Abstract
Sequence-to-sequence models with soft attention have been successfully applied to a wide variety of problems, but their decoding process incurs a quadratic time and space cost and is inapplicable to real-time sequence transduction. To address these issues, we propose Monotonic Chunkwise Attention (MoChA), which adaptively splits the input sequence into small chunks over which soft attention is computed. We show that models utilizing MoChA can be trained efficiently with standard backpropagation while allowing online and linear-time decoding at test time. When applied to online speech recognition, we obtain state-of-the-art results and match the performance of a model using an offline soft attention mechanism.

In document summarization experiments where we do not expect monotonic alignments, we show significantly improved performance compared to a baseline monotonic attention-based model.

Sequence-to-sequence framework

Attention is often roughly monotonic

Decoding with MoChA

Training
\[ a_{t,j} = \text{MonotonicEnergy}(h_{t-1}, h_j) \]
\[ p_{t,j} = \alpha_t(t+1) \]
\[ a_{t,j} = p_{t,j} \ \text{cumprod}(1 - p_{t,j}) \ \text{cumprod}(1 - p_{t,j}) \]
\[ s_{t,j} = \text{ChunkEnergy}(h_{t-1}, h_j) \]
\[ \beta_t = \exp(a_{t,j}) \ \text{MovingSum}(\exp(a_{t,j}), w, 1, w) \]

Test-time
Input: memory $h$ of length $T$, chunk size $w$
State: $a_t = 0, b_t = 0, l_t = 0, p_t = 0$, StartOfSequence
while $p_t < T$ do
    for $j = 1, \ldots, T$ do
        $r_{t,j} = \text{softmax}(s_{t,j})$
        $b_t = \text{chunkstart}(a_{t,j}, h_t, b_t)$
        $b_{t+1} = \text{chunkstart}(a_{t,j}, h_t, b_t)$
        $b_{t+1} = \text{chunkstart}(a_{t,j}, h_t, b_t)$
        $p_{t,j} = 0$
    end for
end while

Results

Synthetic speed benchmark

From “Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition”, Chen, Jaitly, Le, Vinyals.