

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模型

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

文本生成模型

**Multimodal Controllable Text Generation with attention 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

多模态可控文本生成，作为计算机视觉和自然语言处理交叉领域的新技术，其研究价值得到了充分的肯定，并被广泛应用于各个领域。本文以图像到文本为例研究多模态数据转换，并要求在文本端生成可控的文字描述，因此对于模型的构建本身也是一项巨大的挑战，其次还需要融合多个模态的数据。由于当下对此研究甚少，因此本文将从可控文本生成和图像文本生成的新角度出发，研究一套融合注意力机制的多模态可控文本生成框架。

针对可控文本的生成过程，本文提出一种基于生成对抗网络（Generative Adversarial Network，GAN）模型的多模态可控文本生成框架（Multimodal Controllable Text Generator Framework with GAN）。该模型框架以视觉特征和控制信息为输入，利用生成器Generator实现文本的可控生成，而判别器实现对生成文本质量的可控判别。同时，为了解决离散数据分布不可微的问题，我们提出一种有效的梯度估计方法使用Gumbel-Softmax实现对传统Softmax（归一化指数函数）进行改造，使其能够在有限的离散型数据上达到很高的拟合效果，从而提高了提高可控文本生成的准确性和流畅性。

为了进一步提高可控生成框架对多模态信息的融合感知生成能力，提出将注意力机制融入可控文本生成框架，实现了文本过程中能够在最大程度上遵从原始多模态特征。传统解码网络由于只使用编码网络的最终输出，这导致了大量上下文信息没有被保留，而融入注意力机制后，解码器将会得到每个时间步的编码器输出，进一步提高了可控文本生成的质量。

在本文中我们提出了一种多模态可控文本生成框架，并实现了图像可控文本生成，并在COCO数据集和Yelp数据集上验证了模型的可行性和有效性，最终我们取得了显著的效果。目前模型对文本的可控能力还很有限，后续将进一步提升可控生成的质量。

关键词：可控文本生成；多模态；注意力；GAN；Gumbel-Softmax

# Abstract

Multi-modal controllable text generation, as a new technology in the field of computer vision and natural language processing, has been fully affirmed and widely used in various fields. In this paper, image-to-text is taken as an example to study multi-modal data conversion, and it is required to generate controllable text description on the text side. Therefore, the construction of the model itself is also a huge challenge. Secondly, it needs to integrate multiple modal data. Because there is very little research on this at present, this paper will study a set of multi-modal controllable text generation framework that integrates attention mechanism from the new perspective of controllable text generation and image text generation.

Aiming at the process of generating controllable text, this paper proposes a Multimodal Controllable Text Generator Framework with GAN based on the Generative Adversarial Network (GAN) model. The model framework takes the visual features and control information as input, and uses the generator to realize the controllable generation of the text, and the discriminator realizes the controllable discrimination of the generated text quality. At the same time, in order to solve the problem that the discrete data distribution is not divisible, we propose an effective gradient estimation method to transform the traditional Softmax (normalized exponential function) with Gumbel-Softmax, so that it can be achieved on finite discrete data. High fitting effect, which improves the accuracy and fluency of controllable text generation.

In order to further improve the fusion-aware generation ability of the controllable generation framework for multi-modal information, it is proposed to integrate the attention mechanism into the controllable text generation framework, which can achieve the maximum compliance with the original multi-modal features in the text process. Traditional decoding networks use only the final output of the encoding network, which results in a large amount of context information not being retained. After incorporating the attention mechanism, the decoder will get the encoder output for each time step, further improve the quality of controllable text generation.

In this paper, we propose a multi-modal controllable text generation framework, and realize image to controllable text generation, and verify the feasibility and effectiveness of the model on the COCO dataset and Yelp dataset. Finally, we have achieved Significant effect. But at the same time, because the controllability is still limited, we will further improve the quality of controllable generation.

**Key words: Controllable text generation; Multimodal; Attention; GAN; Gumbel-Softmax**

# **目 录**

[中文摘要 I](#_Toc10401167)

[Abstract II](#_Toc10401168)

[**目 录** III](#_Toc10401169)

[第1章 绪论 1](#_Toc10401170)

[1.1 选题背景及研究意义 1](#_Toc10401171)

[1.2 国内外研究 2](#_Toc10401172)

[1.3 研究内容 5](#_Toc10401173)

[1.4 研究难点与解决方案 6](#_Toc10401174)

[1.5 论文的组织结构 6](#_Toc10401175)

[第2章 相关理论研究及技术基础 7](#_Toc10401176)

[2.1研究现状 7](#_Toc10401177)

[2.2 相关基础理论模型简述 7](#_Toc10401178)

[2.2.1 Encoder-Decoder网络结构 7](#_Toc10401179)

[2.2.2 Unidirectional RNN Encoder 9](#_Toc10401180)

[2.2.3 LSTM网络结构 10](#_Toc10401181)

[2.2.4 CNN网络结构 11](#_Toc10401182)

[2.3 小结 12](#_Toc10401183)

[第3章 Multimodal with GAN 13](#_Toc10401184)

[3.1 本文理论模型论述 13](#_Toc10401185)

[3.1.1 Multimodal with GAN 13](#_Toc10401186)

[3.1.2 DCNN网络结构 14](#_Toc10401187)

[3.1.3 GAN网络结构 15](#_Toc10401188)

[3.1.4 Gumbel-Softmax 15](#_Toc10401189)

[3.1.5 Attention RNN Decoder 16](#_Toc10401190)

[3.1.6 Encoder-Decoder With Attention网络结构 18](#_Toc10401191)

[3.2 实验流程设计 21](#_Toc10401192)

[3.3 小结 22](#_Toc10401193)

[第4章 实验与分析 23](#_Toc10401194)

[4.1 实验总体设计 23](#_Toc10401195)

[4.2 实验数据以及实验配置 24](#_Toc10401196)

[4.2.1 实验数据 24](#_Toc10401197)

[4.2.2 实验配置 26](#_Toc10401198)

[4.3 实验评测方案及指标 26](#_Toc10401199)

[4.4 实验过程 28](#_Toc10401200)

[4.4.1 训练模型生成图片描述 28](#_Toc10401201)

[4.4.2 准备可控文本生成训练语料并嵌入词向量矩阵 28](#_Toc10401202)

[4.4.3 训练过程 30](#_Toc10401203)

[4.4.4 对比试验 32](#_Toc10401204)

[4.5 实验结果分析 33](#_Toc10401205)

[4.5.1 Batch-size的影响 33](#_Toc10401206)

[4.5.2 Gamma-decay的影响 34](#_Toc10401207)

[4.5.3 Lambda-g的影响 36](#_Toc10401208)

[4.6 小结 36](#_Toc10401209)

[第5章 实际场景运用 38](#_Toc10401210)

[5.1 文本生成任务之英中文本翻译 38](#_Toc10401211)

[5.2 中文情感对话子任务 40](#_Toc10401212)

[5.3 小结 41](#_Toc10401213)

[第6章 总结与展望 42](#_Toc10401214)

[参考文献 43](#_Toc10401215)

[致 谢 46](#_Toc10401216)

# 第1章 绪论

随着机器学习和深度学习的不断发展，人们的研究领域从简单的自然语言标注任务，文本生成任务扩大到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。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主要任务。

## 1.1 选题背景及研究意义

文本自动生成技术在诸多场景中都已经得到了广泛应用。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。文本自动生成已经越来越多的融入我们的生活，与我们息息相关。

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的生成。

可控生成图像的描述，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。

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楚一件事，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。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。

## 1.2 国内外研究

同传统文本生成任务相比，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将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并且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]能够识别出图像中的物体，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。

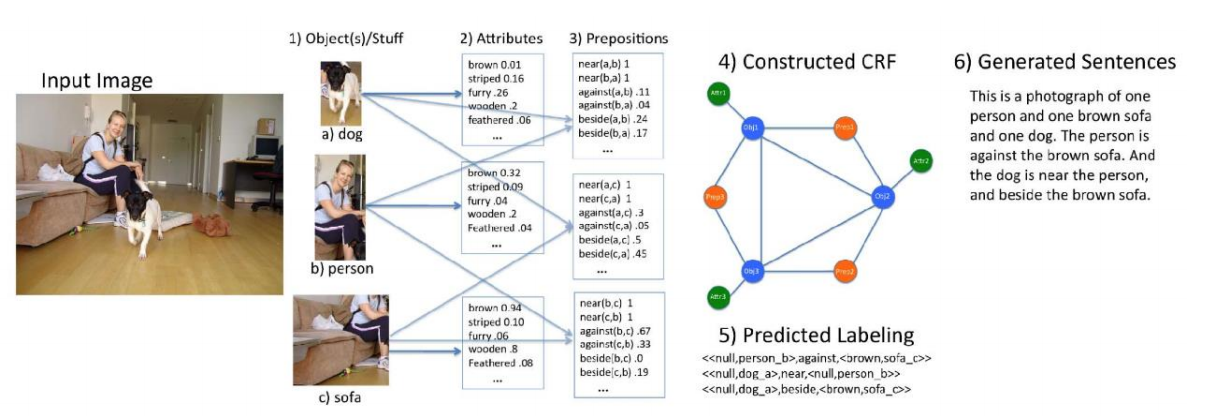


图 1 - 2 一种典型的流水线模型

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；与此同时，文本生成、图像文本生成、智能问答等生成任务也在不断发展。因此，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应运而生。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，使得文本端的生成不再考虑图像端，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。

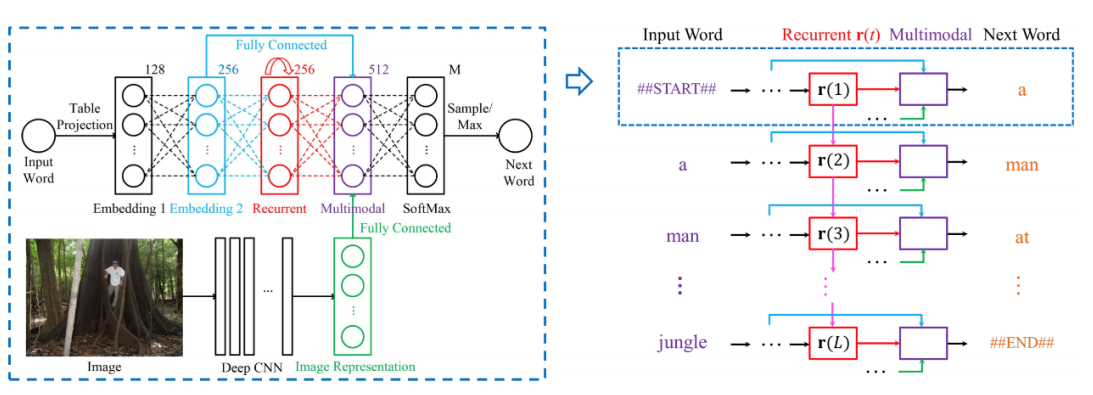


图 1 - 3 m-RNN 模型

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任务研究的进展[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。

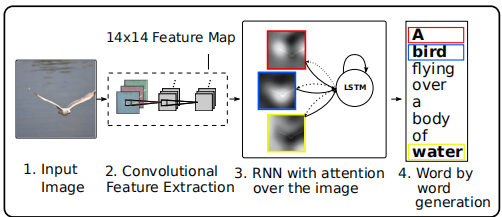


图 1 - 4 视觉“注意”引导的图像标题生成过程

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范围，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]和另一支微软研究院队伍[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。

## 1.3 研究内容

可控文本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文本自动生成任务，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。

我们定义图像可控文本生成技术是给定任意一张图片，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。

综上所述我们希望以图像可控文本生成为例，构建一种多模态可控文本生成框架。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. 深入研究分析相关理论模型，总结分析难点和解决方案。
3. 构建面向多模态的可控文本生成的神经网络模型。
4. 在数据集上验证其有效性

## 1.4 研究难点与解决方案

通过对上述可控文本生成任务和图像文本生成任务的分析总结，实现一种融合注意力机制的多模态可控文本生成主要有以下几个难点：

1. 我们如何有效的融合两个模型，解决多模态问题

针对模型融合问题我们借助图像文本生成技术，使用DCNN（深度卷积网络）处理图片，将图片的描述语句送入GAN网络以此来实现多模态数据的相互转换。

1. 我们如何构建模型实现可控，并提供高质量的生成文本

在实现可控文本生成时，我们使用GAN，它由一个判别器和一个生成成器组成，在分类器与生成不断的对抗学习中，使生成器适应分类器，最终实现可控，同时在GAN网络中融合Attention和Gumbel-Softmax来提高文本生成的质量。

## 1.5 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，从而帮助构建图像到本文的可控文本生成框架。
3. 论文模型，本章主要介绍本文所用的模型，以及实现的思路和过程。
4. 实验与分析，本章主要介绍针对模型实验结果所做的各种分析，对比实验，并优化模型的各个部分。
5. 实际场景运用，本章主要介绍相关技术在测评比赛中的应用。
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提出建议。

# 第2章 相关理论研究及技术基础

本章节主要分析国内外图像文本自动生成和可控文本生成的研究现状，归纳分析现有相关技术，以此帮助构建图像到本文的可控文本生成的框架。

## 2.1研究现状

图像文本生成技术在经历了模板匹配，关键词查询，RNN（循环神经网络），m-RNN（多模态循环神经网络），Encoder-Decoder（编码-解码结构），Encoder-Decoder with Attention（带有注意力的编码-解码结构）的发展过程后已经逐渐壮大，而值得注意的是Kelvin Xu等人在2015年ICML会议上创造性的提出将m-RNN with Attention模型算法用于图像文本生成任务，这种创造性的思想，使图像文本生成任务得到了相当可观的进展，本文同样效仿Kelvin Xu等人，希望通过图像文本生成任务解决多模态融合的问题。

可控文本生成技术由Zhiting Hu等人在2017年ICML会议上提出通过结合变分自动编码器（VAE）和整体属性分解器，有效地加强了语义结构，并取得了显著成果。本文效仿Zhiting Hu使用GAN实现可控，并希望在GAN中融入Attention和Gumbel-Softmax提高文本生成的质量。

多模态可控文本生成的相关研究甚少，但其重大应用前景决定了多模态可控文本生成的研究价值，本文将提出使用Multimodal with GAN模型实现多模态可控文本生成，并加入了Attention和Gumbel-Softmax。

## 2.2 相关基础理论模型简述

### 2.2.1 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，两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。

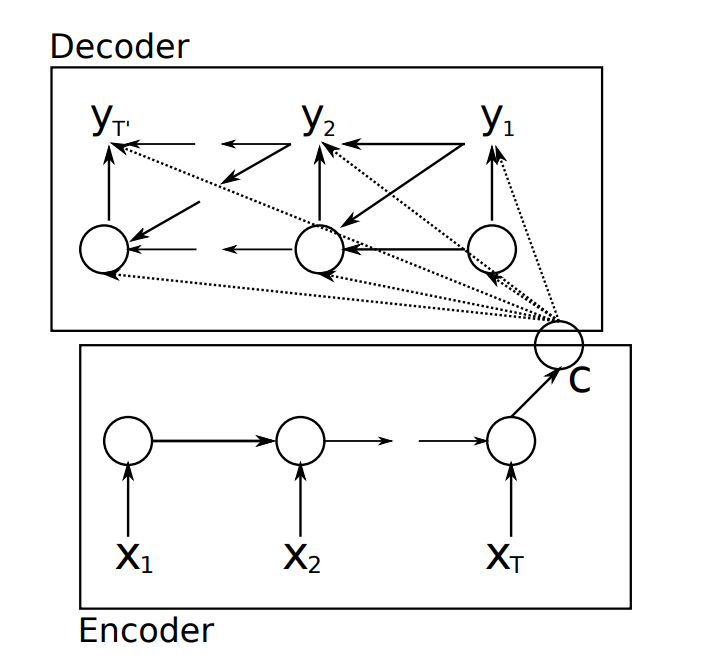


图 2 - 2 Encoder - Decoder递归神经网络模型

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，因此解码器我们使用带有Attention机制的DeepRNN神经网络（多层LSTM循环神经）。

### 2.2.2 Unidirectional RNN Encoder

单向循环神经网络[27]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 2 - 3 单向RNN展开图

循环神经网络的计算公式如下：

 公式2-1

 公式2-2

公式2-1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2-2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2-2式代入2-1式，我们将得到公式2-3：

公式2-3

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。

### 2.2.3 LSTM网络结构

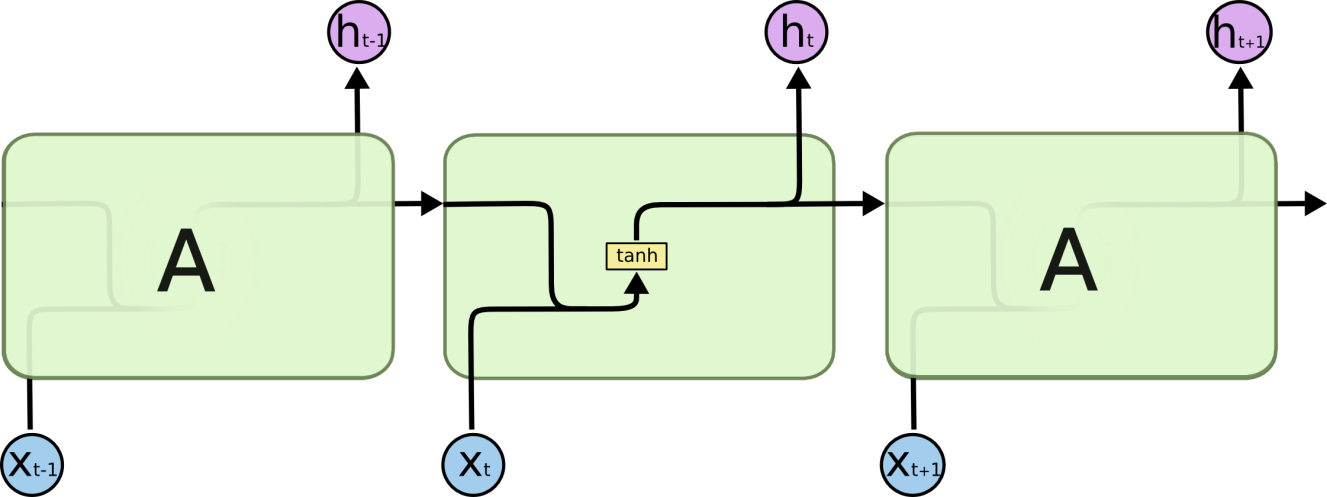


图 2 - 4 RNN-Cell 结构

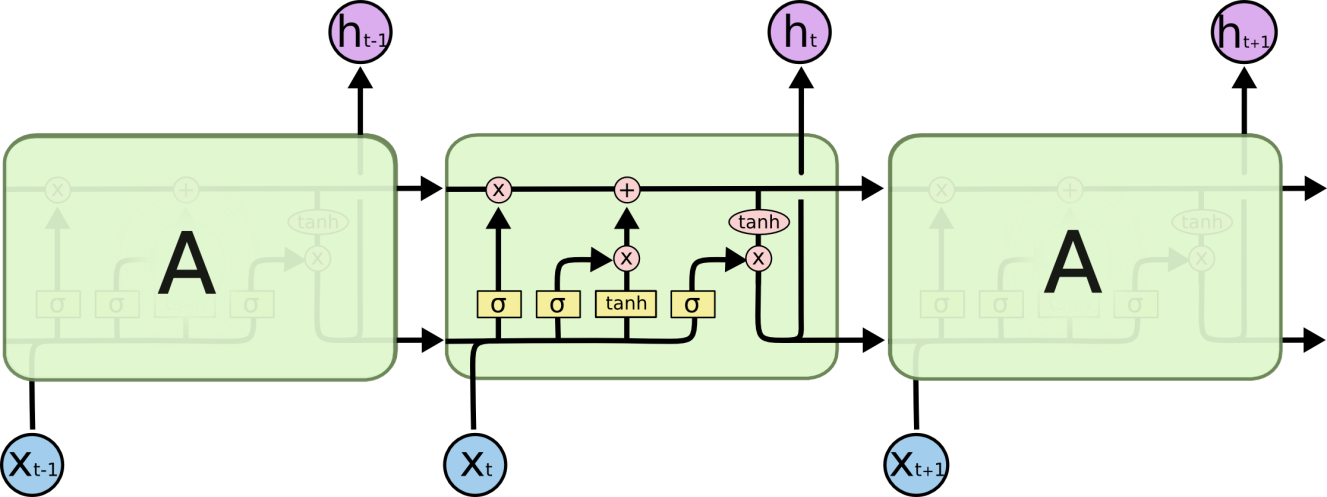


图 2 - 5 LSTM-Cell 结构

LSTM（Long Short Term Memory）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。

### 2.2.4 CNN网络结构

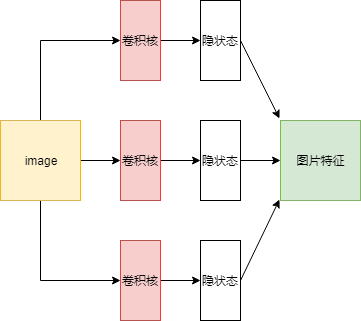


图 2 - 6 CNN网络结构

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2-4，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2-5由于只和已有的信息 ， 有关，因此这些参数也可以从数据中进行end-to-end的自动学习。

公式2-4

公式2-5

## 2.3 小结

本章主要介绍了国内外图像文本自动生成和可控文本生成的研究现状，以及相关基础模型简述，通过分析总结图像文本自动生成有助于我们解决多模态融合的问题，分析可控文本生成技术有助于实现情感可控，同时通过对相关基础模型的分析，有助于我们选择合适的模型算法构建框架，也有助于我们分析基础模型算法的不足，为我们改进提供了理论基础。

# 第3章 Multimodal with GAN

本章节主要介绍了目前国内外文本自动生成和可控文本生成的模型框架，其主要模型结构为Encoder-Decoder网络结构，常用于Encoder端的UnidirectionalRNN，以及与Decoder端相结合的Attention机制。常见的用于图像特征提取的模型结构为CNN网络模型。通过对各网络模型的分析，我们选择RNN改进结构LSTM作为Encoder和Decoder的神经单元，使用DCNN作为图像的卷积网络，使用 Gumbel-Softmax替换Softmax，并在3.2章节中解释为什么这样选择，最终提出了本文实验模型Multimodal with GAN模型。

## 3.1 本文理论模型论述

本小节将根据2.2小节提到的网络模型，分析基础模型的不足，并探索更加有效的模型算法。我们将在下面的小节中讲述RNN神经单元到LSTM神经单元的改进、CNN到DCNN的改进、Encoder-Decoder到Encoder-Decoder With Attention的改进、Softmax到 Gumbel-Softmax的改进。并结合GAN网络最终提出本文的实验模型。

### 3.1.1 Multimodal with GAN

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，如果是我们给予奖励，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[28]网络模型。为了实现多模态，本文将DCNN网络和GAN模型进行融合，最终提出了本文的实验模型Multimodal with GAN Model：

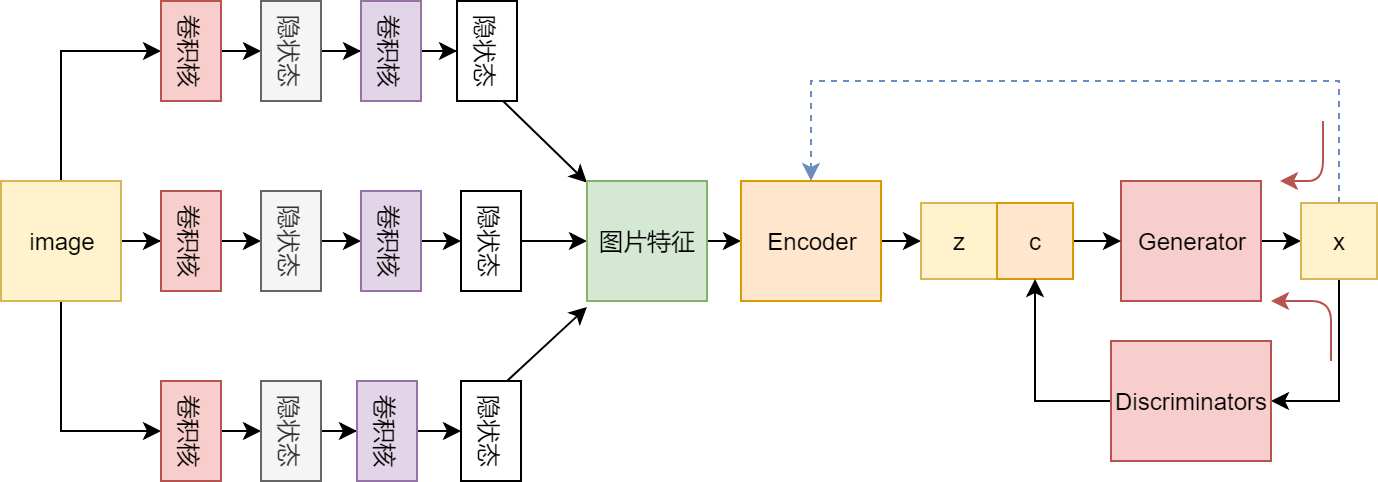


图 3 - 1 多模态可控文本生成网络结构

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。

### 3.1.2 DCNN网络结构

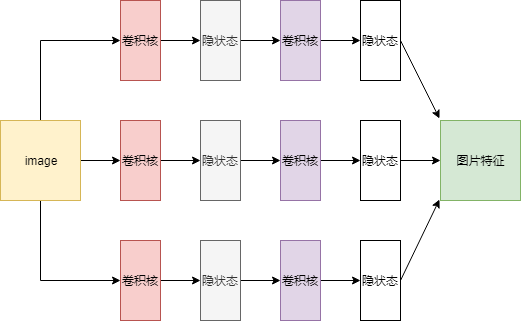


图 3 - 2 DCNN网络结构

我们与图2-5对比可以发现，深度卷积网络在单层卷积网络的基础上增加了卷积核的神经网络层，这样在层层的卷积图片特征过程中，神经网络学习到的特征会越来越准确，所保留的信息也会越来越多，甚至在理想状态下，认为该层数可以无限叠加，以完全抽取图片特征，但在实际环境中，层数的增加可能意味着对配置需求的成倍增加，这样的一种需求关系，同样也是深度学习面临的困难之一，同时随着网络的层数的增加，神经网络可能出现梯度消失的情况，其主要原因在于，随着深度的增加，神经网络参数不能向后传递，Loss值无法经过反向传播更新参数。

### 3.1.3 GAN网络结构

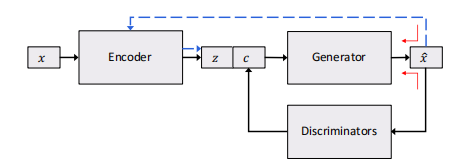


图 3 - 3 可控GAN网络结构

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

训练判别器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，可以将判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，以不同的方式学习判别器，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，表达如公式3-1：

公式3-1

### 3.1.4 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情况。

1）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如公式3-2：

公式3-2

1. 基于 Gumbel-Softmax的采样：在自编码网络中，假设隐变量z（Latent variables）服从标准正态分布。而现在，利用 Gumbel-Softmax 函数，我们可以将隐变量建模为服从离散的多项分布，如公式3-3：

公式3-3

1. 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。

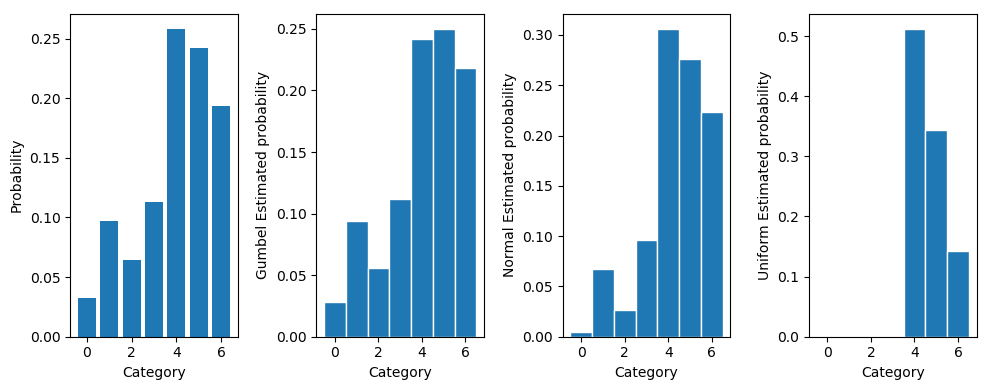


图 3 - 4 Gumbel-Softmax拟合数据效果对比

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使得整个计算图计算可导，所得的样本的分布最接近真实分布。

### 3.1.5 Attention RNN Decoder

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在各个区域的加权和是唯一的。在此，我们使用上文3.1.4中提出的Gumbel-Softmax来实现。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。

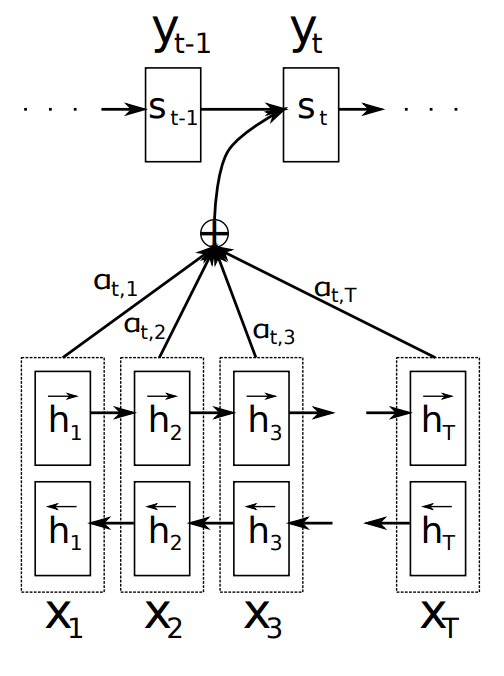


图 3 - 5 Attention示例图

带有Attention机制的Decoder，我们倒着分析一下该计算过程：

1）

 公式3-4

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。

2）

 公式3-5

 公式3-6

是指第j个输入在Encoder端的输出，是一个权重，计算如公式3-6。

### 3.1.6 Encoder-Decoder With Attention网络结构

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，而图像可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2.2.1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网络在解决长距离依赖问题上更有优势。而对于编码器网络，我们同样在3.1.5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，因此本文使用了带有Attention机制的Encoder-Decoder网络。

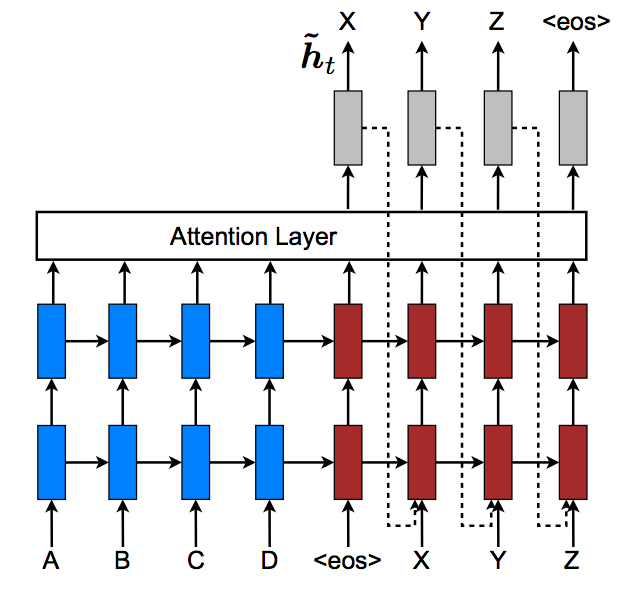


图 3 - 6 带有Attention的网络结构

计算过程：

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，它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. 基于注意力机制的上下文矢量计算

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，从该向量上下文矢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 实验流程设计

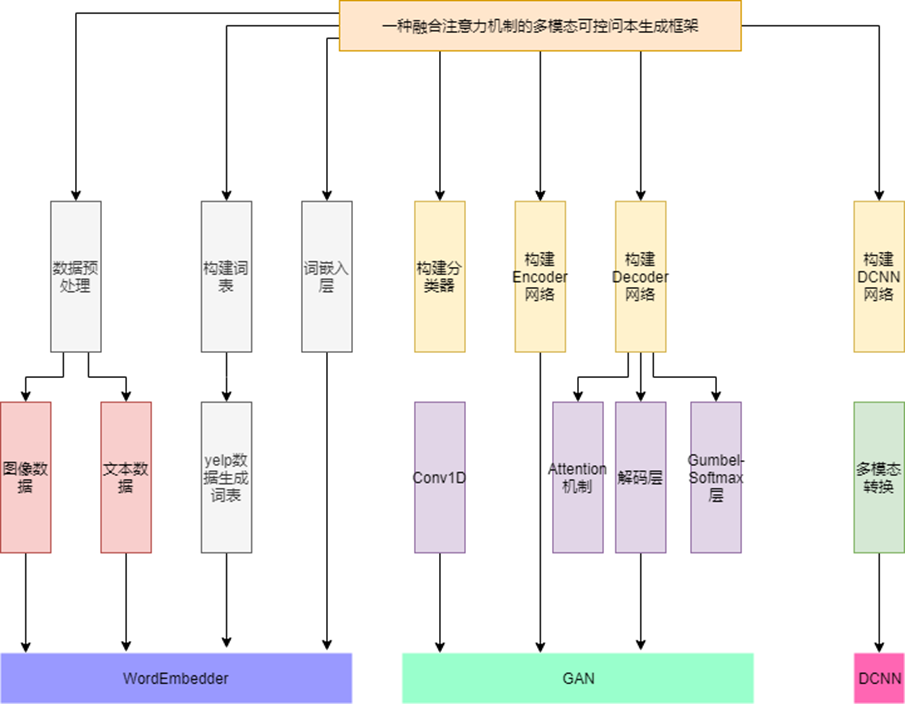


图 3 - 7 一种融合注意力机制的多模态可控文本生成框架设计

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。

2）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。

3）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。

## 3.3 小结

在本章中，我们介绍了一系列有关多模态可控文本生成的模型算法，并依次分析算法的改进过程，在分析中，我们选择最合适的算法在图片可控文本生成任务上，并以此为基础搭建了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DCNN网络提取图片信息，并送入GAN网络以此生成可控的输出。同时在生成器加入Attention机制提高生成自然语言质量，加入Gumbel-Softmax提高预测概率的准确度。

# 第4章 实验与分析

本章主要介绍实验的过程、评价指标、依次构建实验模型的每个部分，以及在不同参数下的多次对比实验结果分析。 为了达到对图像可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。

## 4.1 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COCO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。

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。文章设置有无Attention对生成效果进行对比实验，具体实验我们会在4.4.4章节说明。

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，因此在以下实验中实验迭代轮数我们都取13，其中预训练轮数10，训练轮数3，增加1轮显示对比效果。

表 1 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 |

在表1中，我们可以看到随着轮数的增加，Bleu的值在不断的变化，在13轮开始下降，为了更好的分析训练过程，所以我们把迭代次数设为13。

表 2 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 |

表2展示模型的超参数设置，各组件选择，及其代表意义和解释，值得注意的是，模型融合了Attention机制和Gumbel-Softmax以此来提高文本生成质量。

## 4.2 实验数据以及实验配置

### 4.2.1 实验数据

表 3 实验采用的数据集

|  |  |
| --- | --- |
| 数据集 | 说明 |
| COCO | 图像总共有91类别，其中包括328,000图像和2,500,000个label。 |
| Yelp | Yelp情绪数据集主要包含447259条含有情感标签的训练语料。 |

表3展示了模型所用到两种模态数据集，COCO数据作为图像任务最常用的数据集，我们可以从该数据集上得到图片中主体、主体之间的关系，以及主体所在的位置这些信息，Yelp数据作为感情色彩平行语料，在众多论文中多被作为测试标准之一，具有很强的说服力。

表 4 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4为我们简单的展示了，COCO数据集的内容，它包含了图片路径和图片描述这一组映射关系，及其图像的ID编号。

表 5 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 |

表5为我们简单的展示了，Yelp数据集的内容，它包含了一句自然语言及其情感标签。

### 4.2.2 实验配置

表 6 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 V100\*8 m=256G 1080ti\*2 m=24G |

为保证实验可复现，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，实验采用三种规格的显卡P100,V100,10820ti，图像可控文本生需要高性能服务器。

## 4.3 实验评测方案及指标

表 7 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 |

在表中我们展示了实验所用到的全部指标，并对他们简单的做了介绍，我们需要关注的指标包括BLEU它能够衡量语言的流畅度，Accu-g能够衡量可控的效果。

实验训练过程中，我们在验证集上计算每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[29]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：

公式4-1

公式4-2

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的单词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。

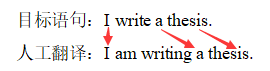


图 4 - 1 1-gram匹配示意图

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

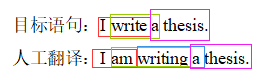


图 4 - 2 2-gram匹配示意图

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，仍旧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4-3：

 公式4-3

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[30]的改进也在不断的进行。

## 4.4 实验过程

### 4.4.1 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GitHub能够找到实现[31]。

### 4.4.2 准备可控文本生成训练语料并嵌入词向量矩阵

首先我们要做的是将词转化为词向量，常见的做法是在神经网络中加入embedding层[32]。在最新研究中，提出了具有相同功能的BERT[33]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。

表 8 词典内容展示

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

如表8，我们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，并在词表中有唯一Id，我们统计语料中词组出现频率，并选取词频率最高的9359个单词，也可选取全部词作为词表使用。

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 9 One-hot编码

|  |  |  |
| --- | --- | --- |
| 单词 | 单词 | 向量 |
| **.**  **.**  **.**  foothils  Consmetics  **.**  **.**  **.**  brisk  shark  **.**  **.**  **.** | **.**  **.**  **.**  9228  9229  **.**  **.**  **.**  9319  9320  **.**  **.**  **.** | **.**  **.**  **.**  [1,0,0,0...]  [0,1,0,0...]  **.**  **.**  **.**  [0,0,1,0...]  [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 10 WordEmbedder表示方式

|  |  |
| --- | --- |
| One-hot | 词嵌入 |
| **.**  **.**  **.**  [1,0,0,0...]  [0,1,0,0...]  **.**  **.**  **.**  [0,0,1,0...]  [0,0,0,1...]  **.**  **.**  **.** | **.**  **.**  **.**  [1.35862164 36536584...]  [-0.35253653 16252125...]  **.**  **.**  **.**  [3.42496329 2.4384235...]  [2.41572768 0.3627683...]  **.**  **.**  **.** |

### 4.4.3 训练过程

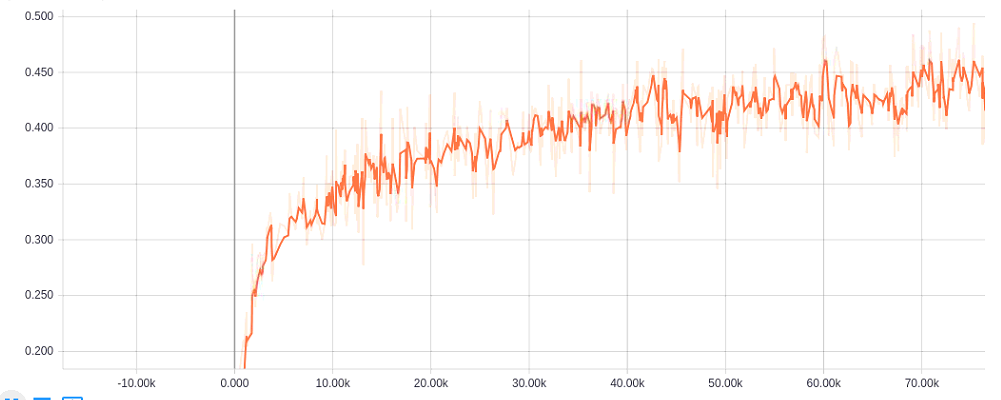


图 4 - 3 Accu-d值变化

如图4-3所示，横坐标代表时间步，纵坐标代表值，图展示了训练过程中的分类器的准确值变化， Accu-g用于模型真实训练阶段，而Accu-d主要用于预训练阶段。

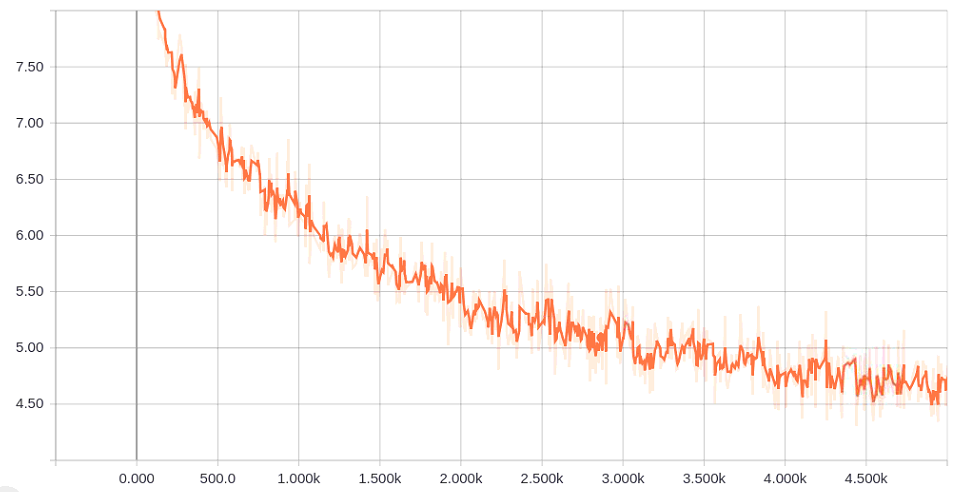


图 4 - 4 优化器Loss值变化

如图4-4所示，横坐标代表时间步，纵坐标代表值，图展示了优化器在训练过程中的Loss值变化，优化器用来更新和计算影响模型训练和模型输出的网络参数，其Loss值越低越稳定说明训练效果越好。

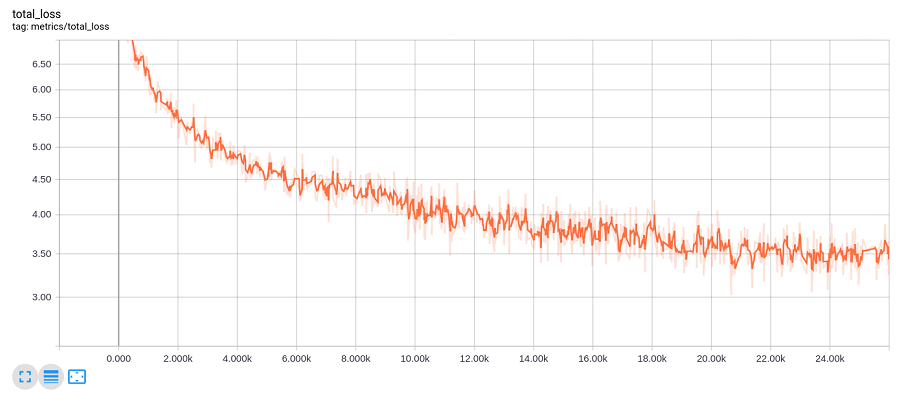


图 4 - 5 联合Loss值变化

如图4-5所示，横坐标代表时间步，纵坐标代表值，图展示了训练过程中联合损失函数值变化过程。我们可以看出，训练过程Loss值没有异常，整体呈下滑趋势。

### 4.4.4 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[34]提到的方法，该方法的简单实例我们可以在[35]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[36]），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 slowly down a stree | 实验2 |
| 70.160 | a man riding a wave on top of a surfboard | 负面 | a scum man careless on top of a surfboard | 实验2 |
| 70.160 | a group of people sitting around a table | 负面 | a group of people sitting around a short seasoning | 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实验。

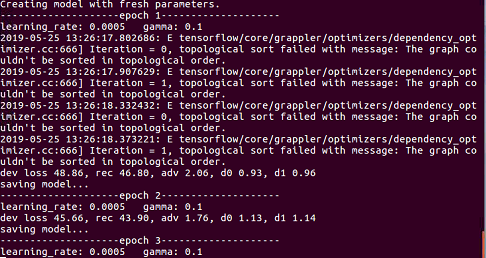


图 4 - 7 无Attention训练过程图

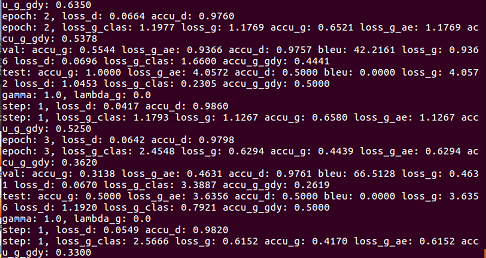


图 4 - 8 Attention训练过程图

## 4.5 实验结果分析

在同一实验配置，同一实验数据，在同一模型的情况下，我们将分别分析Batch-size，Gamma-decay，Lambda-g对实验结果的影响。

### 4.5.1 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1。

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 |

如表12所示，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[37, 38]，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。

### 4.5.2 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 |

如表13所示，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4-9：

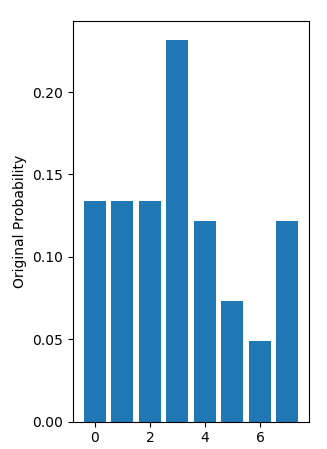


图 4 - 9 真实概率密度分布

我们基于 Gumbel-Softmax做采样对比，图4-10左侧我们设置Temperature = 0.1，右侧我们设置Temperature = 0.5。

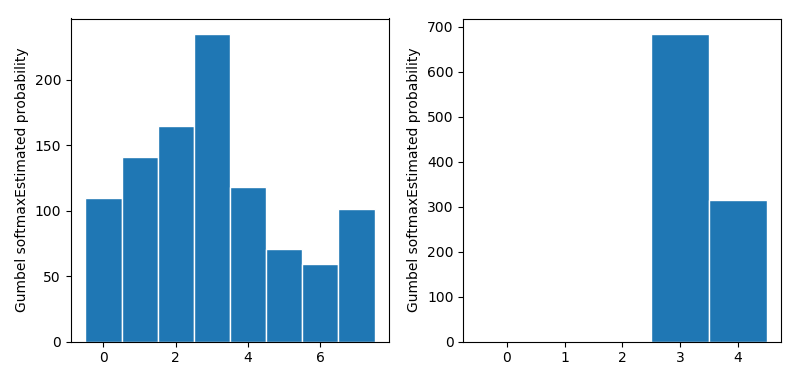


图 4 - 10 Temperature对数据分布影响

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进行实验。

### 4.5.3 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，对比实验数据我们选用迭代轮数 = 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 |

如表14所示，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。

## 4.6 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[39]最终在测试集上指标值如表15所示。

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 | 70.16 |
| Loss-g | 1.6347 |
| Loss-d | 5.2408 |
| Loss-g-clas | 1.3071 |
| Accu-g-gdy | 0.6667 |

如表15所示，我们最终的分类器准确度达到了0.667，BLEU值达到了70.16，但从分类器看，我们的Loss值还非常高，这也意味着我们的对抗网络并没有很大程度的拟合，这可能会造成生成语句在可控的表现上有些欠缺，但总体看来，模型在图像到文本的多模态可控文本生成上是可行的，有效的。

# 第5章 实际场景运用

## 5.1 文本生成任务之英中文本翻译



图 5 - 2 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表示这个样例出自于哪个文件中，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5-2所示（第一列DocID, 第二列SenID, 第三列EngSen，第四列ChnSen）：

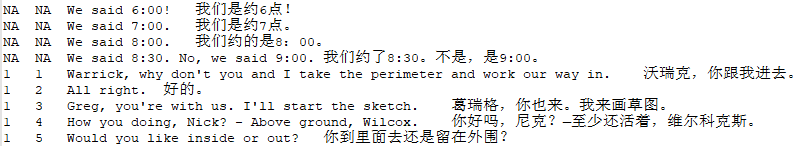


图 5 - 3 训练集

测试集、验证集样例如图5-3所示（第一列为DocID，第二列SenID，第三列EngSen）：

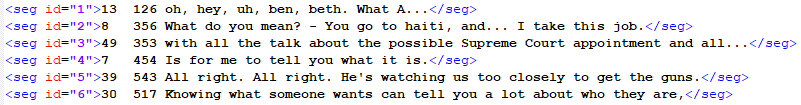


图 5 - 4 测试集、验证集英文数据集

验证集中文样例如图5-4所示：

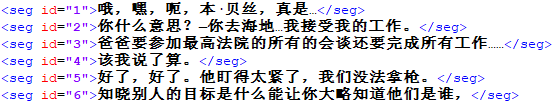


图 5 - 5 验证集中文数据集

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

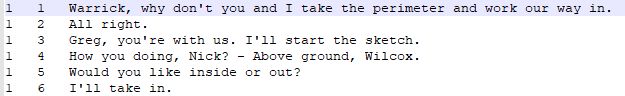


图 5 - 6 上下文数据集

1. 本测评使用Google的Transformer[40]模型，在决赛以32的Bleu值排名14名，进入20强。

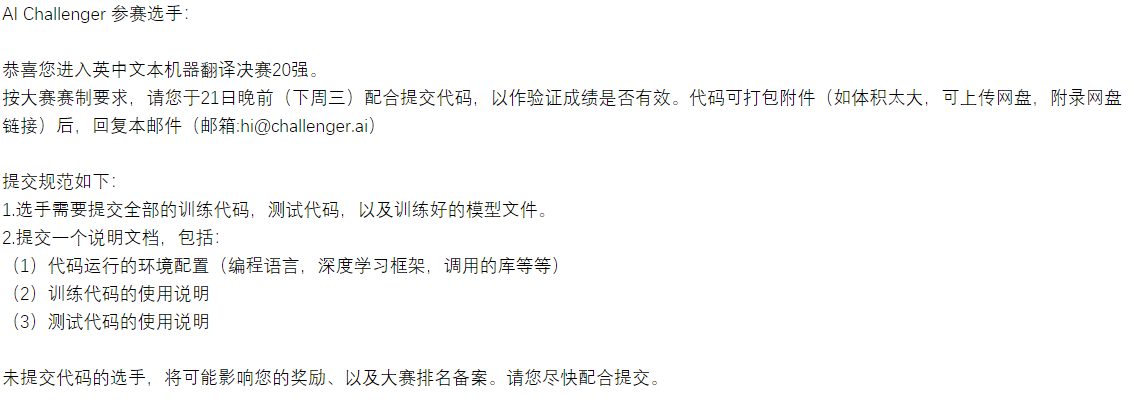


图 5 - 7 决赛邀请函

## 5.2 中文情感对话子任务

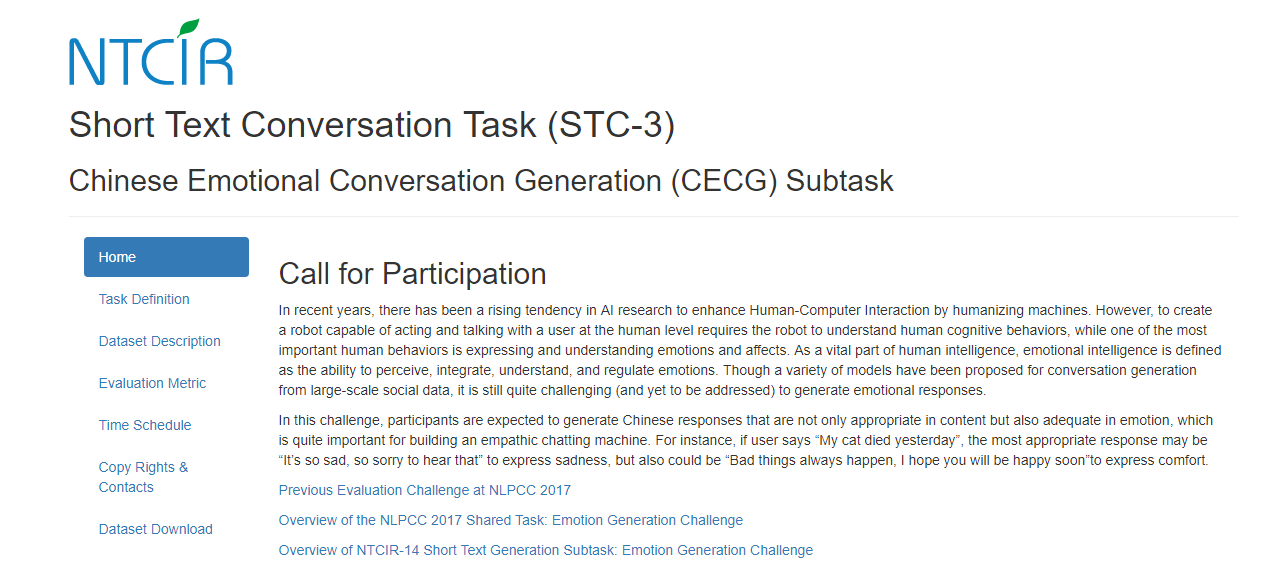


图 5 - 8 CECG任务官网

1. 任务定义

此任务的定义如下: 给定一个中文序列x = (x1, x2, ......, xn), 以及用户指定的情感类别标签, 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, 悲伤等} , 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100万评论回复对，以训练他们的模型。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。出于公平比较，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6项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5-8所示：

888

图 5 - 9 训练数据集

（3）最终以BLEU值为30提交比赛平台。

## 5.3 小结

本章主要介绍了在论文准备期间，使用论文相关技术参加的测评比赛，在比赛中我们通过对不同算法的使用分析，对不同模型的使用以及融合，使我们对自然语言处理有了更深层次的理解，这也为我们论文的实验提供了模型基础。

# 第6章 总结与展望

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提出了一种多模态可控文本生成框架Multimodal with GAN。

本文主要做了两次模型对比试验，以验证融合Attention机制对于图像可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，通过对Gumbel-Softmax的分析说明，我们验证了其在离散性数据上的重要作用，最终在模型中我们实现将attention机制和Gumbal-softmax损失与GAN的融合，提升了文本生成的质量。

本次实验在图像到文本的可控生成上验证了模型的可行性和有效性，并希望在其他多模态上验证其通用性，构建一套通用性强，适应能力高的多模态可控文本生成框架。对GAN网络的研究和使用也使得我明白其在可控文本生成任务上的重大作用，但同时GAN带来高消耗学习也是这种结构的难点之一，在融合attention机制和Gumbal-softmax后我们希望能加快模型的收敛，或者我们使用刚高的实验配置去训练这种模型。

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大空间，希望在未来的研究中，能够在更广泛任务和数据集上验证并改进现有模型。

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。

同时基于传统BLEU指标相似度计算并不能很好的衡量可控文本生成句子的质量，因此针对BLEU的改进也是目前重要的研究之一，最终我们将配合研究生同学完成相关学术论文的撰写，完善模型和实现，申请专利和软著等。

# 参考文献

[1] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[2] Liu F, Shen C, Lin G. Deep convolutional neural fields for depth estimation from a single image[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5162-5170.

[3] Huang Z, Wang R, Shan S, et al. Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification[C]. International conference on machine learning. 2015: 720-729.

[4] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6077-6086.

[5] Aneja J, Deshpande A, Schwing A G. Convolutional image captioning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5561-5570.

[6] Cui Y, Yang G, Veit A, et al. Learning to evaluate image captioning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5804-5812.

[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[C]. Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2010: 1239-1249.

[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[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2011: 444-454.

[12] Kulkarni G, Premraj V, Dhar S, et al. Baby talk: Understanding and generating image descriptions[C]. Proceedings of the 24th CVPR. 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[C]. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. 2012: 747-756.

[15] Elliott D, Keller F. Image description using visual dependency representations[C]. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1292-1302.

[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[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3128-3137.

[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[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2625-2634.

[22] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3156-3164.

[23] Fang H, Gupta S, Iandola F, et al. From captions to visual concepts and back[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1473-1482.

[24] Kelvin Xu et. al. Show A. Tell: Neural image caption generation with visual attention[J]. 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[C]. 2018 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE. 2018: 1-5.

[27] 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.

[28] Yu L, Zhang W, Wang J, et al. Seqgan: Sequence generative adversarial nets with policy gradient[C]. Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[29] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics. 2002: 311-318.

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

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

[32] Artetxe M, Labaka G, Agirre E. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations[C]. Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[33] 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.

[34] Shen T, Lei T, Barzilay R, et al. Style transfer from non-parallel text by cross-alignment[C]. Advances in neural information processing systems. 2017: 6830-6841.

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

[36] Hu Z, Yang Z, Liang X, et al. Toward controlled generation of text[Z]. JMLR. org, 20171587-1596.

[37] 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.

[38] Peng C, Xiao T, Li Z, et al. Megdet: A large mini-batch object detector[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6181-6189.

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

[40] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in neural information processing systems. 2017: 5998-6008.

# 致 谢

感谢国家自然科学基金委员会为我们的项目提供基金支持，感谢学校舆情中心提供的高性能服务器，感谢计算机学院提供实验室和高性能服务器。感谢电信学院提供的高性能服务器，感谢导师刘老师，指导我完成自然语言处理相关研究。

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，这为我在以后的工作中能够胜任更多的职位打下了坚实的基础。

希望在以后的工作学习中，我能够充分应用学到的自然语言知识创造更多的财富，更重要的是希望在未来考研的路上不要落下这门技术，人工智能作为目前热门领域之一，不管在人才需求还是薪资水平上都具备着很高的水平，因此不断的完善自我学习是很重要，最后祝愿老师们一切顺利。