Streaming End-to-End Speech Recognition with Joint CTC-Attention Based Models 笔记

1. 提出了基于Transformer的流式ASR
2. 在encoder端的注意力机制上，使用了 time-restricted self-attention
3. 在encoder-decoder端的注意力机制上，使用了 triggered attention
4. 是当时最好的流式ASR模型

Introduction

1. 常见的基于注意力的流式ASR：
    - neural transducer
    - monotonic chunkwise attention
    - triggered attention
2. 不同的底层神经网络架构：
    - LSTM
    - BLSTM
    - LC-BLSTM
    - PTDLSTM
    - …
3. 本文贡献：
    - 在encoder端的注意力机制上，使用了 time-restricted self-attention
    - 在encoder-decoder端的注意力机制上，使用了 triggered attention
    - Transformer 和 CTC-loss 联合训练

模型细节

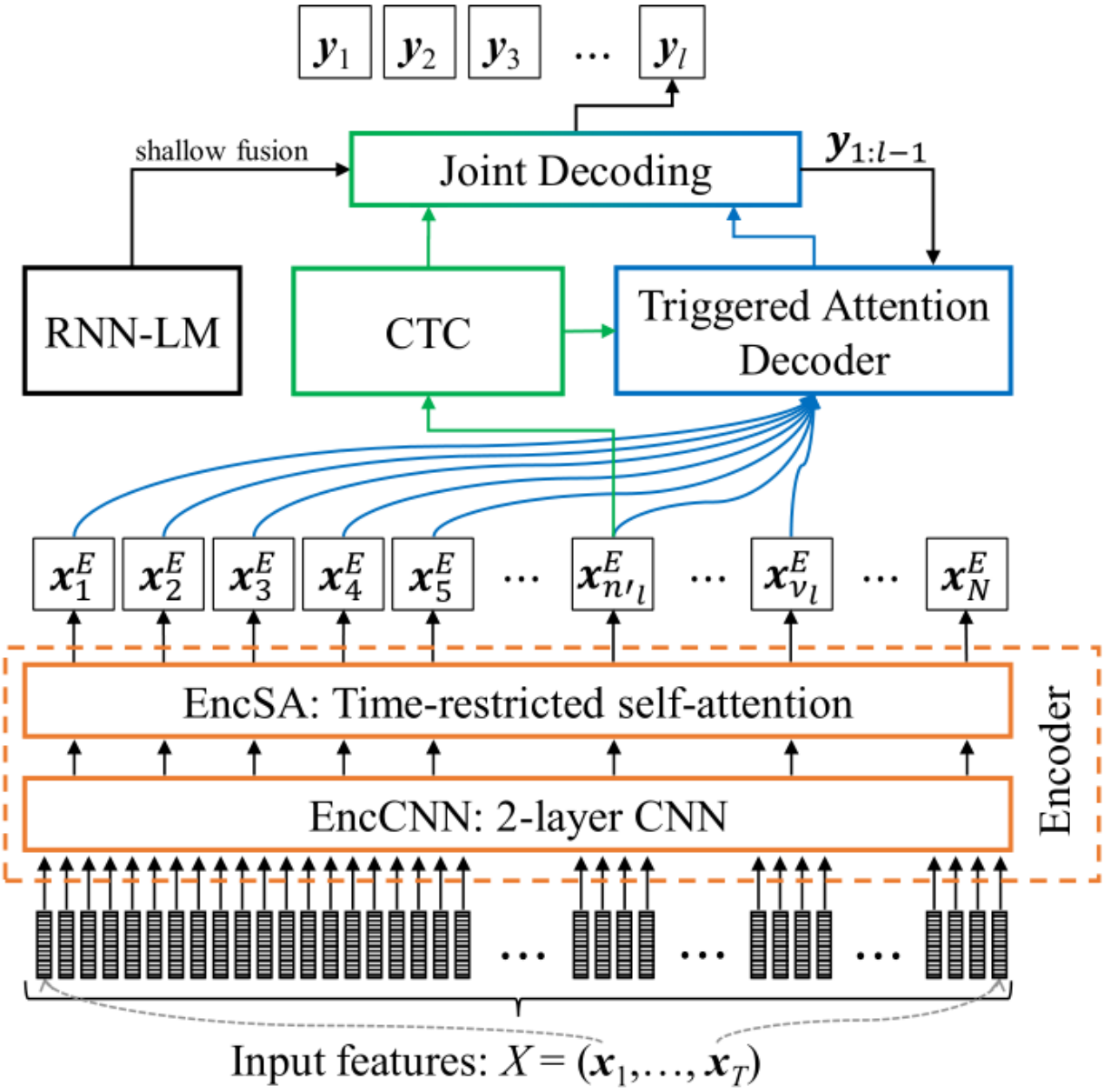

Fig. 1. Joint CTC-TA decoding scheme for streaming ASR with a transformer-based architecture.

1. 编码器：Time-restricted self-attention 编码器包含两层的CNN和 多层 self-attention 层叠，为了控制延迟，输入序列的 future context 限制为固定长度，即所谓的 time-restricted self-attention。

2. 解码器: Triggered attention encoder和decoder之间的attention采用了 Triggered attention 来实现流式。TA训练需要编码器状态序列 X_e 和标签序列 Y=(y_1,\dots, y_L) 之间的对齐，以仅在过去的编码器帧加上固定数量的 look-ahead 帧 ϵ^{dec} 上调节解码器的注意机制。TA 的目标函数定义为：

p_{\mathrm{ta}}\left(Y \mid X_{E}\right)=\prod_{l=1}^{L} p\left(y_{l} \mid \boldsymbol{y}_{1: l-1}, \boldsymbol{x}_{1: \nu_{l}}^{E}\right)