自动化所提出基于神经元整合发放的语音识别新机制

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/8385.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

基于注意力机制的端到端模型正深刻影响着语音识别技术的发展。但经典的注意力识别模型因“要对整句语音编码后投入注意力”的特点面临着无法支持在线（流式）识别、无法提供语音边界时间戳等问题。

中国科学院自动化研究所博士董林昊、研究员徐波将脉冲神经网络中的整合发放思想进行连续化，提出一种低复杂度并具有单调一致性的序列转换机制——连续整合发放（Continuous Integrate-and-Fire，CIF）。CIF会对先后到来的声学信息不断进行整合，当整合的信息量达到识别阈值，将整合后的信息发放以用作后续识别。基于CIF的模型不仅有效地支持了在线识别、边界定位及声学Embedding提取，而且在两个中文基准语音识别集（HKUST、AISHELL-2）上创造了SOTA的性能，有效地解决了目前主流注意力机制模型存在的上述问题。相关成果近期被ICASSP 2020录用为Oral论文。

连续整合发放（CIF）应用于编解码框架。在每一个编码时刻，CIF分别接收编码后的声学编码表示及其对应的权重（表征了蕴含的信息量）。之后，CIF不断地积累权重并对声学编码表示进行整合（加权求和的形式）。当积累的权重达到阈值后，意味一个声学边界被定位到。此时，CIF模拟了整合发放模型的处理思想，将当前权重分为两部分（如图1所示）：一部分用来完成当前标签的声学信息整合（构建一个完整分布），并将整合后的声学信息（声学Embedding）发放到解码器以预测对应的标签；另一部分用作下一个相邻标签的声学信息整合。该过程一直执行到编码后序列的末尾。论文还提出了若干支撑策略来进一步精炼CIF模型的性能，如规整策略、数量损失等。

该研究工作在多个语音识别基准数据集上对CIF模型的性能进行了验证，这些数据集涵盖了不同的语种和不同的语音类型。其中，在英文朗读数据集Librispeech上，虽然采用的输出标签是没有明确声学边界的子词单元，但CIF仍然获得了有竞争力的2.86%的词错误率表现（如图2所示）。在中文朗读数据集AISHELL-2上，由于输出标签间的声学边界较为清楚，基于CIF的模型获得了突出的性能表现，显著地超过了Chain模型的性能，创造了该数据集上state-of-the-art的字错误率结果（如图3所示）。在中文电话数据集HKUST上，虽然语音上具有很多非正式的口语现象，而且数据集规模相对较小，但是基于CIF的模型仍然展现了良好的泛化性，创造了该数据集上state-of-the-art的字错误率结果（如图4所示）。

CIF模型不但可以高准确度提供序列转化结果，而且把语音认知中最重要的发音边界进行了精确定位，为语音识别融合各种知识模型提供了新的手段和路径。CIF将整合发放进行连续化思想可

推广应用到其它序列转换任务中。据悉，该论文工作在研究团队万级小时大规模训练数据的语音识别中，也超过了团队目前CTC、Transformer等主流模型的已有结果，达到了最好性能，意味着该方法具有工业界大规模应用的极大潜能。

图1. CIF机制与注意力机制的对齐形态对比

图2. 在英文朗读数据集Librispeech上，CIF模型与已发表模型的词错误率对比

图3. 在中文朗读数据集AISHELL-2上，CIF模型与已发表模型的字错误率对比

图4. 在中文电话数据集HKUST上，CIF模型与已发表模型的字错误率对比

研究团队单位：自动化研究所

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发