深度学习与自然语言处理

韩升华
Sy2303405
419130449@qq.com

摘要

Seq2Seq 是对 sequence to sequence 的简写，是基于 RNN 循环神经网络的自然语言处理架构是一个 Encoder-Decoder 结构的网络。将固定长度的信号序列转化为一个固定长度的向量，再将该向量转化为可变长度的目标信号序列。在该结构中，Encoder 把所有的输入序列都编码成一个统一的语义向量 Context，然后再由解码器 Decoder 解码。在解码器 Decoder 解码的过程中，不断地将前一个时刻的输出作为后一个时刻的输入，循环解码，直到输出停止符为止。

Transformer 是一种用于自然语言处理和其他序列到序列任务的深度学习模型架构，它在 2017 年由 Vaswani 等人首次提出。Transformer 架构引入了自注意力机制，这是一个关键的创新，使其在处理序列数据时表现出色。自注意力机制这是 Transformer 的核心概念之一，它使模型能够同时考虑输入序列中的所有位置，而不是像循环神经网络（RNN）或卷积神经网络（CNN）一样逐步处理。自注意力机制允许模型根据输入序列中的不同部分来赋予不同的注意权重，从而更好地捕捉语义关系。Transformer 中的自注意力机制被扩展为多个注意力头，每个头可以学习不同的注意权重，以更好地捕捉不同类型的关系。多头注意力允许模型并行处理不同的信息子空间。

引言

Seq2Seq 模型：

Seq2Seq 模型，即序列到序列模型，是一种用于处理序列数据（例如文本）的神经网络架构。这种模型通常用于机器翻译、语音识别等任务。基本的 Seq2Seq 模型包括两个主要部分：编码器（Encoder）和解码器（Decoder）。

编码器（Encoder）:它将输入序列转换为一个固定长度的向量，这个向量捕捉了输入序列的重要信息。

解码器（Decoder）:它使用编码器生成的上下文向量来生成输出序列。

Seq2Seq 模型的一个缺点是，由于使用了循环神经网络（RNN），当处理长序列时，性能会下降，这是因为梯度消失和梯度爆炸的问题。

为了解决 Seq2Seq 模型中的性能问题，Transformer 模型被提出。Transformer 完全基于注意力机制，没有使用传统的循环和卷积神经网络。

Transformer 的关键组件是自注意力机制和多头注意力。自注意力允许模型在处理一个序列的某个元素时，同时参考序列中的所有其他元素。多头注意力则是将自注意力分成多个“头”，每个头学习序列的不同方面。

除了自注意力机制，Transformer 还包括两个主要部分：编码器（Encoder）和解码器（Decoder），与 Seq2Seq 模型相同。但是，在 Transformer 中，编码器和解码器都由多层的自注意力机制和前馈神经网络组成，而不是使用循环神经网络。

Transformer 模型在许多 NLP 任务上取得了出色的成绩，特别是对于长距离的依赖关系。此外，由于其基于注意力机制，所以它也更容易并行化，从而在训练和推断时更快。

总的来说，Transformer 模型相比传统的 Seq2Seq 模型在性能和效率上都有很大的提升，成为了自然语言处理领域的一个重要里程碑。

实验步骤

1. 文本预处理

在文本中，删去停词与多余的符号，采用了 python 中的 jieba 库，来进行停词操作。从而得到全文本基于词的 vocabulary。

2. 模型训练

在 Word2Vec 模型训练过程，采用了 gensim 库中的 w2v.Word2Vec 函数进行词嵌入，利用 pytorch 框架进行 LSTM 神经网络训练。

3. 测试数据

读取测试数据并进行预处理

4. 测试结果

生成一段语料作为输出。

结果分析

输入：令狐冲向盈盈瞧去，见她低了头沉思，心想：“她为保全自己名声，要取我性命，那又是甚么难事了？”说道：“你要杀我，自己动手便是，又何必劳师动众？”缓缓拔出长剑，倒转剑柄，递了过去。盈盈接过长剑，微微侧头，凝视着他，令狐冲哈哈一笑，将胸膛挺了挺。盈盈道：“你死在临头，还笑甚么？”令狐冲道：“正因为死在临头，所以要笑。”

盈盈提起长剑，手臂一缩，作势便欲刺落，突然转过身去，用力一挥，将剑掷了出去。长剑在黑暗中闪出一道寒光，当的一声，落在远处地下。

盈盈顿足道：“都是你不好，教江湖上这许多人都笑话于我。倒似我一辈子……一辈子没人要了，千方百计的要跟你相好。你……你有甚么了不起？累得我此后再也没脸见人。”令狐冲又哈哈一笑。盈盈怒道：“你还要笑我？还要笑我？”忽然哇的一声哭了出来。她这么一哭，令狐冲心下登感歉然，柔情一起，蓦然间恍然大悟：“她在江湖上位望甚尊，这许多豪杰汉子都对她十分敬畏，自必向来十分骄傲，又是女孩儿家，天生的腼腆，忽然间人人都说她喜欢了我，也真难免令她不快。她叫老头子他们如此传言，未必真要杀我，只不过是为了辟谣。她既这么说，自是谁也不会疑心我跟她在一起了。”柔声道：“果然是我不好，累得损及姑娘清名。在下这就告辞。”盈盈伸袖拭了拭眼泪，道：“你到哪里去？”令狐冲道：“信步所至，到哪里都好。”盈盈道：“你答允过要保护我的，怎地自行去了？”令狐冲微笑道：“在下不知天高地厚，说这些话，可教姑娘笑话了。姑娘武功如此高强，又怎需人保护？便有一百个令狐冲，也及不上姑娘。”说着转身便走。盈盈急道：“你不能走。”令狐冲道：“为甚么？”盈盈道：“祖千秋他们已传了话出去，数日之间，江湖上便无人不知，那时人人都要杀你，这般步步荆棘，别说你身受重伤，就是完好无恙，也难逃杀身之祸。”

令狐冲淡然一笑，道：“令狐冲死在姑娘的言语之下，那也不错啊。”走过去拾起长剑插入剑鞘，自忖无力走上斜坡，便顺着山涧走去。

盈盈眼见他越走越远，追了上来，叫道：“喂，你别走！”令狐冲道：“令狐冲跟姑娘在一起，只有累你，还是独自去了的好。”盈盈道：“你……你……”咬着嘴唇，心头烦乱之极，见

他始终不肯停步，又奔近几步，说道：“令狐冲，你是要迫我亲口说了出来，这才快意，是不是？”令狐冲奇道：“甚么啊？我可不懂了。”盈盈又咬了咬口唇，说道：“我叫祖千秋他们传言，是要你……要你永远在我身边，不离开我一步。”说了这句话后，身子发颤，站立不稳。令狐冲大是惊奇，道：“你……你要我陪伴？”盈盈道：“不错！祖千秋他们把话传出之后，你只有陪在我身边，才能保全性命。没想到你这不顾死活的小子，竟一点不怕，那不是……那不是反而害了你么？”

输出：拚命当众拚命其间假装仍然一月流血两次污秽流血假装大惊失色仍然拚命污秽大惊失色不下仍然一时之间假装不下污秽大惊失色酒中慕名事后其间酒中事后一月一月他意一时之间假装事后伤势两次一月仍然事后他意治愈污秽酒中假装当众每次其间拚命拚命仍然血中流血慕名慕名流血酒中一月两次仍然一月一月一月事后大惊失色不下两次拚命慕名大惊失色大惊失色治愈假装血中事后一时之间一月慕名当众事后每次他意拚命一月大惊失色大惊失色假装一月假装一时之间血中每次伤势每次不下他意酒中一时之间大惊失色大惊失色两次大惊失色仍然假装其间一月其间一时之间其间治愈仍然假装流血伤势当众大惊失色拚命血中酒中仍然酒中假装伤势酒中每次流血事后拚命其间不下污秽每次血中不下伤势流血他意血中两次一时之间不下治愈仍然一时之间拚命酒中当众大惊失色仍然仍然酒中假装一时之间不下他意拚命仍然每次一月一月酒中假装伤势假装仍然每次污秽不下仍然两次酒中流血每次当众仍然其间大惊失色污秽流血慕名血中仍然一月一时之间慕名血中慕名一时之间一时之间每次一时之间其间大惊失色酒中一时之间当众两次事后每次污秽当众伤势污秽慕名血中假装污秽酒中两次酒中流血酒中事后拚命大惊失色假装流血伤势一时之间大惊失色两次拚命不下两次仍然流血事后慕名流血血中治愈污秽血中每次伤势仍然污秽其间一时之间假装治愈假装一时之间每次两次仍然拚命慕名污秽不下污秽大惊失色他意仍然流血慕名事后慕名血中慕名治愈仍然假装拚命伤势仍然慕名拚命两次两次一月大惊失色大惊失色事后事后大惊失色其间慕名污秽每次慕名当众仍然仍然他意每次事后酒中慕名拚命治愈两次两次每次大惊失色拚命仍然污秽流血拚命每次不下他意治愈他意他意治愈酒中治愈当众不下治愈污秽污秽污秽他意治愈两次其间拚命血中两次一月拚命血中伤势事后事后假装治愈大惊失色流血一月不下酒中不下不下伤势假装假装拚命污秽不下一月每次一月当众其间一时之间伤势他意污秽两次慕名仍然伤势他意伤势血中流血

Seq2seq 模型生成的结果不太令人满意，只能在个别词附近语义相对符合逻辑，在整体上语义并不通顺，与循环神经网络的特征符合。