cslt – roobo 语音识别项目总结报告

1. 项目主要阶段概述
2. 第一阶段：ci1006芯片上的命令词识别系统；
3. 第二阶段：通用嵌入式芯片上的命令词识别系统；
4. 第三阶段：通用嵌入式芯片上的关键词唤醒技术。
5. 主要技术难点及解决方案
6. ci1006芯片上的命令词识别系统：
   1. 适配ci1006芯片的90维声学特征提取；
   2. 适配ci1006芯片的DNN声学模型的训练及转换；
   3. 解码器的工程化实现及移植；
   4. 解码图优化。

表1 第一阶段主要技术点及解决方案

|  |  |  |
| --- | --- | --- |
| 技术点 | 主要解决者 | 解决方案 |
| 1.1 | roobo | roobo提供了特征提取程序/脚本。 |
| 1.2 | cslt / roobo | cslt提供模型及训练脚本，roobo负责模型转换及移植。 |
| 1.3 | roobo | roobo负责解码器的工程化。 |
| 1.4 | cslt | cslt通过实验验证了几种优化方法，对roobo现有模型进行了调优。cslt还向roobo提供了调优方法，供roobo后续调优。 |

1. 通用嵌入式芯片上的命令词识别系统：
   1. 语音识别引擎的工程化实现；
   2. 声学模型训练方法；
   3. 解码图构造方法；
   4. 矩阵运算的定点化；
   5. 解码图优化；
   6. 置信度的计算方法。

表2 第二阶段主要技术点及解决方案

|  |  |  |
| --- | --- | --- |
| 技术点 | 主要解决者 | 解决方案 |
| 2.1 | roobo | roobo负责语音识别引擎的工程化。 |
| 2.2 | cslt | cslt首先提供了基于kaldi-nnet1的声学模型训练脚本（单机单GPU）；而后提供了基于kaldi-nnet3的声学模型训练脚本（多机多GPU），并提供了nnet3-nnet1的转换程序。 |
| 2.3 | cslt | cslt提供了解码图构造的脚本。 |
| 2.4 | roobo | roobo完成了矩阵运算的定点化。 |
| 2.5 | cslt | Cslt通过实验验证了几种优化方法，对roobo现有模型进行了调优。Cslt还向roobo提供了调优方法，供roobo后续调优。 |
| 2.6 | roobo | roobo进行了实验验证，并完成了工程化实现。 |

1. 通用嵌入式芯片上的关键词唤醒技术：
   1. 关键词唤醒技术方案的确定；
   2. 声学模型训练方法；
   3. 解码图构造方法；
   4. 解码图优化；
   5. 语音端点检测（VAD）方法。

表3 第三阶段主要技术点及解决方案

|  |  |  |
| --- | --- | --- |
| 技术点 | 主要解决者 | 解决方案 |
| 3.1 | cslt | cslt通过实验确定了关键词唤醒的技术方案。 |
| 3.2 | cslt | cslt通过实验确定了声学模型训练方法，并提供了训练脚本。 |
| 3.3 | cslt | cslt提供了针对关键词唤醒模型的解码图构造脚本。 |
| 3.4 | cslt | cslt通过实验验证了几种优化方法，对roobo现有模型进行了调优。cslt还向roobo提供了调优方法，供roobo后续调优。 |
| 3.5 | cslt / roobo | cslt提供了VAD的方法，并通过实验进行了初步的性能测试；roobo负责VAD具体工程实现。 |

1. 关键技术的具体实现细节：
2. 适配ci1006芯片的DNN声学模型的训练：

由于ci1006芯片中，特征提取和神经网络计算部分由硬件实现，不可改动，因此有许多需要注意的地方。

* 1. 声学特征不能更换，必须使用90维特征，但在生成训练用的alignment时，则可以利用现有模型和其他声学特征，如mfcc、fbank。但需要注意的是，芯片中特征提取部分的首帧、末帧处理方式与kaldi不同，因此在利用kaldi生成alignment之前，必须利用程序对kaldi提取出的特征进行处理，每句话砍掉句首6帧，句尾1帧，这样才能与90帧特征相匹配。
  2. 网络结构必须为标准DNN，不能有其他复杂的网络结构，并且在特征与DNN输入层中间，不能有复杂的变换。

1. 命令词识别系统的解码图优化：

解码图优化的目的，主要是减少误识别。解决此问题的基本思路是在解码图中加入一些silence loop，使其能够吸收“非命令词”。在此过程中，亦有许多需要注意的问题：

2.1为了避免解码图增大，也方便调整silence loop的权重，我们一般不在jsgf当中加入silence loop，而是在生成HCLG之后，利用脚本手动加入silence loop；

2.2 silence loop当中，一般包含sil和spn两个phone；

2.3 加入silence loop后，silence路径的权重需要根据测试集调整。一般情况下，最优权重在0附近。

1. 关键词唤醒方法：

我们确定的关键词唤醒技术方法，大致原理是，在现有连续识别模型的基础之上，保留所有与关键词相关的pdf和sil、spn对应的pdf，而把其余pdf全部合并为一个——garbage，然后重新训练。这样就可以的到一个输出层节点经过合并的声学模型。此过程有一些注意事项：

* 1. 此方法的优势在于，它可以利用连续语音的训练数据，而不仅仅局限于特定关键词数据，因此在更换关键词表时，不需要重新录音。但需要注意的是，如果更换关键词表，模型必须重新训练；
  2. 因为此模型的声学模型源自连续识别模型，所以此模型对应的解码图，亦需要从连续识别模型相对应的解码图修改而来；
  3. 在获取与关键词相关的pdf时，不要使用类似phone-to-pdf的方式。因为这种方式会保留所有context-dependent-phone对应的pdf，会造成冗余。应该首先使用未合并pdf的声学模型构建HCLG，然后从HCLG中改变pdf对应关系来获取最终要使用的HCLG；
  4. 得到获取HCLG之后，同样应该加入silence-loop来减少误识别，这一点跟命令词识别相同。但这里不仅应该加入sil和spn，还应该加入garbage。
  5. 实验证明，silence-loop中是否加入spn影响不大，因此也可以不单独保留spn对应的pdf，而把它们合并到garbage中。

1. 资源及成果的汇总