(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116416967 B
(45)授权公告日 2024.09.24

(21)申请号 202111651840.0
(22)申请日 2021.12.30
(65)同一申请的已公布的文献号
申请公布号 CN 116416967 A
(43)申请公布日 2023.07.11
(73)专利权人 重庆大学
地址 400044 重庆市沙坪坝区沙正街174号
专利权人 重庆医科大学
(72)发明人 张美伟 余娟 吕洋 李文沅 余维华 王香霖
(74)专利代理机构 重庆缙云专利代理事务所(特殊普通合伙) 50237
专利代理师 王翔
(51)Int.Cl.
G10L 15/00(2013.01)
G10L 15/06(2013.01)
G10L 15/26(2006.01)
G10L 25/18(2013.01)
(56)对比文件
CN 107679491 A,2018.02.09
CN 112151030 A,2020.12.29
审查员 高莹

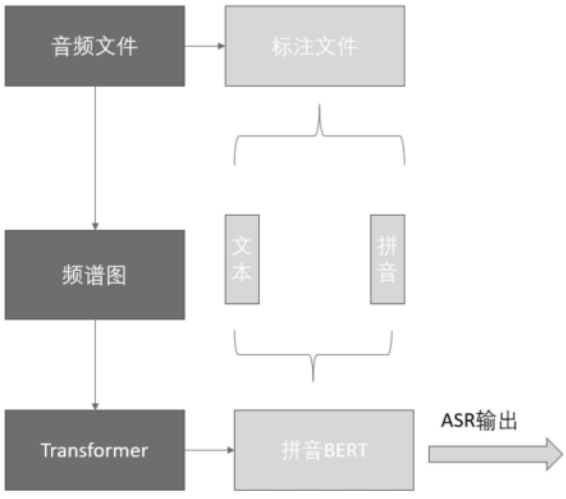
权利要求书2页 说明书6页 附图3页

(54)发明名称

一种通过迁移学习提升重庆方言语音识别的方法

(57)摘要

本发明公开一种通过迁移学习提升重庆方言语音识别的方法,步骤为:1)获取语音数据;2)得到语音频谱图;3)对语音频谱图向量化,得到向量V;4)获取transformer模型的输入X;5)将参数Q、参数K、参数V输入到transformer模型的编码器中,得到编码器输出Y1和编码器输出Y2;6)将编码器输出Y1和编码器输出Y2输入到transformer模型的解码器中,得到语音识别文本;8)确定拼音BERT模型的输入x;9)将输入x输入到拼音BERT模型中,得到语音识别结果。本发明通过pipeline设计模式,将ASR中的声学模型,语言模型独立开,增强了ASR模型选择的多样性。



CN 116416967 B

1.一种通过迁移学习提升重庆方言语音识别的方法,其特征在于,包括以下步骤:

1)获取语音数据;

2)对语音数据进行傅里叶转换,得到语音频谱图;

3)利用VGG网络对语音频谱图向量化,得到向量v;

4)获取transformer模型的输入X;所述transformer模型包括编码器encoder1、编码器encoder2和解码器decoder;

5)对输入X进行转化,得到参数Q、参数K、参数V;

6)将参数Q、参数K、参数V输入到transformer模型的编码器encoder1和编码器encoder2中,分别得到编码器输出Y1和编码器输出Y2;

7)将编码器输出Y1和编码器输出Y2输入到transformer模型的解码器中,得到语音识别文本;

8)基于语音识别文本,确定拼音BERT模型的输入x;

9)将输入x输入到拼音BERT模型中,得到语音识别结果;

向量v如下所示:

$v=VGG(DFT(A))$ (1)

式中,A为语音数据;

transformer的输入X如下所示:

$X=PE(DFT(A))+Fbank(v)$ (2)

式中,PE为位置编码函数;

参数Q、参数K、参数V如下所示:

$Q=XW^{Q},K=XW^{K},V=XW^{V}$ (3)。

2.根据权利要求1所述的一种通过迁移学习提升重庆方言语音识别的方法,其特征在于:所述语音数据包括方言。

3.根据权利要求1所述的一种通过迁移学习提升重庆方言语音识别的方法,其特征在于:编码器encoder包括多头注意力层、前向传播层;

多头注意力层的输出MultiHead(Q,K,V)如下所示:

$MultiHead(Q,K,V)=Concat(head_1,...,head_h)W^{O}$ (4)

其中,参数$head_i$如下所示:

$head_i=Attention(QW_i^{Q},KW_i^{K},VW_i^{V}),i=1,2,...,h$ (5)

式中,h为attention层的层数;W_i^{Q}、W_i^{K}、W_i^{V}为第i层权重;

注意力Attention(Q,K,V)如下所示:

$$Attention(Q,K,V)=\mathrm{softmax}(\frac{QK^{T}}{\sqrt{d_k}})V \quad (6)$$

式中,$\sqrt{d_k}$为归一化参数;

前向传播层的输出FFN(x')如下所示:

$FFN(x')=\max(0,x'W_1+b_1)W_2+b_2$ (7)

前向传播层的输入x'如下所示:

$x'=norm(X+MultiHead(Q,K,V))$ (8)

encoder编码器的输出Y如下所示：

$$Y=FFN(x') \quad (9)$$

。

4.根据权利要求1所述的一种通过迁移学习提升重庆方言语音识别的方法，其特征在于，拼音BERT模型的输入x如下所示：

$$x=Concat(CE,GE,PYE)W_F+PE' \quad (10)$$

式中，CE表示字嵌入；GE表示字形嵌入；PYE表示拼音嵌入；PE'表示位置嵌入；W_F表示全连接层；Concat表示向量拼接。

5.根据权利要求4所述的一种通过迁移学习提升重庆方言语音识别的方法，其特征在于，字形嵌入GE如下所示：

$$GE=Concat(flatten(I_1),flatten(I_2),flatten(I_3))W_G \quad (11)$$

式中，I_1、I_2、I_3表示字形图像；W_G表示全连接层；flatten表示将二维图像转化为一维向量。

6.根据权利要求4所述的一种通过迁移学习提升重庆方言语音识别的方法，其特征在于，拼音嵌入PYE如下所示：

$$PYE=max\text{-}pooling(CNN(S)) \quad (12)$$

式中，S表示拼音序列；max-pooling表示最大池化；CNN表示卷积计算。

7.根据权利要求1所述的一种通过迁移学习提升重庆方言语音识别的方法，其特征在于，语音识别结果$p(x_1,x_2,x_3,\cdots,x_n)$如下所示：

$$\begin{aligned}&p(x_1,x_2,x_3,\cdots,x_n)=p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)\cdots p(x_n|x_1,x_2,\cdots,x_{n-1})\\&=p(x_3)p(x_1|x_3)p(x_2|x_3,x_1)\cdots p(x_n|x_3,x_1,\cdots,x_{n-1})\ldots p(x_{n-1})\\&=\ldots\\&=p(x_1|x_{n-1})p(x_n|x_{n-1},x_1)\cdots p(x_2|x_{n-1},x_1,\ldots,x_3)\end{aligned} \quad (13)$$

式中，$p(x_2|x_1)$表示语音识别文本概率分布。

一种通过迁移学习提升重庆方言语音识别的方法

技术领域

[0001] 本发明涉及领域,具体是一种通过迁移学习提升重庆方言语音识别的方法。

背景技术

[0002] 语音识别技术起步于上世纪五十年代,现如今已取得了不错成绩,同样,自然语言处理技术伴随深度学习技术的发展,也逐渐从统计模型逐渐发展成为了深度语义模型,在一些经典的NLP场景得到了广泛的应用,比如NLG任务,命名体识别等任务。

[0003] 人工智能产品在各个IT领域应用广泛,ASR技术是人工智能的重要组成部分,实现了计算机能“听懂”人的语音,ASR技术的发展有助于人与更多的人工智能产品进行交流,实现“人机交互”,从而让人们享受到科技发展给生活带来的便利与高效。

[0004] ASR的实现可分为pipeline或者end2end思路,其中主要区别在于声学模型的识别单元上。模型识别单元大小(词发音模型、字发音模型、半音节模型或音素模型)对语音训练数据量大小、语音识别率,以及灵活性有较大的影响。对中等词汇量以上的语音识别系统来说,识别单元小,则计算量也小,所需的模型存储量也小,要求的训练数据量相对也小,但带来的问题是对应语音段的定位和分割困难,以及更复杂的识别模型规则。通常大的识别单元易于包括协同发音在模型中,这有利于提高系统的识别率,但要求的训练数据相对增加。

[0005] 综上看,基于统计的语言模型受预料大小影响,效果有限,且统计信息在语义层面表达能力有限。现有技术在声学模型中没有融入语言模型,且大部分基于深度学习的声学模型采用CNN或者类RNN结构,计算效率有限。BERT等模型在文本生成任务中由于其双向attention机制,在NLG任务中效果有限。

发明内容

[0006] 本发明的目的是提供一种通过迁移学习提升重庆方言语音识别的方法,包括以下步骤:

[0007] 1)获取语音数据。所述语音数据包括方言。

[0008] 2)对语音数据进行傅里叶转换,得到语音频谱图。

[0009] 3)利用VGG网络对语音频谱图向量化,得到向量v。

[0010] 向量v如下所示:

[0011] $v=VGG(DFT(A))$ (1)

[0012] 式中,A为语音数据。

[0013] 4)获取transformer模型的输入X。所述transformer模型包括编码器encoder1、编码器encoder2和解码器decoder。

[0014] transformer的输入X如下所示:

[0015] $X=PE(DFT(A))+Fbank(v)$ (2)

[0016] 式中,PE为位置编码函数。

[0017] 5)对输入X进行转化,得到参数Q、参数K、参数V。

[0018] $Q=XW^{Q},K=XW^{K},V=XW^{V}$ (3)

[0019] 6)将参数Q、参数K、参数V输入到transformer模型的编码器encoder1和编码器encoder2中,分别得到编码器输出Y1和编码器输出Y2。

[0020] 参数Q、参数K、参数V如下所示:

[0021] 编码器encoder包括多头注意力层、前向传播层。

[0022] 多头注意力层的输出MultiHead(Q,K,V)如下所示:

[0023] $\mathrm{MultiHead}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_h)W^{O}$ (4)

[0024] 其中,参数head_i如下所示:

[0025] $\mathrm{head}_i=\mathrm{Attention}(QW_i^{Q},KW_i^{K},VW_i^{V}),i=1,2,\ldots,h$ (5)

[0026] 式中,h为attention层的层数;W_i^{Q}、W_i^{K}、W_i^{V}为第i层权重。

[0027] 注意力Attention(Q,K,V)如下所示:

[0028] $$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}(\frac{QK^{T}}{\sqrt{d_k}})V \quad (6)$$

[0029] 式中,$\sqrt{d_k}$为归一化参数;

[0030] 前向传播层的输出FFN(x)如下所示:

[0031] $\mathrm{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2$ (7)

[0032] 前向传播层的输入x如下所示:

[0033] $x=\mathrm{norm}(X+\mathrm{MultiHead}(Q,K,V))$ (8)

[0034] encoder编码器的输出Y如下所示:

[0035] $Y=\mathrm{FFN}(x)$ (9)

[0036] 7)将编码器输出Y1和编码器输出Y2输入到transformer模型的解码器中,得到语音识别文本。

[0037] 8)基于语音识别文本,确定拼音BERT模型的输入x。

[0038] 拼音BERT模型的输入x如下所示:

[0039] $x=\mathrm{Concat}(\mathrm{CE},\mathrm{GE},\mathrm{PYE})W_F+\mathrm{PE}$ (10)

[0040] 式中,CE表示字嵌入。GE表示字形嵌入。PYE表示拼音嵌入。PE表示位置嵌入。W_F表示全连接层。Concat表示向量拼接。

[0041] 字形嵌入GE如下所示:

[0042] $\mathrm{GE}=\mathrm{Concat}(\mathrm{flatten}(I_1),\mathrm{flatten}(I_2),\mathrm{flatten}(I_3))W_G$ (11)

[0043] 式中,I_1、I_2、I_3表示字形图像。W_G表示全连接层。flatten表示将二维图像转化为一维向量。

[0044] 拼音嵌入PYE如下所示:

[0045] $\mathrm{PYE}=\text{max-pooling}(\mathrm{CNN}(S))$ (12)

[0046] 式中,S表示拼音序列。max-pooling表示最大池化。CNN表示卷积计算。

[0047] 9)将输入x输入到拼音BERT模型中,得到语音识别结果。

[0048] 语音识别结果$p(x_1,x_2,x_3,\ldots,x_n)$如下所示:

[0049] $$p(x_1,x_2,x_3,\cdots,x_n)=p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)\ldots p(x_n|x_1,x_2,\ldots,x_{n-1})$$

[0050] $$=p(x_3)p(x_1|x_3)p(x_2|x_3,x_1)\ldots p(x_n|x_3,x_1,\ldots,x_{n-1})\ldots p(x_{n-1})$$

[0051] $=\ldots$

[0052] $=p(x_1|x_{n-1})p(x_n|x_{n-1},x_1)\ldots p(x_2|x_{n-1},x_1,\ldots,x_3)$ (13)

[0053] 式中,$p(x_2|x_1)$表示语音识别文本概率分布。

[0054] 本发明的技术效果是毋庸置疑的,本发明将ASR技术中的语言模型由统计模型换位基于大规模预料的预训练模型,能更全面的捕捉到语义层面信息,并通过pipeline设计模式,将ASR中的声学模型,语言模型独立开,增强了ASR模型选择的多样性。

[0055] 本发明通过将位置的embedding放在声学模型中,使声学模型具备了一定的语言模型的能力,同时增强了声学模型提取到声学信息并完成解码的有效性。

[0056] 本发明通过引入拼音,字形等embedding,全方位捕捉语言的信息,这和ASR在中文中存在的一些难点,例如声母相同,韵母相同,发音相同等特点是相匹配的,同时也提高了语言模型在解码过程中的准确性。

[0057] 本发明将UniLM模型应用到ASR场景,借助UniLM算法在文本生成任务中的有效性,提高了ASR解码的准确率。

[0058] 针对近年来NLP技术采用预训练方法在大量NLP任务上的卓越表现,本发明提出利用transformer充当声学模型得到初步的ASR结果,再根据语言场景的预料,结合其拼音预训练得到的预训练模型(UniLM)充当语言模型,最终得到ASR结果输出。

附图说明

[0059] 图1为语音识别流程;

[0060] 图2为语音特征处理流程;

[0061] 图3为transformer结构图;

[0062] 图4为输入整理;

[0063] 图5为输入信息融合;

[0064] 图6为拼音embedding。

具体实施方式

[0065] 下面结合实施例对本发明作进一步说明,但不应该理解为本发明上述主题范围仅限于下述实施例。在不脱离本发明上述技术思想的情况下,根据本领域普通技术知识和惯用手段,做出各种替换和变更,均应包括在本发明的保护范围内。

[0066] 实施例1:

[0067] 参见图1、图2、图3、图4、图5、图6,一种通过迁移学习提升重庆方言语音识别的方法,包括以下步骤:

[0068] 1)获取语音数据。所述语音数据包括方言。

[0069] 2)对语音数据进行傅里叶转换,得到语音频谱图。

[0070] 3)利用VGG网络对语音频谱图向量化,得到向量v。

[0071] 向量v如下所示:

[0072] $v=VGG(DFT(A))$ (1)

[0073] 式中,A为语音数据。

[0074] 4)获取transformer模型的输入X。所述transformer模型包括编码器encoder1、编

码器encoder2和解码器decoder。

[0075] transformer的输入X如下所示：

[0076] $X=PE(DFT(A))+Fbank(v)$ (2)

[0077] 式中,PE为位置编码函数。Fbank()表示语音特征提取操作。

[0078] 5)对输入X进行转化,得到参数Q、参数K、参数V。

[0079] $Q=XW^{Q},K=XW^{K},V=XW^{V}$ (3)

[0080] 6)将参数Q、参数K、参数V输入到transformer模型的编码器encoder1和编码器encoder2中,分别得到编码器输出Y1和编码器输出Y2。

[0081] 参数Q、参数K、参数V如下所示：

[0082] 编码器encoder包括多头注意力层、前向传播层。

[0083] 多头注意力层的输出MultiHead(Q,K,V)如下所示：

[0084] $MultiHead(Q,K,V)=Concat(head_1,...,head_h)W^{O}$ (4)

[0085] 其中,参数$head_i$如下所示：

[0086] $head_i=Attention(QW_i^{Q},KW_i^{K},VW_i^{V}),i=1,2,...,h$ (5)

[0087] 式中,h为attention层的层数;W_i^{Q}、W_i^{K}、W_i^{V}为第i层权重。

[0088] 注意力Attention(Q,K,V)如下所示：

[0089] $$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}(\frac{QK^{T}}{\sqrt{d_k}})V \quad (6)$$

[0090] 式中,$\sqrt{d_k}$为归一化参数;

[0091] 前向传播层的输出FFN(x)如下所示：

[0092] $FFN(x)=\max(0,xW_1+b_1)W_2+b_2$ (7)

[0093] 前向传播层的输入x如下所示：

[0094] $x=norm(X+MultiHead(Q,K,V))$ (8)

[0095] encoder编码器的输出Y如下所示：

[0096] $Y=FFN(x)$ (9)

[0097] 7)将编码器输出Y1和编码器输出Y2输入到transformer模型的解码器中,得到语音识别文本。

[0098] 8)基于语音识别文本,确定拼音BERT模型的输入x。

[0099] 拼音BERT模型的输入x如下所示：

[0100] $x=Concat(CE,GE,PYE)W_F+PE$ (10)

[0101] 式中,CE表示字嵌入。GE表示字形嵌入。PYE表示拼音嵌入。PE表示位置嵌入。W_F表示全连接层。Concat表示向量拼接。

[0102] 字形嵌入GE如下所示：

[0103] $GE=Concat(flatten(I_1),flatten(I_2),flatten(I_3))W_G$ (11)

[0104] 式中,I表示字形图像。W_G表示全连接层。flatten表示将二维图像转化为一维向量。

[0105] 拼音嵌入PYE如下所示：

[0106] $PYE=max\text{-}pooling(CNN(S))$ (12)

[0107] 式中,S表示拼音序列。max-pooling表示最大池化。CNN表示卷积计算。

[0108] 9)将输入x输入到拼音BERT模型中,得到语音识别结果。

[0109] 语音识别结果$p(x_1,x_2,x_3,...,x_n)$如下所示:

[0110] $p(x_1,x_2,x_3,...,x_n)=p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)...p(x_n|x_1,x_2,...,x_{n-1})$

[0111] $=p(x_3)p(x_1|x_3)p(x_2|x_3,x_1)\cdots p(x_n|x_3,x_1,...,x_{n-1})...p(x_{n-1})$

[0112] $=...$

[0113] $=p(x_1|x_{n-1})p(x_n|x_{n-1},x_1)...p(x_2|x_{n-1},x_1,...,x_3)$ (13)

[0114] 式中,$p(x_2|x_1)$表示语音识别文本概率分布。

[0115] 实施例2:

[0116] 一种通过迁移学习提升重庆方言语音识别的方法,包括以下步骤:

[0117] 1)根据音频,采用信号处理技术以及傅里叶变换得到单个音频文件的频谱图,通过VGG网络结构提取整个结构图的向量表达。

[0118] 其公式可表示为:

[0119] V=VGG(DFT(A))

[0120] A:音频文件;DFT:离散傅里叶变换;VGG:VGG网络;V:VGG输出的向量表达

[0121] 2)根据频谱图,得到每个频谱单元在原图的位置信息,并将其embedding向量化之后,与Fbank一并输入到transformer里。

[0122] encoder计算流程及公式:

[0123] transformer输入X由位置编码和Fbank两部分组成,PE为位置编码函数:

[0124] X=PE(DFT(A))+Fbank(V)

[0125] 将输入X转化为Q,K,V:

[0126] $Q=XW^Q,K=XW^K,V=XW^V$

[0127] 注意力计算公式:

[0128] $$\text{Attention}(Q,K,V)=\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

[0129] 多头注意力层:

[0130] $\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1,...,\text{head}_h)W^O$

[0131] 其中:

[0132] $\text{head}_i=\text{Attention}(QW_i^Q,KW_i^K,VW_i^V),i=1,2,...,h$

[0133] 前向传播层:

[0134] $FFN(x)=\max(0,xW_1+b_1)W_2+b_2$

[0135] 其中:

[0136] x=norm(X+MultiHead(Q,K,V))

[0137] encoder的输出:

[0138] Y=FFN(x)

[0139] decoder计算过程与encoder类似,具体参照图3,便不再赘述。

[0140] 3)汉字的最大特性有两个方面:一是字形,二是拼音。汉字是一种典型的意音文字,从其起源来看,它的字形本身就蕴含了一部分语义。比如,“江河湖泊”都有偏旁三点水,这表明它们都与水有关。而从读音来看,汉字的拼音也能在一定程度上反映一个汉字的语

义,起到区别词义的作用。比如,“乐”字有两个读音,yuè与lè,前者表示“音乐”,是一个名词;后者表示“高兴”,是一个形容词。而对于一个多音字,单单输入一个“乐”,模型是无法得知它应该是代表“音乐”还是“快乐”,这时候就需要额外的读音信息进行去偏。从汉字本身的这两大特性出发,将汉字的字形与拼音信息融入到中文语料的预训练过程。一个汉字的字形向量由多个不同的字体形成,而拼音向量则由对应的罗马化的拼音字符序列得到。二者与字向量一起进行融合,得到最终的融合向量,作为预训练模型的输入。模型使用全词掩码(Whole Word Masking)和字掩码(Character Masking)两种策略训练,使模型更加综合地建立汉字、字形、读音与上下文之间的联系。

[0141] $X=Concat(CE,GE,PYE)W_F+PE$

[0142] CE:字嵌入,GE:字形嵌入,PYE:拼音嵌入,PE:位置嵌入,WF:全连接层,X:BERT的输入,Concat:向量拼接。

[0143] 底层的融合层(Fusion Layer)融合了除字嵌入(Char Embedding)之外的字形嵌入(Glyph Embedding)和拼音嵌入(Pinyin Embedding),得到融合嵌入(Fusion Embedding),再与位置嵌入相加,就形成模型的输入。字形嵌入使用不同字体的汉字图像得到。每个图像都是24*24的大小,将仿宋、行楷和隶书这三种字体的图像向量化,拼接之后再经过一个全连接W_G,就得到了汉字的字形嵌入。

[0144] 该过程如图5所示:

[0145] $GE=Concat(flatten(I_1),flatten(I_2),flatten(I_3))W_G$

[0146] I:字形图像,WG:全连接层,GE:字形嵌入,flatten:将二维图像转化为一维向量。

[0147] 拼音嵌入首先使用pypinyin将每个汉字的拼音转化为罗马化字的字符序列,其中也包含了音调。比如对汉字“猫”,其拼音字符序列就是“mao1”。对于多音字如“乐”,pypinyin能够非常准确地识别当前上下文中正确的拼音。

[0148] 该过程如图6所示:

[0149] PYE=max-pooling(CNN(S))

[0150] S:拼音序列,max-pooling:最大池化,CNN:卷积计算,PYE:拼音嵌入。

[0151] 4)结合预训练模型UniLM生成最终ASR识别结果,相比较于基于语言模型的生成模型,BERT由于其双向解码无法满足语言模型要求,但是通过Mask attention手动控制解码方向,从双向变为单向即可:

[0152] $p(x_1,x_2,x_3,\cdots,x_n)=p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)\cdots p(x_n|x_1,x_2,\ldots,x_{n-1})$

[0153] $=p(x_3)p(x_1|x_3)p(x_2|x_3,x_1)\cdots p(x_n|x_3,x_1,\ldots,x_{n-1})\cdots p(x_{n-1})$

[0154] $=\cdots$

[0155] $=p(x_1|x_{n-1})p(x_n|x_{n-1},x_1)\cdots p(x_2|x_{n-1},x_1,\ldots,x_3)$

[0156] $x_1,x_2,\cdots,x_n$任意一种“出场顺序”都有可能。原则上来说,每一种顺序都对应着一个模型,所以原则上就有n!个语言模型.实现一种顺序的语言模型,就相当于将原来的下三角形式的Mask以某种方式打乱。正因为Attention提供了这样的一个n×n的Attention矩阵,本发明才有足够多的自由度去以不同的方式去Mask这个矩阵,从而实现多样化的效果。从而满足了语言模型的要求。

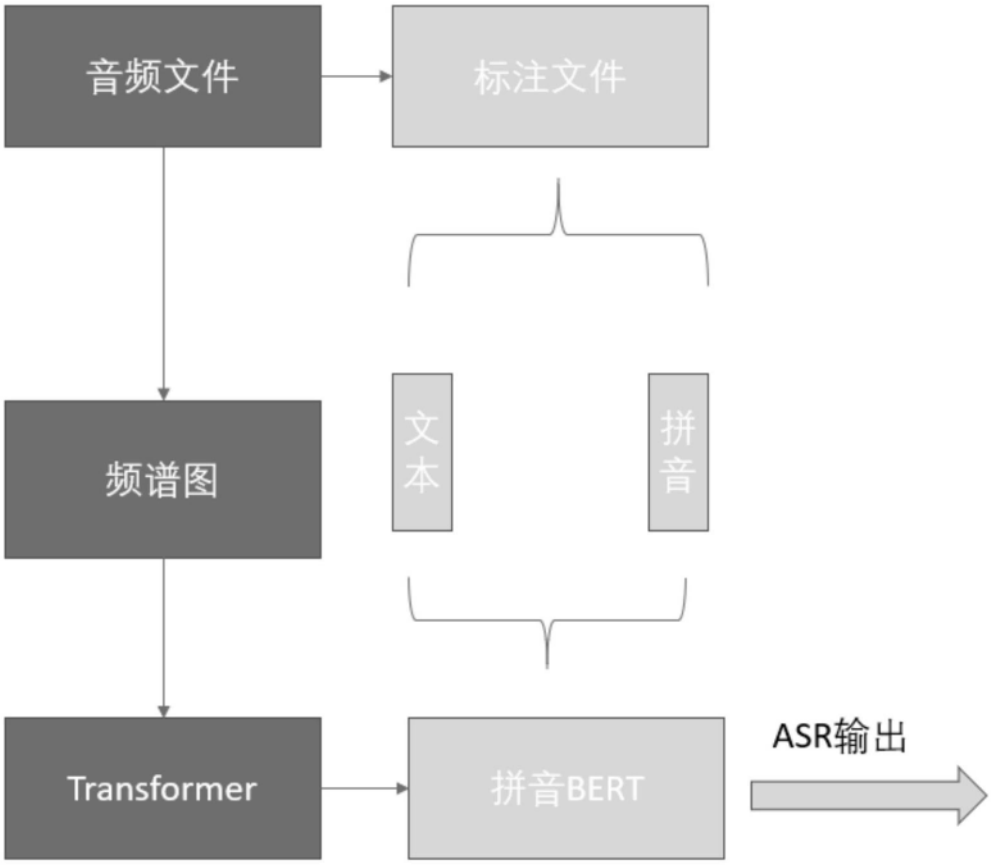

音频文件
标注文件
频谱图
文本
拼音
Transformer
拼音BERT
ASR输出

图1

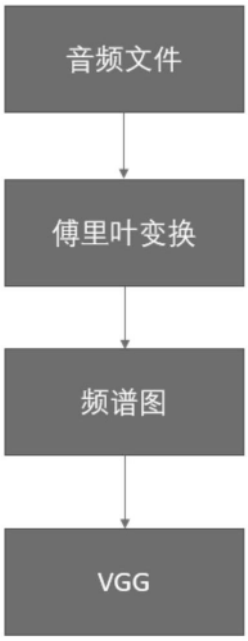

音频文件
傅里叶变换
频谱图
VGG

图2

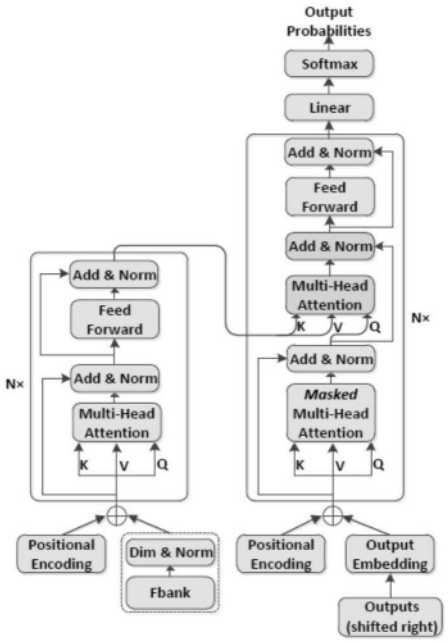

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
K V Q
Add & Norm
Masked Multi-Head Attention
K V Q
N×
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
K V Q
N×
Positional Encoding
Dim & Norm
Fbank
Positional Encoding
Output Embedding
Outputs (shifted right)

图3

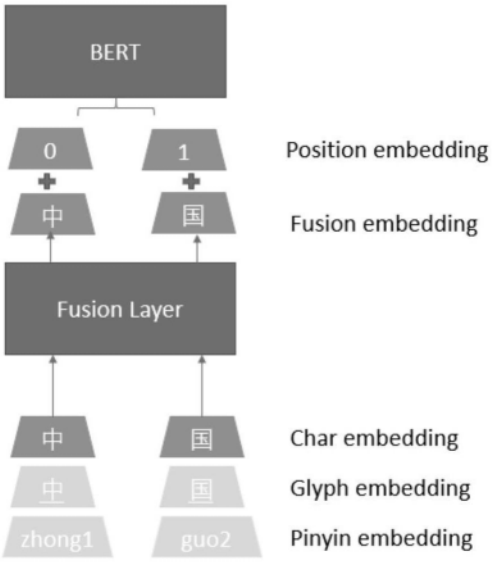

BERT
0
1
Position embedding
中
国
Fusion embedding
Fusion Layer
中
国
Char embedding
中
国
Glyph embedding
zhong1
guo2
Pinyin embedding

图4

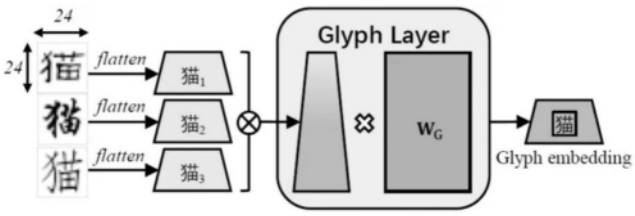

24
24
flatten
flatten
flatten
Glyph Layer
W_G
Glyph embedding

图5

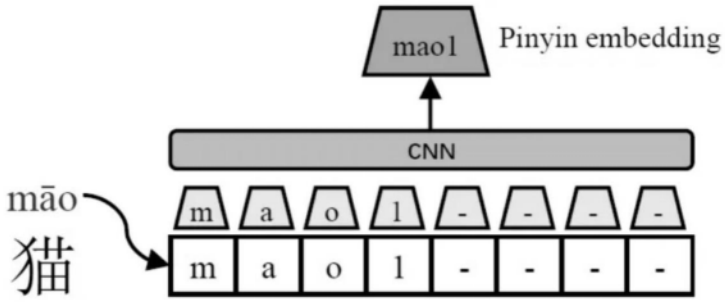


图6