Transformer versus LSTM Language Models Trained on Uncertain ASR Hypotheses in Limited Data Scenarios

Imran Sheikh, Emmanuel Vincent, Irina Illina*
Université de Lorraine, CNRS, Inria, Loria F-54000 Nancy, France
*irina.illina@loria.fr

Background
- Domain/Application specific ASR LMs require:
  - in-domain text, and/or
  - manually verified in-domain speech transcriptions
- Such domain specific text resources are scarce
- In-domain speech data also limited, e.g., in
  - the early development stages of a new application
  - in privacy-critical applications
  - for under-resourced languages

Goal: LMs from a limited amount (25-50h) of in-domain speech.
Problem: Training neural LMs on ASR decoded graphs.

Prior Art
- n-gram LMs on ASR lattices [Kuznetsov et al., 2016]
- Adaptation of RNN LMs on 1-best hypotheses [Li et al. 2018]
- Train LSTM LMs on confusion networks [Sheikh et al. 2021]
- Transformers on ASR lattices and confusion networks
  - for machine translation [Zhang et al., 2019]
  - and language understanding [Liu et al., 2020]

This work: LSTM vs Transformer LMs on ASR confusion networks.

Training LSTM LMs on Confusion Nets

Given LSTM with L layers and weights \( \Theta = \{ \theta_{in}, \theta_{hid}, \theta_{out} \} \)

**Sampling based training:**
Sample one arc at a time from each confusion bin: \( w \sim p(w_t|S) \). Cross entropy loss for training with sampled path:

\[
\hat{\theta} = \arg \min_{\theta} \sum_t -\log q(w_{t+1} = w|h^1_{t+1}).
\]

**KL divergence based training:**
Compute hidden state \( h^1_{t,i} \) for all arcs \( i \) in a confusion bin and pool:

\[
h^1_{t,i} = \alpha(h^1_{hid}h^1_{t-1} + \theta_{in}n_{t,i})
\]

\( h^1_t = \text{pool}(h^1_{t,i}) \).

Handle multiple outputs and uncertainty using KL loss:

\[
\hat{\theta} = \arg \min_{\theta} \sum_t D_{KL}(p(w_t|S) || q(w_t|h^1_t))
\]

\[
= \arg \min_{\theta} \sum_t \sum_i p(w_t = v|S) \log \frac{p(w_{t+1} = v|S)}{q(w_{t+1} = v|h^1_t)}.
\]

Hierarchical Scheme for Transformer LMs

- Collapse the confusion bins into a sequence.
- Separate arc-level and bin-level layers.
- Self-attention mask based on confusion bin posteriors

\[
M_{t,t'} = \begin{cases} 
\log p(w_{t'}) & t' \leq t \\
-\infty & \text{otherwise} 
\end{cases}
\]