Efficient Encoders for Streaming Sequence Tagging

Ayush Kaushal♣∗ Aditya Gupta♠ Shyam Upadhyay♠ Manaal Faruqui♠
♠Google Assistant
♣The University of Texas at Austin
ayushk4@utexas.edu, {gaditya, shyamupa, mfaruqui}@google.com

Abstract

A naive application of state-of-the-art bidirectional encoders for streaming sequence tagging would require re-encoding all tokens from scratch whenever a new token appears in an incremental streaming input (like transcribed speech). The lack of re-usability of previous computation leads to a higher number of Floating Point Operations (or FLOPs) and higher number of unnecessary label flips. Increased FLOPs consequently lead to higher wall-clock time and increased label flipping leads to poorer streaming performance. In this work, we present Hybrid Encoder with Adaptive Restart (HEAR) that addresses these issues while maintaining the performance of bidirectional encoders over offline (or complete) inputs and improving performance on streaming (or incomplete) inputs. HEAR uses a HYBRID unidirectional-bidirectional encoder architecture to perform sequence tagging, along with an Adaptive Restart Module (ARM) to selectively guide the restart of bidirectional portion of the encoder. Across four sequence tagging tasks, HEAR offers FLOPs savings in streaming settings up to 71.1% and also outperforms bidirectional encoders for streaming predictions by up to +10% streaming exact match.

1 Introduction

State-of-the-art text encoding methods assume the offline setting, where the entire input text is available when encoding it. This differs from the streaming setting where the input text grows over time (such as transcribed speech or a typed query) (Cho and Esipova, 2016; Gu et al., 2017; Chang et al., 2022). Processing streaming input incrementally can enable suggestions on partial inputs (Iranzo Sanchez et al., 2022), reduced final latency (Zhou et al., 2022) and lead to more interactive NLU agents (Cai and de Rijke, 2016).

Existing state-of-the-art bidirectional encoders (such as Devlin et al. (2019)) are computationally expensive for streaming processing. When a new token is received, these models require a restart, i.e., re-computation of representation of each token by re-running the bidirectional layer (Kahardipraja et al., 2021). This adds to the computational cost (i.e., Higher FLOPs) during streaming and leads to higher wall clock time. Another limitation of these encoders is poorer generalization to partial inputs (Madureira and Schlangen, 2020), stemming from these models being trained only on complete (and offline) inputs. Despite these disadvantages, such bidirectional models offer better offline performance than unidirectional models across several NLP tasks like sequence tagging (Kahardipraja et al., 2021).

We address these issues by proposing HEAR – Hybrid Encoder with Adaptive Restarts, where a separate Adaptive Restart Module (ARM) predicts when to restart the encoder. The encoder in HEAR is a HYBRID encoder that reduces the computational cost of running the models in streaming settings. In a HYBRID encoder, the earlier layers are unidirectional and the deeper layers are bidirectional. This design allows early contextualization, and limits the need for restart to the later layers. While HYBRID encoder reduce streaming computational overhead, restarting them at every new token may not be required. Thus, we propose a lightweight Adaptive Restart Module (ARM) to guide restarts, by predicting whether restarting the

<table>
<thead>
<tr>
<th>Model</th>
<th>GFLOPs</th>
<th>Offline F1</th>
<th>Streaming EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDi encoder</td>
<td>74.7</td>
<td>93.1</td>
<td>75.1</td>
</tr>
<tr>
<td>HYBRID encoder</td>
<td>43.4</td>
<td>93.0</td>
<td>72.5</td>
</tr>
<tr>
<td>HEAR (HYBRID + ARM)</td>
<td>21.6</td>
<td>93.0</td>
<td>85.1</td>
</tr>
</tbody>
</table>

Table 1: Computation cost (Giga FLOPs), Offline F1 and Streaming Exact Match accuracy of standard bidirectional (BiDi) encoder, HYBRID encoder model and HEAR (HYBRID encoder with ARM for guiding Adaptive Restarts) on SNIPS test set.
bidirectional layers of the HYBRID encoder will benefit the model performance. This module is also compatible with fully bidirectional encoders.

Table 1 showcases the strength of HEAR on one of the sequence tagging dataset we consider. In terms of streaming computation, measured in terms of FLOPs (lower is better), HYBRID encoder offers significant saving from purely bidirectional (BiDi) encoders and FLOPs savings improve upon incorporating ARM in HEAR. In terms of offline performance, HYBRID encoder and HEAR achieves parity with BiDi encoders and in terms of streaming performance, HEAR significantly outperforms BiDi encoders.

Following are our key contributions:

- We introduce HYBRID encoder for computationally cheaper streaming processing (§3.1), that maintains the offline F1 score of bidirectional encoders, while reducing FLOPs by an average of 40.2% across four tasks.

- We propose the ARM module (§3.2) to decide when to restart. The ARM reduces FLOP of HYBRID encoder by 32.3% and improves streaming predictions by +4.23 Exact Match.

- Our best model HEAR combines HYBRID encoder with ARM (§5) to achieve strong streaming performance while saving FLOPs and offering competitive offline performance across four sequence tagging tasks.

2 Streaming Sequence Tagging

In the streaming sequence tagging task, we assume that at time $1 \leq t \leq n$, we have received the first $t$ tokens $x_t = (x_1, \ldots, x_t)$ as input from a stream of transcribed speech or user-typed input.\footnote{We assume that new tokens get added without changing previous tokens (contrary to some ASR systems), even though, our method can be used in such settings.} The model then predicts the tags for all $t$ tokens $\hat{y}_t = (\hat{y}_{1,t}, \hat{y}_{2,t}, \ldots, \hat{y}_{t,t})$. The models predictions $\hat{y}_n$ over the offline input $x = x_n$ is the offline tag sequence prediction which is evaluated against the ground truth $y^*_n = (y^*_1, y^*_2, \ldots, y^*_n)$. However, in the streaming settings, we are concerned with predicted label sequences over all of the prefixes $Y = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$.

During training we only have access to the offline ground-truth label sequence over the offline input sequence $y^*_n$, even though ground truth labels for tokens may change as additional context is received in the future timesteps.

3 Streaming Sequence Tagging with HEAR

In order to motivate HEAR, consider how existing BiDi encoders would be used in streaming sequence tagging. At time $t$ and the input sequence $x_t$, the model restarts to predicts the label sequence $\hat{y}_t$ i.e. it re-computes all its layers for all tokens, without leveraging any of its previous computation or any previously predicted labels. Consequently, running a typical $O(n^2)$ encoder would have $O(n^3)$ computations in streaming settings, over $n$ input tokens. Therefore, naively using existing BiDi encoders for streaming settings is highly inefficient. Previous works tackle inefficiency by modifying the streaming model to infer only once for each word, after $k$ future tokens have been received (Oda et al., 2015; Grissom II et al., 2014). This leads to a $k$-delayed output with possibility for revisions after additional future words have been received. However, this may lead to poor performance for tasks involving long-range dependency (e.g., SRL) and a higher output lag.

HEAR is a system consisting of a trained HYBRID encoder model and an ARM that is trained over the HYBRID encoder. The early unidirectional layers in HYBRID encoder, reduces the computational cost of restart of the encoder. The ARM guides when to restart the HYBRID encoder. It saves computational cost by keeping the model from restarting at each timestep and also improves streaming performance by avoiding unnecessary label flips stemming from avoidable restarts. Figure 2 shows running of HEAR in streaming settings over an example. For each new token in the streaming input, all the unidirectional layers and a part of first bidirectional layer is ran for the token, to obtain its unidirectional encodings and updated cache. These are then used by ARM to predict whether to restart the bidirectional layer or not. If the bidirectional layer is to be restarted, then we obtain updated labels for all the tokens received in the stream. Otherwise, the auxiliary predictor is ran over the unidirectional encoder for the current token to obtain its label and the labels from previous timestep is copied for all other tokens.

We now formally introduce HYBRID encoder (§3.1), followed by ARM (§3.2).
3.1 HYBRID Encoder

The HYBRID encoder is a combination of unidirectional and bidirectional encoding layers, where the early layers are unidirectional and the later layers are bidirectional as shown in Figure 2. The earlier unidirectional layers do not require a restart as the previous tokens’ embeddings do not need to be updated for the newly received tokens. Thus, the HYBRID encoder only require a restart for its later bidirectional layers. Formally, the HYBRID encoder has \( u \) unidirectional and \( b \) bidirectional layers where total layers in the model are \( u + b = l \).

Let \( \mathcal{L} \) and \( \mathcal{L}_u \) denote unidirectional and bidirectional layers in the model respectively. Each of the \( \mathcal{L} \) layers are one of the existing unidirectional layers - RNN, GRU, transformer with causal masking etc. with the first of these being the static embedding layer. Each of the \( \mathcal{L}_u \) layers are bidirectional layers, being one of bi-GRU, bi-LSTM, transformer etc.

3.1.1 HYBRID Encoder in Offline Setting:

First, the input tokens \( \mathbf{x} \) are fed to \( u \) unidirectional layers to get the unidirectional encodings \( \mathbf{h} \),

\[
\mathbf{h} = [\mathbf{h}_1, \ldots, \mathbf{h}_n] = \mathcal{L}_u \circ \cdots \circ \mathcal{L}(\mathbf{x})
\]

where \( \circ \) denotes the functional composition and \( \mathbf{h}_i \) is the unidirectional encoding for the token \( x_i \).

Then the bidirectional encodings \( \mathbf{H} \) are computed and used to predict the label sequence:

\[
\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n] = \mathcal{L} \circ \mathcal{L}_u \circ \cdots \circ \mathcal{L} \circ \mathcal{L}(\mathbf{H})
\]

The final offline labels are obtained by passing \( \mathbf{H} \) through a feed-forward neural network layer.

3.1.2 HYBRID Encoder in Streaming Setting

In the streaming setting, the unidirectional layers’ computation can be cached.\(^2\) These cached intermediate representations for the previously received tokens \( \mathbf{x}_{t-1} \) are used in computing the unidirectional encoding \( \mathbf{h}_t \) of the new token \( x_t \). This along with the cached unidirectional encoding of \( \mathbf{x}_{t-1} \) gives us \( \mathbf{H}_t = [\mathbf{H}_{t-1}; \mathbf{h}_t] \). The bidirectional encoding, however, for each token in \( x_t \) is restarted at time \( t \) as \( \mathbf{H}_t = [\mathbf{h}_1, \ldots, \mathbf{h}_{t,t}] \) using the obtained final unidirectional encoding \( \mathbf{H}_t \) as input.

3.1.3 Training and Inference of HYBRID Encoder

We predict labels over both \( \mathbf{H}_t \) and \( \mathbf{H}_{t,t} \) of the HYBRID encoder using a linear layer with softmax at each timestep. Let \( \mathbf{y}_t = [y_{1,t}, \ldots, y_{t,t}] \) denote the predictions over bidirectional embeddings at time \( t \) for all the tokens received so far. Let \( \mathbf{y}_t = [\mathbf{y}_{1,t}, \ldots, \mathbf{y}_t] \) be predictions over unidirectional embeddings at time \( t \). While the unidirectional predictions do not perform as well as bidirectional predictions, these auxiliary predictions enables waited restarts (§3.2).

---

\(^2\) Cache for RNNs and causally masked transformer consists of hidden states and keys-values respectively. This is similar to implementations of auto-regressive models (Wolf et al., 2020; Heck et al., 2020).

---

Table 2: Notations used in HYBRID encoder for layers, token representations, and predicted labels.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L, \mathcal{L} )</td>
<td>Unidirectional and bidirectional layers, respectively.</td>
</tr>
<tr>
<td>( \mathbf{h}<em>i, \mathbf{h}</em>{i,t} )</td>
<td>Encoding of ( x_i ) token when ( t ) tokens have been received, using ( \mathcal{L} ) and ( \mathcal{L}_u \circ \cdots \circ \mathcal{L}(\mathbf{H}) ) respectively.</td>
</tr>
<tr>
<td>( \mathbf{h}<em>i, \mathbf{h}</em>{i,t} )</td>
<td>of first ( t ) tokens, respectively.</td>
</tr>
<tr>
<td>( y_i, \mathbf{y}_{i,t} )</td>
<td>Predicted label of ( x_i ) token using ( \mathbf{h}<em>i ) and ( \mathbf{h}</em>{i,t} ) respectively.</td>
</tr>
</tbody>
</table>
HYBRID encoder is trained over the offline input-output sequence pairs and optimize both the bidirectional predictions and unidirectional predictions. For the standard softmax cross entropy loss against the offline ground truth label sequence $y^*$. In order to preserve the strong performance of the BiDi encoders, we inhibit backward flow of gradient from the parameters $\theta_{uni}$ of unidirectional prediction head (consisting only of a linear layer with softmax) to the remaining parameters - $\theta_{bi}$. Following are the losses of these two portions of the model, optimized together with equal weight.\(^3\)

\[ L_{uni}(\theta_{uni}, \hat{y}_{n}, y^*) = CE(\hat{y}_{n}, y^*) \quad (3) \]

\[ L_{bi}(\theta_{bi}, \hat{y}_{n}, y^*) = CE(\hat{y}_{n}, y^*) \quad (4) \]

\[ L(\theta) = L_{bi}(\theta_{bi}, \hat{y}_{n}, y^*) + L_{uni}(\theta_{uni}, \hat{y}_{n}, y^*) \quad (5) \]

3.2 Waited Restarts

The HYBRID encoder reduces the computational overhead of each restart by limiting the recomputation over previous tokens only to the later bidirectional layers. However, restarting at each step is not required and the auxiliary predictions from the unidirectional layers can suffice.

We define a variable $\text{RESTART}(t)$ to decide whether to restart the bidirectional layers at time $t$. When $\text{RESTART}(t) = 1$, we restart the model to get the bidirectional predictions $\hat{y}_t$ as the final predicted label sequence $\hat{y}_t = \hat{y}_t$. If $\text{RESTART}(t) = 0$, the model does not run the bidirectional encoder, but uses the unidirectional encoding to predict the current token's label $\hat{y}_t$ and copy the label sequence from the previous step $\hat{y}_{t-1}$ to obtain $\hat{y}_t$. Formally, the predicted sequence $\hat{y}_t = (\hat{y}_{t, t}, \ldots \hat{y}_{t, t})$ at time $t$ is

\[ \hat{y}_t = \begin{cases} \hat{y}_t & \text{if } \text{RESTART}(t) = 1 \text{ or } t = n \\ [\hat{y}_{t-1}; \hat{y}_t] & \text{otherwise} \end{cases} \quad (6) \]

where ‘;’ denotes the concatenation. We always restart in the final timestep (i.e., $\text{RESTART}(n) = 1$) to preserve the offline performance.

3.2.1 Adaptive Restart Module (ARM)

A baseline restart strategy for waited restarts would be to restart the bidirectional layers every fixed $k$ steps, i.e., $\text{RESTART}(t) = 1$ whenever $t$ is a multiple of $k$. Note that $k = 1$ reduces to a HYBRID encoder model without waited restarts. We refer to this as $\text{RESTART}-k$.

Rather than having a heuristic function for $\text{RESTART}(t)$, we can let a lightweight parametric module determine when to do the restart by determining $\text{RESTART}_{\theta}(t)$ for each $t$. We refer to this as Adaptive Restart Module (ARM). We first discuss the set of features for ARM (§3.2.2), followed by its architecture (§3.2.3) and training (§3.2.4).

3.2.2 Input Features for ARM

We use HYBRID encoder’s intermediate representations that do not require restart as the features for ARM. This includes the unidirectional layers as well as from the pre-softmax features from the first $L$ layer. The main motivation to use these features is to incorporate more information without incurring any restarts.

The following features are stored in unidirectional cache and used without any restart for ARM:

(i) unidirectional encodings $\overrightarrow{h}_t$ from the last unidirectional layer computed once and cached for capturing the backward flow,

(ii) query $q_t$ and key $k_t$ from the first bidirectional layer computed only from $h_t$,

(iii) unnormalized forward attention scores $a_t = [q_1 \cdot k_t, \ldots, q_{t-1} \cdot k_t]$ from the first bidirectional layer, where ‘$\cdot$’ is the dot product. These scores are concatenated across the prefix tokens (restricted and padded to latest $m$ tokens) across all heads.

These features are then concatenated and provided as input to the ARM as $\bar{f}_t = [\overrightarrow{h}_t, q_t, k_t, a_t]$.

3.2.3 ARM Architecture

We use a single-layered GRU (Cho et al., 2014) to predict restart. The probability of restart $p_{\theta}(\text{RESTART}(t) = 1)$ is modeled as,

\[ (o_t, s_t) = \text{GRU}(\bar{f}_t, s_{t-1}) \quad (7) \]

\[ p_{\theta}(\text{RESTART}(t) = 1) = \sigma(\text{LINEAR}(o_t)) \quad (8) \]

Here $s_t$ and $o_t$ are the GRU hidden state and output at time $t$, respectively. These outputs can be post-processed to handle too frequent or too infrequent waiting and obtain the final selector predictions. Details can be found in Appendix (§A.2).
Figure 2: The HEAR model in streaming setting. At $t=1$, ARM decides not to run the bidirectional layers and exit early through the auxiliary predictor over unidirectional encoding. Similarly at $t=2$, ARM decides not to restart and considers the auxiliary predictions of the latest token and the predictions from the previous steps. At $t=3$ and $t=4$, ARM decides to restart and runs the bidirectional layers to get prediction for all the tokens from scratch. Note that in all the steps, for unidirectional layers and ARM, the computation is performed only for the latest token received.

3.2.4 Training the ARM

Ideally, HYBRID encoder should restart only when it would lead to improved predictions over the prefixes i.e., more number of prefix inputs should have the same output as the the offline ground truth for those prefixes. We use this to define our ground truth RESTART sequence for training the ARM.

Given the ground truth offline tag sequence $y^*$, unidirectional predictions $\hat{y} = [\hat{y}_1, \cdots, \hat{y}_n]$ and final bidirectional prediction sequence over each of the $n$ prefixes $\tilde{Y} = [\tilde{y}_1, \cdots, \tilde{y}_n]$, we define the ground truth policy $\text{RESTART}^*(t) = \pi^*_t$ for ARM, at time $t$, as follows.

$$\pi^*_t = \begin{cases} 1 & \text{if } |y^*_t = \hat{y} \leq t| < |y^*_t = \tilde{y} \leq t| \\ 0 & \text{otherwise} \end{cases}$$ (9)

The policy decides to restart at time $t$ if more tokens in bidirectional predictions match with the ground truth ($|y^*_t = \hat{y} \leq t|$) than those from unidirectional predictions ($|y^*_t = \tilde{y} \leq t|$).\(^4\) ARM is trained against this policy with features from a frozen and trained HYBRID encoder model with a standard binary cross entropy loss.

4 Experimental Setup

Datasets and Tasks. We consider four common sequence tagging tasks – Slot Filling over SNIPS (Coucke et al., 2018), Semantic Role Labelling (SRL) over Ontonotes (Pradhan et al., 2013), and Named Entity Recognition (NER) and Chunking (CHUNK) over CoNLL-2003 (Sang and Meulder, 2003). The standard train, dev and test splits are used for all datasets.

Models and Baselines. For our models, we consider a layer budget of 4 for the encoder. Naturally, our baselines are the UniDi encoder and BiDi encoder with all unidirectional and bidirectional predictions ($|y^*_t = \hat{y} \leq t|$).\(^4\) This policy is greedy. The optimal policy obtained via dynamic programming relies on the future tokens.

\(^4\)This policy is greedy. The optimal policy obtained via dynamic programming relies on the future tokens.
Table 3: Example computation for the metrics described in §4.1. EO is calculated as fraction of edits that were unnecessary. RC measures the fraction of prefix predictions that matches the model’s final predictions. Streaming EM measures the fraction of prefix predictions that are correct with respect to the ground truth labels.

<table>
<thead>
<tr>
<th>Current Token</th>
<th>Prefix Predictions</th>
<th>Total Edits</th>
<th>Unnecessary Edits</th>
<th>Exact Match w/ ground truth</th>
<th>Exact Match w/ final prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>find</td>
<td>O</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>new</td>
<td>O</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>york</td>
<td>LOC LOC LOC</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>times</td>
<td>ORG ORG ORG</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>square</td>
<td>ORG ORG ORG LOC</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

EO (Edit Overhead) is calculated as the fraction of total edits that were unnecessary with the ground truth. For example in Table 3, only the first label sequence (“O”) for input “find” and third label sequence (“LOC”) for input “york” were unnecessary as it differs from its final label “ORG”.

RC measures the relative correctness of prefixes, i.e. correctness of prefix prediction sequence with respect to the final label sequence over the complete input. For the example in Table 3, only the label sequence in the first timestep (“O”), fourth timestep (“ORG ORG ORG”) and final (fifth) timestep are prefixes of the label sequence in the final timestep (“ORG ORG ORG ORG LOC”). Thus 3 prefixes of a total 5 prefixes were correct. So, the Relative Correctness is 3/5, where higher is better.

While EO and RC capture consistency and stability in streaming predictions, neither of these measures performance with respect to the ground truth label over offline input. Relying on these metrics alone is not sufficient to measure streaming performance. For example, a UniDi encoder achieves perfect EO and RC scores, despite having poor predictions with respect to the offline ground truth. Thus, we consider Streaming Exact Match (Streaming EM), which is the streaming setting analogue of the Exact Match metric. Streaming EM calculates the percentages of prefix label sequence which are correct with respect to the offline ground truth labels. For example in Table 3, only the first label sequence (“O”) for input “find” and third label se-
### Table 4: Offline F1 over test set of datasets described in §4. On all datasets, HYBRID encoder achieves performance parity with the BiDi models.

<table>
<thead>
<tr>
<th>Model</th>
<th>SNIPS</th>
<th>CHUNK</th>
<th>NER</th>
<th>SRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniDi encoder</td>
<td>86.8</td>
<td>88.1</td>
<td>73.8</td>
<td>56.4</td>
</tr>
<tr>
<td>HYBRID encoder</td>
<td>93.0</td>
<td>89.4</td>
<td>86.8</td>
<td>80.0</td>
</tr>
<tr>
<td>BiDi encoder</td>
<td>93.1</td>
<td>89.4</td>
<td>86.0</td>
<td>80.1</td>
</tr>
</tbody>
</table>

HYBRID Encoder’s Offline Performance is Competitive to BiDi Encoders. Table 4 shows the offline F1 scores of the HYBRID encoder, UniDi and BiDi encoders across the four tasks. The HYBRID encoder has similar offline F1 as BiDi encoders. In fact, on NER, it even outperforms it by a margin of 0.8 F1 score. As expected, the UniDi model performs poorly compared to the BiDi encoder across all the tasks. From here on, we omit UniDi encoder.

HYBRID Encoder Improves Streaming Efficiency. Table 5 shows the GFLOP (per input instance) in streaming settings across the four datasets for the HYBRID encoder and BiDi encoders, with a trivial restart at every new token to get predictions over streaming text. We observe that HYBRID encoder consistently offers lower GFLOP count than BiDi encoder across the datasets, and in three of the four datasets, offering more than 40% FLOP reduction. However, HYBRID encoder does not improve on streaming performance (Streaming EM, EO, and RC) over BiDi, and its FLOP can be further reduced. We next see how incorporating HEAR addresses this.

HEAR Improves Streaming Efficiency and Performance. Across all datasets, we observe HEAR further reduces GFLOP count from the already computationally lighter HYBRID encoder, giving us up to 71% total FLOP reduction from BiDi. HEAR has better streaming predictions, as its Streaming EM is much higher (upto +10.0) than BiDi, improving on the shortcoming of Hybrid. This highlights that naively restarting at each timestep can worsen the streaming performance, as evident through the lower performance of Hybrid and BiDi. HEAR also offers reduced number of avoidable label flips, as signified by a lower EO. Its high RC score signifies that its streaming predictions are more consistent with its final prediction.

All these results demonstrate that HEAR preserves the offline performance of BiDi encoders, while being computationally lighter by 58.9% on average across tasks. HEAR has more consistent streaming predictions which are accurate with respect to offline ground truth.

6 Analysis and Ablations

In this section, we perform various analysis pertaining to the HYBRID encoder and ARM.

6.1 HYBRID Encoder’s Unidirectional Layers Performance

Table 6 compares the performance of the prediction over the intermediate (unidirectional) and final (bidirectional) encodings for the best HEAR model. While the intermediate ones lag in comparison to the final, we get decent performance from intermediate ones across all except for SRL dataset. This shows that, when used selectively, UniDi intermediate predictions can serve as a good source of auxiliary predictions.

6.2 ARM vs RESTART-\(k\).

Table 6 shows the performance of baseline - the best RESTART-\(k\) from \{2, 3, 5, 8\} over dev set Streaming EM and the ARM. We observe that across all the datasets, on both Streaming EM and EO metric, ARM performs significantly better than the heuristic RESTART-\(k\) mechanism. This shows that using a fixed length wait is not sufficient.

6.3 Intrinsic Performance of ARM

Table 6 shows the performance of the ARM on its binary classification task. Given the lightweight ARM architecture and the task complexity, the
Table 5: Streaming performance over test set of datasets. HEAR significantly outperforms BiDi on all metrics improving upon efficiency and performance. Direction of arrow indicates whether higher or lower is better.

Table 6: Development set ablation and analyses. From left to right - comparing performance of HYBRID encoder over sequence tagging using its bidirectional (final) vs unidirectional (intermediate) encoding, comparison of Streaming EM and EO of RESTART-k vs ARM, and the performance of ARM w.r.t. its ground truth policy labels.

6.4 ARM’s Architecture Ablation

Table 7: Performance of HEAR with different ARM model architectures over SNIPS development set.

Table 7 shows the Streaming EM and EO scores of HEAR with different ARM model architectures for the SNIPS development set. We observe that HEAR having either a Linear layer or MLP as ARM does much better than having no ARM and restarting at each timestep. However, modeling ARM as a GRU recurrent model gives the best scores in both metrics.

7 Related Work

Streaming (or incremental) setting has been widely studied in machine translation and parsing, dating as far back as two decades (Larchevêque, 1995; Lane and Henderson, 2001). Specifically, for incremental parsing, there are two broad approaches that have been studied: transition-based and graph-based. Transition-based incremental parsers allow for limited backtracking and correcting parsing over prefixes by keeping track of multiple parse segments (Buckman et al., 2016) or via beams search (Bhargava and Penn, 2020). Such methods can fail on garden-path sentences and long-range dependencies and only work proficiently with a large beam. Graph-based parsers incrementally assign scores to edges of the graph, discarding those edges that cause conflicts to tree-structure of the graph. (Yang and Deng, 2020) proposed an attach-juxtapose system to grow the tree, requiring restart at each new token over streaming input. (Kitaev et al., 2022) improved on its efficiency by proposing a information bottleneck. However, these methods rely on the structured nature of parsing output can not be adapted to incremental sequence tagging tasks without restarting at each token.

Recently, Madureira and Schlangen (2020) benchmarked the modern encoders on streaming sequence labelling and observed poor streaming performance of pretrained transformer models. They explored improvement strategies by adopting techniques from other streaming streams like chunked training (Dalvi et al., 2018), truncated training (Köhn, 2018) (training model on heuristically-aligned partial input-output pairs), and prophecy (Alinejad et al., 2018) (autocompleting input using a separate language model). They observed performance degradation from truncated training unless used with prophecies. However, running a language model at each timestep for prophecy is computationally infeasible. Therefore, unlike our approach, neither methods can improve
performance feasibly.

Previous works have attempted to improve computation efficiency in BiDi encoders. Monotonic attention moves away from $O(n^2)$ soft attention overhead by restricting attention to monotonically increase across timesteps (Raffel et al., 2017; Chiu and Raffel, 2018; Ma et al., 2020). However, unlike HEAR, such attempts can’t maintain offline performance. Recently, Kahardipraja et al. (2021) used linear transformer (Katharopoulos et al., 2020) as unidirectional model using masking for streaming sequence tagging and classification. This approach performs well under the assumption of delayed output (Grissom II et al., 2014; Oda et al., 2015), a relaxation, where the model waits for additional tokens before predicting. Furthermore, the unidirectional model could not revise its output, rendering the model incapable of handling long-range dependency or tasks that go backward like SRL — an ability common to any model with some bidirectionality, such as HEAR. Similar drawbacks were in the partial bidirectional encoder, a bidirectional attention with restricted window, proposed by Iranzo Sanchez et al. (2022).

Improving efficiency through adaptive computing has been independently studied for reasoning-based tasks (Eyzaguirre and Soto, 2020), text generation models (Arumae and Bhatia, 2020; Eyzaguirre et al., 2022) and diffusion models (Ye et al., 2022). These works are restricted to offline settings and can be readily incorporated within the proposed overall approach of HEAR.

8 Conclusion

We propose HEAR for sequence tagging in streaming setting where the input is received one token at a time to the model. The encoder in our model is HYBRID encoder where early layers are unidirectional and later are bidirectional. It reduces the computational cost in streaming settings, by reducing the need of restart only to the later bidirectional layers while preserving the offline performance of the model. HEAR additionally consists of an ARM to predict when to restart the model. Using ARM leads to reduced number of restart of the encoder, leading to better streaming performance and further savings in computation. Compared to BiDi encoders, our model, HEAR (HYBRID encoder + ARM) reduces the computation by up to 71% in streaming settings while maintaining the performance of the BiDi encoders across various sequence tagging tasks. HEAR improved streaming EM by up to +10.0% and reduced unnecessary edits by up to -12.0%.

Limitations

An additional but small training cycle is required to train the lightweight ARM module of HEAR in order to reap the benefits of extra savings in efficiency and streaming performance. Also, even though we do not assume any language specific-design choices, we benchmarked on the standard streaming sequence labelling benchmark datasets, all of which were in English.

References


Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation?


Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. 2020. Flax: A neural network library and ecosystem for JAX.


Peter C. R. Lane and James B. Henderson. 2001. Incremental syntactic parsing of natural language corpora.


### A Experiment Details

#### A.1 Implementation Details

All the experiments were done using the Flax ([Heek et al., 2020](#)) and Jax ([Bradbury et al., 2018](#)) with Adam optimizer ([Kingma and Ba, 2014](#)) on TPUs. We measure flops in terms of XLA’s HLO ([Sabne, 2020](#)). The optimizer’s learning rate is set to 1e-3, betas are set to default at (0.9, 0.999), batch size is 256 and feedforward hiden is 2048 with 512 transformer dimension and 8 heads.

For all the datasets we use the standard splits, the links to which can be found in their respective papers along with statistics. Following ([Ratinov and Roth, 2009](#)), we use the BIOES tagging scheme for the NER task and BIO scheme for the rest. In SRL, following ([He et al., 2017](#)) the indicator for predicate verb is also used as input along with the sentence.
All the values over test sets are averaged across four seeds. For experiments involving ARM we average across four different HYBRID encoder trained from different seeded initialization. We initialize static embeddings with 300 dimensional glove embedding (Pennington et al., 2014).

A.2 Postprocessing ARM Predictions

We post process the ARM predictions to prevent too frequent or too infrequent restarts. Specifically, for hyperparameters $\alpha$ and $\beta$ where $0 \leq \alpha < \beta$, at any time $t$, if the ARM hasn’t restarted once since $\max(0, t - \beta)$ timesteps, then ARM’s prediction is set to 1, else, if the ARM has restarted at least once since $\max(0, t - \alpha)$ timesteps, then it is set to 0. We tune for the values of $\alpha$ and $\beta$ in the range $\{0, 1, 2, 3, 5, 10\}$ over the development set.

We also observe that even if bidirectional layers improve predictions over previous tokens, for the latest token only, the unidirectional label $\hat{y}_t$ is often better than $\hat{\gamma}_t$. If this is observed, then, we exclude the most recent token from getting updated during restart. We tune for this binary postprocessor over the development set as well.

B Analysis and Ablation

B.1 Benefits of Optimizing Unidirectional Predictor Separately

In HYBRID encoder the performance we inhibit backward gradient flow from the auxiliary predictor over the unidirectional layers unidirectional. This is because we observed offline F1 performance drop of bidirectional predictor when trained with unidirectional predictor, dropping from 81.20 to 71.62 on CoNLL development set. Even after tuning for loss scaling for the to predictors in ratio $\{1:1, 3.3:1, 10:1, 33:1, 100:1\}$, the performance was only increased upto 75.06 F1.

B.2 Oracle Policy for ARM

We also experimented with alternate policy for ARM where it’s label is conditioned on the last bidirectional restart. However, it led to a performance drop in terms of Streaming EM from 86.7 to 83.0, as observed across SNIPS development set. We attribute this poor performance on alternate policy, due to the lightweight nature of ARM. Thus, we did not proceed with this policy.