

基于循环生成对抗网络的机器翻译方法研究

夏 珺¹, 周湘贞², 隋 栋³

(1. 黔南民族师范学院外国语学院, 贵州 都匀 558000)

(2. 马来西亚国立大学信息科学与技术学院, 马来西亚 雪兰莪 43600)

(3. 北京建筑大学电气与信息工程学院, 北京 102406)

[摘要] 近几年来,智能语言处理在语言学习方面已经得到了广泛的应用,但是由于在处理语言中往往会存在网络模型优化困难、强制对其的标记数据会出现精度偏差,与以往大多数使用判别模型结合 HMM 混合模型进行声学模型训练的系统相比,本文提出了一种基于循环生成对抗网络的机器翻译方法,该方法主要结合生成对抗网络来训练机器翻译模型。首先,将一段语音输入神经机器翻译模块进行离散,预先变换得到 MFCC 特征;然后,将经过预处理的语音输入到特征提取模块并结合长时短时记忆网络循环提取语音特征;最后,将网络模型输出的语音与人工翻译的语音进行对比,并判别网络模型输出的语音特征与人工翻译的语音是否匹配,如果不匹配则继续优化生成网络。实验结果表明,我们的网络与传统的高斯核混合模型相比有明显的提升。本文方法在 CSDN 口令集、Rockyou 口令集、Tianya 口令集和 Yahoo 口令集中均取得了优越的结果,其中在 Yahoo 口令集中单词错误率降至 19.5%。

[关键词] 语音识别,语言翻译,循环对抗网络,长短时记忆模块

[中图分类号] TP391 **[文献标志码]** A **[文章编号]** 1001-4616(2022)01-0104-06

Research on Machine Translation Method Based on Cyclic Generation Countermeasure Network

Xia Jun¹, Zhou Xiangzhen², Sui Dong³

(1. School of Foreign Languages, Qiannan Normal University for Nationalities, Duyun 558000, China)

(2. Faculty Information Science and Technology, National University of Malaysia, Selangor 43600, Malaysia)

(3. School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102406, China)

Abstract: In recent years, intelligent language processing has been widely used in language learning. However, due to the difficulty of network model optimization and the accuracy deviation of its labeled data, compared with most previous systems using discriminant model combined with HMM hybrid model for acoustic model training, This paper proposes a machine translation method based on cyclic generation countermeasure network. This method mainly combines generation countermeasure network to train machine translation model. Firstly, a speech is input into the neural machine translation module for discrete pre transformation to obtain MFCC features; Then, the preprocessed speech is input to the feature extraction module, and the speech features are extracted circularly combined with the long-term and short-term memory network; Finally, the speech output from the network model is compared with the artificially translated speech, and whether the speech features output from the network model match the artificially translated speech is judged. If not, the network is optimized. The experimental results show that our network is significantly improved compared with the traditional Gaussian kernel mixture model. This method has achieved excellent results in CSDN password set, rockyou password set, Tianya password set and Yahoo password set, and the word error rate in Yahoo password set is reduced to 19.5%.

Key words: speech recognition, language translation, cyclic countermeasure network, long-short memory module

语言作为一种高级符号系统,本身就非常复杂,并且语言是一个复杂网络的观点已经被人们广泛地接受。然而,智能语言处理^[1-4]作为一种语言学习任务,通常是需要从一段未被预处理的语音中获得一系列可能的语音标签,并将语音信号转换为单词和子词单元,并将处理后的语音单元转换为我们需要的输出以

收稿日期:2021-10-11.

基金项目:贵州省教育厅人文社科项目(2019zcl16)、国家自然科学基金青年基金项目(61702026)。

通讯作者:夏珺,副教授,研究方向:机器学习,智能翻译,自然语言处理。E-mail:273976230@qq.com

实现语音识别和语言翻译的效果。

2001年,人们发现语言中连接单词的图与其他复杂网络具有相同的统计特征。在这之后,不同语言单元组成的语言网络及其在不同语言中的关系受到了学者们的关注。2012年,微软研究人员将前馈深度神经网络(FFDNN)应用于大词汇连续语音任务的声学建模,使用DNN而不是GMM-HMM。该混合高斯模型可以提供更好的观测概率,引发了一波混合神经网络和隐马尔可夫混合建模^[5-6]。复杂网络的构成要素主要是网络节点和节点间边,而如何确定语言网络的节点和边呢?学者们提出了不同的构建语言网络的方法,主要包括可以根据同义词表确定原始词与其同义词之间的网络连接;可以根据词汇表进行语义连接;根据词在句子中的共现情况,可以构建语言的共现网络;通过标注依存句法的语料库,可以得到语言网络连接。

在声学建模过程中,DNN-HMM^[7]通常使用左右相邻特征拼接在一起作为卷积神经网络的输入。上下文窗口可以反映两帧之间的关系,这更符合实际情况。为了获得更好的建模结果,输出在GMM-HMM中聚类或状态绑定后使用三音素(senone)来减少参数爆炸^[8]的问题。相邻音素中语音的每一个单词都相互影响。当每个特征帧的长度之间有相关性时,必须考虑声学模型。在以往的工作中发现,深层变压器是很难训练的,例如超过12层的变压器^[9]。这是由于优化网络模型的挑战:每一层的输出随着深度的增加而变化,导致不稳定的梯度,最终导致训练的收敛度不理想。训练数据必须在卷积神经网络之间进行处理,即每一帧的标记数据、输入的特征序列和标记的特征序列必须具有相同的长度。当使用大数据训练模型时,标记训练数据需要专业知识和大量的工作,而现有的模型需要在获取注释^[10]时强制输入数据对齐和标记序列对齐。强制对齐所使用的模型会有一定程度的精度偏差,导致标记错误。对训练数据注释的依赖性和强制对齐问题限制了语音识别的进一步发展。

本文提出了一种将长期短期记忆网络和循环网络相结合的基于循环生成对抗网络的机器翻译方法。首先,将LSTM顶层的Softmax向量输出连接到神经机器翻译模型上,并使用神经机器翻译解码方法减少了整个序列的损失,因此,我们可以在预测LSTM输出的预测概率中,正确地预测该序列的标签。然后,将经过预处理的语音输入到特征提取模块并结合长时短时记忆网络循环提取语音特征;最后,将网络模型输出的语音与人工翻译的语音进行对比,并判别网络模型输出的语音特征与人工翻译的语音是否匹配,如果不匹配则继续优化生成网络。实验结果表明,本文方法既关注了每个特征帧的长度之间的相关性,减少了标记数据间的偏差,又解决了优化网络模型困难的问题。本文设计的基于循环生成对抗网络的机器翻译网络模型,该模型主要有生成网络模型和对抗网络组成,并通过对抗网络优化生成网络的输出从而达到预期的结果。

1 网络模型

为了解决处理语言中存在的网络模型优化困难、强制对其标记数据会出现精度偏差等问题,本文设计了基于循环生成对抗网络的机器翻译网络模型,网络模型如图1所示。

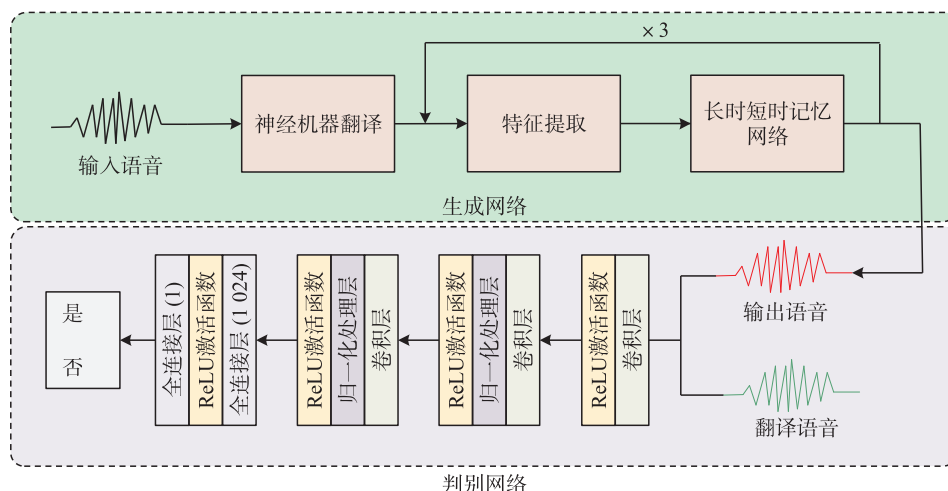


图1 基于循环生成对抗网络的机器翻译网络

Fig. 1 Machine translation network based on cyclic generation countermeasure network

模型主要包括两大部分,分别为生成网络和判别网络. 生成网络是对未经过预处理的语音进行特征提取并翻译成需要的单词. 首先,模型主要利用了对抗网络的特点来优化生成网络,解决网络模型优化困难的问题;其次,在生成网络中使用了长期和短期记忆模块,缓解了由于强制标记数据而出现精度偏差问题;最后,模型经过优化训练保存生成网络的网络模型来处理自然语言. 判别网络则是通过判别生成网络的输出语音和人工翻译的语音是否相符,达到相符则输出“是”,否则输出“否”. 如果不符合则反馈给生成网络以优化生成网络,直到生成网络输出的语音与人工翻译的语音达到一定的相符度就会自动保存生成网络.

1.1 神经机器翻译模块

到目前为止,Dewangan 等^[11]和 Shterionov 等^[12]提出了各种 NMT 框架. 其中,基于自我注意的框架(称为变压器)实现了最先进的翻译性能.

变压器遵循编码器-解码器架构,其中编码器将源句子 X 转换为一组上下文向量 C . 解码器从上下文向量 C 中生成目标句子 Y . 给定一个并行的句子对数据集 $D=\{(X,Y)\}$,其中 X 为源句子, Y 为目标句子,损失函数可以定义为:

$$L(D;\theta)=\sum_{(X,Y)\in D}\log p(Y|X;\theta). \tag{1}$$

1.2 长期短期记忆模块

长期短期记忆模块,又称 LSTM,是一种改进的时间递归神经网络,可以有效地处理时间序列中的长期依赖问题,该模块在语音识别上有着强大的优越性^[13-16].

图 2 显示了 LSTM 网络的结构,其中重复的模块代表了每次迭代中的隐藏层. 每个隐藏层都包含许多神经元. 每个神经元对输入向量进行线性矩阵计算,然后在激活函数^[17-20]的非线性作用后输出相应的结果. 在每次迭代中,上一次迭代的输出都会与文本的下一个字向量进行交互,这决定了信息的保存或放弃,以及当前状态的更新. 图中 x_t 是本次迭代中隐藏层的输入. 根据当前的状态信息,得到隐藏层 \hat{y} 的预测输出值,并同时为下一层隐藏层提供输出向量 h_t . 每当网络中有一个新的词向量输入时,下一时刻的输出与最后时刻的输出一起计算. 隐藏层循环并保持最新的状态^[21-24].

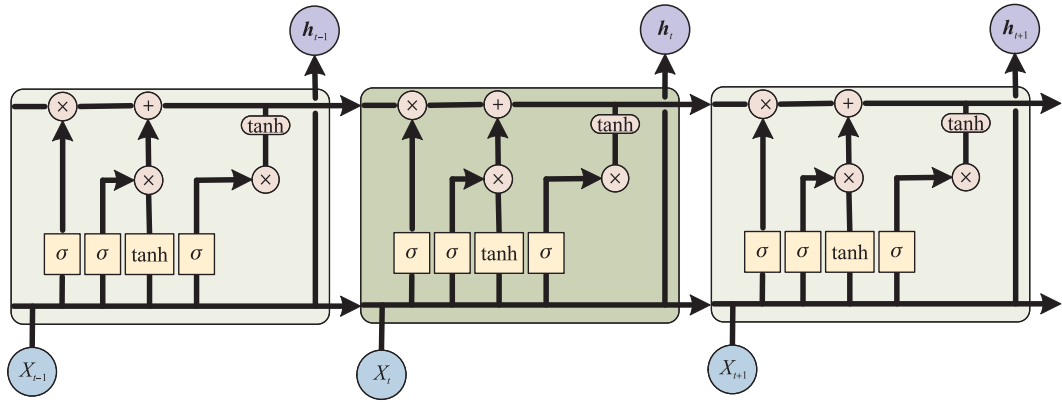


图 2 长期短期记忆网络
Fig. 2 Long term short term memory network

最后将隐藏层与传统的前馈网络作为输出层进行连接. 输出层中的每个节点 y_i 对应于下一时刻的未归一化对数概率,然后通过 softmax 函数对输出值 y 进行归一化. 其公式如下:

$$\hat{y}_i = \text{softmax}(W^{(s)}h_i) \tag{2}$$

式中, \hat{y}_i 是基于每次迭代中所有词汇量计算的隐藏层的概率分布. 也就是说,当模型预测以下单词时,将确定文档中所有预定的单词和观察词向量 $x(t)$ 的权重.

2 实验

2.1 数据集

本文使用了来自不同服务类型和规模网站的 4 个大规模的真实口令集. 并且这 4 种口令集的语言和

文化背景也有一定的差别,它们的服务类型分别为程序员论坛、游戏、社交网站和互联网门户,它们分别来自中国和美国包括中文和英语两种类型的语言,并且每种口令集的口令总数也不同.这4种口令集的详情如表1所示.

表1 口令集

Table 1 Password set

口令集	服务类型	地区	语言	口令总数
CSDN	程序员论坛	中国	中文	6 428 287
Rockyou	游戏	美国	英文	32 603 388
Tianya	社交网站	中国	中文	30 233 633
Yahoo	互联网门户	美国	英文	5 626 485

如表1所示,我们的口令集分别是本文中的语音库是CSDN口令集、Rockyou口令集、Tianya口令集和Yahoo口令集.其中,口音集均分为训练集和测试集,训练集和测试集有语音紧凑的句子和语音上不同的句子,并且训练集和测试集的数据结构不一致.

2.2 实验结果

我们分别在上文中提到的口令集中做了对比实验,分别将CSDN口令集、Rockyou口令集、Tianya口令集和Yahoo口令集中的测试集口令输入DNN-HMM、LSTM-CTC和本文模型中得到的单词错误率如表2和图3所示:

表2 在4个数据集上的实验结果

Table 2 Experimental results on four data sets

口令集	模型	单词错误率/%
CSDN	DNN-HMM	20.9
	LSTM-CTC	20.1
	本文模型	19.9
Rockyou	DNN-HMM	21.9
	LSTM-CTC	20.7
	本文模型	19.7
Tianya	DNN-HMM	20.7
	LSTM-CTC	20.5
	本文模型	19.8
Yahoo	DNN-HMM	20.6
	LSTM-CTC	20.4
	本文模型	19.5

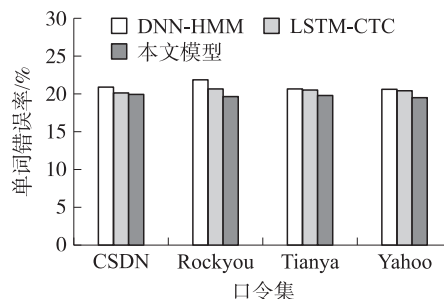


图3 不同口令集上的错误率对比

Fig. 3 Comparison of error rates on different password sets

如表2所示,本文模型在口令集中均取到了最优的结果.其中,与DNN-HMM模型和LSTM-CTC模型相比本文模型在Yahoo的口令集中的单词错误率将至19.5%,分别比DNN-HMM模型、LSTM-CTC模型的错误率降低了5%和4.5%,结果达到了最优.

如图3所示,根据直方图可以更清楚直接地看出我们的方法与其他两种方法(DNN-HMM\LSTM-CTC)相比,在CSDN口令集、Rockyou口令集、Tianya口令集和Yahoo口令集中均取到了最优的结果,其中在Rockyou口令集中我们的方法与DNN-HMM的方法相比,我们的方法的单词错误率明显低于DNN-HMM方法的单词错误率.因此我们的方法更适合用于机器翻译.

虽然本文设计的网络模型已经取得了比较优秀的结果,但是本文模型还存在一些不足,表3展示了本文模型在测试结果中经常出错的例子.

如表3所示,可以看出一些读音相同的字经常会被错误识别.针对以上问题我们还需继续实验,继续研究读音相同的字的识别方法,这也是一个值得挑战的困难,我们未来将从分析语境方面入手.

2.3 隐藏层中神经元消融的影响分析

隐藏层中存在大量的神经元,适当的隐藏层层数有利于对语音特征的提取,但是过多的设计隐藏层会对网络模型带来巨大的计算开销,所以合理设计隐藏层的层数非常重要.因此,本文对隐藏层的层数做了

表3 部分识别错误示例

Table 3 Examples of partial identification errors

实际语音	误识结果	实际语音	误识结果
文	人	树	书
Like	Lack	Think	Drink
向	乡	点	电

消融实验,实验结果如表 4 所示.

表 4 显示,隐藏层数对语音识别系统的准确性有很大的影响. 当隐藏层数增加时,网络的识别能力就会增加,但当隐藏层继续增加时,识别效果就会回归. 随着层数的增加,训练时间也会变长,从而导致系统效率的降低. 因此,通过将隐藏层设置为 4 层,可以得到最佳的结果.

隐藏层中存在大量的神经元,适当的神经元个数有利于对语音特征的提取,但是过多的神经元个数会对网络模型带来巨大的计算开销,所以合理设计隐藏层中的神经元个数非常重要. 因此,本文对隐藏层的神经元的个数做了消融实验,实验结果如表 5 所示.

表 4 隐藏层层数对网络模型的影响		表 5 隐藏层神经元个数对网络模型的影响	
Table 4 Influence of hidden layers on network model		Table 5 Influence of number of hidden layer neurons on network model	
隐藏层	单词错误率/%	隐藏层的神经元个数	单词错误率/%
3	23.4	120	56.3
4	19.9	240	21.6
5	22.7	480	19.9
6	23.0	600	20.7
		1024	22.6

为了研究每层神经元数量对识别结果的影响,本文选择了不同数量的神经元:120、240、480、600、1 024. 对比表 3 的结果表明,神经单元的数量太少,网络的拟合能力不足. 从而导致系统的音素的错误率过高. 然而,当隐藏层数的数量继续增加时,音素错误率逐渐降低,但是神经元数量增加,会导致系统效率下降,训练所需时间的增加. 所以,在长期短期记忆模块中,每层的单位数被设置为 480 个.

3 结论

本文设计了一种基于循环生成对抗网络的机器翻译网络模型,使用生成网络处理自然语言,并通过判别网络优化生成网络,两个网络相互作用最终得到一个比较理想的机器翻译网络模型,我们也分别在 CSDN 口令集、Rockyou 口令集、Tianya 口令集和 Yahoo 口令集等口令集中做了大量的对比实验,实验结果表明,在每个口令集中我们的结果均达到了最优.

[参考文献]

[1] DAS D,DAS A K,PAL A R,et al. Meta-heuristic algorithms-tuned elman vs. jordan recurrent neural networks for modeling of electron beam welding process[J]. Neural processing letters,2021,53(2):1647-1663.

[2] BRUNO J H,JARVIS E D,LIBERMAN M,et al. Birdsong learning and culture:analogies with human spoken language[J]. Annual review of linguistics,2021,7(1):89-97.

[3] BOER B D,THOMPSON B,RAVIGNANI A,et al. Evolutionary dynamics do not motivate a single-mutant theory of human language[J]. Scientific reports,2020,10(1):22-31.

[4] MUKHERJEA A,ALI S,SMITH J A. A human rights perspective on palliative care:unraveling disparities and determinants among asian american populations[J]. Topics in language disorders,2020,40(3):278-296.

[5] LI K,PAN W,LI Y,et al. A method to detect sleep apnea based on deep neural network and hidden Markov model using single-lead ECG signal[J]. Neurocomputing,2018,294(6):94-101.

[6] 张旭东,黄宇方,杜家浩,等. 基于离散型隐马尔可夫模型的股票价格预测[J]. 浙江工业大学学报,2020,48(2):148-153.

[7] OUISAADANE A,SAFI S. A comparative study for Arabic speech recognition system in noisy environments[J]. International journal of speech technology,2021,11(3):1-10.

[8] PIKHART M. Human-computer interaction in foreign language learning applications:applied linguistics viewpoint of mobile learning[J]. Procedia computer science,2021,184:92-98.

[9] YAO Q,UBALE R,LANGE P,et al. Spoken language understanding of human-machine conversations for language learning applications[J]. Journal of signal processing systems,2020,92(3):78-89.

[10] HINTON G E. A practical guide to training restricted boltzmann machines[J]. Momentum,2012,9(1):599-619.

- [11] DEWANGAN S, ALVA S, JOSHI N, et al. Experience of neural machine translation between Indian languages[J]. Machine translation, 2021, 35(1): 71–99.
- [12] SHTERIONOV D, SUPERBO R, NAGLE P, et al. Human versus automatic quality evaluation of NMT and PBSMT[J]. Machine translation, 2018, 32(3): 217–235.
- [13] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1743.
- [14] GRAVES A, JAITLY N. Towards end-to-end speech recognition with recurrent neural networks[C]//International Conference on Machine Learning, Beijing, 2014: 1764–1772.
- [15] DENG L, YU D. Deep learning for signal and information processing[J]. Now publishers, 2013, 12(8): 218–227.
- [16] WU S, LI G, DENG L, et al. Li-norm batch normalization for efficient training of deep neural networks[J]. IJEE transactions on neural and learning systems, 2018, 30(7): 2043–2051.
- [17] MAAS A L, QI P, XIE Z, et al. Building DNN acoustic models for large vocabulary speech recognition[J]. Computer speech & language, 2014, 41(C): 195–213.
- [18] CUI X, ZHANG W, FINKLER U, et al. Distributed training of deep neural network acoustic models for automatic speech recognition: a comparison of current training strategies[J]. IEEE signal processing magazine, 2020, 37(3): 39–49.
- [19] TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[J]. Computer science, 2015, 5(1): 36–41.
- [20] XU Y, DU J, DAI L R, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE signal processing letters, 2013, 21(1): 65–68.
- [21] SALMELA L, TSIPINAKIS N, FOI A, et al. Predicting ultrafast nonlinear dynamics in fibre optics with a recurrent neural network[J]. Nature machine intelligence, 2021, 12(8): 1–11.
- [22] POLIAK A, RASTOGI P, MARTIN M P, et al. Efficient, compositional, order-sensitive n-gram embeddings[C]//Conference of the European Chapter of the Association for Computational Linguistics, London, 2017: 503–508.
- [23] 黄光许, 田垚, 康健, 等. 低资源条件下基于 i-vector 特征的 LSTM 递归神经网络语音识别系统[J]. 计算机应用研究, 2017, 34(2): 392–396.
- [24] 舒帆, 屈丹, 张文林, 等. 采用长短期记忆网络的低资源语音识别方法[J]. 西安交通大学学报, 2017, 51(10): 120–127.

[责任编辑: 陆炳新]