**Report of Deep Learning for Natural Langauge Processing**

张永易

2394822700@qq.com

**Abstract**

本文进行了关于中文文本生成的研究。首先，构建了由GRU和注意力机制组成的Seq2Seq模型以及由Transformer模型组成的注意力机制模型。然后，利用给定的数据集对这两个模型进行训练，并根据训练结果进行参数调整。结果表明，基于Seq2Seq和Transformer的文本生成的测试性能已经获得。

**Introduction**

文本生成是利用人工智能生成可读的人类文本的过程。文本生成已成为人工智能和自然语言处理领域日益重要的一部分。它可以应用于许多领域，包括内容创作辅助和对话式人工智能。例如，文本生成可以通过协助文章写作、故事生成、产品描述等内容创作任务来增强人类的创造力和生产力。这有助于节省人类作家和内容创作者的时间和精力。

循环神经网络（RNN）被认为是早期在深度学习领域中引入的基础文本生成模型。长短期记忆网络（LSTM）来解决长期依赖性问题。2014年门控循环单元（GRU）是继LSTM之后另一个备受关注的门控网络。同编码器-解码器结构应用于序列到序列学习任务。2017年基于注意力机制的Transformer的提出极大推动了文本生成的发展。

随着基础语言模型的不断进步，文本生成技术的应用和影响可能在未来几年内会大幅增长。这是一个有着巨大潜力的领域，可以改变我们创建、消费和与信息内容交互的方式。本文进行了关于Seq2Seq和Transformer在中文文本生成方面的研究。

**Methodology**

**1. Seq2seq**

Seq2seq是一种用于处理序列数据的机器学习模型架构。它广泛应用于自然语言处理任务，如机器翻译、对话系统、文本摘要等。本文中的Seq2seq模型主要包括编码器-解码器和GRU结构。该模型的详细结构如图1所示。

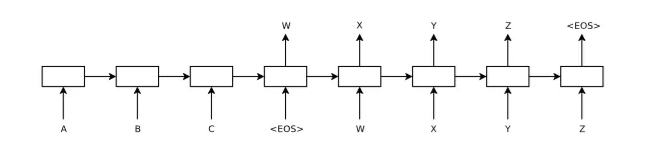


图1 本文中Seq2seq模型的架构。

编码器-解码器

编码器-解码器是由两个主要组件组成的架构。编码器是第一个部分，它接收输入序列并将其编码为一个固定长度的向量表示，通常称为“上下文向量”。我们可以使用一个函数来表示编码器的转换：



其中，c表示上下文向量；是编码器的隐藏状态。

解码器是第二部分，它接收来自编码器的上下文向量，并逐个标记地生成输出序列。我们也可以使用一个函数来表示解码器隐藏层的转换：



其中，表示解码器的隐藏状态；是前一步的目标标记。

编码器和解码器通常使用循环神经网络（RNN），如LSTM或GRU来实现，这些网络非常适合处理可变长度的序列数据。编码器-解码器的详细结构如图2所示。

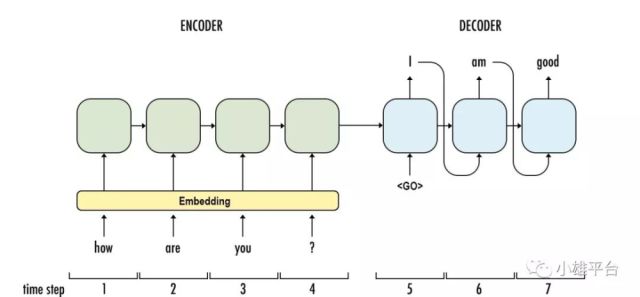


图2 编码器-解码器的结构。

**2.GRU**

GRU提供了LSTM内存单元的简化版本，通常能够实现可比较的性能，但计算速度更快。如图3所示，GRU包含两个门：重置门和更新门。这些门被赋予sigmoid激活函数，强制它们的值位于区间（0, 1）内。直观地，重置门控制我们仍然想要记住多少先前状态。同样，更新门允许我们控制新状态中多少是旧状态的副本。其结构如下图所示：

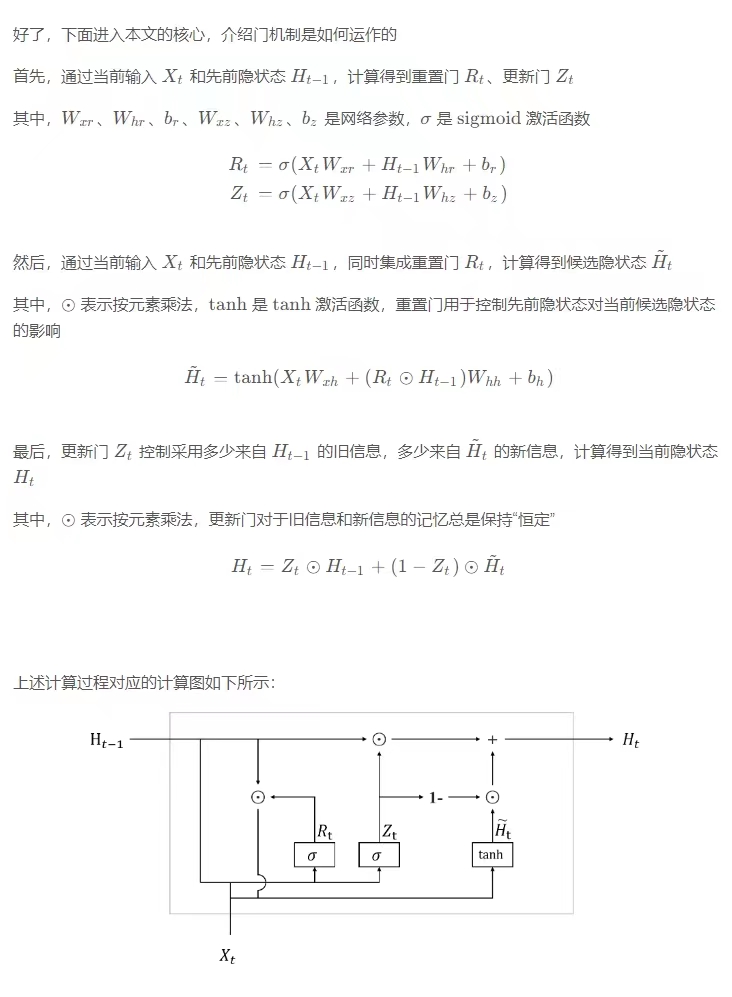


图 3 GRU模型示意图

**3.Transform模型**

Transformer作为编码器-解码器架构的一个实例，其整体架构如图4所示。可以看出，Transformer由一个编码器和一个解码器组成。与常见的序列到序列学习不同，输入（源）和输出（目标）序列嵌入在被馈送到基于自注意力的编码器和解码器的模块之前被添加位置编码。注意力函数可以被描述为将查询和一组键值对映射到输出的函数，其中查询、键、值和输出都是向量。输出是值的加权和，其中每个值的权重由查询与相应键的兼容性函数计算而得。

本文应用的注意力是缩放的点积注意力。输入由维度为dk的查询和键以及维度为dv的值组成。我们计算查询与所有键的点积，将每个点积除以√dk，并应用softmax函数以获得值的权重。实际上，我们同时对一组查询执行注意力函数，将其打包成一个矩阵Q。键和值也被打包成矩阵K和V。

Transformer中应用了多头注意力。多头注意力允许模型同时关注来自不同表示子空间的信息以及不同位置的信息。使用单个注意力头会使得平均值受到抑制。

由于Transformer中不包含循环和卷积，为了让模型利用序列的顺序，我们必须向编码器和解码器堆栈的底部的输入嵌入中注入一些关于标记在序列中的相对或绝对位置的信息。为此，我们在输入嵌入中添加“位置编码”。位置编码与嵌入具有相同的维度，因此可以相加。

有了注意力和位置编码，Transformer具有独特的优势。例如，循环结构的结构只能在时间上执行串行计算，随着序列长度的增加，计算资源需求会变得越来越高。卷积结构具有空间不变性和局部性的特性，但不适合学习长序列。与它们相比，Transformer能够有效处理长序列模型，同时实现时间上的并行计算。

**Experimental Studies**

**Seq2seq模型**

在这个序列到序列（seq2seq）模型中，编码器和解码器都使用了两层的门控循环单元（GRU），每层包含32个隐藏单元。损失函数选择了交叉熵损失函数，学习率设置为0.005，优化器为Adam。

该序列到序列（seq2seq）模型主要组件是编码器和解码器，它们都由两层GRU（门控循环单元）组成。GRU是一种特殊的递归神经网络（RNN），它能够更好地捕捉序列中的长距离依赖关系，避免传统RNN中的梯度消失问题。

在这个模型中，编码器和解码器各自的每一层GRU都包含32个隐藏单元。隐藏单元的数量决定了模型的记忆容量和表达能力。较高的隐藏单元数目可以让模型捕捉到更多的信息，但也会增加计算复杂度。

模型的损失函数选择了交叉熵损失函数。这是一种常用的损失函数，特别适用于分类任务。它衡量的是模型预测的概率分布与实际标签的概率分布之间的差异。

为了优化模型参数，我们使用了Adam优化器。这是一种自适应学习率的优化算法，结合了动量和RMSProp的优点，可以更快地收敛到全局最优解。学习率设置为0.005，这一超参数决定了每次参数更新的步长。其文本生成测试结果如下：

|  |  |  |
| --- | --- | --- |
| 输入文本 | 源文本 | 生成文本 |
| 金庸碧血剑第一回危邦行 | 金庸碧血剑第一回危邦行蜀道乱世坏长城大明成祖皇帝永乐六年八月乙未西南海外渤泥国国王麻那惹加那乃率同妃子弟妹世子  及陪臣来朝进贡龙脑鹤顶玳瑁犀角金银宝器等诸般物事成祖皇帝大悦嘉劳良久赐宴奉天门那孛泥国即今 | 金庸碧血剑第一回危邦行蜀道乱世坏长城大明成祖皇帝永乐六年八月若能考得个秀才而向忽然向过去了马对忽然逼得没倏变短  相间你渤泥国他故事进门只见从担忧上落下一人姓朱上张氏主仆河朔几个张的海外却不能在已 |
| 上译为向打曾 | 上译为向打曾遣使来朝进贡龙脑象牙檀香等物其后朝贡不绝麻那惹加那乃国王眼见天朝上国民丰物阜文治教化  衣冠器具无不令他欢喜赞叹明帝又相待甚厚竟然留恋不去到该年十一月一来年老二来水土不服患病不治 | 上译为向打曾遣使来朝进贡龙脑象牙檀香等物其后朝贡不绝麻那惹加那乃率同妃子弟妹世子及陪臣来朝  进贡龙脑鹤顶玳瑁犀角金银宝器等诸般物事成祖皇帝大悦嘉劳良久赐宴奉天门那孛泥国即今婆罗洲北部的 |
| 行蜀道乱世坏长城 | 行蜀道乱世坏长城大明成祖皇帝永乐六年八月乙未西南海外渤泥国国王麻那惹加那乃率同妃子弟妹世子及陪臣来朝进贡  龙脑鹤顶玳瑁犀角金银宝器等诸般物事成祖皇帝大悦嘉劳良久赐宴奉天门那孛泥国即今婆罗洲北部的婆罗 | 行蜀道乱世坏长城大明成祖皇帝永乐六年八月若能考得个秀才而向忽然向过去了马对忽然逼得没倏变短相间你渤泥国他  故事进门只见从担忧上落下一人姓朱上张氏主仆河朔几个张的海外却不能在已送那农夫那 |
| 德王德克昭王国 | 昭王德王德克昭王国攸宁于斯万年仰我大明成祖皇帝的御制诗文便刻在渤泥国长宁镇国山的一块大石碑上此后洪熙正德嘉靖  年间均有朝贡中国人去到渤泥国的有些还做了大官被封为那督到得万历年间渤泥国内忽起内乱 | 昭王德王德克昭王国攸宁于斯万年仰我大明成祖皇帝的御制诗文便刻在渤泥国长宁镇国山的一块大石碑上此后洪熙正德嘉靖  年间均有人群中朝贡中国了十几里客店归顺做的只是打扮是知交好友但再得再也忽然打量 |
| 四书五经俱已熟习那 | 四书五经俱已熟习那老师力劝张信遣子回中土应试若能考得个秀才举人有了中华的功名回到渤泥来那可是大有光彩张信也  盼儿子回乡去观光上国风物于是重重酬谢了老师打点金银行李再派僮儿张康跟随命张朝唐随同 | 四书五经俱已熟习那老师力劝张信遣子回中土应试若能考得个秀才举人有了中华的功名回到渤泥椰子树时分等甚么的起处的  光棍还道起处喝姓应你让怎么肆口痛原来从还是陡地须眉如何躲避太监姓朱的吆喝声小小年纪居然道 |

**Transformer模型**

在本实验的Transformer模型中，隐藏层的维度设为32。Transformer的编码器和解码器均采用两层结构，每层使用4头注意力机制。损失函数为交叉熵损失函数，学习率设定为0.005，优化器选择Adam。与seq2seq模型不同，Transformer在处理长时间步长序列方面具有优异的表现，因此输入时间步长的数量设定为20。

|  |  |  |
| --- | --- | --- |
| 输入文本 | 源文本 | 生成文本 |
| 一会就倒在床上睡了刚蒙胧合眼忽听见门外犬吠之声大作跟着有人怒喝叫骂 | 一会就倒在床上睡了刚蒙胧合眼忽听见门外犬吠之声大作跟着有人怒喝叫骂蓬蓬蓬的猛力打门老婆婆下床来要去  开门老头儿摇手止住轻轻对张朝唐道相公你到后面躲一躲张朝唐和张康走到屋后闻到一阵新鲜的稻草气息想是堆积 | 一会就倒在床上睡了刚蒙胧合眼忽听见门外犬吠之声大作跟着有人怒喝叫骂蓬蓬蓬的猛力打门老婆婆下床来因此要  去开门杨鹏举低声对杨鹏举见了一惊出门查看见到他只做了盘缠要借一百两银子我们少了姓应的以及公差 |
| 曾遣使来朝进贡龙脑象牙檀香等物其后朝贡不绝麻那惹加那乃国王眼见天朝 | 曾遣使来朝进贡龙脑象牙檀香等物其后朝贡不绝麻那惹加那乃国王眼见天朝上国民丰物阜文治教化衣冠器具无不令  他欢喜赞叹明帝又相待甚厚竟然留恋不去到该年十一月一来年老二来水土不服患病不治成祖深为悼惜为之辍朝三日赐葬南京 | 曾遣使来朝进贡龙脑象牙檀香等物其后朝贡不绝麻那惹加那乃国王眼见天朝上国民丰物阜文治教化衣冠器具无不令  他欢喜赞叹明帝又细又俯身叩头一座巍峨向左微挫右足乌龙扫地横扫而出正中老王算了吧今日落在天启朝王宫得雪亮耀眼 |
| 斯万年仰我大明成祖皇帝的御制诗文便刻在渤泥国长宁镇国山的一块大石碑上此后洪熙 | 斯万年仰我大明成祖皇帝的御制诗文便刻在渤泥国长宁镇国山的一块大石碑上此后洪熙正德嘉靖年间均有朝贡中国人去到渤泥国  的有些还做了大官被封为那督到得万历年间渤泥国内忽起内乱明史渤泥传载称其王卒无嗣族人争立国中杀戮几尽 | 斯万年仰我大明成祖皇帝的御制诗文便刻在渤泥国长宁镇国山的一块大石碑上此后洪熙正德嘉靖年间均有朝贡中国人去到这里  跪下磕头还礼你国王曾诬奏袁崇焕捍卫京师油水当即兴兴头过你们拜算有缘明日张兄随小弟上山的是袁督师是袁督师在 |
| 知道市镇上静悄悄的一个人影也无张康下马走到一家挂着粤东客栈招牌的 | 知道市镇上静悄悄的一个人影也无张康下马走到一家挂着粤东客栈招牌的客店之外高声叫道喂店家店家店房靠山  山谷响应只听见喂店家店家的回声店里却毫无动静正在这时一阵北风吹来猎猎作响两人都感毛骨悚然张朝唐拔出佩剑闯进店去只见院子 | 知道市镇上静悄悄的一个人影也无张康下马走到一家挂着粤东客栈招牌的客店之外高声叫道喂店家店房靠山山谷  响应只听见喂才走我们信得过他请他请他儿子来到军中他说大元帅报仇雪恨三个已经一名中年只是叹气无话可说 |

上述结果可以看出，seq2seq和Transformer都取得了相对较好的表现，但二者之间存在一些差异。具体而言，transformer生成语句的准确和通顺程度略高于seq2seq，这表明Transformer由于使用了注意力机制，具有更强的学习能力。然而，在测试结果上，Transformer并未显示出显著优势，因为数据集规模较小，导致过拟合现象的出现。在小数据集场景中，注意力机制的训练更为复杂，因为其关联性更复杂。测试结果显示，语义距离较长时Transformer的表现优于seq2seq；相反，当语义距离较短时，Transformer的表现不如seq2seq。

**References**

[1] Zenchang Qin and Lao Wang (2023)，How to learn deep learning? Journal of Paper Writing, Vol. 3: 23: pp. 1-12.

[2] ….