Low Latency End-to-End Streaming Speech Recognition with a Scout Network 笔记

1. 基于Transformer，提出了一种新的流式语音识别模型
2. 包含一个 scout network 和 一个 recognition network
3. scout network 不看未来的帧来检测整个 word boundary
4. recognition network 预测下一个sub-word

本文提出的模型和之前的两种方法的对比

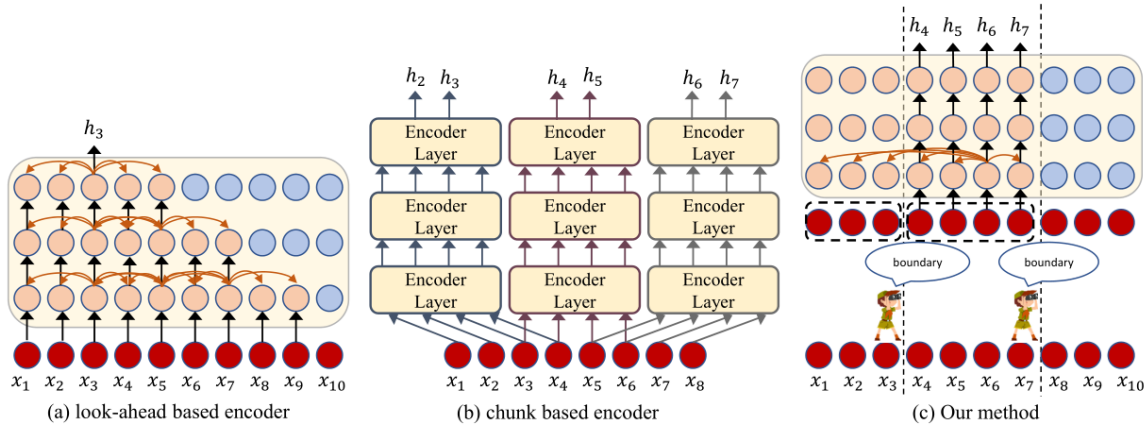

Figure 1: *A comparison between three Transformer based streaming models.*

Introduction

1. MoChA 和 TA 的提出，用于替代encoder-decoder之前的全局注意力
2. 之前的流式识别方法：
    - look-ahead 法
    - chunk-based 法
3. 本文提出一种自适应 look head 的方法，动态修改 context 窗口
4. 引入一个神经网络检测语音中词汇开始和结束的边界，即检测网络（scout network），且仅向前看帧直到检测到的字边界。因此识别延迟不是固定的，而是

取决于单词的持续时间和网络的分割性能。
    - 训练过程中，将每个帧看成二元分类问题（边界\非边界）
    - 由于不看未来帧，因此没有额外的延迟
5. 使用基于TA的Transformer作为 recognition network
6. 模型不仅可以降延迟，识别效果还是SOTA

模型细节

1. Scout Network 采用 ASR encoder 相似的结构（CNN+Attention）
2. 使用了 Montreal Forced Aligner 来实现单词级别的强制对齐并获得标签
3.