，生成一句以标签为前提的符合图片描述的自然语言。本实验结合了图像生成任务和文本生成任务，因此在难点问题上也继承了两者的核心难点：

（1）要求计算机能够准确切割出图片中的主体，还要能够理解主体之间的相互关；

（2）在于我们希望在文本生成阶段，文本能基于句子的生成，而不是基于词的生成；

（3）在于如何有效的解决跨模态问题；

（4）在于我们如何构建模型来实现可控；

通过对实验难点分析，以及现有深度学习算法的探索，我们提出了DCNN+GAN（深度卷积网络+对抗网络）这样的模型结构解决上述难点，同时我们在Decoder端加入Attention机制和Gumbel-SoftMax希望提高文本生成的准确性和流畅性。在本次实验中我们充分证明了这种模型结构和算法的可行性，这也为我们下一步转向基于中文的多模态可控文本生成提供模型基础。

关键词：多模态；注意力机制；可控文本生成

# **Abstract**

In recent years, artificial intelligence and natural language processing have continued to grow and develop in academia and industry. It has become one of the hottest areas in the world. This paper implements a multi-modality that integrates attention mechanisms based on traditional text generation tasks in this field. Control text generation framework, this experiment takes image generation text as an example to realize cross-modal text generation, and requires to generate a controllable text description on the text side. For example: we give any picture and give a label at the same time, we hope The machine can learn through the neural network to generate a natural language that conforms to the picture description based on the label. This experiment combines image generation tasks and text generation tasks, so it also inherits the core difficulties of the two on difficult issues:

1. Requiring the computer to accurately cut out the subjects in the picture, but also to understand the interrelationship between the subjects;
2. It is in the hope that during the text generation phase, text can be based on the generation of sentences rather than word-based generation;
3. It is how to effectively solve the cross-modal problem;
4. It is how we build the model to achieve controllable;

Through the analysis of experimental difficulties and the exploration of existing deep learning algorithms, we propose a model structure such as DCNN+GAN (deep convolution network + confrontation network) to solve the above difficulties. At the same time, we add Attention mechanism and Gumbel-SoftMax on the Decoder side. I hope to improve the accuracy and fluency of text generation. In this experiment, we have fully proved the feasibility of this model structure and algorithm, which also provides a model basis for us to turn to Chinese-based multimodal controllable text generation.

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

# **目 录**

[中文摘要 I](#_Toc24195)

[Abstract II](#_Toc13773)

[目 录 III](#_Toc4145)

[图表目录 1](#_Toc28715)

[第1章 绪论 3](#_Toc4858)

[1.1 选题背景及研究意义 3](#_Toc29125)

[1.2 研究现状 4](#_Toc4458)

[1.3 主要任务 4](#_Toc2393)

[1.4 论文的组织结构 5](#_Toc15395)

[第2章 相关研究 6](#_Toc13939)

[2.1 国内外研究 6](#_Toc10280)

[2.2 实验模型难点 9](#_Toc9996)

[2.3 本章小结 9](#_Toc3211)

[第3章 理论模型 10](#_Toc23294)

[3.1 相关基础理论简述 10](#_Toc17026)

[3.1.1 Encoder-Decoder网络结构 10](#_Toc25672)

[3.1.2 LSTM网络结构 11](#_Toc17202)

[3.1.3 Soft-Attention网络结构 12](#_Toc12850)

[3.2 本文理论模型论述 13](#_Toc30889)

[3.2.1 Sequence-to-Sequence With Attention网络结构 14](#_Toc1744)

[3.2.2 GAN网络结构 16](#_Toc9188)

[3.3 本章小结 17](#_Toc29055)

[第4章 实施方案 18](#_Toc26582)

[4.1 实验分析过程 18](#_Toc15943)

[4.2 WordEmbedder 19](#_Toc18886)

[4.3 CNN卷积网络 21](#_Toc2373)

[4.4 UnidirectionalRNNEncoder 22](#_Toc14721)

[4.5 AttentionRNNDecoder 23](#_Toc24213)

[4.6 Gumbel-softmax 24](#_Toc2096)

[4.7 本章小结 25](#_Toc21071)

[第5章 实验与分析 26](#_Toc10964)

[5.1 实验总体设计 26](#_Toc6540)

[5.2 实验数据以及实验配置 28](#_Toc19898)

[5.2.1 实验数据 28](#_Toc13469)

[5.2.2 实验配置 29](#_Toc10335)

[5.3 实验评测方案及指标 29](#_Toc31692)

[5.4 实验过程 32](#_Toc15770)

[5.4.1 训练模型生成图片描述 32](#_Toc11110)

[5.4.2 准备可控文本生成训练语料 32](#_Toc21005)

[5.4.3 构建embedding层 32](#_Toc12515)

[5.4.4 构建自动编码器AE 32](#_Toc29485)

[5.4.5 构建Gumbel-sofmax 34](#_Toc19037)

[5.4.6 构建分类器以及优化器 34](#_Toc31575)

[5.4.7 对比试验 34](#_Toc21177)

[5.5 实验结果分析 37](#_Toc9171)

[5.5.1 Batch-size的影响 37](#_Toc10292)

[5.5.2 Gamma-decay的影响 37](#_Toc31671)

[5.5.3 Lambda-g的影响 39](#_Toc18476)

[5.6 实际场景运用 40](#_Toc4864)

[5.6.1 文本生成任务之英中文本翻译 40](#_Toc21086)

[5.6.2 中文情感对话子任务 44](#_Toc25440)

[5.7 本章小结 46](#_Toc30455)

[第6章 总结与展望 47](#_Toc3216)

[6.1 总结 47](#_Toc26668)

[6.2 展望 48](#_Toc19672)

[参考文献 49](#_Toc22875)

[致 谢 52](#_Toc27309)

## 图表目录

[图 2 - 1 一种典型的流水线模型 7](#_Toc25227)

[图 2 - 2 多模态 m-RNN 模型 8](#_Toc11315)

[图 2 - 3 视觉“注意”引导的图像标题生成过程 8](#_Toc16919)

[图 3 - 1 Encoder - Decoder递归神经网络模型 11](#_Toc1415)

[图 3 - 2 RNN-Cell 结构 12](#_Toc21200)

[图 3 - 3 LSTM-Cell 结构 12](#_Toc28506)

[图 3 - 4 Attention示例图 13](#_Toc27388)

[图 3 - 5 带有Attention的网络结构 14](#_Toc10499)

[图 3 - 6 GAN网络结构 16](#_Toc11802)

[图 4 - 1 多模态可控文本生成网络结构 18](#_Toc11136)

[表 1 词典内容展示 19](#_Toc28353)

[表 2 One-hot编码 20](#_Toc3668)

[表 3 WordEmbedder表示方式 20](#_Toc24711)

[图 4 - 2 CNN网络结构 21](#_Toc20570)

[图 4 - 3 单向RNN展开图 22](#_Toc8722)

[图 4 - 4 Attention作用过程 23](#_Toc3078)

[图 4 - 5 Gumbel-softmax拟合数据效果对比 24](#_Toc10434)

[表 4 实验轮数对比效果展示 27](#_Toc6903)

[表 5 实验参数表 27](#_Toc8850)

[表 6 实验采用的数据集 28](#_Toc2752)

[表 7 COCO数据集展示 28](#_Toc18471)

[表 8 Yelp数据集展示 28](#_Toc5458)

[表 9 实验配置 29](#_Toc10272)

[表 10 各评测指标 29](#_Toc8566)

[图 5 - 1 1-gram匹配示意图 30](#_Toc32295)

[图 5 - 2 2-gram匹配示意图 30](#_Toc13352)

[表 11 模型效果测试对比 35](#_Toc31577)

[图 5 - 3 无Attention训练过程图 36](#_Toc16863)

[图 5 - 4 Attention训练过程图 36](#_Toc21230)

[表 12 Batch-size对实验的影响 37](#_Toc1799)

[表 13 Gamma-decay对实验的影响 38](#_Toc13661)

[图 5 - 5 真实概率密度分布 38](#_Toc2346)

[图 5 - 6 Temperature对数据分布影响 39](#_Toc30020)

[表 14 Lambda-g对实验的影响 39](#_Toc11944)

[图 5 - 7 AI Challenger比赛官网 40](#_Toc11478)

[图 5 - 8 训练集 41](#_Toc2285)

[图 5 - 9 测试集、验证集英文数据集 41](#_Toc12209)

[图 5 - 10 验证集中文数据集 41](#_Toc3454)

[图 5 - 11 上下文数据集 42](#_Toc13367)

[图 5 - 12 Transform模型 42](#_Toc7262)

[图 5 - 13 决赛邀请函 43](#_Toc14565)

[图 5 - 14 CECG任务官网 44](#_Toc26255)

[表 16 数据说明 45](#_Toc3394)

[图 5 - 15 训练数据集 45](#_Toc7621)

[表 15 最终实验结果 46](#_Toc25430)

# 第1章 绪论

随着对机器学习和深度学习的不断发展，人们从简单的自然语言标注任务，文本生成任务以及后来结合计算机视觉的图像文本生成任务。随着人们需求不断增高，人们对计算机的要求也在不断增高，人们将不再简单要求计算机能够准确完成任务，同时还希望计算机能够特定条件下的进行特殊的智能操作，这也是人们对神经网络可控的需求，本文将延续这种需求，力求构建一种融合注意力机制的多模态可控文本生成框架。

在当前多模态可控文本生成资源稀少的情况下，我们面临着图片主体提取不准确、文本生成语言不流畅、多模态模型融合困难、可控如何实现等诸多问题，我们希望通过分析现有模型和算法，以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；这些都将作为文本生成的前提或约束，模型的核心目的在于实现约束下自然语言文本。

本章节希望通过对背景意义、研究现状的分析总结能够明确本实验的组要任务。

## 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。

自动文本生成技术我们了解不多，但实际在诸多场景中都已经得到了应用，小米旗下的小爱同学，苹果手机的智能语音助手等都是智能问答与对话系统的典型代表，除此之外我们使用的有道翻译，百度翻译等均是文本生成机器翻译任务的经典运用，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。自动文生成技术已经越来越多的融合在我们的生活中，与我们息息相关。

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。

受控生成图像的描述，作为标题生成模型不仅必须足够强大，准确的分割出图片中的主体，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，它将作为本文的难点之一着重解决。图像到文本的过程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

我们定义图像受控文本生成技术是指给定任意一张图片，能够生成描述该图片主题的自然语言文本。例如微信朋友圈，微博发文等。根据生成自然语言文本的长度以及程度的不同，Image Caption任务又分为图像标题自动生成和图像说明自动生成。标题自动生成要求根据应用场景的不同，突出图像的主题，例如，在新闻媒体播报的新闻就需要标注能说明图片内容的主题，而且要求在语言形式上能吸引读者；

说明自动生成要求讲明图像的主要内容，要求内容具体，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，并使用自然语言表述出来；但对于计算机而言，想要按照一定逻辑表述清一件事，它往往需要运用包括自然语言处理在内的多领域知识。图像文本生成作为一项交叉领域任务，集合了各领域的知识人才。在2010年，图像文本生成开始活跃在TACL、ACL和 EMNLP等知名会议和期刊；在2013年，包括IEEE TPAMI（模式识别与人工智能领域顶级国际期刊）、IJCV（计算机视觉领域顶级国际期刊）在内的期刊也开始收录相关内容，在2015年，在 CVPR（计算机视觉领域的知名国际会议）发表了[1-3]等10篇相关论文，同时在ICML（机器学习领域知名国际会议）中也有2篇相关论文发表。在不断的发展中，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[4-6]。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

文本可控生成的定义是指计算机通过深度学习后，能够模仿人类的思维（构词，感情等）生成符合语言逻辑的文本。我们希望对包括文本风格、文本时态、文本内容等特征实现可控，并结合传统的文本生成任务和图像自动生成任务，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，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COCO数据上验证其有效性
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，主要介绍了课题的来源，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，总结分析成果和不足，并对以后的多模态可控文本生成任务进行建议。

# 相关研究

本章节主要分析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，分析其瓶颈和难点，从中帮助构建图像到本文的可控文本生成的框架。

## 2.1 国内外研究

同传统文本生成任务，图像到文本的自动生成过程同样需要三个阶段[7]，不同的是图像文本生成需要充分理解图像主体以及主体之间的关系：

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。内容抽取的是否完整取决于计算机视觉算法是否强大，模式识别是否准确，这也是当代研究者不断努力的方向。

在内容生成方面，我们希望计算机能够特定的主题场景生成相应的文本描述，而且要求意义表述连贯。内容生成的是否流畅以及符合标准这取决于自然语言处理技术和计算机视觉技术。

最后，在句子实现部分，选取条件概率分布的结果，依次生成描述。

早期的文本生成模型都遵循以上三个过程。例如,在 Yao等人[8]的工作中，图像被细分并标记为对象、组成部分，以及场景，我们选择合适的模板匹配识别到的结果，这类似于一个填槽过程，当槽位填充完毕，我们根据模板生成结果。Feng与Lapata[9, 10]使用概率图模型同时建模文本信息和图像信息，并从新闻图像所在的文本报告中选择适当的关键词作为体现图像内容的关键词，语言模型将所选内容关键字和必要的功能词汇链接到基本上符合语法的图像标题中。还有一些工作[11-15]能够识别出图像中物体，并找到物体获得它的上下位置关系，然后依靠概率图模型和语言模型来选择适当的描述顺序，将这些对象概念和介词短语块连接成一个完整的句子。Hodosh 等人[16]则利用核函数的典型关联分析（Kernel Canonical Correlation Analysis， KCCA）去寻找文本与图像之间的关联，并根据图像信息对候选句子进行排序，以获得最佳描述句子。值得注意是，Hodosh等人的工作[16]和Feng与Lapata[10, 11]的工作并不依赖于现有的物体识别技术。

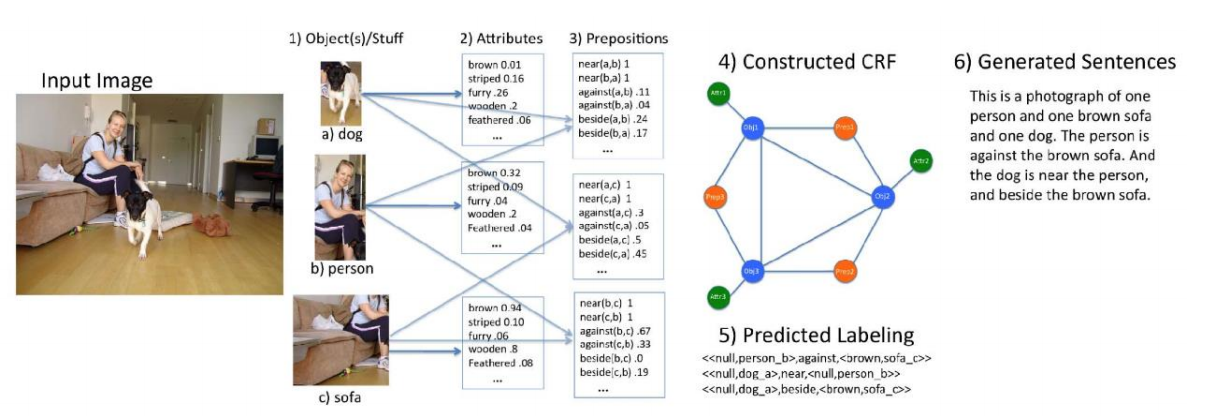


图 2 - 1 一种典型的流水线模型

随着机器学习和深度学习的不断创新和改进，文本语义标注、文本分类、命名实体识别等分类任务得到了快速发展；于此同时，文本生成、图像文本生成、智能问答等生成任务也在不断提高。于是乎，借助语义标注，知识图谱等作为知识特征网络的文本生成模型也运用而生。我们以图像生成为例，我们在图像端采用多层深度卷积神经网络（Deep Convolution Neural Network，DCNN）卷积图像的特征，在文本端采用循环神经网络（Recurrent Neural Network，RNN）或自注意力机制生成自然语言的描述[17]。传统图像语义标注工作主要致力于准确切割主体以及抽取主体之间的关系，而对一些诸如表情，手势等抽象概念的关注甚少。在Socher等人[18]提出的利用递归神经网络生成句子，并利用句法解析树标记表情，手势等抽象概念，进而联合图像端与文本端，这使得图像文本生成变得更清晰准确。为了融合图像与文本这两个多模态数据Chen与Zitnick[19]使用RNN网络将文本信息与图像信息相融合，利用图像信息作为Memory（记忆模块）指导文本句子的生成，同时构建一个全连接层用以重构图像信息，从而实现了图像到文本、文本到图像的双方向表示。在Mao等人[20]的改进下，使用m-RNN网络替代之前的RNN网络将图像端信息和文本端信息融合，并将图像信息融入到了自然语言句子生成的序列过程中，这使得图像文本生成的效果进一步得到了提高。同样的算法思想也在Donahue 等人[21]的论文中提到。但我们仔细观察m-RNN网络的句子生成过程，我们发现由于图像端约束的缺乏，使得文本段的生成不在考虑图像端，例如：当生成单词“jungle”的时候，图像标注的信息并没有在文本端进行使用。

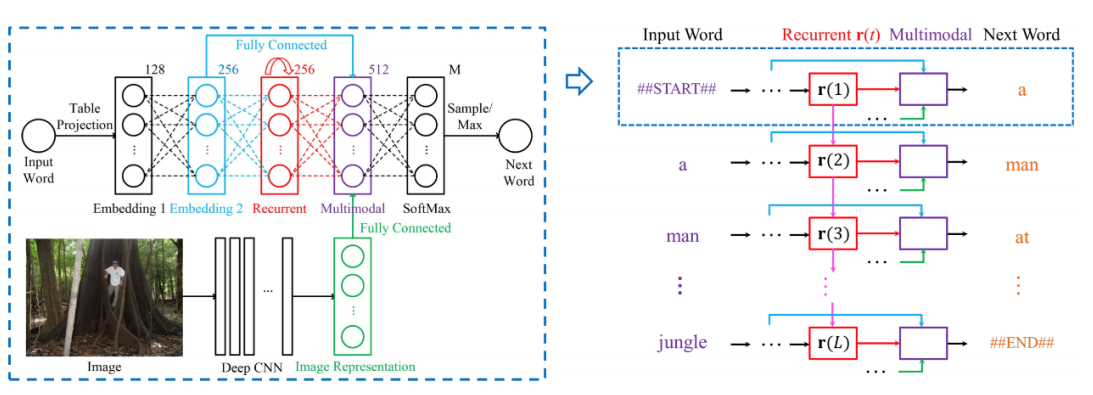


图 2 - 2 多模态 m-RNN 模型

在2015年，谷歌通过借鉴文本生成的最新研究成果推进了图像到文本自动生成的进展[22, 23]。Google利用深层卷积神经网络DCNN，将图像特征经过Encoder神经网络层编码后，在经过Decoder神经网络层（该神经网络层由LSTM（Long-Short Term Memory Network，LSTM）构成）解码成自然语言文本。Encoder-Decoder网络结构与传统模型相比，不再需要图像-词对齐、调序等繁琐步骤。与此同时，加拿大蒙特利尔大学和多伦多大学通过机器翻译领域的新技术开创性的提出在借鉴计算机视觉领域中的“注意力”(Attention)机制来获取图像块和词语之间的关联，从而使文本段在句子生成时，能够有效的参考图像标注信息，进而模仿人类视觉到句子生成的过程，使生成的句子更符合人的表述习惯。

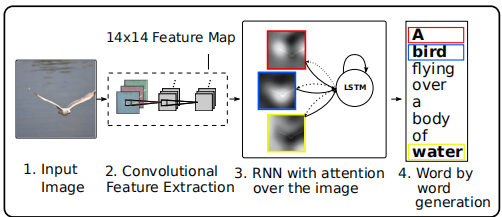


图 2 - 3 视觉“注意”引导的图像标题生成过程

此外，微软研究院[23]利用CNN卷积神经网络和多示例学习（Multiple Instance Learning，MIL）对图像建模，并使用判别语言模型生成候选句子，使用机器翻译模型中的经典最小错误率训练（Minimum Error Rate Training，MERT）来发掘文本和图像级别的特征以及对候选句子进行排序。 虽然图像到文本的生成技术还处在起步阶段，与实际产业应用存在一定的距离，但工业界积极配合学术界扩大研究方向，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[21]获得第五名，蒙特利尔-多伦多联队[24]和另一只微软研究院队伍[25]总成绩并列第三名，谷歌公司[22]和微软研究院[23]取得了总成绩并列第一名。

而反观国内学术界由于对图像到文本的生成技术研究开展较晚，大部分科研团队还停留在基础自然语言任务，只有少数研究团队开展了相关图像研究，这其中包括阿里研究院，联想研究院，百度研究院等工业界巨头。

## 2.2 实验模型难点

图像到文本的生成技术包含领域之多，需要联合多领域知识，具有极高的理论研究价值和实用前景。

对于本文而言我们由于结合了传统图像到文本生成模型和文本生成模型，除去两者自身困难点，我们的核心难点在于：

1）我们如何有效的融合两个模型，解决多模态问题；

2）我们如何构建模型实现可控，并提供高质量的文本转换内容；

只有有效的解决以上两个核心难点，我们才能保证在实验结果正确的基础上改进传统文本生成任务和图像文本生成任务的难点。

## 2.3 本章小结

本章主要介绍了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 理论模型

本章节主要介绍了目前国内外文本自动生成和文本受控生成的模型框架主要模型结构为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。我们可以认为Encoder-Decoder是一个通用生成框架。对于Encoder端的输入通常称为Source语句，Decoder端的输出我们通常称为Target语句。

Encoder：编码器处理输入序列并压缩该序列信息转换成固定长度的上下文向量（语义编码/语义向量Context）。我们通常使用该向量代表整个输入序列的信息特征。在图像模型中编码器被用来卷积图片特征，模型使用DCNN[26]来提取L个D维向量，其中每个向量都对应图像的一个块：，CNN与DCNN网络不同，我们通过多层网络卷积得到的特征会比单层得到的特征更准确，而且因为多层的缘故，我们可以在多层网络卷积过程中，让Decoder得到正确的图像特征。这好比自然语言处理中的Seq2Seq任务，只是由词序列转换成了词向量的序列。

Decoder：我们将Encoder的输出作为Decoder的输入之一，通过一系列的计算我们得到了输出序列的条件概率分布，在经过Softmax后得到输出。在Attention机制使用之前，我们只使用Encoder的最终输出，在使用Attention机制之后，我们选用每个词序列的输出作为Attention计算权重的数据，这使得Encoder的输出更多的包含了源语句的信息，因此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）。

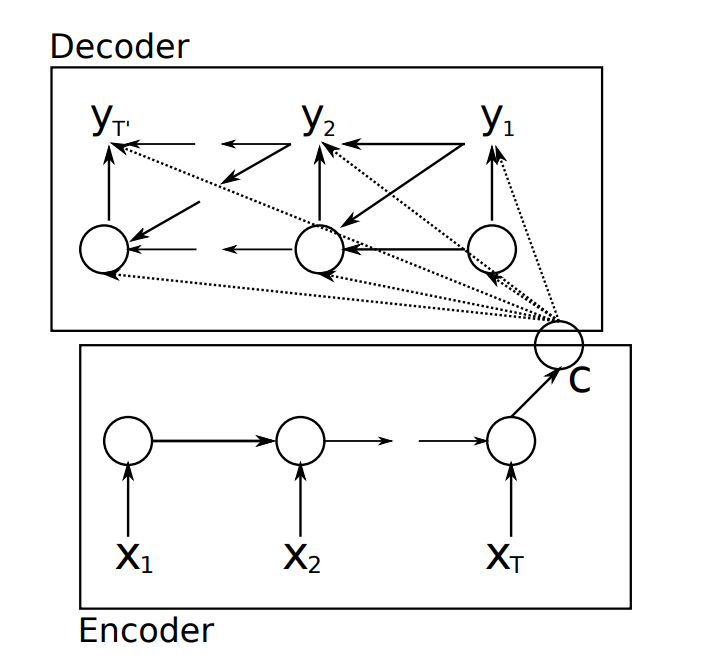


图 3 - 1 Encoder - Decoder递归神经网络模型

### 3.1.2 LSTM网络结构

LSTM（Long Short Term）网络作为RNN网络的改进，其主要优点在于它有效的解决了普通RNN网络无法解决长距离依赖这一问题。LSTM由[Hochreiter & Schmidhuber（1997）](http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf)提出，近几年由[Alex Graves](https://scholar.google.com/citations?user=DaFHynwAAAAJ&hl=en)改良了新版本，这使得LSTM网络在包括文本生成，文本分类等任务获得了广泛的应用。  
 LSTM由于其独特的神经结构，使得它在解决长距离依赖问题上并不需要耗费过多的资源，与标准的RNN 循环神经网络相比，LSTM网络的重复模块有四个，而RNN只有一个（通常情况下为一个Tanh层），它们以一种特殊的方式进行交互。

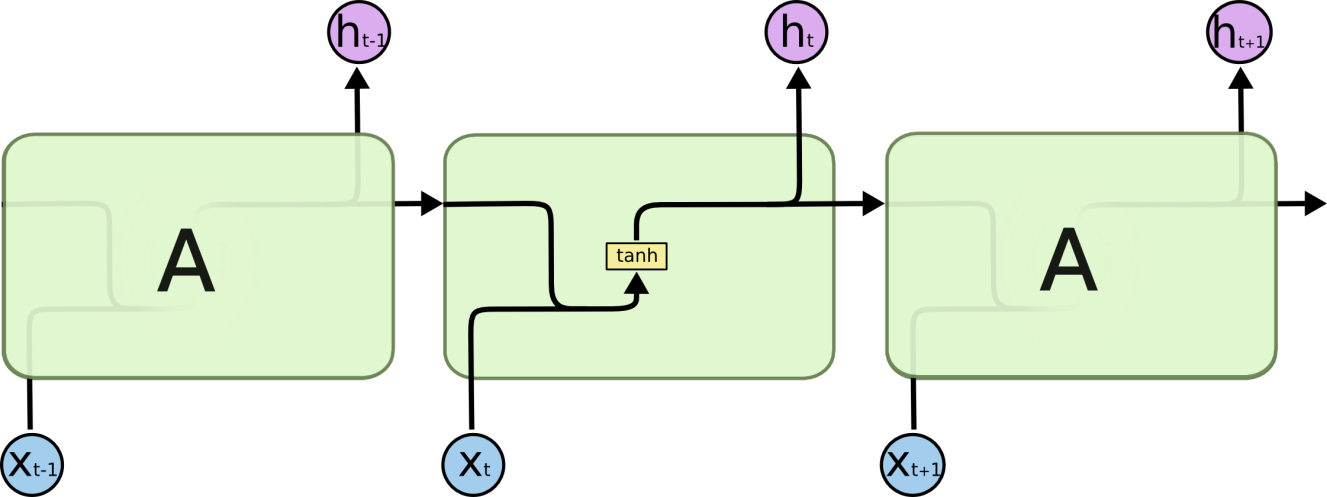


图 3 - 2 RNN-Cell 结构

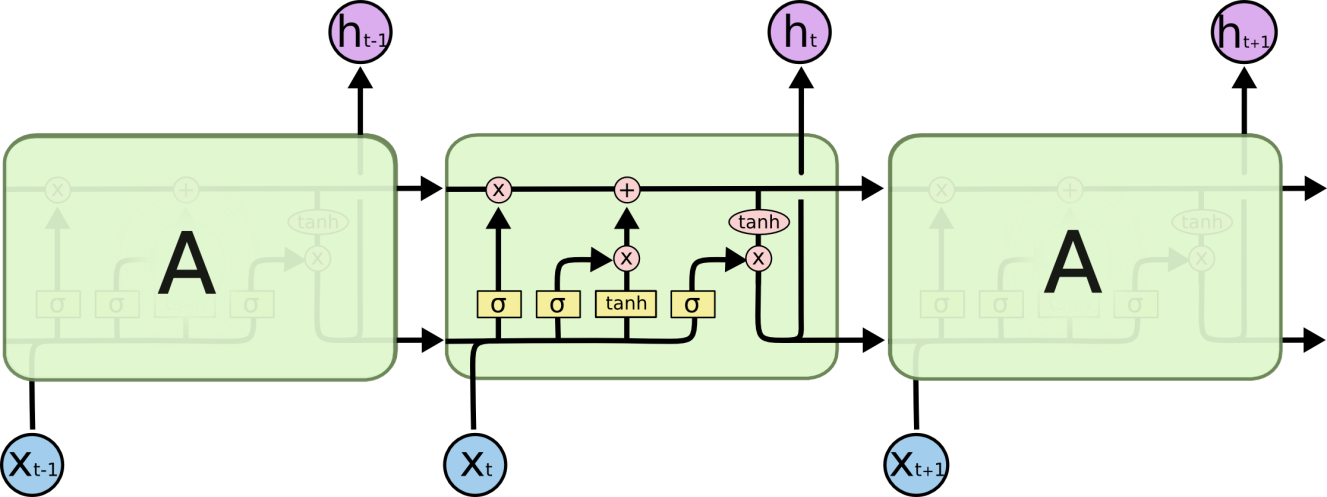


图 3 - 3 LSTM-Cell 结构

### 3.1.3 Soft-Attention网络结构

所谓Soft，其实就是Softmax函数，该函数通常使用在求注意力分配概率分布的时候，通过该函数我们能够预测当前时间步的输出。

首先我们需要知道，Attention机制通常作用于Decoder端，用以解码不是时间步长下不同的图像区域，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。因为需要满足输入序列的各个区域的权重的加和是唯一的，我们使用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。

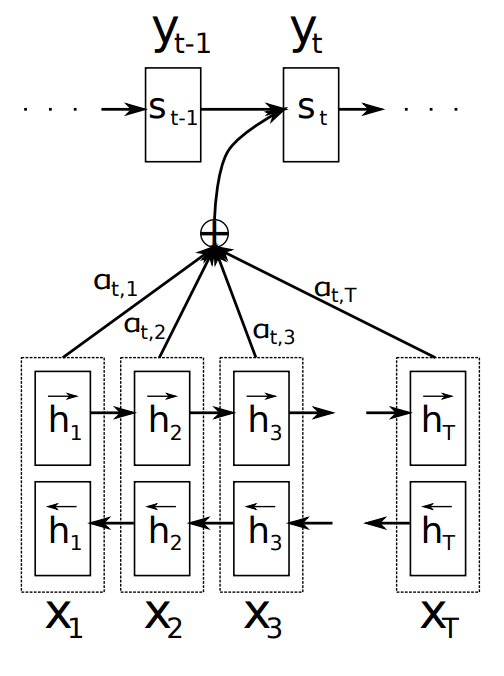


图 3 - 4 Attention示例图

## 3.2 本文理论模型论述

通过对Encoder-Decoder网络结构的分析，我们知道该结构常用于文本生成任务，而图像受控文本生成其本质还是文本生成，因此我们也应该采用这种结构。如3.1章节中我们提到的，编码器，解码器作为神经网络层，我们可以定义编码器使用LSTM神经网络，因为它显然比普通RNN网络在解决长距离以来问题上更有优势。

而对于编码器网络，我们同样在3.1章节中提到了关于Attention机制，这种从计算机视觉继承到文本生成的技术，很大程度上提高了文本生成的质量，因此我们提出了带有Attention机制的Sequence-to-Sequence网络。

当我们完成了文本生成的模型构建时，我们发现，实验的难点在于如何实现可控，当我们接触文本分类任务的时候，我们了解到如果我们训练一个分类器，使它能够完成给定一句话，输出它属于哪一类，这样我们把类别设置为0和1，其中0代表正面情感，1代表负面情感。这样我们在编码器生成自然语言后，用分类器生成句子的分类类别，并去判断它是否是我们指定的类别，如果是我们基于奖励，如果不是我们会反馈，在分类器，生成器这样的组合网络中，我们发现了SeqGAN[27]网络模型。

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

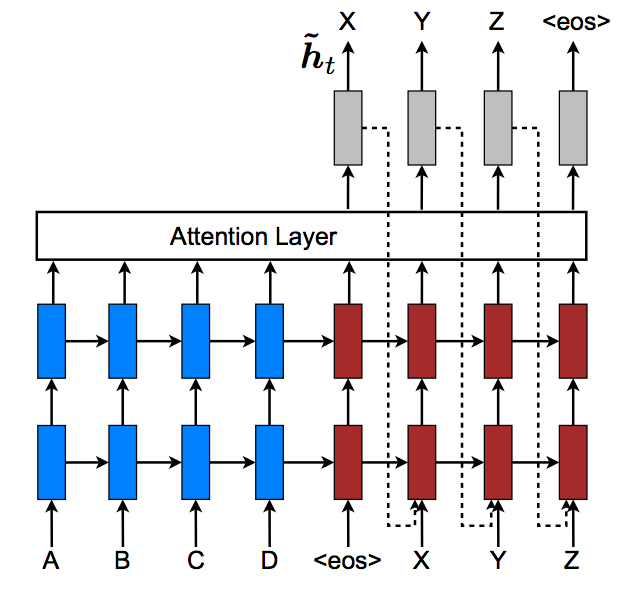


图 3 - 5 带有Attention的网络结构

计算过程：如图3 - 4参数

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，我们编码输入序列，并得到Final State这包含输入序列的全部信息。



在注意模型中，由于我们需要每个时间步的输出，因此Final State也是多个：



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。

分数的计算需要当前时间步Encoder的输出，以及上一时间步Decoder的输出。

使用函数计算每个时间步的注意力权重：



我们对这些权重使用两个下标，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，如果计算两个时间步，我们将得到以下结果：



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

同时，在我们得到每个词序列的权重后，我们希望知道它在整个句子中所占的权重，因此我们必须使用[softmax函数](https://en.wikipedia.org/wiki/Softmax_function)进行归一化处理，这也使得我们能对归一化处理的结果直接计算。



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。



最终我们通过一个全连接层得到结果Y1。

### 3.2.2 GAN网络结构

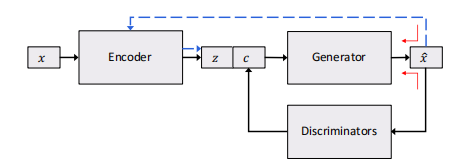


图 3 - 6 GAN网络结构

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：。

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：

## 3.3 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 实施方案

本章主要介绍本实验所用到的模型，以及实现的思路和过程。通过对实验的逐步分析，依次构建实验模型的每个部分。

## 4.1 实验分析过程

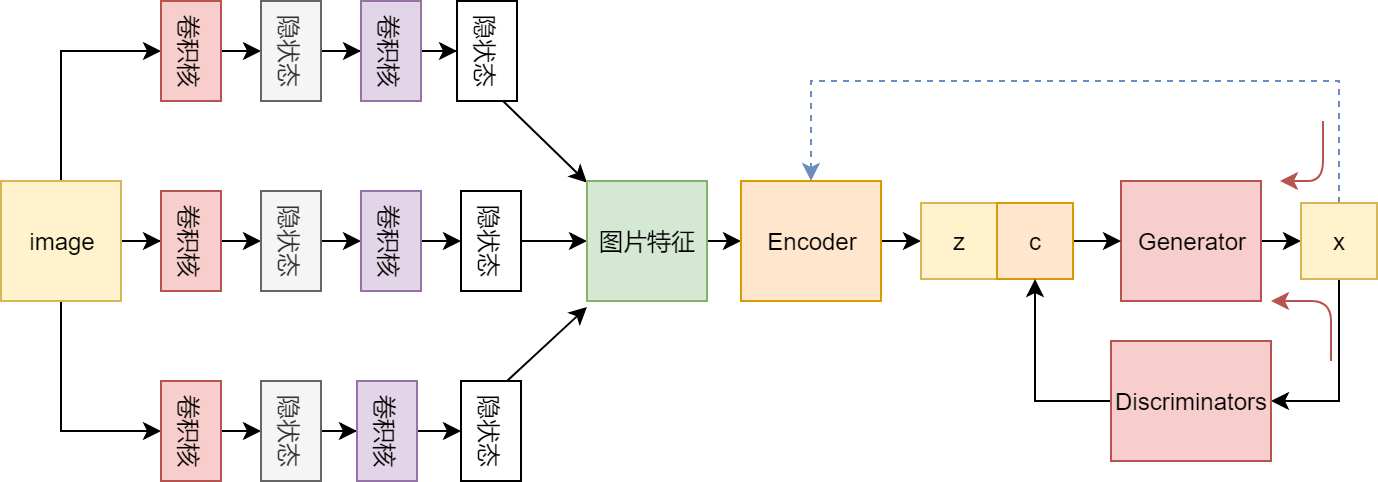


图 4 - 1 多模态可控文本生成网络结构

将图像神经网络得来的句子特征经过GAN网络，与指定的感情标签绑定一起送入Encoder端，与此同时我们的分类器也得到了一对含有感情标签的句子对，我们将Encoder的结果送入Decoder，Decoder在Attention的帮助下充分学习Encoder的结果，并将学到的结果经过Decoder生成自然语言文本，这时候我们将生成的文本经过分类器去判别生成是否满足要求，满足给予奖励，不满足就会反馈学习，这就是对抗网络对抗学习的一个过程，在实验中我们充分验证了其模型的有效性，但同时对抗学习也对服务器的性能配置要求极高。

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[28]。在最新研究中，提出了具有相同功能的BERT[29]神经网络层。

（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。

表 1 词典内容展示

|  |  |  |
| --- | --- | --- |
| 单词 | 编号 | 大小 |
| **.**  **.**  **.**  foothils  Consmetics  **.**  **.**  **.**  brisk  shark  **.**  **.**  **.** | **.**  **.**  **.**  9228  9229  **.**  **.**  **.**  9319  9320  **.**  **.**  **.** | 9359 |

（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。

表 2 One-hot编码

|  |  |  |
| --- | --- | --- |
| 单词 | 单词 | 向量 |
| **.**  **.**  **.**  foothils  Consmetics  **.**  **.**  **.**  brisk  shark  **.**  **.**  **.** | **.**  **.**  **.**  9228  9229  **.**  **.**  **.**  9319  9320  **.**  **.**  **.** | **.**  **.**  **.**  [0,0,...,1,...,0,0]  [0,1,...,0,...,0,0]  **.**  **.**  **.**  [0,0,...,0,...,1,0]  [0,0,...,0,...,0,1]  **.**  **.**  **.** |

（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。

表 3 WordEmbedder表示方式

|  |  |
| --- | --- |
| One-hot | 词嵌入 |
| **.**  **.**  **.**  [0,0,...,1,...,0,0]  [0,1,...,0,...,0,0]  **.**  **.**  **.**  [0,0,...,0,...,1,0]  [0,0,...,0,...,0,1]  **.**  **.**  **.** | **.**  **.**  **.**  [-4.42496777e-01 2.49782610e...]  [1.42921007 0.4299607824281...]  **.**  **.**  **.**  [-3.424963293795 2.43842345...]  [2.415727685 0.362761246283...]  **.**  **.**  **.** |

## 4.3 CNN卷积网络

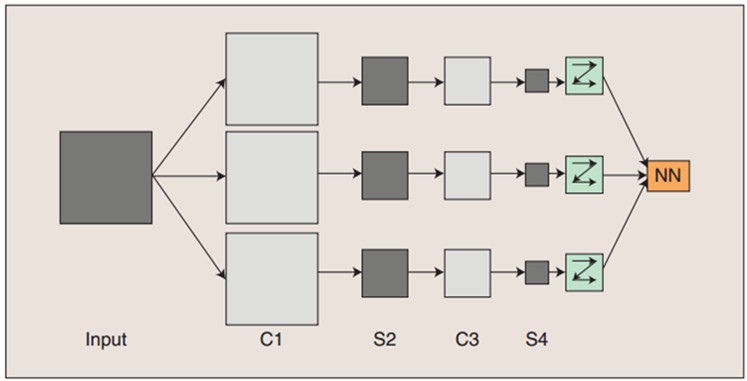


图 4 - 2 CNN网络结构

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：如公式（1），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：，公式（2）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。

 （1）

 （2）

## 4.4 UnidirectionalRNNEncoder

单向循环神经网络[30]由三部分组成，分别为输入层、隐藏层和输出层。X是一个向量，它代表输入层的值；S是一个向量，它代表隐藏层的值；U是一个矩阵，用以计算输入层到隐藏层的权重，O是一个向量，它代表输出层的值；V是一个矩阵，用以计算隐藏层到输出层的权重。S向量的取值不仅取决于当前时间步输入值X，还受制于上一层传递的S。权重矩阵 W从上一层隐藏层不断地传递到下一层作为输入。



图 4 - 3 单向RNN展开图

循环神经网络的计算公式如下：

 （1）

 （2）

公式(1)用以计算输出层，输出层作为一个全连接层，它的每个节点和隐藏层的每个节点都是相连的。V是一个矩阵，被用来计算输出层的权重，g是激活函数（常用的激活函数用RELU、Sinmod等）。公式(2)用以计算隐藏层，U也是一个矩阵，用以计算输入x的权重，W是一个矩阵，继承于上一次s的计算，并被传递到下一次输入权重的计算，f同g都是激活函数。

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

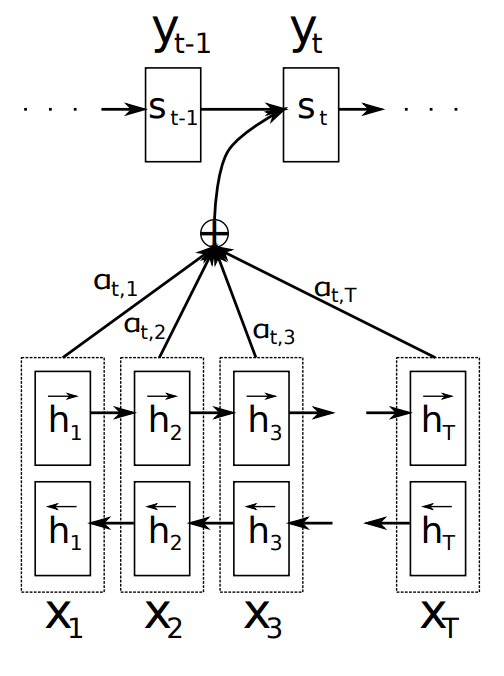


图 4 - 4 Attention作用过程

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，计算如公式（1）

 （1）

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作为Softmax的退火率，它决定着Softmax是否能够拟合数据概率分布，当Temperature值越低，拟合效果越好，但不能低于0。

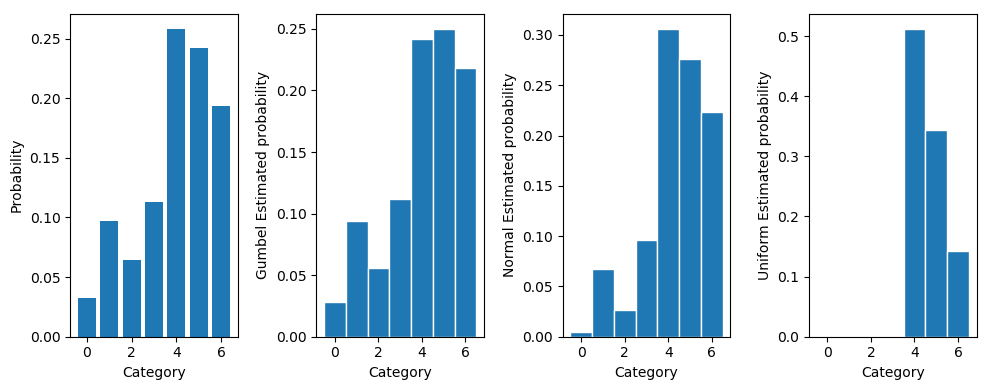


图 4 - 5 Gumbel-softmax拟合数据效果对比

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。

## 4.7 本章小结

本章通过对实验模型可能会用到的模型算法加以分析，通过模型每个部分的分析选择最有的算法模型搭建一种融合注意力机制的多模态可控文本生成。

# 第5章 实验与分析

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。

实验轮数效果对比展示如表4：

表 4 实验轮数对比效果展示

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

实验模型参数配置如表5：

表 5 实验参数表

|  |  |  |
| --- | --- | --- |
| 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 |

## 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

表 6 实验采用的数据集

|  |  |
| --- | --- |
| 数据集 | 说明 |
| COCO2014 | 图像总共又91各类别，其中包括328,000影像和2,500,000个label。从该数据集上我们能得到图片中主体、主体之间的关系，以及主体所在的位置。 |
| Yelp | Yelp情绪数据集主要包含447259条含有情感标签的训练语料。 |

COCO数据集内容：

表 7 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 |

Yelp数据集展示（0：负面情感，1：整面情感）

表 8 Yelp数据集展示

|  |  |
| --- | --- |
| 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 |

### 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。

表 9 实验配置

|  |  |
| --- | --- |
| 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 |

## 5.3 实验评测方案及指标

表 10 各评测指标

|  |  |
| --- | --- |
| 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 |

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[31]年提出，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：

 （1）

 （2）

Pn被用来计算多元词组的精度，所谓多元词组（N-gram），就是比较目标语句和人工翻译之间N个词组的相似占比。我们以英中文本翻译为例解释BLEU算法的过程：

源语句：我在写论文

目标语句：I write a thesis.

人工翻译：I am writing a thesis.

如果我们计算1-gram，我们可以发现目标语句中单个词组的在人工翻译的句子中出现的次数为3，因此1-gram的匹配度就是3/4。

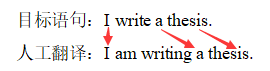


图 5 - 1 1-gram匹配示意图

如果计算2-gram，目标语句可能的2元词组组合有3组，匹上的组合次数为1，因此2-gram的匹配度就是1/3.

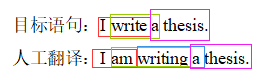


图 5 - 2 2-gram匹配示意图

依次类推，我们很容易的计算出N-gram的匹配度，但同样我们会忽略一个问题，假设我们目标语句为：

源语句：我在写论文

目标语句：a a a a

人工翻译：I am writing a thesis.

如果按照上述方法计算我们会发现它的匹配度为1，但事实上它的流畅度几乎为0。因此我们加入了召回率这个概念，也就是，我们计算N个词组在人工翻译语句中出现的最大次数，并取两者的最小值，经过改进后匹配度为1/2，这避免了重复词组对精度的影响。

在不断的发展中，我们发现现有的精度计算方法，仍就存在精度值随着句子长度变短而值激增的情况，假设我们的目标语句为：

源语句：我在写论文

目标语句：am a

人工翻译：I am writing a thesis.

按照上述计算方法我们得到1-gram的匹配度为1，这显然也是不合理的。因此我们加入了“惩罚因子”的概念。公式如下：



从公式可以看出当目标语句的长度小于等于人工翻译的语句时，我们使用惩罚因子，降低精度值，随着不断的发展变化，我们有了BLEU的现在计算方式，但现行计算方法不足之处还有很多，业界对BLEU[32]的改进也在不断的进行。

## 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[24]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[33]在能够找到实现。

### 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

自动编码器作为VAE[34, 35]的简化版本，常用于文本生成任务。

|  |
| --- |
|  |

5.4.4.1 构建Encode

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[36]提到的方法，该方法的简单实例我们可以在[37]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机制的网络（以下我们统称实验2[38]），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。

表 11 模型效果测试对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

如表11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。

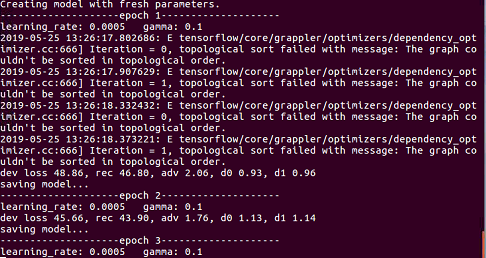


图 5 - 3 无Attention训练过程图

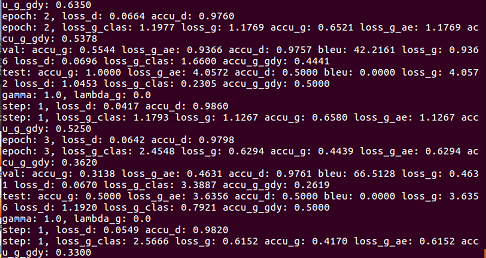


图 5 - 4 Attention训练过程图

## 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。

表 12 Batch-size对实验的影响

|  |  |  |
| --- | --- | --- |
| 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 |

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[39, 40]，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。

表 13 Gamma-decay对实验的影响

|  |  |  |
| --- | --- | --- |
| Gamma-decay | BLEU | Accu-g |
| 1.0 | 63.100 | 0.6563 |
| 0.5 | 67.514 | 0.5102 |
| 0.1 | 68.903 | 0.5534 |

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。

实验3：我们定义一个多项分布，其真实密度函数如图5-3：

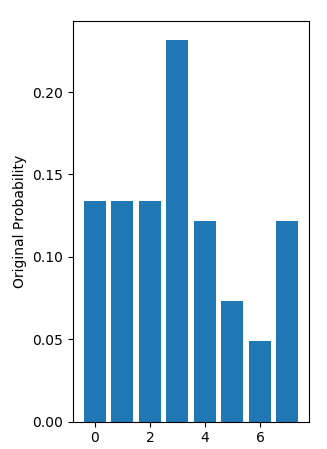


图 5 - 5 真实概率密度分布

我们基于Gumbel-softmax做采样对比，图5-4左侧我们设置Temperature = 0.1，右侧我们设置Temperature = 0.5。

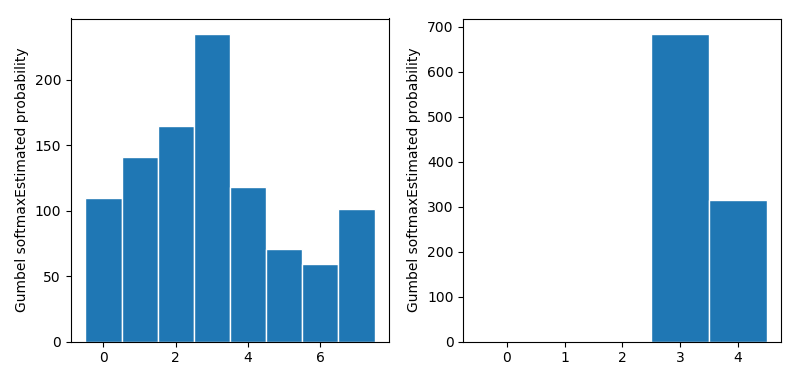


图 5 - 6 Temperature对数据分布影响

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。

表 14 Lambda-g对实验的影响

|  |  |  |
| --- | --- | --- |
| Lambda-g | BLEU | Accu-g |
| 1.0 | 64.609 | 0.5800 |
| 0.5 | 68.514 | 0.6249 |
| 0.1 | 70.016 | 0.6562 |

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 实际场景运用

### 5.6.1 文本生成任务之英中文本翻译



图 5 - 7 AI Challenger比赛官网

1. 赛道

英中文本机器翻译

1. 任务介绍

英中机器文本翻译作为此次比赛的任务之一，目标是评测各个团队机器翻译的能力。本次机器翻译语言方向为英文到中文。测试文本为口语领域数据。参赛队伍需要根据评测方提供的数据训练机器翻译系统，可以自由的选择机器翻译技术。例如，基于规则的翻译技术、统计机器翻译及神经网络机器翻译等。

本次竞赛将利用机器翻译的客观考核指标（BLEU、NIST score、TER）进行评分，BLEU得分会作为主要的机器评价指标。组委会将通过客观指标，并结合答辩表现，综合评估参赛者的算法模型。

1. 数据说明

训练集文件名train.txt，其中每个训练样例包含自左至右4个元素：DocID， SenID， EngSen，ChnSen。DocID表示这个样例出现在哪个文件中，DocID用来提供训练集中句子出现的场景和上下文情景。SenID表示这个样例在DocID中出现的位置，比如，如果SenID为94，那么这个样例就是DocID的第94句话。若无上下文信息，则DocID和SenID均为NA。EngSen和ChnSen分别对应英文句子和中文句子，二者互译。

验证集和测试集为.sgm文件，句子格式和训练集相同。其中测试集没有与英文句子EngSen对应的中文句子ChnSen。

训练集和测试集、验证集的上下文文件包含所有语句的上下文的信息，其中每行包含自左至右三个元素：DocID， SenID， EngSen

训练集样例如图5-8所示（第一列DocID, 第二列SenID, 第三列EngSen，第四列ChnSen）：

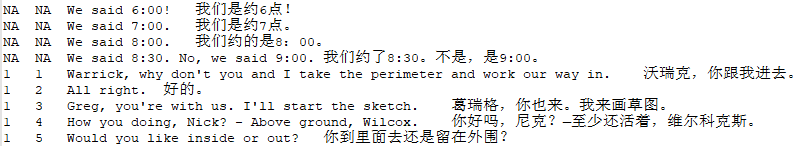


图 5 - 8 训练集

测试集、验证集样例如图5-9所示（第一列为DocID，第二列SenID，第三列EngSen）：

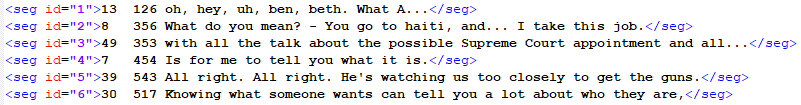


图 5 - 9 测试集、验证集英文数据集

验证集中文样例如图5-10所示：

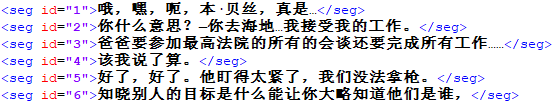


图 5 - 10 验证集中文数据集

上下文文件样例如图5-11所示（第一列为DocID，第二列SenID，第三列EngSen）：

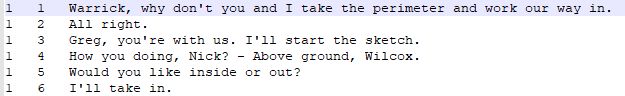


图 5 - 11 上下文数据集

1. 本测评使用Google的Transform[41]模型，在决赛以32的Bleu值排名14名，进入20强。

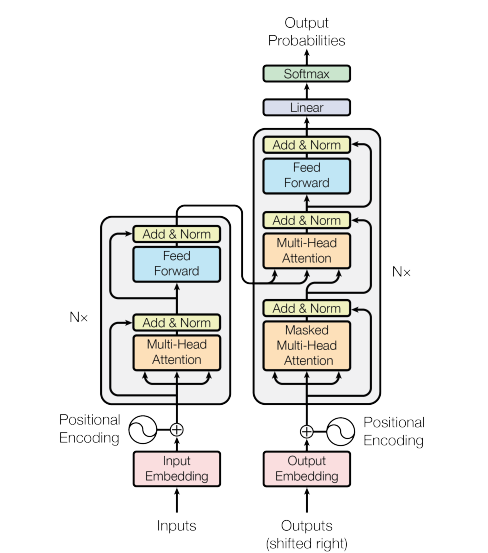


图 5 - 12 Transform模型

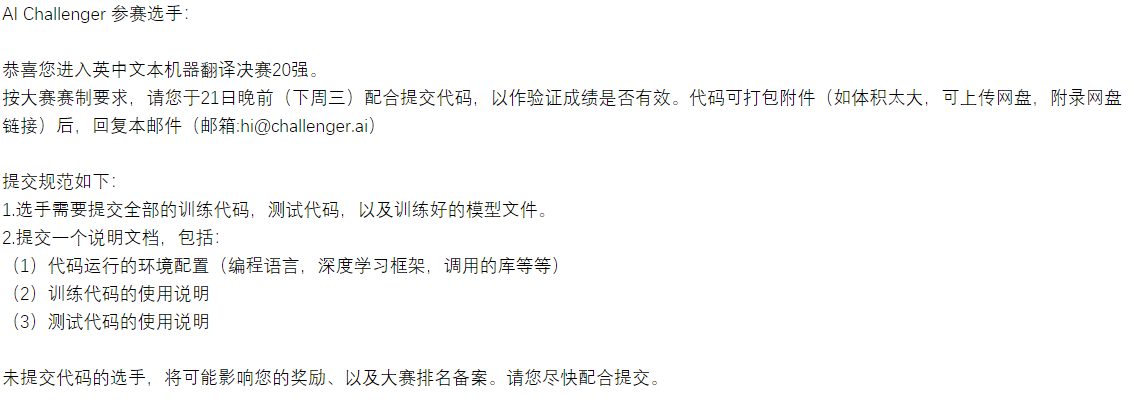


图 5 - 13 决赛邀请函

### 5.6.2 中文情感对话子任务

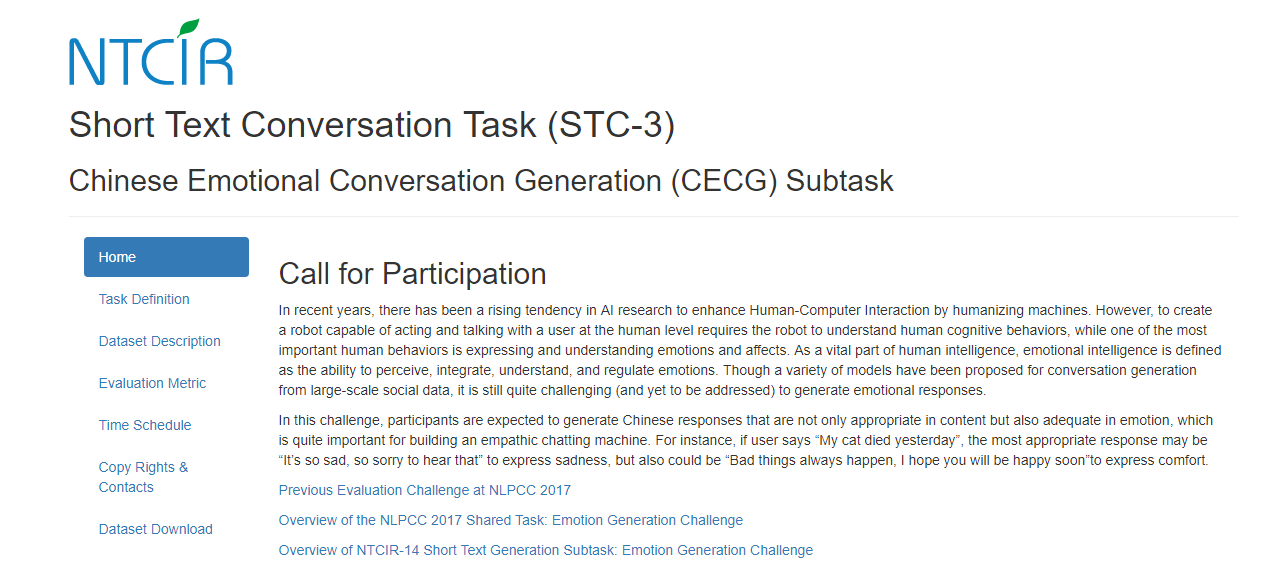


图 5 - 14 CECG任务官网

1. 任务定义

此任务的定义如下: 给定一个中文后 x = (x1, x2, ......, xn), 以及用户指定的情感类别的响应要生成, 目标是生成响应

y = (y1,y2, ......, ym), 这与情绪类别是一致的。 情绪范畴为 {愤怒、厌恶、幸福, 比如悲伤, 其他} , 与NLPCC情感分类挑战中定

义的相同。

1. 数据集描述

数据集由微博帖子和回复/评论构成。将向参与者提供超过100万个微博后响应对，以培训他们的模型。为了确保公平比较，不允许额外的训练数据用于对话生成，但是参与者可以使用其他数据来训练用于情绪分类的补充分类器。应在提交结果时报告此类详细信息。

测试数据集由大约5000个帖子组成，而100~200个帖子将被手动评估，并且对于每个帖子，将手动指定最多3个情绪类别以指示所生成的响应的情绪类别。参与系统应该为每个情绪类别生成响应。请注意，参与者应为具有适当情绪类别的所有帖子生成回复。对于公平比较，参与者不知道将手动检查哪些部分的帖子。

数据集将包括每个帖子和响应的标签。这些标签仅供参考，它们是通过基于双向LSTM模型的简单分类器获得的。分类器接受了来自NLPCC的数据的训练 。我们的六向分类分类器的准确率约为64％。换句话说，这些数据的情感标签是NOISY。标签和情感类之间的对应关系可以看作如下表：

表 16 数据说明

|  |  |
| --- | --- |
| 便签 | 含义 |
| 0 | 其他 |
| 1 | 喜欢 |
| 2 | 悲伤 |
| 3 | 厌恶 |
| 4 | 愤怒 |
| 5 | 幸福 |

训练数据集如图5-15所示：

888

图 5 - 15 训练数据集

（3）最终以BLEU值为30提交比赛平台。

## 5.7 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，除此之外我们还可以在其他方面进行改进[42]最终在测试集上指标值如表15。

表 15 最终实验结果

|  |  |
| --- | --- |
| 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 |

# 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] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition, 2016[C].2016.

[2] Liu F, Shen C, Lin G. Deep convolutional neural fields for depth estimation from a single image, 2015[C].2015.

[3] Huang Z, Wang R, Shan S, et al. Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification, 2015[C].2015.

[4] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering, 2018[C].2018.

[5] Aneja J, Deshpande A, Schwing A G. Convolutional image captioning, 2018[C].2018.

[6] Cui Y, Yang G, Veit A, et al. Learning to evaluate image captioning, 2018[C].2018.

[7] Reiter E, Dale R. Building natural language generation systems[M]. Cambridge university press, 2000.

[8] Yao B Z, Yang X, Lin L, et al. I2t: Image parsing to text description[J]. Proceedings of the IEEE, 2010,98(8):1485-1508.

[9] Feng Y, Lapata M. How many words is a picture worth? automatic caption generation for news images, 2010[C]. Association for Computational Linguistics, 2010.

[10] Feng Y, Lapata M. Automatic caption generation for news images[J]. IEEE transactions on pattern analysis and machine intelligence, 2012,35(4):797-812.

[11] Yang Y, Teo C L, Daumé III H, et al. Corpus-guided sentence generation of natural images, 2011[C]. Association for Computational Linguistics, 2011.

[12] Kulkarni G, Premraj V, Dhar S, et al. Baby talk: Understanding and generating image descriptions, 2011[C]. Citeseer, 2011.

[13] Kulkarni G, Premraj V, Ordonez V, et al. Babytalk: Understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013,35(12):2891-2903.

[14] Mitchell M, Han X, Dodge J, et al. Midge: Generating image descriptions from computer vision detections, 2012[C]. Association for Computational Linguistics, 2012.

[15] Elliott D, Keller F. Image description using visual dependency representations, 2013[C].2013.

[16] 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.

[17] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions, 2015[C].2015.

[18] Socher R, Karpathy A, Le Q V, et al. Grounded compositional semantics for finding and describing images with sentences[J]. Transactions of the Association for Computational Linguistics, 2014,2:207-218.

[19] Chen X, Zitnick C L. Learning a recurrent visual representation for image caption generation[J]. arXiv preprint arXiv:1411.5654, 2014.

[20] Mao J, Xu W, Yang Y, et al. Deep captioning with multimodal recurrent neural networks (m-rnn)[J]. arXiv preprint arXiv:1412.6632, 2014.

[21] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description, 2015[C].2015.

[22] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator, 2015[C].2015.

[23] Fang H, Gupta S, Iandola F, et al. From captions to visual concepts and back, 2015[C].2015.

[24] Show A. Tell: Neural image caption generation with visual attention[J]. Kelvin Xu et. al.. arXiv Pre-Print, 2015,23.

[25] Devlin J, Cheng H, Fang H, et al. Language models for image captioning: The quirks and what works[J]. arXiv preprint arXiv:1505.01809, 2015.

[26] Nguyen L D, Lin D, Lin Z, et al. Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation, 2018[C]. IEEE, 2018.

[27] Yu L, Zhang W, Wang J, et al. Seqgan: Sequence generative adversarial nets with policy gradient, 2017[C].2017.

[28] Artetxe M, Labaka G, Agirre E. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations, 2018[C].2018.

[29] Devlin J, Chang M, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[30] Al-Sabahi K, Zuping Z, Kang Y. Bidirectional Attentional Encoder-Decoder Model and Bidirectional Beam Search for Abstractive Summarization[J]. arXiv preprint arXiv:1809.06662, 2018.

[31] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation, 2002[C]. Association for Computational Linguistics, 2002.

[32] Zhukov V, Golikov E, Kretov M. Differentiable lower bound for expected BLEU score[J]. arXiv preprint arXiv:1712.04708, 2017.

[33] https://github.com/DeepRNN/image\_captioning.

[34] Song J, Zhang H, Li X, et al. Self-supervised video hashing with hierarchical binary auto-encoder[J]. IEEE Transactions on Image Processing, 2018,27(7):3210-3221.

[35] Chen T Q, Li X, Grosse R B, et al. Isolating sources of disentanglement in variational autoencoders, 2018[C].2018.

[36] Shen T, Lei T, Barzilay R, et al. Style transfer from non-parallel text by cross-alignment, 2017[C].2017.

[37] https://github.com/shentianxiao/language-style-transfer.

[38] Hu Z, Yang Z, Liang X, et al. Toward controlled generation of text, 2017[C]. JMLR. org, 2017.

[39] Yao Z, Gholami A, Keutzer K, et al. Large batch size training of neural networks with adversarial training and second-order information[J]. arXiv preprint arXiv:1810.01021, 2018.

[40] Peng C, Xiao T, Li Z, et al. Megdet: A large mini-batch object detector, 2018[C].2018.

[41] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need, 2017[C].2017.

[42] Freitag M, Al-Onaizan Y. Beam search strategies for neural machine translation[J]. arXiv preprint arXiv:1702.01806, 2017.

致 谢

在这美好的大学四年中，我收获了很多。回顾大学时光，想起自己刚入学的时候上的第一堂专业课，这也是我第一次接触网络工程专业，在随后的2年中，在院学生会的培养下，我当上了新闻部部长，这是我大学几年认识朋友最多时间，我也很庆幸有这段经历，当然我特别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。

感谢在大学时光里遇到的全部人，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！