Google

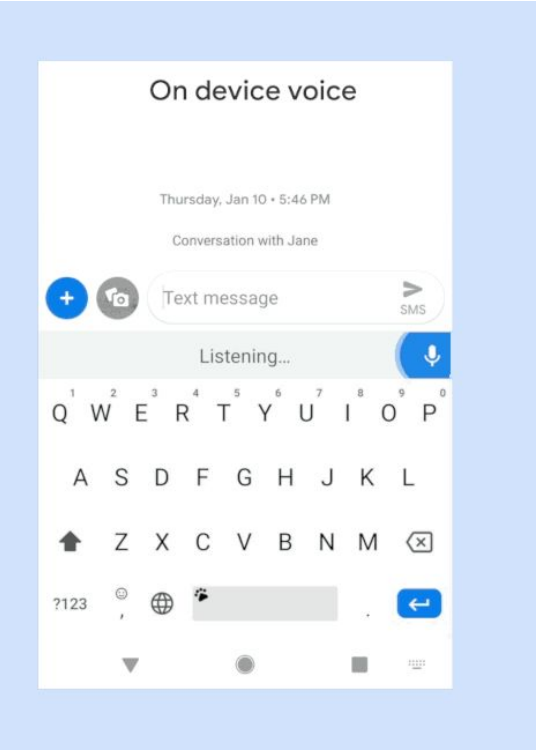
Dual-mode ASR: Unify and Improve Streaming ASR with Full-context Modeling

Jiahui Yu, Wei Han*, Anmol Gulati*, Chung-Cheng Chiu, Bo Li,
Tara N. Sainath, Yonghui Wu, Ruoming Pang

March 13, 2021

* Equal Contribution

Streaming (Online) ASR vs. Full-context (Offline) ASR



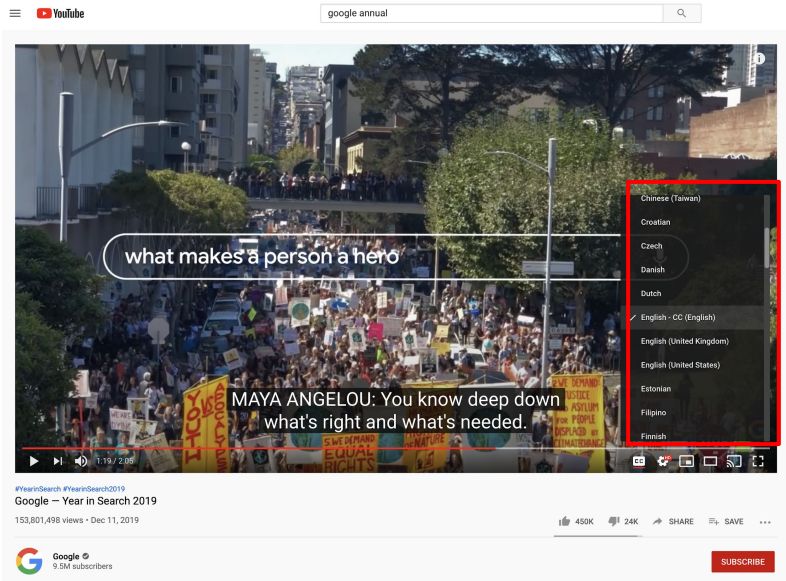
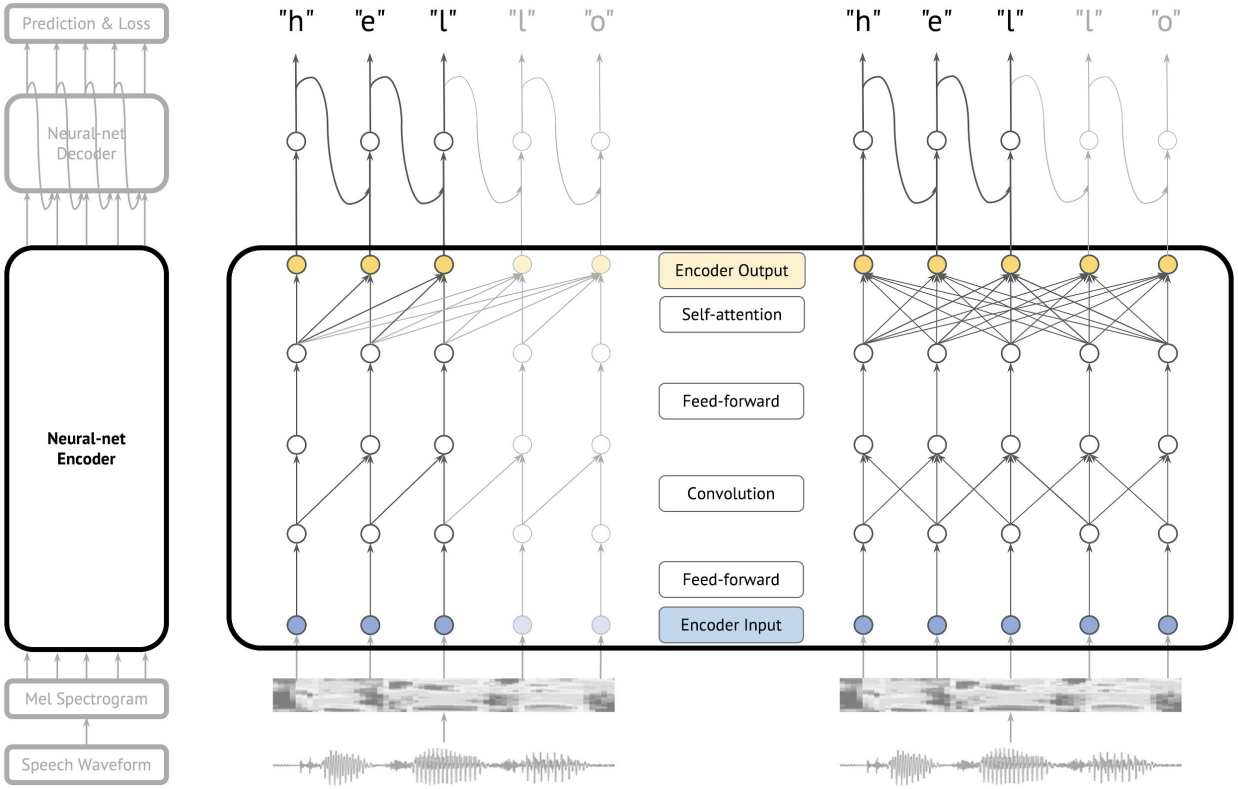


Photo Credit: ai.googleblog.com

Motivation: Unify Steaming and Full-context ASR

- Obvious benefits:
  1. Simplified development and deployment workflow.
  2. Reduced model download and storage on devices.
- What's more:
  1. Explore if unification might help model quality:
     a. Word Error Rate
     b. Emission latency of streaming ASR

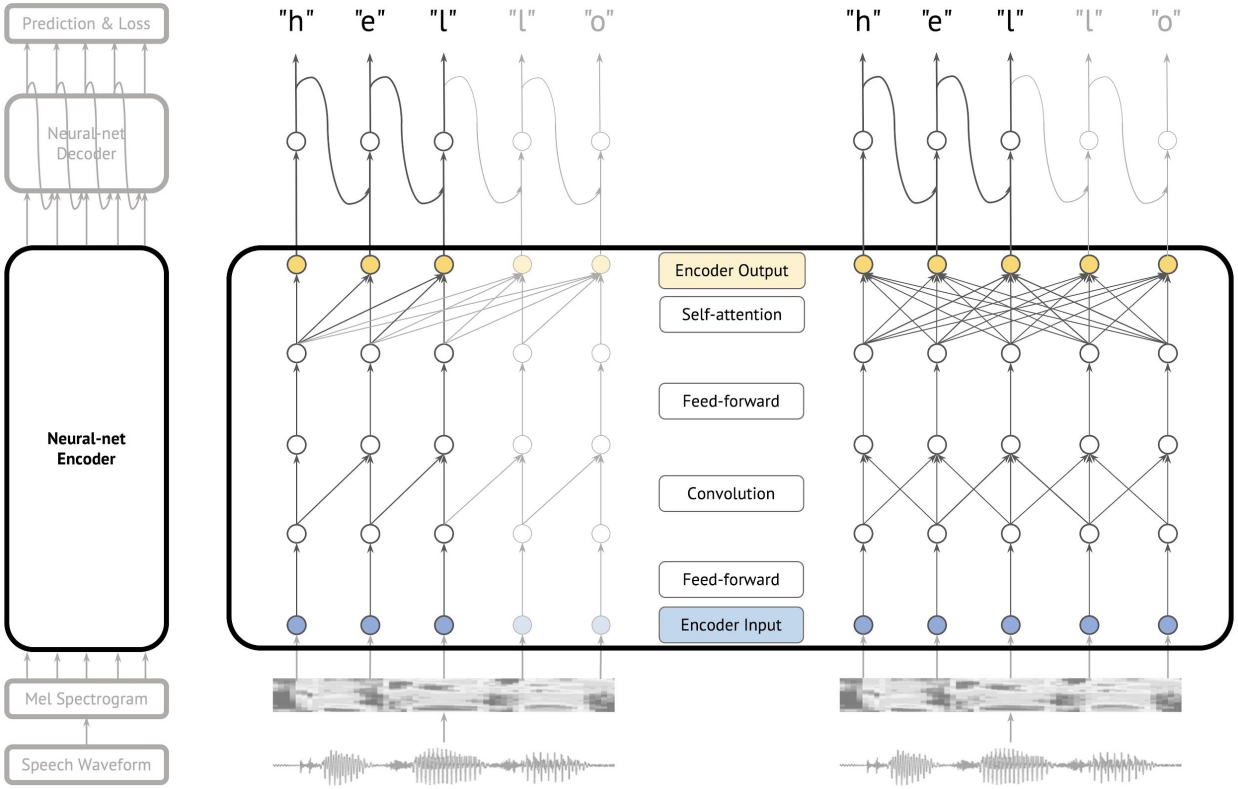
A Review of Streaming / Full-context ASR Models



Streaming ASR with
Auto-regressive Encoder

Full-context ASR with
Full-context Encoder

A Review of Streaming / Full-context ASR Models



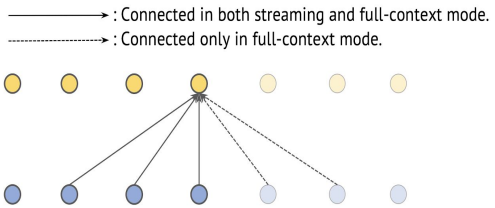
Streaming ASR with Auto-regressive Encoder

Full-context ASR with Full-context Encoder

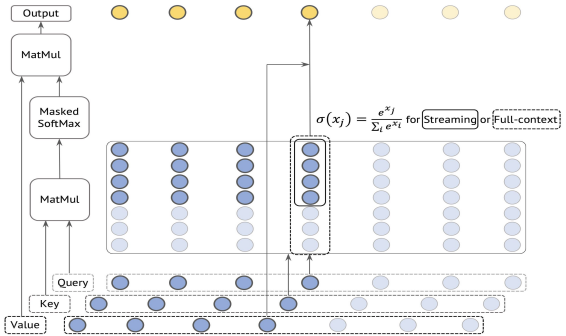
Observation: Modern end-to-end streaming and full-context ASR models **share** most of the neural architectures and training recipes in common, with the most significant **difference in the ASR encoder**.

Challenge: How to unify ASR encoders (dual-mode encoders)?

- We propose two **Design Principles** for dual-mode encoders:
  1. Each layer in a dual-mode encoder should be either dual-mode or streaming (a.k.a., causal).
  2. The design of a dual-mode layer should not introduce significant amount of additional parameters, compared with its streaming model.



Dual-mode Temporal-wise Convolution
(kernel size: 3 or 5)



Dual-mode Self-attention

Challenge: How to unify ASR encoders (dual-mode encoders)?

- Applying two design principles to all neural-net layers:
  1. Pointwise operators
  2. Dual-mode convolution
  3. Dual-mode average pooling
  4. Dual-mode self-attention
  5. Dual-mode normalization
  6. ...

Inplace Distillation:

We use full-context mode as teacher, to distill the streaming mode (student) within the SAME model:

NO additional computation / memory cost.

Algorithm 1 Pseudocode of training Universal ASR networks.

```
# Requires: data_loader; context manager with support of mode switching by network.mode();
    universal_network with support of running both modes under context manager;

for x, y in data_loader: # Load a minibatch of speech input x and text label y.
   with universal_network.mode('fullcontext'): # Switch context to 'fullcontext' mode.
      # Compute full-context prediction given speech input x and text label y.
      fullcontext_pred = universal_network.forward_encoder_decoder(x, y)
      # Compute RNN-T loss of full-context mode.
      fullcontext_loss = rnnt_loss(fullcontext_pred, y)

   with universal_network.mode('streaming'): # Switch context to 'streaming' mode.
      # Compute streaming prediction given speech input x and text label y.
      streaming_pred = universal_network.forward_encoder_decoder(x, y)
      # Compute RNN-T loss of streaming mode.
      streaming_loss = rnnt_loss(streaming_pred, y)

   # Add inplace knowledge distillation loss (full-context prediction as teacher).
   distill_loss = inplace_distill_loss(streaming_pred, stop_gradient(fullcontext_pred))

   # Compute total loss as a sum of full-context, streaming and distillation losses.
   loss = fullcontext_loss + streaming_loss + distill_loss
   loss.backward() # Update weights.
```

Experiments -- Dataset

Table 1: Summary of datasets we used in our experiments.

Dataset Name	**# Hours**	**# Utterances**	**Speech Domain**
LibriSpeech (Panayotov et al., 2015)	~ 970	$\sim 281,000$	Single domain of English reading speech.
MultiDomain (Narayanan et al., 2018)	$\sim 413,000$	$\sim 287,000,000$	Multiple domains including: Voice Search, Farfield Speech, YouTube and Meetings.

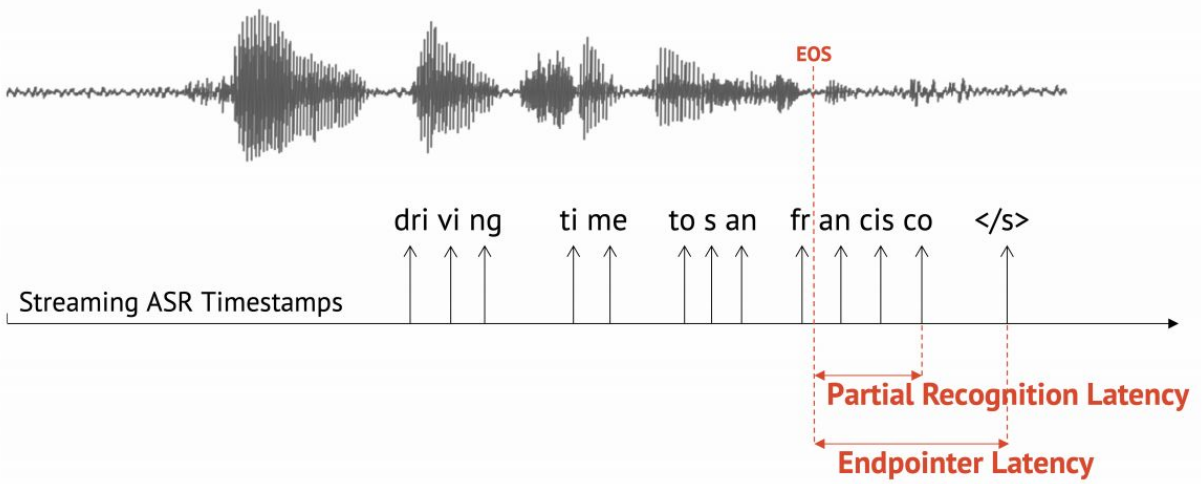
Experiments -- Evaluation Metrics

WER: Word Error Rate (recognition quality)

Partial Recognition Latency: the timestamp difference of two events:

1) when the last token is emitted in the finalized recognition result;

2) the end of the speech when a user finishes speaking



Experiments -- Main Results on **MultiDomain**

Table 2: Summary of our results on MultiDomain dataset (Narayanan et al., 2018). We report WER on Voice Search test set. Compared with standalone ContextNet and Conformer models, Dual-mode ASR models have slightly higher accuracy and much better streaming latency. ASR models that capture stronger contexts can emit the full hypothesis even slightly before they are spoken, leading to a *negative latency*.

Method	**Mode**	**# Params (M)**	**VS Test** WER(%)	**Latency@50** (ms)	**Latency@90** (ms)
ContextNet	Full-context	133	5.1	——	——
Conformer	Full-context	142	5.2	——	——
LSTM (Sainath et al., 2020)	Streaming	179	6.4	190	350
ContextNet (Han et al., 2020)	Streaming	133	6.1	160	310
Conformer (Gulati et al., 2020)	Streaming	142	6.1	160	300
Dual-mode ContextNet	Full-context	133	4.9	——	——
	Streaming		6.0 (-0.1)	10 (-150)	220 (-90)
Dual-mode Conformer	Full-context	142	5.0	——	——
	Streaming		6.0 (-0.1)	-50 (-210)	130 (-170)

Experiments -- Main Results on **LibriSpeech**

Table 3: Summary of our results on Librispeech dataset (Panayotov et al., 2015). We report WER on TestClean and TestOther (noisy) set. Compared with standalone ContextNet and Conformer models, Dual-mode ASR models have both higher accuracy in average and better streaming latency.

Method	**Mode**	**# Params (M)**	**Test Clean/Other** WER(%)	**Latency@50** (ms)	**Latency@90** (ms)
LSTM-LAS	Full-context	360	2.6 / 6.0	——	——
QuartzNet-CTC	Full-context	19	3.9 / 11.3	——	——
Transformer	Full-context	29	3.1 / 7.3	——	——
Transformer	Full-context	139	2.4 / 5.6	——	——
ContextNet	Full-context	31.4	2.4 / 5.4	——	——
Conformer	Full-context	30.7	2.3 / 5.0	——	——
Transformer	Streaming	18.9	5.0 / 11.6	80	190
ContextNet	Streaming	31.4	4.5 / 10.0	70	270
Conformer	Streaming	30.7	4.6 / 9.9	140	280
ContextNet Look-ahead	Streaming	31.4	4.1 / 9.0	150	420
Dual-mode Transformer	Full-context	29	3.1 / 7.9	——	——
	Streaming		4.4 (-0.6) / 11.5 (-0.1)	-50 (-130)	30 (-160)
Dual-mode ContextNet	Full-context	31.8	2.3 / 5.3	——	——
	Streaming		3.9 (-0.6) / 8.5 (-1.5)	40 (-30)	160 (-110)
Dual-mode Conformer	Full-context	30.7	2.5 / 5.9	——	——
	Streaming		3.7 (-0.9) / 9.2 (-0.7)	10 (-130)	90 (-190)

Summary

1. We propose a unified framework, Dual-mode ASR, to train a single end-to-end ASR model with shared weights for both streaming and full-context speech recognition.
2. We show that the latency and accuracy of streaming ASR significantly benefit from weight sharing and joint training of full-context ASR, especially with inplace knowledge distillation.
3. We present extensive experiments with two state-of-the-art ASR networks, ContextNet and Conformer, on two datasets, a widely used public dataset LibriSpeech and an internal large-scale dataset MultiDomain.