软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(6):1747−1760 [doi: 10.13328/j.cnki.jos.005709]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于端到端句子级别的中文唇语识别研究*

张晓冰, 龚海刚, 杨 帆, 戴锡笠

(电子科技大学 计算机科学与工程学院,四川 成都 611731)

通讯作者: 戴锡笠, E-mail: daixili_cs@163.com

摘 要: 近年来,随着深度学习的广泛应用,唇语识别技术也取得了快速的发展.与传统的方法不同,在基于深度学习的唇语识别模型中,通常包含使用神经网络对图像进行特征提取和特征理解两个部分.根据中文唇语识别的特点,将识别过程划分为两个阶段——图片到拼音(P2P)以及拼音到汉字(P2CC)的识别.分别设计两个不同子网络针对不同的识别过程,当两个子网络训练好后,再把它们放在一起进行端到端的整体架构优化.由于目前没有可用的中文唇语数据集,因此采用半自动化的方法从 CCTV 官网上收集了 6 个月 20.95GB 的中文唇语数据集 CCTVDS,共包含 14 975 个样本.此外,额外采集了 269 558 条拼音汉字样本数据对拼音到汉字识别模块进行预训练.在 CCTVDS 数据集上的实验结果表明,所提出的 ChLipNet 可分别达到 45.7%的句子识别准确率和 58.5%的拼音序列识别准确率.此外,ChLipNet 不仅可以加速训练、减少过拟合,并且能够克服汉语识别中的歧义模糊性.

关键词: 中文唇语识别;深度学习;中文汉语言的特征;数据集采集及处理;端到端模型

中图法分类号: TP18

中文引用格式: 张晓冰,龚海刚,杨帆,戴锡笠.基于端到端句子级别的中文唇语识别研究.软件学报,2020,31(6):1747−1760. http://www.jos.org.cn/1000-9825/5709.htm

英文引用格式: Zhang XB, Gong HG, Yang F, Dai XL. Chinese sentence-level lip reading based on end-to-end model. Ruan Jian Xue Bao/Journal of Software, 2020,31(6):1747−1760 (in Chinese). http://www.jos.org.cn/1000-9825/5709.htm

Chinese Sentence-Level Lip Reading Based on End-to-End Model

ZHANG Xiao-Bing, GONG Hai-Gang, YANG Fan, DAI Xi-Li

(School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: In recent years, with the widely application of deep learning, lip reading recognition technology has achieved rapid development. Different from traditional methods, lip reading recognition methods based on the deep learning usually use the neural network model both for the feature extraction and comprehension. According to the characteristics of Chinese language, a two-step end-to-end architecture is implemented, in which two deep neural network modules are applied to perform the recognition of picture-to-pinyin (P2P) and pinyin-to-hanzi (P2CC) respectively. After the two modules are trained with convergence, they are then jointly optimized to improve the overall performance. Due to the lack of Chinese lip reading dataset, the 6-month daily news broadcasts are collected from China Central Television (CCTV), and they are semi-automatically labelled into a 20.95 GB dataset CCTVDS with 14 975 samples. In addition, the supplementary dataset with 269 558 samples are collected during the pre-training of P2CC. According to experimental results trained on the CCTVDS, the proposed ChLipNet can achieve 45.7% sentence-level and 58.5% Pinyin-level accuracies. In addition, ChLipNet can not only accelerate training, reduce overfitting, but also overcome syntactic ambiguity in the recognition of Chinese language.

Key words: Chinese lip reading recognition;deep learning; characteristics of Chinese language; data collecting and preprocessing; end-to-end model

* 基金项目: 国家自然科学基金(61572113)
Foundation item: National Natural Science Foundation of China (61572113)
收稿时间: 2018-05-10; 修改时间: 2018-09-04; 采用时间: 2018-11-16

唇语识别主要通过观察说话者嘴唇的运动变化序列从而识别出相应的文本信息,其研究内容涉及到模式识别、图像处理、语音识别及自然语言处理等多个领域,具有广阔的应用场景.例如在高噪环境中,由于说话者音频受到环境的干扰,导致识别率降低,而视觉信息相对很稳定,因此,通过唇语识别利用视觉信息从而能够极大地辅助提高语音识别的准确率.在非噪声环境下,当进行语音识别时,辅助观察说话者的脸部表情变化、嘴唇运动以及人体肢体动作等信息,能够更加准确地理解对方所要表达的内容.此外,嘴唇同虹膜、鼻子等一样,作为人脸的一项重要生物特征,在人脸身份检测中发挥了重要作用.例如在人脸活体检测应用中,通过核查说话者嘴唇运动,可进一步提高活体识别的安全性,从而排除了传统人脸识别中使用其他工具造假的可能.此外,唇语识别可与手语识别相互依存,一起促进聋哑人在日常生活中的正常交流.

目前为止,唇语识别研究已经取得了一定的成果.然而,由于日常应用场景及条件的多样化和复杂化,使得唇语识别技术在实际应用中依然面临巨大的挑战:(1) 人的嘴唇是一个三维的非刚性物体,不同的说话者对象、不同的语句内容,都会使得人的嘴唇运动在视频中显示不同的变化,这给识别带来了很大的困扰;(2) 光源照射和人脸角度的不同等因素,使得人的嘴唇在视频中有不同的形态,从而对识别率造成很大的影响.

近几年来,深度学习在各个领域取得的显著成果,也促进了应用神经网络来解决唇语识别的研究.随着技术的成熟,唇语识别率也在不断提高,例如 DeepMind 的 WLAS[1]和 LipNet[2].然而,已有的这些研究都是基于单词分类或者英文句子的识别,与中文唇语识别的内容截然不同.汉语与英语不同:英语是由 26 个字母组成的字母语言,所有的单词都是由字母拼读而成,通过拼读可以准确地确定某个单词;而汉语不同,汉语的发音是由 23 个元音字母和 24 个辅音字母组成,去掉一些不可能的拼读组合,再加上 4 种不同的音调,拼音总共大约有 1 000 种,然而汉语中的汉字总数超过 90 000 个,其中有 3 000 个是经常使用的,也就是说,每个拼音平均对应 3~90 个汉字.据统计,汉语是信息熵含量最大的语言.因此,从汉语这种高模糊性语言中提取具有显著区别的特征信息,是中文唇语识别中的一个重要并且富有挑战性的任务.

本文根据中文的特点,首次提出了句子级别的中文唇语识别模型 ChLipNet,该模型由两个子模块组成,即嘴唇图片序列映射到拼音字符序列的拼音序列识别模块和拼音字符序列转换为汉字序列的汉字序列识别模块,如图 1 所示.其中,

(1) 拼音序列识别主要利用卷积神经网络 Convolutional Neural Network(CNN)作为嘴唇图片帧序列的特征提取器,然后,使用循环神经网络 Recurrent Neural Network(RNN)理解并分析提取的特征,最后利用 Connectionist Temporal Classification(CTC)损失函数匹配输入输出序列.该过程简称为 P2P 过程,生成的拼音序列识别网络简称为 P2P 网络;

(2) 汉字序列识别是一个基于语言模型的 Encoder-Decoder 网络框架,这个过程简称为 P2CC 过程.P2CC 网络的输入是拼音字符序列,其中 Encoder 网络负责对拼音字符序列进行编码,而 Decoder 网络则对 Encoder 的输出进行解码,从而生成汉语句子.

当 P2P 和 P2CC 两个子模块分别训练好后,把它们联合在一起组成中文唇语识别网络 ChLipNet 并进行最终的端到端训练.

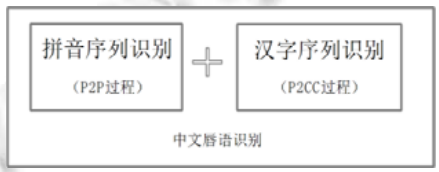

Fig.1 Module division of Chinese lip reading recongnition framework

图 1 中文唇语识别框架的模块划分

由于现有的唇语数据集都是针对字符、单词、数字或者短语的,且都是关于非中文的.为此,我们采集了 6 个月的 CCTV 新闻联播视频及其对应的文稿,使用半自动化技术,通过视频剪辑、文本和时间戳生成以及嘴唇检测等操作生成包含 14 975 条中文句子及其对应嘴唇序列的中文唇语数据集 CCTVDS.此外,在汉字序列识别

P2CC 的预训练过程中,还额外统计了近 30 个月的新闻联播文稿内容作为辅助数据.

1 相关工作

Petajan 等人在 1984 年最初提出了唇语识别系统[3],该系统把单个词作为最小的识别单元,通过计算输入的嘴唇图片序列,得到能够表示的特征向量,并与数据集中所有词的特征模板进行相似度匹配,最后将相似度最高的词作为预测结果并输出.之后,在 1988 年,Petajan 等人在原唇语识别系统上引入矢量化和动态时间规整等算法,主要用于解决训练和识别过程中说话人语速变化较大的问题,从而对唇语识别系统进行改进[4],极大地提高了唇语识别的正确率.

近几年也出现了很多尝试用深度学习解决唇语识别的工作,例如:

1) Noda[5]利用 VGGNet 对人嘴唇图片进行单词和短语的预训练,然后通过 RNN 网络,可分别实现 44.5%短语识别和 56.0%单词分类准确率;
2) Chung 等人提出了利用 VGG 时空卷积神经网络在 BBCTV 数据集上进行单词分类[6],并再次提出了一种用于学习嘴部特征的视听最大边缘匹配模型[7],并将其作为一个 LSTM 的输入,从而用于 OuluVS2 数据集上的 10 个短语的分类;
3) Wand 等人[8]在 2016 年将 LSTM 递归神经网络引入用于唇语识别的研究,虽然该研究没有包含句子序列的预测和说话者的独立性,但在 GRID 语料库上,模型识别讲话人的准确率可达 79.6%;
4) Chung 等人在 2017 年提出由卷积神经网络和循环神经网络组成的 WLAS 模型,其在含有 1 万条样本句子的 LRC 数据集上可取得 46.8%的句子准确率,是目前句子级别的英语唇语识别中较好的成绩.

2 汉语发音规则及中文特征

汉语是一种非形态语言.汉语中的词语没有严格意义的形态变化,只有音节符号.汉语中的音节是由声母、韵母和声调按照拼音规则组合而成,其中,声母有 23 个,韵母有 24 个,另外还包括 4 种声调,见表 1.因此,粗略计算得音节种类不超过 2 208 个(包括不合理拼音).

Table 1 Classification of initial, vowel and intonation mark

表 1 声母、韵母以及声调分类

声母	b	p	m	韵母	单韵母	a	o	e	i	声调	-
	f	d	t			u	ü				ˊ
	n	l	g		复韵母	ai	ei	ui	ao		ˇ
	k	h	j			ou	iu	ie	üe		
	q	x	y			er					ˋ
	w	zh	ch		前鼻韵母	an	en	in	un		
	sh	r	z			ün					˙
	c	s			后鼻韵母	ang	eng	ing	ong		

自 1955 年开始,汉语拼音被用作是辅助汉字发音的一种工具.它和英语的音标类似,但又大不相同.其中,声母又叫做辅音字母,用在韵母之前,并跟韵母一起构成完整的音节.辅音的主要特点是发音时气流在口腔中会受到各种不同的阻碍,因此可以说,声母发音的过程也就是气流受阻和克服阻碍的过程.除声母、音调之外的部分,就是韵母,包括韵头、韵腹和韵尾这 3 部分.音节是人类听觉系统能感受到的最小语音单位,在汉语中,单个汉字就是单个音节(但汉字不是音节文字).然而,在 1994 年出版的《中华字海》中,约有 87 019 个汉字(其中重复字 320 个),因此汉字数量空间是远远大于音节数量空间的(最多 2 208 维).单个汉字具有丰富的信息和意义,例如,“中”可解释为“里面”,或者“适于”“合适”等含义.当汉字与汉字通过语言规则相互组合成词、短语或者句子时,才有具体含义,例如“中间”“中计”等.如图 2 为“中国人”短语示例,其中,绿色框代表拼音序列,黄色框代表声调,蓝色虚线框代表单个汉字.

据统计,汉字中超过 85%的是同音字,即:同一个发音(嘴型)可至少对应 2 个,至多对应 120 个汉字.这也是中文比其他语言更难识别的一个重要原因之一.如表 2 所示为部分音节和对应的常见汉字,其中的数字表示该音

节可对应不同汉字的总个数.

Fig.2 Examples of Chinese, Pinyin, Hanzi and intonation mark

图 2 汉语、拼音、汉字和音调的示例

Table 2 List of syllables and the corresponding common Hanzi

表 2 音节及对应常见汉字示例

辅音 / 音节 / 元音	a			ou			eng		
m	mā	妈,麻	12	mōu	哞	1	mēng	蒙	5
	má	麻,吗	24	móu	谋,眸	38	méng	萌,檬	108
	mǎ	马,码	22	mǒu	某	10	měng	蒙,猛	28
	mà	骂,蚂	30	–	–	–	mèng	梦,孟	26
zh	zhā	扎,渣	46	zhōu	周,州	71	zhēng	争,征	81
	zhá	闸,扎	58	zhóu	轴,妯	8	–	–	–
	zhǎ	眨	34	zhǒu	肘,帚	21	zhěng	整,拯	16
	zhà	炸,乍	42	zhòu	咒,昼	79	zhèng	挣,郑	42

3 ChLipNet 整体框架

3.1 拼音序列识别模型(P2P)

与大多数图像识别任务不同,拼音序列识别中需要网络能够捕捉图像最细微的特征,尤其是嘴唇图片间的运动变化.图片到拼音的拼音序列识别 P2P 模型的结构如图 3 所示.

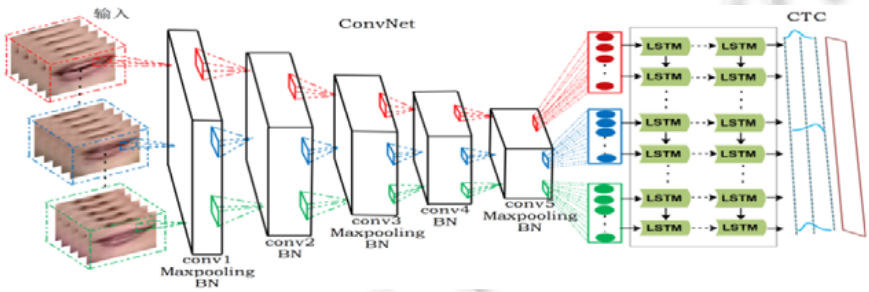

Fig.3 Model architecture of Pinyin-to-Hanzirecognition (P2P)

图 3 图片到拼音识别(P2P)网络模型

在 P2P 网络模型中,嘴唇图片先转换成灰度图,其中,每隔两帧的连续 5 张 120×120 灰度图片作为输入,因此,若一个样本有 n 张嘴唇图片,则输入长度为$\lfloor(n-4)/2\rfloor$,然后用卷积网络 ConvNet 提取图像特征信息.假设 P2P 模型的输入为 $X=x_1,x_2,\ldots,x_k$,其中,x_i 包含连续 5 张嘴唇图片序列,则 ConvNet 将图片序列按公式(1)转换为特征向量:

$$y_i = CNN_{\theta_c}(x_i) \tag{1}$$

其中,CNN_{θ_c} 表示 ConvNet 将嘴唇图片 x_i 映射成一个 512 维度的特征向量 y_i,θ_c 表示 ConvNet 的模型参数.公式(2)表示对 ConvNet 提取的嘴唇特征向量 y_i 进行降维,生成维度为 d 特征向量 v_i,如下:

$$v_i=K[y_i]+b \tag{2}$$

其中,K 是维度为 d×512 的矩阵,d 是 LSTM 单元中 embedding 空间的大小,b 是维度为 d 的误差参数.

最终,连续的特征向量 v_i 序列作为 n-LSTM 的输入,通过 n-LSTM 输出一个维度为 d'的向量 v_i'.在 n-LSTM 中,其 LSTM 单元的运算过程见公式(3):

$$\begin{aligned}&(i_i, f_i, o_i, g_i) = W_x v_i + W_h h_{i-1} + b_{lstm}\\&\begin{bmatrix}\overline{i}_i\\ \overline{f}_i\\ \overline{o}_i\\ \overline{g}_i\end{bmatrix} = \begin{bmatrix}sigm(i_i)\\ sigm(f_i)\\ sigm(o_i)\\ \tanh(g_i)\end{bmatrix}\\&c_i = \overline{f}_i \times c_{i-1} + \overline{i}_i \times \overline{g}_i\\&h_i = o_i \times \tanh(c_i)\end{aligned} \tag{3}$$

其中,b_{lstm} 表示误差参数矩阵,维度为 $4d'$;W_x 和 W_h 分别表示 LSTM 单元中与输入 x 和隐藏状态 h 有关的参数矩阵,维度为 $4d\times d'$;$\overline{i}_i$ 表示在 i 时刻 LSTM 单元输入门的输出;$\overline{f}_i$ 表示 i 时刻 LSTM 单元遗忘门的输出;$\overline{g}_i$ 表示在 i 时刻 LSTM 单元的细胞更新信息;$\overline{c}_i$ 表示在 i 时刻 LSTM 单元的细胞状态;$\overline{h}_i$ 表示在 i 时刻 LSTM 单元的隐藏状态,同时也表示在 i 时刻 LSTM 的输出.最后,经过全连接网络生成一个 26 维的向量并传递给 CTC 函数.整个 P2P 模型的损失函数如公式(4):

$$L(S) = -\ln\prod_{(X,Z)\in S} P(Z \mid X) = -\sum_{(X,Z)\in S} \ln P(Z \mid X) \tag{4}$$

其中,S 表示整个数据集,(X,Z)为数据样本,Z 为输入图片对应的真实拼音字符序列标签,$P(Z|X)$为输入 X 得到 Z 的概率模型.

3.1.1 Connectionist Temporal Classification (CTC)

CTC[9]是一种通用的损失函数,主要用于解决未知输入序列和输出序列对齐的网络系统.由于 CTC 损失函数比传统的隐马尔科夫模型等具有更好的性能,因此在 P2P 模型中,本文我们采用 CTC 损失函数来实现输入图片和输出拼音序列的自动划分与对齐.给定一个拼音序列的分布并且用空白字符进行增强,那么 CTC 通过最大化所有与之等价的序列来定义生成该序列的可能性.假设拼音序列的标签为 L,$\overline{L} = L \cup -$ 表示空白字符增强后的拼音序列,其中,"–"表示空白字符.定义映射函数 $\Gamma: \overline{L}^* \to L^*$ 表示移除空白字符并且删除相邻相同的拼音字符,对于一个序列 $y\in L^*$,CTC 计算如下:

$$P(y \mid x) = \sum_{v\in\Gamma^{-1}(y)\ \text{s.t.}\ |v|=step} P(v_1, ..., v_k \mid x) \tag{5}$$

其中,$step$ 是序列模型中步长.例如,假设 $step$=3,CTC 计算'qi'的概率为:$p(qqi)+p(qii)+p(-qi)+p(q-i)+(qi-)$.

3.2 汉字序列识别模型(P2CC)

3.2.1 P2CC 辅助数据

在 P2CC 模型中,训练数据为拼音字符序列和对应的汉字语句.在 CCTVDS 数据集中,共有 14 975 个样本.为了克服汉语的语义模糊性,同时保证辅助数据跟原始样本同源,从而减少冗余信息,加速模型的训练,我们在 CCTV 官网上额外下载了近 30 个月的新闻文稿,并生成对应的拼音序列以对 P2CC 进行预训练.最后训练中,删除新增数据集中过短(字数少于 4)和过长(字数大于 25)的句子,得到的辅助数据共约 304 223 条拼音汉字样本.

3.2.2 Encoder-Decoder 模型

自从 Bengio 提出使用神经网络来训练语言模型[10]后,越来越多的基于语言模型的网络框架[11−13]被用来解决自然语言处理中的各种问题.其中,Encoder-Decoder 框架[14]被广泛应用.

在一个 Encoder-Decoder 框架中,Encoder 和 Decoder 可选用循环神经网络 RNN(LSTM 或者 GRU)的组合,如图 4 所示为 LSTM 组成的 Encoder-Decoder 模型示例.

在 RNN 网络中,t 时刻的隐藏状态是由上一个时刻 t–1 的隐藏状态和输入数据 X_t 共同决定的,见公式(6):

$$h_t=f(h_{t-1},X_t) \tag{6}$$

在 RNN 组成的 Encoder 网络中,实际上的语义编码往往用最后时刻的隐藏状态代替,见公式(7):

$$C=h_T \tag{7}$$

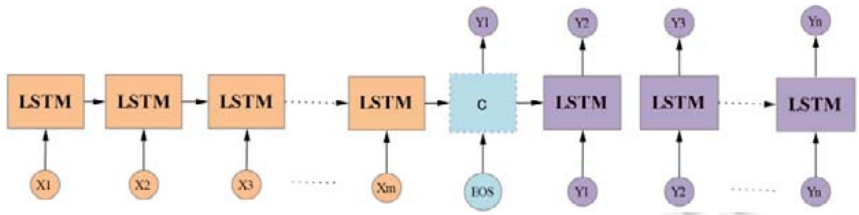

Fig.4 Example of Encoder-Decoder model composed of LSTM

图 4 LSTM 组成的 Encoder-Decoder 网络示例

在图 4 中,若 Encoder 的输入是 $X=\{X_1,X_2,X_3,\ldots,X_m\}$序列,得到 Encoder 的输出为 $C=h_m$.Decoder 是通过当前已经输出的序列 $Y_1,Y_2,\ldots,Y_{t-1}$ 来预测当前 Y_t 的输出,那么 Decoder 输出序列为 $Y=\{Y_1,Y_2,Y_3,\ldots,Y_n\}$的联合概率见公式(8):

$$\left.\begin{aligned} P(Y_t\mid\{Y_1,Y_2,...,Y_{t-1}\},C)=g(Y_{t-1},h_t,C)\\ P(Y)=\prod_{t=1}^{n}P(Y_t\mid\{Y_1,Y_2,...,Y_{t-1}\},C)\end{aligned}\right\} \tag{8}$$

其中,$g(\cdot)$是一种多层网络函数,用于计算输出 Y_t 的概率,是一种非线性映射.

3.2.3 Sequence-to-sequence 模型

基于对语言模型框架 Encoder-Decoder 的理解,我们构建了一个将发音(拼音)字符序列转化为汉字序列的模型 P2CC,如图 5 所示.

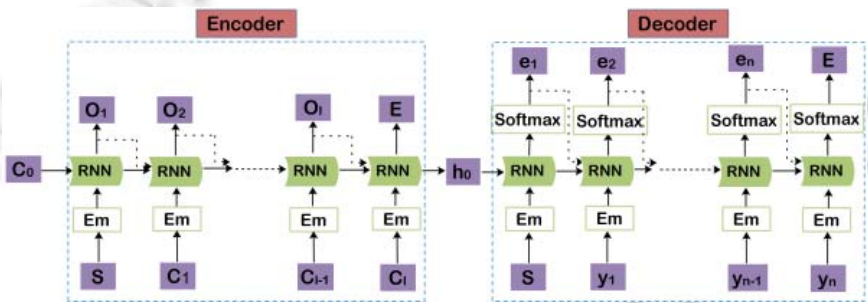

Fig.5 Model architecture of Pinyin-to-Hanzi recognition (P2CC)

图 5 拼音到汉字识别(P2CC)网络模型

在 P2CC 模型中,每一个 RNN 单元表示 2 层 GRU 单元连接(后面的实验分析表明,2 个 GRU 效果最好).在 Encoder 模块的训练阶段,输入是拼音字符序列 $C=C_1,C_2,\ldots,C_l,C_i$ 表示拼音序列中第 i 个字母,为 26 维的向量.Encoder 先将输入序列通过 Embedding(Em 单元)进行升维,然后输入给 RNN,最后,输出序列为 $O=O_1,O_2,\ldots,O$.第 i 时刻的输出向量 O_i 用于参数化下一个时刻输入 C_{i+1} 的预测分布 $\Pr(C_{i+1}|O_i)$.

Encoder 模块的目标函数 $L_{Encoder}(C)$如下公式(9):

$$L_{Encoder}(C)=\max\sum_{i=0}^{l-1}\log\Pr(C_{i+1}\mid C\leqslant i) \tag{9}$$

其中,C 表示拼音语句序列的训练数据.

在 Decoder 模块的训练阶段,输入是汉字序列$\langle S,y_1,y_2,\ldots,y_n\rangle$,$y_i$ 表示输入序列中第 i 个汉字,为 M 维的向量(M 表示数据集 CCTVDS 中汉字的数量).Decoder 先利用 Embedding 对输入序列进行升维,然后再传递给 RNN,最后,通过 softmax 非线性激活运算,输出序列$\langle e_1,e_2,\ldots,e_n,E\rangle$.其中,$S$ 和 E 分别表示输入开始标识和输出结束标识.Decoder 模块的目标函数 $L_{Decoder}(Y)$如公式(10):

$$L_{Decoder}(Y)=\max\sum_{i=0}^{n-1}\log\Pr(y_{i+1}\mid y\leqslant i) \tag{10}$$

当 Encoder 和 Decoder 都训练收敛之后,将 Encoder 的输出作为 Decoder 的输入,从而进行拼音到汉字模型

的整体训练.设定 $F^e(\cdot)$和 $F^d(\cdot)$分别表示已经预训练好的 Encoder 和 Decoder 模型,则整个 P2CC 模型的映射关系如公式(11)所示:

$$\left.\begin{aligned} m_i^{(e)} &= M_{\cdot,c_i}^{(e)}, \\ h_i^{(e)} &= F^e(m_i^{(e)}, h_{i-1}^{(e)}), \\ m_i^d &= M_{\cdot,y_{i-1}}^{(d)}, \\ h_i^d &= F^d(m_i^d, h_{i-1}^{(d)}), \\ p_i^d &= soft\max(W_{he}h_i^d + b_e) \end{aligned}\right\} \tag{11}$$

其中, $m_i^{(e)}$ 表示拼音序列 $C=C_1,C_2,\ldots,C_l$ 中第 i 个发音字符 C_i 对应的 Embedding 向量, $h_i^{(e)}$ 表示第 i 个发音字符 C_i 经过 RNN 运算后的隐藏状态.在训练阶段,y_{i-1} 表示真实的样本标签;在测试阶段,则表示在 i–1 时刻的预测值.

3.3 中文唇语识别整体网络架构

通过对分别已经收敛的拼音序列识别模型 P2P 和汉字序列识别模型 P2CC 进行整体端到端的优化,从而构建中文唇语识别的整体架构 ChLipNet.在 ChLipNet 中,P2P 模型中的 CTC 损失函数被移除,即 P2P 产生的输出序列直接依次作为 P2CC 模块的输入序列.中文唇语识别整体网络模型 ChLipNet 如图 6 所示.

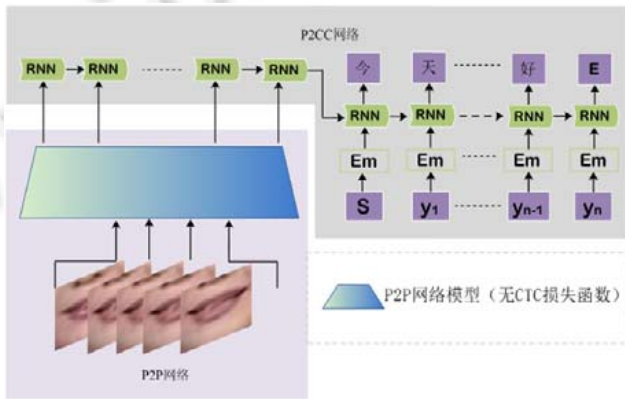

Fig.6 Chinese lip reading architecture of ChLipNet

图 6 中文唇语识别模型 ChLipNet

在 ChLipNet 网络中,输入是一个句子对应的嘴唇图片,相当于普通网络的一个 mini-batch 输入.由于句子长度是可变的,所以 ChLipNet 是一个动态的模型.

4 数据集

4.1 已有数据集介绍

目前,公开常用的唇语识别数据集有:

(1) AVLetter 英语数据集.该数据集由 5 男 5 女录制而成,语料为 26 个字母.每个人要求对每个字母读 3 遍,共计 78 个样本;

(2) AVLetters2 英语数据集.该数据库是对上述 AVLetter 的扩展,由 5 个人录制,每人对 26 个字母重复读 7 遍,共计 182 个样本;

(3) OuluVS1 数据集.本数据集包含了 10 个日常简单的单词或者短语,由 20 个人录制而成,且每人要求对每一个短语重复 5 次;

(4) MIRACL-VC1 数据集.该数据集中的语料为 10 个单词和 10 个短语,每个人对每个单词和短语重复 10

遍,最终获得 3 000 个样本;

(5) GRID 数据集.该数据集是由给定单词根据固定格式组成的“句子”,例如“Place red at J 2, Please”,第 1 个单词为动词,第 2 个单词为颜色词,然后依次分别为介词、字母、数字等,且每个单词有可选的固定候选集,所以这种固定形式构成的数据集从实际意义来说并不是基于“句子”级别的,样本总共大约有 9 000 个;

(6) BBCTV 数据集.该数据集收集了从 2010 年~2016 年期间约 6 年的 BBC 视频,涉及到新闻和讨论内容,约有 118 116 个句子,共 17 428 个单词.

4.2 自建唇语数据集CCTVDS

CCTVDS 来源于 CCTV 官网上 2016 年 4 月~10 月连续 6 个月的新闻联播视频,通过对视频进行半自动化处理,最终生成形如(嘴唇图片序列,中文语句)的数据集.其中,嘴唇图片大小为 120×120.

如图 7 所示为 CCTVDS 数据集中“今天”字段的连续嘴唇图片.

Fig.7 Continuous lip pictures of “today” in CCTVDS dataset

图 7 CCTVDS 数据集中“今天”字段的连续嘴唇图片

下面详细介绍 CCTVDS 数据集的生成过程,具体流程如图 8 所示.

Fig.8 Semi-automatic generation process of CCTVDS dataset

图 8 CCTVDS 数据集的半自动生成过程

4.2.1 视频剪辑

在镜头检测模块,采用图像的全局直方图来判断 CCTV 视频中镜头在主播单独说话和其他场景之间的切换,得到粗略的单人主播视频片段.全局直方图通过统计帧内所有像素点在各个颜色(灰度)等级的个数,按照公

式(12)计算出两帧间的差异值:

$$D(i,i+1)=\sum_{j=1}^{M}|H_i(j)-H_{i+1}(j)| \tag{12}$$

其中,$H_i(j)$表示的是第 i 帧内等级为 j 的直方图的值,M 是直方图的总等级数.然后,再通过人工二次检验核查视频片段的正确性,生成合理可用的视频片段;同时,以 25fps 的频率将视频转换为连续的帧图片,生成包含有人脸的初始样本集.

4.2.2 文本处理和时间戳标记

文本处理主要由人工完成,在保证语句含义合理的前提下,以每行不超过 18 个字的标准规范文本格式.通过 OksrtClient 实现视频和文本语句的自动对齐,同时可获得视频片段中每句话的起始和结束时间戳.根据时间戳和图片的保存频率可将每句话的对应帧图片自动查找出来.

4.2.3 嘴唇检测与分割

1) 人脸检测

在嘴唇检测前,首先对图片进行人脸检测.在所裁剪的图片数据中,均是包含正面角度的人脸,且无显著的光照影响,因此采用经典的人脸检测算法——Viola-Jones 检测器进行检测便可满足需求.该算法在图像矩形区域进行像素处理时,使用积分图方法加速Haar-like特征的计算,再在整张图片上通过滑动窗口提取类Haar特征,并以此通过多个级联弱分类器进行判断.当所有的分类器判定均为正样本时,才把该窗口确定为人脸图片.

2) 嘴唇定位与分割

人脸检测之后,再次用相同算法对人眼睛进行定位,根据人脸的空间几何特征,实现对单张图像中人嘴唇区域的首次定位.然后,根据同一系列的人脸图像嘴唇区域位置信息,利用最小二乘法对图像帧序列中的嘴唇位置进行二次定位,最终将图片裁剪成 120×120 大小的嘴唇区域.如图 9 所示为利用人脸空间几何特征定位嘴唇.

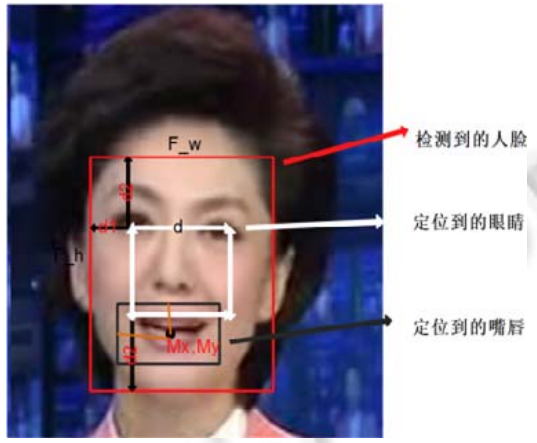

Fig.9 Locate the lips according to the geometric features of face space

图 9 根据人脸空间几何特征定位嘴唇

其中,d 表示眼睛间距,$d1$=0.4d,$d2$=0.7d,F_h=2.4d,F_w=1.8d.嘴唇中点坐标为(Mx,My),其中,Mx=0.5F_w=0.8F_h,嘴唇宽和高分别为 0.3F_w 和 0.67F_h.

4.3 CCTVDS数据集统计

在 CCTVDS 数据集中,每个样本句子对应的嘴唇图片序列长度范围为 10~196,平均包含了 2~25 个汉字,样本数量共计 14 975 个,且样本标签中共计 2 972 355 个汉字.CCTVDS 具体统计见表 3,训练集、测试集和验证集按照 7:2:1 进行划分.

Table 3 Sample statistics of CCTVDS dataset

表 3 CCTVDS 数据集的样本统计数据

标签长度(字数)	嘴唇图片数量范围	样本数量
2~4	10~28	2 136
5~9	32~60	7 452
10~25	65~196	5 387

5 实验结果及分析

对模型进行训练之前,先对数据进行整理,移除标签数据中的特殊字符(例如@,!,?,<等),并将少于 2 个汉字的句子删除.此外,根据句子的长度,将数据分成 3 个子数据集,各自分别包含的汉字个数为 2~9,10~15 和 16~25,样本数量分别为 95 883 619 和 1 768.

实验过程中,使用拼音准确率(PAR)、汉字准确率(HAR)和混淆值来衡量模型的性能.*PAR* 和 *HAR* 定义为 1–错误率=1–(*D*+*S*+*I*)/*N*,其中:*D*,*S*,*I* 分别为从结果序列转换到真实标签时,需要删除、代替和插入的拼音字母或者汉字的数量;*N* 为真实标签中拼音字母或者汉字的数量.混淆度是概率分布的一个衡量,越小的混淆值,表明分布的预测越好.

5.1 P2P网络模型实验分析

5.1.1 训练技巧

批度规范化(batch normalization,简称 BN)在运行过程中需要统计每一个 mini-batch 的一阶统计量和二阶统计量,不适合用于动态的网络结构和循环神经网络 RNN 中.因此在 P2P 模型中,卷积神经网络 ConvNet 使用了 BN 操作,而动态的 RNN 网络在训练过程中尝试使用其他规范算法:层规范化(layer normalization,简称 LN)、参数规范化(weight normalization,简称 WN)、余弦规范化(cosine normalization,简称 CN).

使用 BN 算法后,在训练时,P2P 网络可以设置较高的初始学习率,加速网络的收敛.然而,过高的初始学习率会导致 P2P 模型中的 *n*-LSTM 很难收敛.针对这个问题,我们给 P2P 中不同的模块设置不同的初始学习率.在 ConvNet 网络中,选取较高的初始学习率,如 0.1.而在 *n*-LSTM 网络中,选取较小的初始学习率,如 0.001.通过给两个不同的模块设置不同的学习率,促使 ConvNet 和 *n*-LSTM 尽量同时收敛,进而达到 P2P 模型完全收敛的效果,使其具有更好的拼音序列识别能力.

5.1.2 实验结果

P2P 模型由卷积神经网络 ConvNet、循环神经网络 RNN 和 CTC 组成.实验结果表明:P2P 模型中使用不同的特征提取器 ConvNet,产生的结果也不相同.实验中共尝试了 VGG-M,VGG-16,IncepV2 和 AlexNet 这 4 种不同卷积网络,且网络中的后三层全连接均用一个平均值池化层代替.其中,VGG-M 取得的最高识别准确率为 58.51%,VGG-16 的 41.28%次之,IncepV2 和 AlexNet 的准确率分别为 40.11%和 39.19%,见表 4.

Table 4 Pinyin-level recognition accuracy statistics of P2P network (%)

表 4 P2P 网络的拼音识别准确率统计 (%)

CNN / RNN	AlexNet	IncepV2	Vgg-16	VGG-M
1-256-LSTM	37.18	39.78	40.39	48.27
2-256-LSTM	36.34	38.12	39.83	44.15
3-256-LSTM	35.95	36.37	37.73	40.29
1-256-GRU	35.60	37.51	39.70	41.36
1-512-LSTM	**39.19**	**40.11**	**41.28**	**58.51**
2-512-LSTM	37.64	38.91	40.73	46.70
3-512-LSTM	35.99	36.37	37.64	45.42
1-512-GRU	36.21	37.10	38.51	44.39

在所有实验中,无论采用哪种特征提取器,当循环神经网络为 1-512-LSTM(1 层 512 的 LSTM)时,均取得最好的结果.GRU 表现较差,可能是因为 CTC 损失函数的使用,导致在反向传播的过程中不能找到有效的梯度回

传路径.同时,在训练时发现:以 VGG-16 作为特征提取器时,若对 VGG-16 模型微调,模型的性能会急剧下降;以 IncepV2 或者 ResNet 作为嘴唇图片序列的特征提取器时,由于网络层数太深,训练时很容易发生梯度消失.

此外,实验表明:使用不同的学习率去训练网络,其收敛速度最快;而在 RNN 中分别使用层规范化 LN、参数规范化 WN 和余弦规范化 CN 的速度次之;无任何训练技巧时网络收敛最慢.同时,在对循环神经网络进行规范化(LN、WN 和 CN)时,使用 CN 能够得到比 LN 和 WN 更稳定的损失值.在训练期间,尝试在 P2P 模型中使用双向的 LSTM 神经网络,结果显示:准确率并没有显著提高;与单向的 LSTM 网络相比,反而更加消耗存储空间和计算成本.因此,最终在模型中采用普通的 LSTM 提取图片间的序列信息.

5.2 P2CC网络模型实验分析

5.2.1 Encoder 的训练技巧

大量实验表明:在序列模型中,当时间步数过长时,网络收敛得很慢且很难训练[15].因此,我们将 P2CC 模型分为 Encoder 和 Decoder 两个模块分别进行预训练.

我们从 CCTV 官网上额外下载了从 2016 年 1 月 1 日~2017 年 6 月 15 日的文本数据作为辅助数据集,并采用相同的分组方式,将句子按照长度分别分为 3 个子数据集.辅助数据集中的汉字总数和句子总数分别为 2972355,215697.受 curriculum learning 的启发,模型先用短的数据集进行训练,然后再不断加长训练数据的长度.实验中可观察到:这样训练模型的收敛速度会快很多,并且可以大幅度地减少过拟合.猜测可以把这种训练方式看作是数据增强的一种,所以模型表现出更好的性能.

在预训练过程中,Encoder 的输入是拼音字符序列.而在整个唇语识别 ChLipNet 网络中,拼音序列识别模型 P2P 的输出作为 P2CC 的输入序列.由于 P2P 网络存在损失,无法输出完全正确的拼音字符序列.为了模拟 P2P 模型的生成序列,我们对 P2CC 模型中 Encoder 的输入进行随机增加、删除以及替换等错误处理,并且字符的随机增加、删除和替换率保证在 0 到 25%之间.

例如,将拼音字符序列“jintiantianqihenhao”变为“jingtiantanqqihenghhao”.

5.2.2 Decoder 的训练技巧

在训练循环神经网络时,通常将前一时刻的真实输出作为下一时刻的输入,这有助于模型学习一种超过预期目标的语言模型.然而在推断过程中,样本的真实标签是不可用的,均使用前一时刻的预测输出,但是模型还无法容忍之前时刻的错误预测输出,从而导致较差的性能.于是,我们采用 Bengio 等人提出的预定抽样方法[16],弥补 Decoder 在训练和推断过程中的差异.如图 10 所示为预定抽样方法示例图.

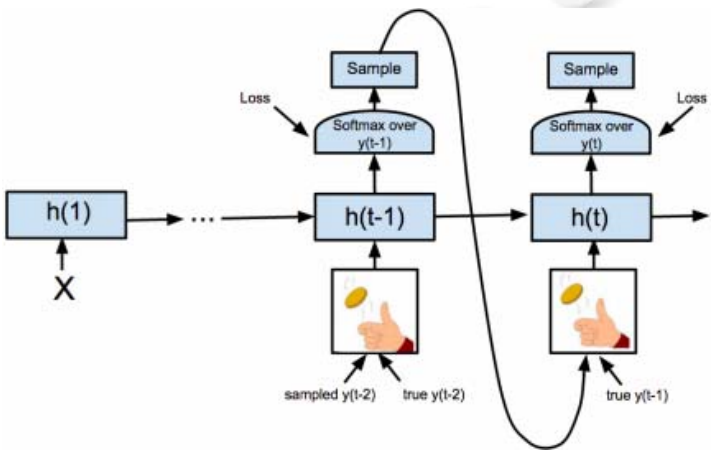

Fig.10 Example of a predetermined sampling method

图 10 预定抽样方法示例

在训练 Decoder 网络时,若输入序列较短,则直接使用真实的样本标签;当输入序列较长时,则从前一时刻的

输出中随机采样,而不是始终使用真实的样本标签作为下一个 RNN 单元的输入,采样概率随着时间从 0 增加到 0.2.实验表明:当抽样概率大于 0.2 时,无法实现 Decoder 网络的稳定学习.

5.2.3 实验结果

P2CC 模型的参数初始化范围为[−0.02,0.02],初始学习率设为 0.001.实验结果显示:P2CC 模型结构直接决定着其性能,在不同的 RNN 配置下,模型混淆度和损失值均不相同.随着随机错误率的增加,网络的损失值也略微增加,混淆度和最小损失值分别为 2.69 和 0.99;错误率为 15%,则最小混淆度和最小损失值分别为 2.77 和 1.02;当 RNN 采用 2-1024-GRU 时,P2CC 在 4 种不同错误率下的混淆度和损失值均为最小.错误率为 10%时,错误率为 20%时,最小混淆度和最小损失值分别为 2.97 和 1.09;错误率为 25%时,最小混淆度和最小损失值分别为 3.00 和 1.10.当采用 2-512-GRU 时,模型的最小混淆度为 2.94,最小损失值为 1.08.由于当采用 1024-GRU 时,其结果并不明显优于 512-GRU 的结果,而且还存在大量的网络计算,因此,从实验结果和网络训练效率层面来看,RNN 采用 2-512-GRU 便可很好地满足汉字序列识别研究.表 5 为 P2CC 模型在随机抽样率为 10%时,不同 RNN 网络的实验结果.

Table 5 Experimental results of P2CC with different RNN networks under 10% sampling rate

表 5 P2CC 在错误率为 10%下不同 RNN 网络下的实验结果

RNN		混淆度		损失值	
单元数	层数	LSTM	GRU	LSTM	GRU
256	1	6.89	12.06	1.93	2.49
	2	6.49	7.54	1.87	2.02
	3	9.12	12.56	2.21	2.5
512	1	5.05	3.29	1.62	1.19
	2	3.97	**2.94**	1.38	**1.08**
	3	5.87	3.71	1.77	1.31
1024	1	3.49	3.06	1.25	1.12
	2	2.77	**2.69**	1.02	**0.99**
	3	3.89	3.32	1.36	1.20

5.3 ChLipNet模型实验分析

通过将 ChLipNet 网络与其他唇语识别模型在 CCTVDS 数据集上进行对比分析,进而来测试 ChLipNet 模型的整体性能.表 6 所示为不同唇语识别模型在 CCTVDS 上的实验对比结果,并且包括模型的各自网络结构以及相关的训练数据集和语言类别,其中,*AiC* 表示模型在各自原文章中的准确率,*AiS* 表示模型在 CCTCDS 上重新训练所取得的最高句子识别准确率,*AiP* 表示在 CCTVDS 上所取得的拼音字母序列的准确率.

Table 6 Experimental results of different lip reading models on CCTVDS

表 6 不同唇语模型在 CCTVDS 上的实验结果

唇语模型	网络结构	数据&语言	*AiC* (%)	*AiS* (%)	*AiP* (%)
WLAS	CNN+LSM+Attention	BBCTV&Eng	46.8	**36.7**	**49.8**
[Assael, *et al*., 2016]	STCNN+BiLSTM+CTC	GRID&Eng	93.4	28.9	41.6
[Wand, *et al*., 2016]	NN+LSTM	GRID&Eng	79.6	16.7	30.5
[Noda, *et al*., 2014]	CNN+GMM+HMM	JAVD&JAP	37	18.6	35.7
[Garg, *et al*., 2016]	CNN+LSTM	MIRACL-VC&Eng	76	29.15	47.3
ChLipNet	CNN+LSTM+GRU+CTC	CCTV&Chin	–	**45.7**	**58.5**

通过实验结果,在基于深度学习的唇语识别模型中,可得出以下结论.

(1) 对比 NN+LSTM 和 CNN+LSTM 的实验结果,发现 CNN 表现出强大的特征提取能力;

(2) 通过和 STCNN+BiLSTMCTC 对比,时空卷积网络的性能不一定比传统的 2D-CNN 更优,尤其是在唇语识别研究中;

(3) 从序列模型中可得,RNN(LSTM/GRU)在语义解码方面具有强大的优势,适用于文本编译和生成.

除了 WLAS 模型外,其他唇语识别模型都是用来预测非中文的单词或者短语的.因此,它们在句子级别的 CCTVDS 数据集上效果较差.另外,WLAS 模型的输入为嘴唇图片和音频数据,由于 CCTVDS 数据集中不包含音

频信息,因此 WLAS 在 CCTVDS 也不能产生较高的准确率.正如实验结果所示:本文提出的中文句子级别的唇语识别模型 ChLipNet 可分别取得 45.7%的句子准确率和 58.5%的拼音序列准确率,而 WLAS 相应的准确率分别为 36.7%和 49.8%.

6 总 结

本文首次提出了端到端的中文唇语识别模型 ChLipNet,该模型可以将输入嘴唇图片不用分割直接自动地转化为汉语句子输出.在训练过程中,两个不同的网络模块分别各自解决图片到拼音和拼音到汉字的识别,当这两个模块分别训练好后,再整体进行端到端的优化调整.通过增加训练技巧,实验结果表明:在中文唇语数据集 CCTVDS 上,ChLipNet 的性能超过之前相关的唇语架构.

在之后的工作中,我们将会进行以下的尝试.

(1) 采集更为丰富的数据集.在深度神经网络中,数据集的样本量直接决定了网络模型参数训练的完成度以及分类回归的准确率;

(2) 对网络输入进行扩充.目前的 ChLipNet 网络只支持视觉信息的输入,我们将尝试增加网络输入数据类型,实现多种类型数据的输入.例如,在输入中增加音频数据,实现视觉和音频的整合,提高网络识别精度;

(3) 将 ChLipNet 应用到不同的方言中;

(4) 在 Encoder-Decoder 模型使用 Attention 机制,Attention 机制能够避免因为输入序列过长造成的错误输出,使得网络输出更加精准的结果.此外,Attention 能够解释和可视化我们的网络模型,加强对网络的理解.

References:

[1] Chung JS, Senior A, Vinyals O, *et al*. Lip reading sentences in the wild. arXiv:1611.05358, 2016. 3444−3453.

[2] Assael YM, Shillingford B, Whiteson S, *et al*. Lipnet: Sentence-level lipreading. arXiv:1611.01599, 2016.

[3] Petajan ED. Automatic lipreading to enhance speech recognition (speech reading) [Ph.D. Thesis]. Universityof Illinois at Urbana-Champaign, 1984.

[4] Petajan E, Bischoff B, Bodoff D, *et al*. An improved automatic lipreading system to enhancespeech recognition. In: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems. ACM, 1988. 19−25.

[5] Noda K, Yamaguchi Y, Nakadai K, *et al*. Lipreading using convolutional neural network. In: Proc. of the 15th Annual Conf. of the International Speech Communication Association, 2014.

[6] Chung JS, Zisserman A. Lip reading in the wild. In: Proc. of the Asian Conf. on Computer Vision. 2016.

[7] Chung JS, Zisserman A. Out of time: Automated lip sync in the wild. In: Proc. of the Workshop on Multi-view lipreading. ACCV, 2016.

[8] Wand M, Koutník J, Schmidhuber J. Lipreading with long short-term memory. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. 2016. 6115−6119.

[9] Graves A, Gomez F. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proc. of the Int'l Conf. on Machine Learning. ACM, 2006. 369−376.

[10] Bengio Y, Ducharme R, Vincent P, *et al*. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3(Feb):1137−1155.

[11] Collobert R, Weston J, Bottou L, *et al*. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011,12(Aug):2493−2537.

[12] Mikolov T, Karafiát M, Burget L, *et al*. Recurrent neural network based language model. In: Proc. of the Conf. of the Int'l Speech Communication Association (INTERSPEECH 2010). Makuhari: DBLP, 2010. 1045−1048.

[13] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Computer Science, 2014.

[14] Mikolov TA. Statistical language models based on neural networks. 2012.

[15] Feichtenhofer C, Pinz A, Zisserman A. Convolutionaltwo-Stream network fusion for video action recognition. In: Proc. of the CVPR. 2016.

[16] Bengio S, Vinyals O, Jaitly N, Shazeer N. Scheduledsampling for sequence prediction with recurrent neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2015. 1171−1179.

张晓冰(1992－),女,河南洛阳人,博士生,主要研究领域为深度学习,视觉处理.

龚海刚(1975－),男,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为计算机网络与系统安全,云计算与大数据处理,深度学习.

杨帆(1993－),女,硕士,CCF 学生会员,主要研究领域为深度学习.

戴锡笠(1990－),男,博士,主要研究领域为机器视觉,机器学习,深度学习.