Google

April 5, 2022 @ Rutgers University, Efficient AI seminar

Reducing Longform Errors in End2End Speech Recognition

Liangliang Cao
Senior Staff Research Scientist, Google Inc
http://llcao.net

Acknowledgement to many colleagues in Google:

Thibault Doutre, **Zhiyun Lu**, Wei Han, Yu Zhang, Chung-Cheng Chiu, Arun Narayanan, Bo Li, Ruoming Pang, Peng Chen, Yanwei Pan, Min Ma, Heng Su, Andrea Chu, Yongqiang Wang, Rohit Prabhavalkar, Chao Zhang, David Qiu, Yanzhang He, Basi Garcia, Hank Liao, Qian Zhang, Han Lu, Hasim Sak, Olivier Siohan, Parisa Haghani, Tara Sainath, James Fan, Yonghui Wu, Trevor Stroman, Françoise Beaufays

Outline

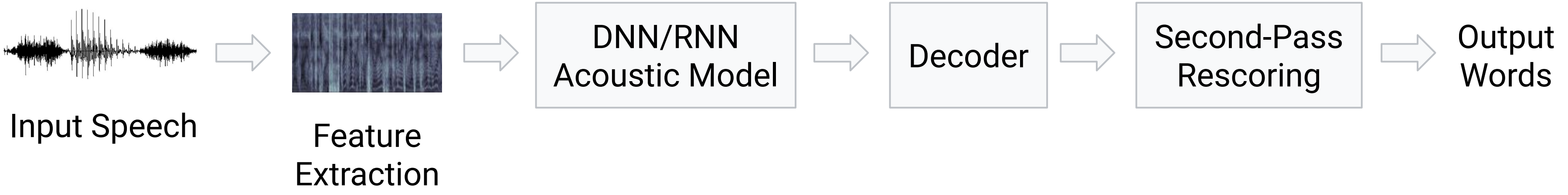
- What is End2End Speech Recognition
- Long form errors and Universal perturbation
- Teacher distillation on Youtube data
- Combining CTC and RNN-T teachers

What is End-to-End ASR?

Slides in this section are borrowed from Bo Li et al's ISCSLP'18 Tutorial

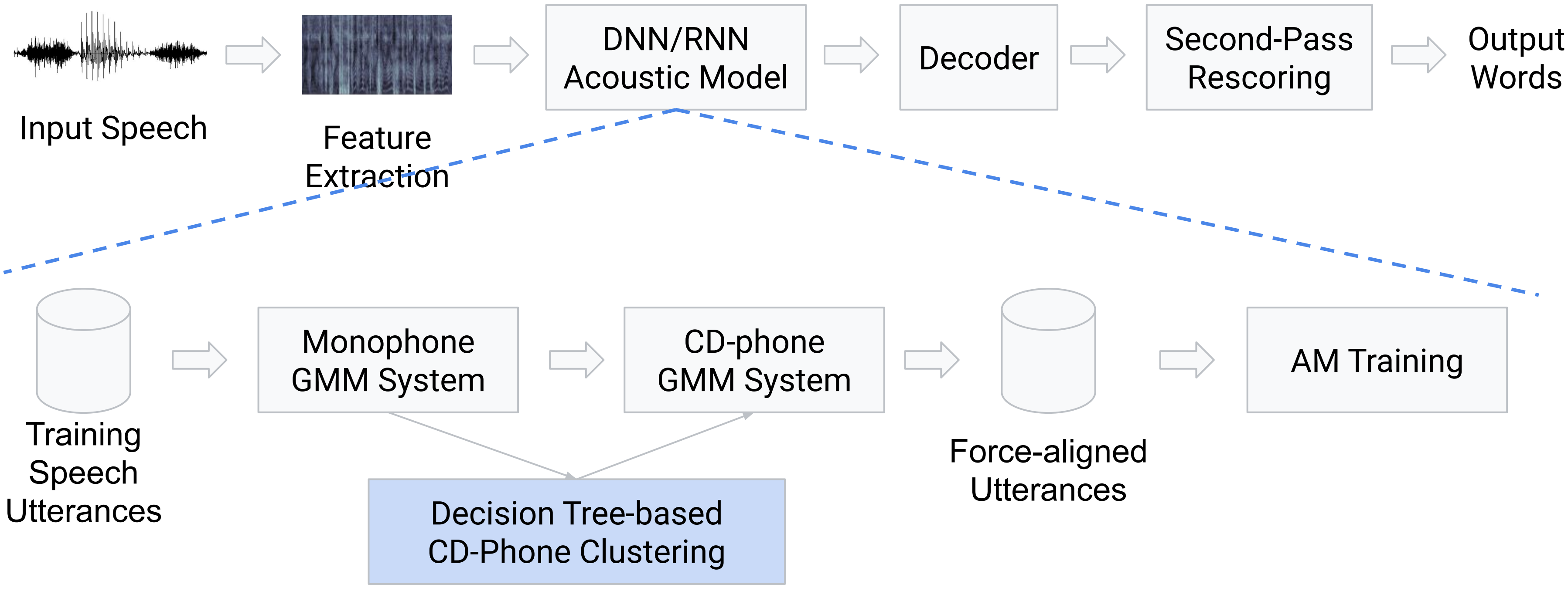
Conventional speech recognition

Pipeline

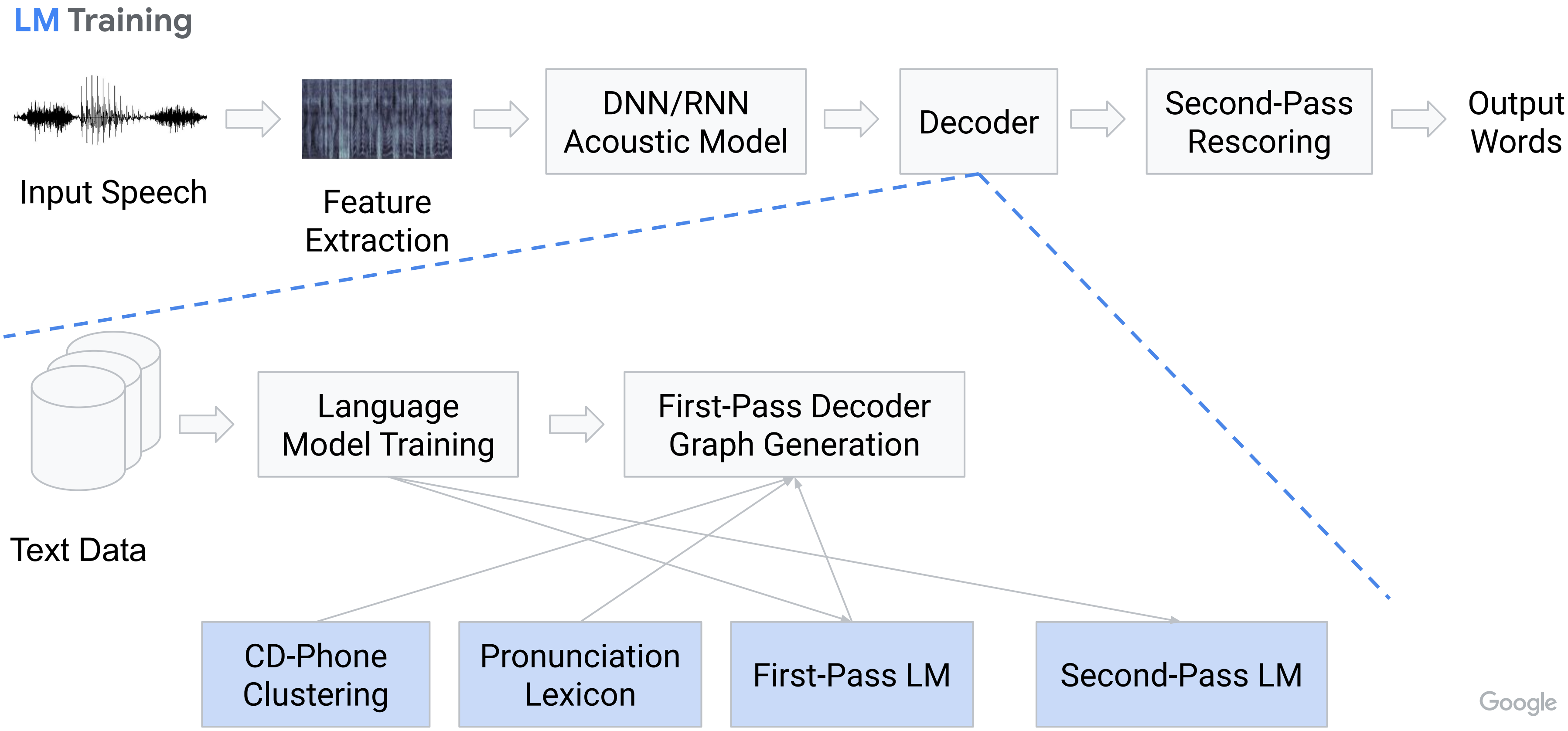


Conventional speech recognition

AM Training



Conventional speech recognition



What is end2end learning?

"A system which is trained to optimize criteria that are related to the final evaluation metric that we are interested in (typically, word error rate)."

Examples of end2end learning:

- End2end speech recognition
- AlexNet (for end2end image classification)
- DETR (for end2end object detection)

From conventional to end2end ASR

Conventional Speech System

Acoustic Model

Pronunciation Model

Verbalizer

Language Model

Endpointer

End2End Trained Sequence-to-Sequence Recognizer

Benefits:

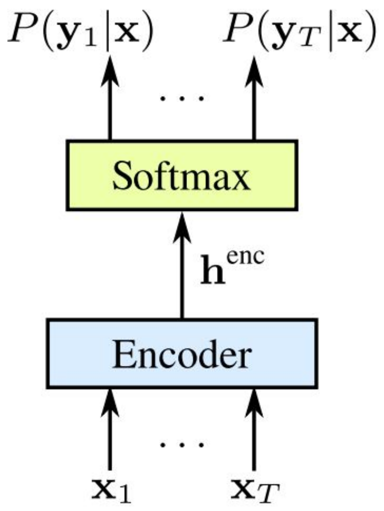
- Simply the speech recognition pipeline
- Speed up the training process (weeks -> days)
- TPUs/GPUs did all the heavy-lifting jobs
- Good with large scale training data

 - Some end2end models (such as RNN-T) is 10x smaller than the conventional models!

Different end2end ASR models

- CTC (Connectionist Temporal Classification)
- Listen Attend and Spell (LAS)
- RNN-Transducer (RNN-T)
  - Transformer and Conformer can be viewed as special cases of RNN-T

Connectionist Temporal Classification (CTC)

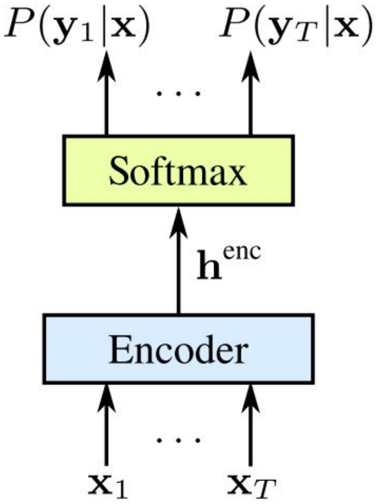


References:
- Alex Graves, Navdeep Jaitly, Towards End-To-End Speech Recognition with Recurrent Neural Networks, 2014
- Amodei et al., DeepSpeech2, 2015

Key Takeaway

CTC allows for training an acoustic model without the need for frame-level alignments between the acoustics and the transcripts.

Connectionist Temporal Classification (CTC)



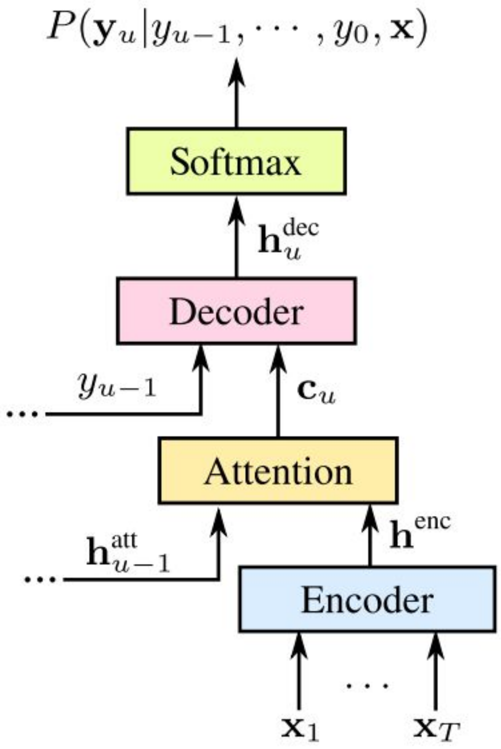

```
B B c B B a a B B t
B c c B a B B B B t
         ...
B c B B a B B t t B
```

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\hat{\mathbf{y}} \in \mathcal{B}(\mathbf{y},\mathbf{x})} \prod_{t=1}^{T} P(\hat{y}_t|\mathbf{x})$$

Key Takeaway

CTC introduces a special symbol - blank (denoted by B) - and maximizes the total probability of the label sequence by marginalizing over all possible alignments

Listen, Attend and Spell (LAS)

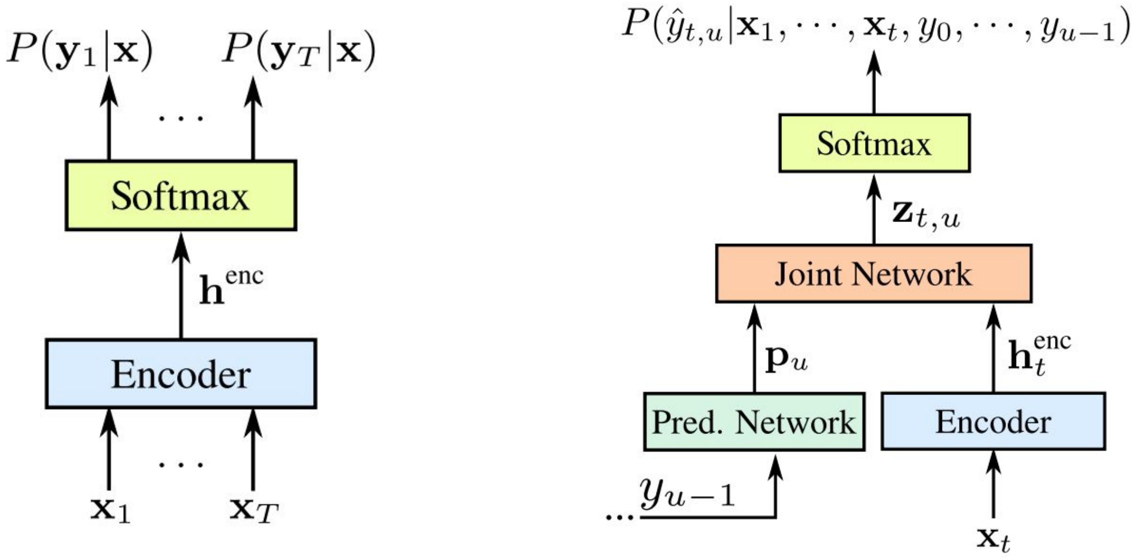


- Encoder (analogous to AM):
  - Transforms input speech into higher-level representation
- Attention (alignment model):
  - Computes a similarity score between the decoder and each frame of the encoder
  - Identifies encoded frames that are relevant to producing current output
- Decoder (analogous to PM, LM):
  - Operates autoregressively by predicting each output token as a function of the previous predictions

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals, “Listen, Attend, and Spell”, ICASSP 2016

Recurrent Neural Network Transducer (RNN-T)

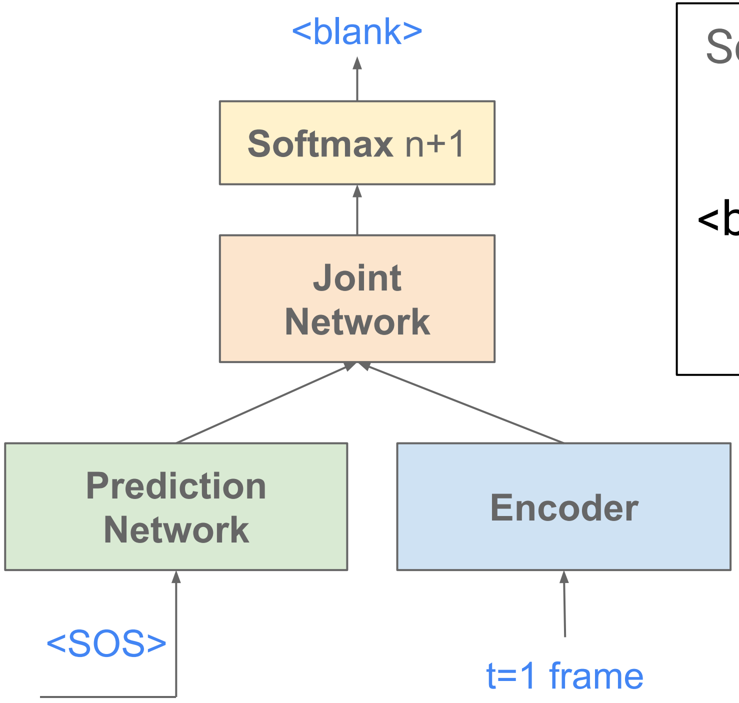
RNN-T augments CTC encoder with a recurrent neural network LM



Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton, Speech Recognition with Deep Recurrent Neural Networks, 2013

Recurrent Neural Network Transducer (RNN-T)

Output:

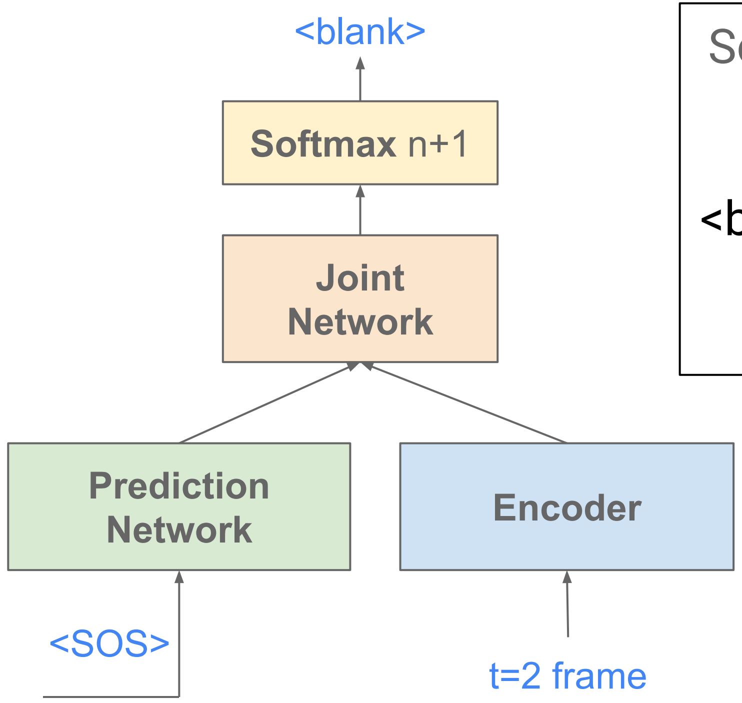


Softmax over n + 1 labels, includes a blank like CTC.

<blank> → advance in Encoder, retain prediction network state

Recurrent Neural Network Transducer (RNN-T)

Output:

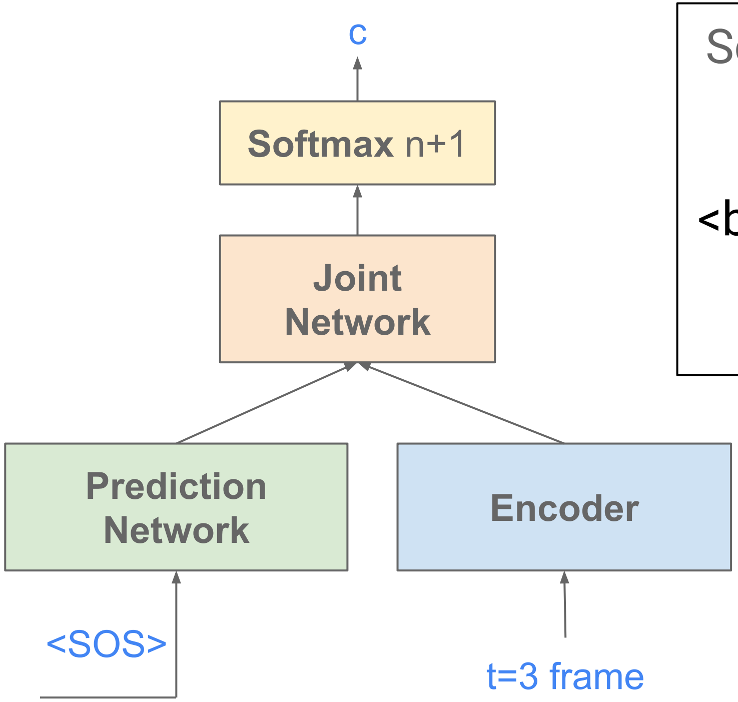


Softmax over n + 1 labels, includes a blank like CTC.

<blank> → advance in Encoder, retain prediction network state

Recurrent Neural Network Transducer (RNN-T)

Output: c

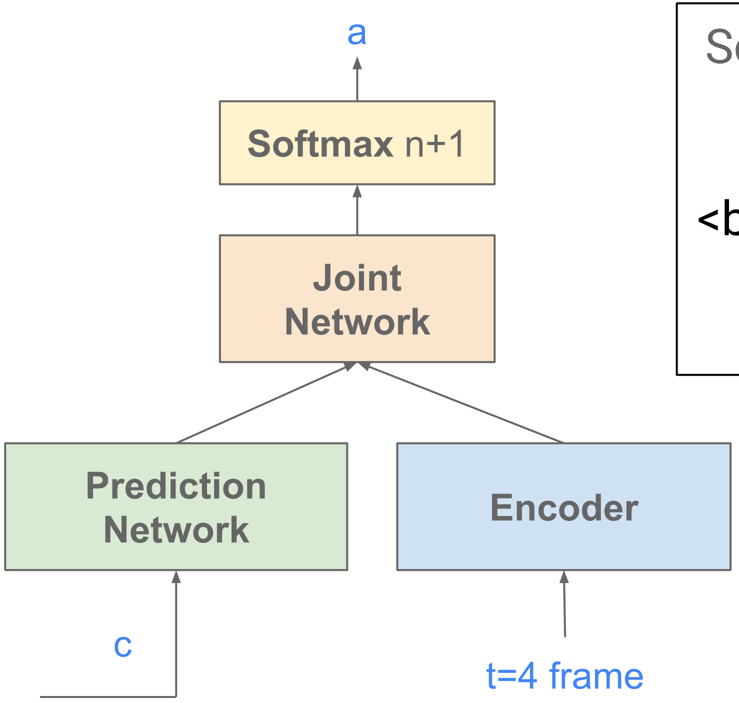


Softmax over n + 1 labels, includes a blank like CTC.

<blank> → advance in Encoder, retain prediction network state

Recurrent Neural Network Transducer (RNN-T)

Output: ca



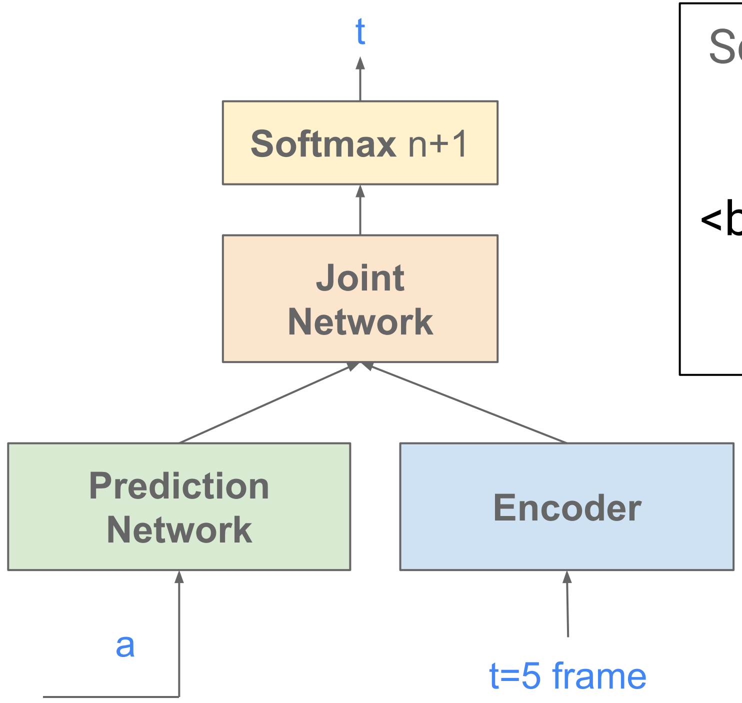
Softmax over n + 1 labels, includes a blank like CTC.

<blank> → advance in Encoder, retain prediction network state

Google

Recurrent Neural Network Transducer (RNN-T)

Output: cat

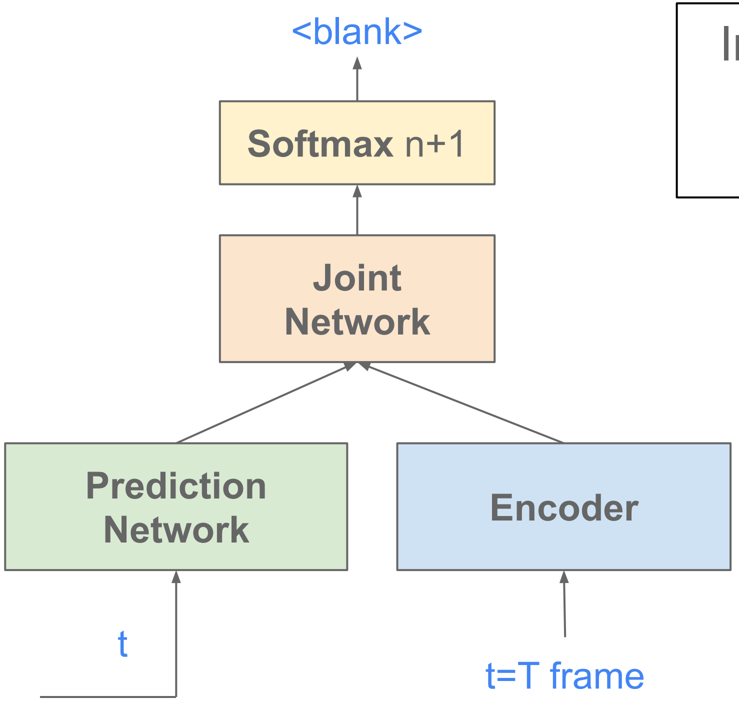


Softmax over n + 1 labels, includes a blank like CTC.

<blank> → advance in Encoder, retain prediction network state

Recurrent Neural Network Transducer (RNN-T)

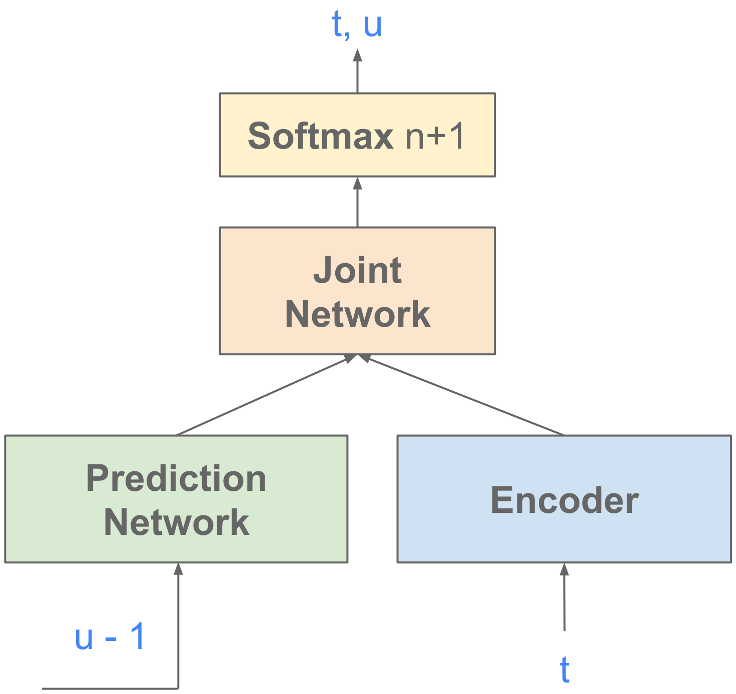
Output: cat



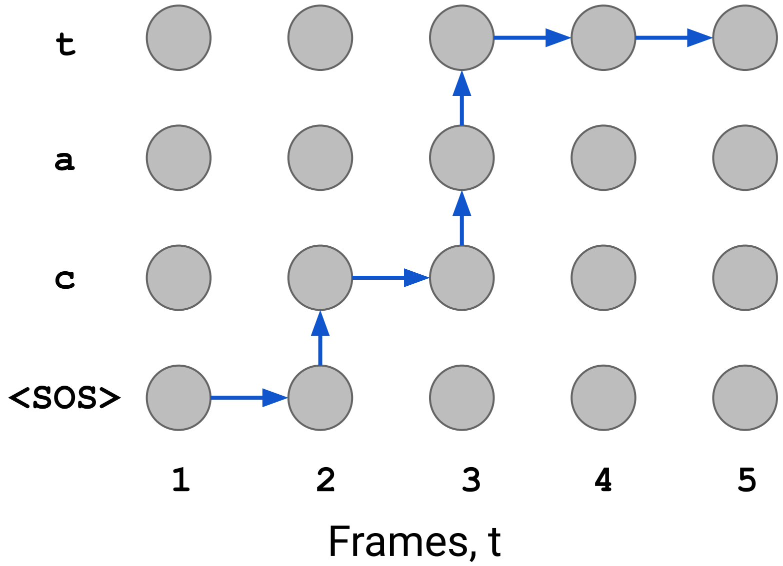
Inference terminates when all input frames have been consumed

Google

Recurrent Neural Network Transducer (RNN-T)



During training feed the true label sequence to the LM.
Given a target sequence of length U and T acoustic frames we generate UxT softmax



Related Google Papers/Blogs

How RNN-T was deployed in the Google products:

- Johan Schalkwyk's Google AI blog: "An All-Neural On-Device Speech Recognizer", 2019

A few recent improvement:

- Sainath and He et al, A Streaming On-Device End-to-End Model Surpassing Server-Side Conventional Model Quality and Latency, ICASSP 2020
- Zhang et al, Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss, ICASSP 2020
- Gulati et al, Conformer: Convolution-augmented Transformer for Speech Recognition, Interspeech 2020

Longform error and Universal Perturbation

RNN-T may suffer from high deletion errors on long form audios

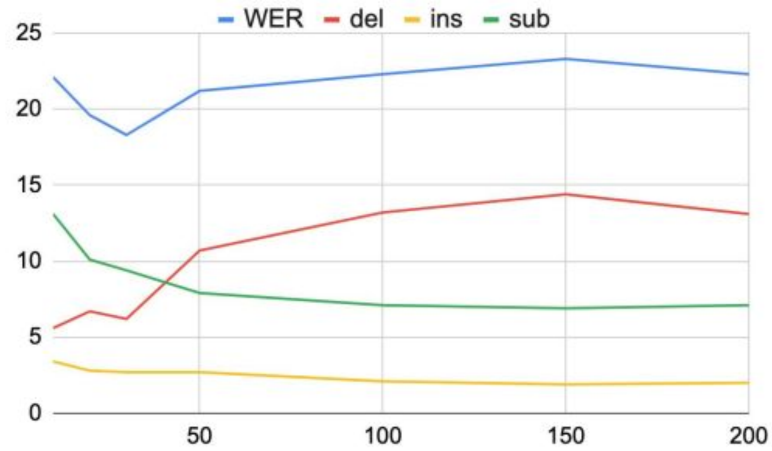


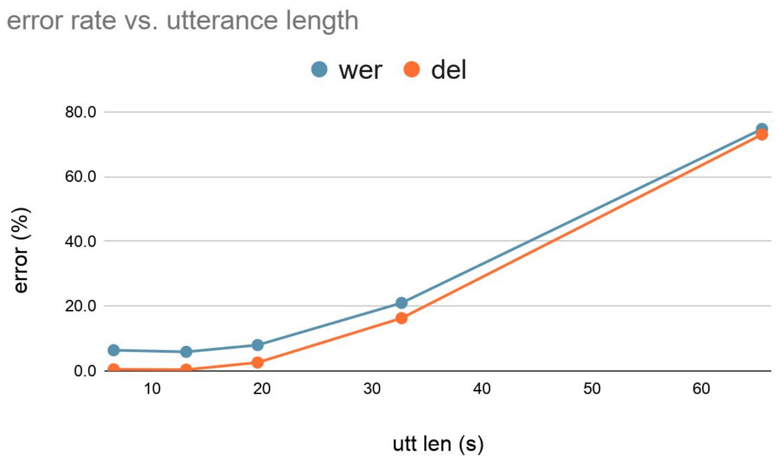
Fig. 3: WERs and the respective deletion, insertion, and substitution errors for non-streaming model on *YT-long* as a function of training steps.

with Chung-Cheng Chiu et al: RNN-T Models Fail to Generalize to Out-of-Domain Audio: Causes and Solutions, SLT 2021

RNN-T may suffer from high deletion errors on long form audios

[Conformer](#) model on concatenated librispeech test-other set.

# concatenation	# seconds	WER (del/ins/sub)
1 (original)	6.5	6.4 (0.5/0.8/5.1)
3	19.6	8.0 (**2.6**/0.7/4.7)
5	32.7	21.0 (**16.3**/0.6/4.1)
10	65.5	74.7 (**73.0**/0.2/1.4)

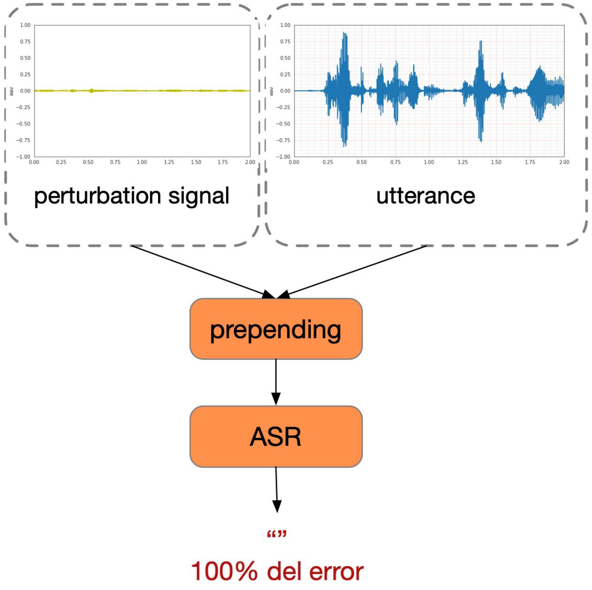


with Zhiyun Lu et al: Exploring Targeted Universal Adversarial Perturbations to End-to-end ASR Models, Interspeech 2021

Create long-form errors by universal perturbation

We find we can intentionally create deletion errors by learning a magic 4 second audio at beginning every audio.

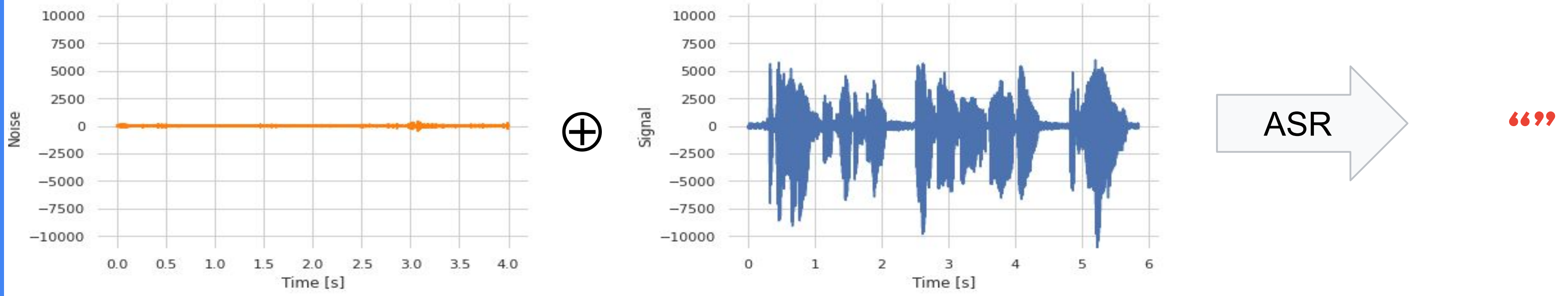
One perturbation that works for all utterances, even unseen one!



with Zhiyun Lu et al: Exploring Targeted Universal Adversarial Perturbations to End-to-end ASR Models, Interspeech 2021

Google

Problem statement



δ $\quad$ $\boldsymbol{x}$ $\quad$ θ $\quad$ y'

x	an audio from some $\mathcal{D}$
$\mathcal{T}_\delta(\boldsymbol{x})$	apply δ to x
y’	the mis-transcription we specify
$\ell(\mathcal{T}_\delta(\boldsymbol{x}), y'; \theta)$	loss (cross-entropy, RNNTLoss, CTCLoss)

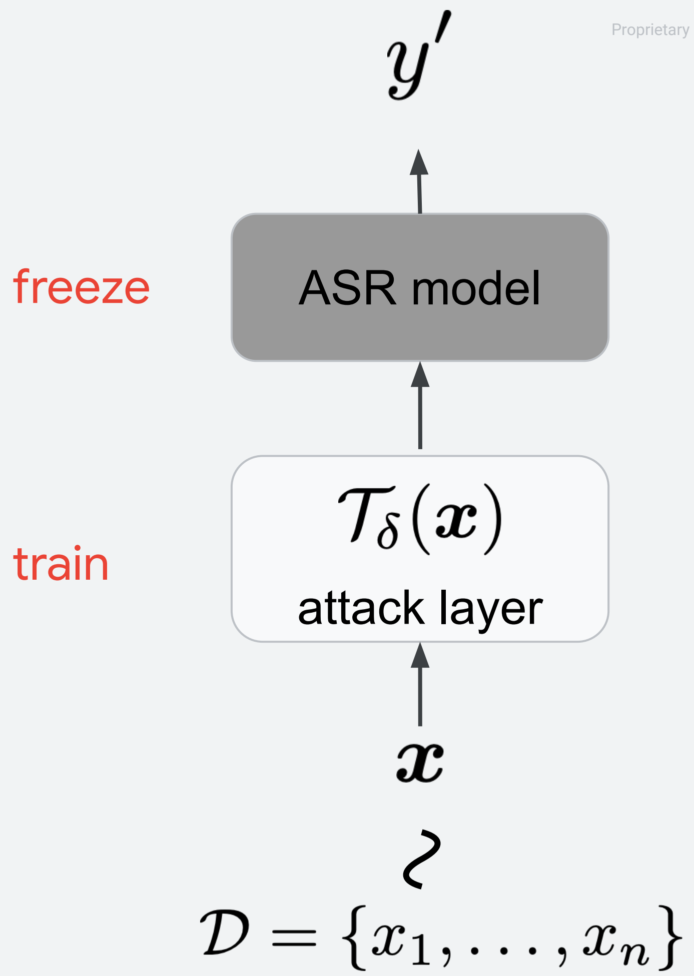
Learning the universal perturbation

$$\min_{\delta} \sum_{\boldsymbol{x} \in \mathcal{D}} \ell(\mathcal{T}_\delta(\boldsymbol{x}), y'; \theta)$$

cf. normal model training

$$\min_{\theta} \sum_{\boldsymbol{x} \in \mathcal{D}} \ell(\theta; \boldsymbol{x}, y)$$

Note: Our experiment only applies for Librispeech models, but NOT Google's production models (for latter we cannot compute gradient with a non-differentiable frontend)

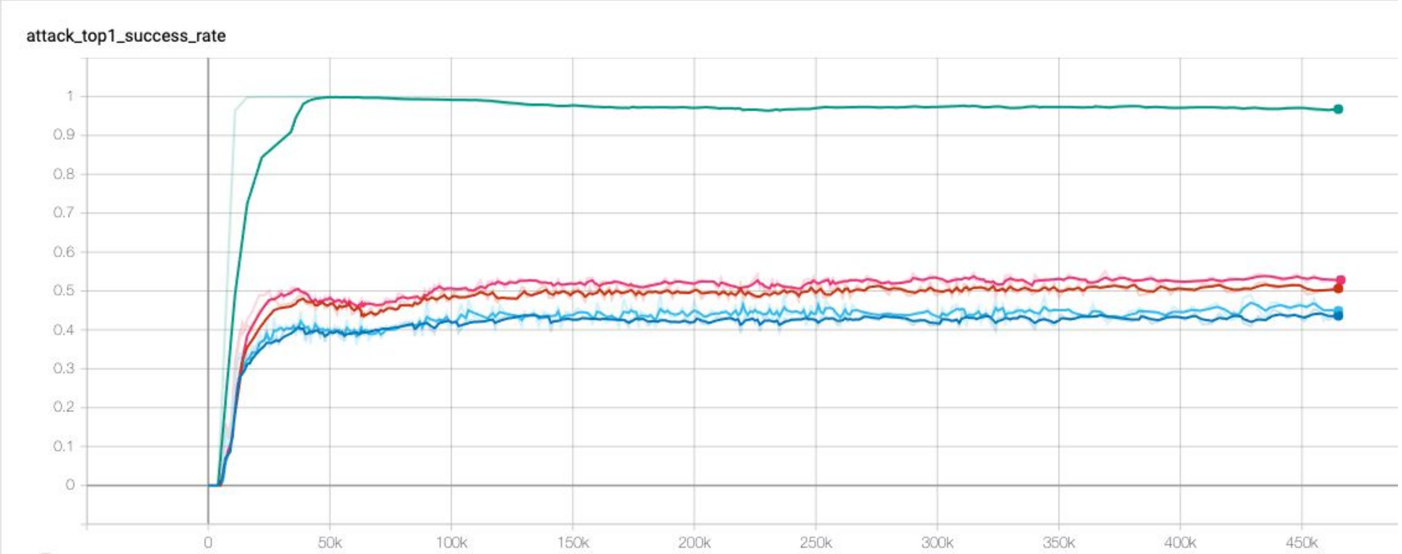


Experiment setup

- dataset: Librispeech
    - train on 960h
    - report on test-clean (2620 audio), test-others (2939 audio)
- evaluation metrics
    - success rate: $\frac{\#(\text{utt outputs} = y')}{\#(\text{utt})}$
    - dB: measure distortion (loudness)

$$D(\delta, \boldsymbol{x}) = \mathrm{dB}(\delta) - \mathrm{dB}(\boldsymbol{x}), \quad \mathrm{dB}(\boldsymbol{x}) = 20 \log_{10}(\max_i(\boldsymbol{x}_i))$$

Experiment



y’ = ""
prepend noise

Listen to the adversarial perturbation (Conformer-LAS)

Using models trained from public Librispeech and an unseen data

Fool the model to predict “ ” on all utterances in Librispeech test sets.

universal perturbation
(4 seconds)

prediction “”

transcript_truth
a cold lucid indifference reigned in his soul

prediction “”

transcript_truth
he hoped there would be stew for dinner turnips and carrots and bruised potatoes and fat mutton pieces to be ladled out in thick peppered flour fattened sauce

Listen to the adversarial perturbation (Conformer-Transducer)

Using models trained from public Librispeech and an unseen data

Fool another model to predict “ ” on the unseen testing set.

universal perturbation (4 seconds)

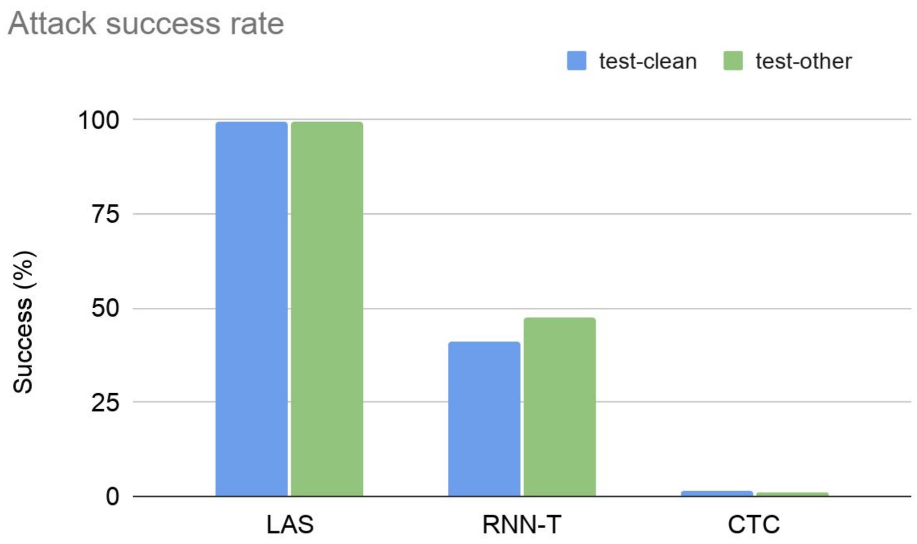
prediction “”

transcript_truth
a cold lucid indifference reigned in his soul

prediction “”

transcript_truth
he hoped there would be stew for dinner turnips and carrots and bruised potatoes and fat mutton pieces to be ladled out in thick peppered flour fattened sauce

Attack easiness: LAS > RNN-T > CTC



y’ = ""
prepend noise

Teacher distillation on Youtube data

Google

What did we learn?

RNN-T may easily suffer from long form deletion errors

We may reduce this problem by

- Learning from a diversified unlabeled data source -> Youtube audios
- Distilled from more powerful teachers
    - Non-streaming models suffer less from deletion errors than streaming
    - CTC suffers less than RNN-T

First Try: Distill non-streaming teacher

Given a strong non-streaming teacher

1. We gather unlabeled utterances from YouTube.
2. We segment utterances, randomly between 5 and 15 seconds.
3. We label the resulting utterances using the teacher model.
4. We train a streaming student on these semi-supervised data.

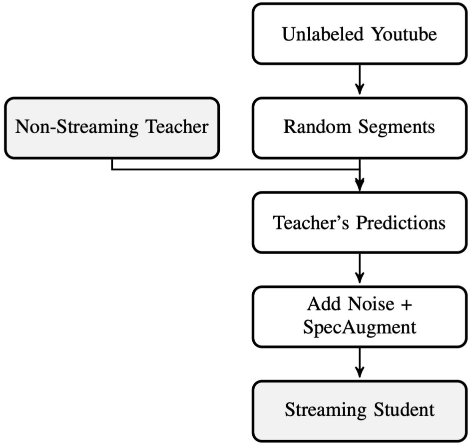


Fig. 1. Our method trains a streaming model, learning from the predictions of a powerful non-streaming teacher model on large-scale unlabeled data via a teacher-student training framework. See Sect. 2.2 for more details.

with Thibault Doutre and Wei Han et al:Improving streaming automatic speech recognition with non-streaming model distillation on unsupervised data, ICASSP 2021

First experiments on Librispeech

We first validate the method on Librispeech

- Non-streaming Conformer teacher labels LibriLight [30].
- We train a streaming Conformer model [29] on
  1. LibriSpeech only
  2. LibriSpeech + LibriLight

Table 1. WERs of different models on LibriSpeech. The streaming baseline model and the non-streaming teacher are trained on LibriSpeech 960h. The streaming student model is trained on both LibriSpeech 960h and the predictions of the non-streaming teacher on LibriLight.

	Streaming baseline [29]	Non-streaming teacher	Streaming student
test-clean	4.6	1.7	3.3
test-other	9.7	3.8	8.1

[29] Jiahui Yu, Wei Han, Anmol Gulati, Chung-Cheng Chiu, et al., "Dual-mode ASR: Unify and Improve Streaming ASR with Full-context Modeling," ICLR, 2021.

[30] Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny ` Kharitonov, et al., "Libri-light: A benchmark for ASR with limited or no supervision," in Proc. ICASSP. 2020, pp. 7669– 7673, IEEE.

First experiments on Youtube

Training data:

- **YT-segments**: unsupervised segments from YouTube
- **Confisland**: YouTube data aligned user-uploaded transcripts [13]

	Confisland	*YT-segments*
Spanish	13,000	41,000
French	10,000	29,000
Portugese	2,500	5,000

Test data:

- **YT-long**: long utterances from had-transcribed YouTube videos

[13] Hank Liao, Erik McDermott, and Andrew Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, 2013, pp. 368–373.

Teacher-student learning:

1. Randomly segment YouTube data into utterances of 5s - 15s: **YT-segments**.
2. **Transcribe** using an ensemble of non-streaming teachers.
3. Train a **streaming student** on the pseudo labels.

Results on Youtube

Table 2. WERs of ASR models trained on *Confisland*.

	Test set	Streaming model on *Confisland*	Non-streaming teacher model on *Confisland*
French	YT-long	34.5	18.6
	Common Voice	36.2	33.2
Spanish	YT-long	35.9	18.6
	Common Voice	22.0	11.2
Portuguese	YT-long	30.8	22.8
	Common Voice	30.9	25.8
Italian	YT-long	24.0	16.2
	Common Voice	30.0	27.3

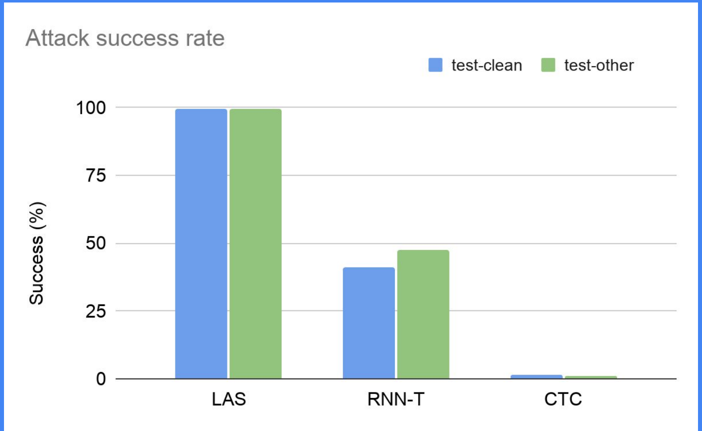
Table 4. Comparing the WERs of streaming RNN-T models trained on *Confisland* with the model from our distillation approach trained on the corresponding random segments.

	Test set	Streaming model on *Confisland*	Streaming student on *YT-segments*
French	YT-long	34.5	25.0
	Common Voice	36.2	34.7
Spanish	YT-long	35.9	28.0
	Common Voice	22.0	16.5
Portuguese	YT-long	30.8	28.3
	Common Voice	30.9	28.9
Italian	YT-long	24.0	20.8
	Common Voice	30.0	23.6

with Thibault Doutre and Wei Han et al:Improving streaming automatic speech recognition with non-streaming model distillation on unsupervised data, ICASSP 2021

Can we do better?

recall CTC is more robust than RNN-T



Expand to multiple teachers

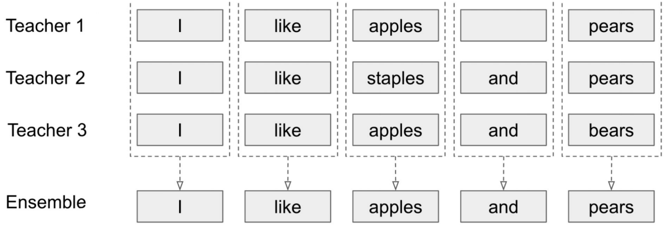
Non-streaming teacher models

We use 3 different teacher models, trained on various types of data.

	Encoder	Decoder	Data
MD-RNNT	17 blocks	1 LSTM	Multi-domain
YT-RNNT	16 blocks	1 LSTM	YouTube
YT-CTC	16 blocks	1 layer	YouTube

Teacher ensemble

Predictions of multiple teacher models are ensemble using **Recognizer Output Voting Error Reduction (ROVER)**.



With Thibault Doutre et al, Bridging the gap between streaming and non-streaming ASR systems by distilling ensembles of CTC and RNN-T models, Interspeech 2021

Results

- The teacher ensemble outperforms all teachers separately
- Student models trained from the teacher ensemble are better

Table 3: *WERs of a streaming Conformer student model trained on YT-segments, distilled from non-streaming teacher models.*

	Teacher model	Teacher WER on *YT-long*	Student WER on *YT-long*
Spanish	MD-RNNT	16.4	33.4
	YT-RNNT	18.6	23.4
	YT-CTC	20.2	16.9
	Teacher ensemble	18.1	16.4
Portuguese	MD-RNNT	29.1	31.9
	YT-RNNT	22.8	26.7
	YT-CTC	24.8	23.0
	Teacher ensemble	21.9	20.5
French	MD-RNNT	31.9	42.8
	YT-RNNT	18.8	23.6
	YT-CTC	21.0	16.6
	Teacher ensemble	20.2	16.7

CTC vs RNN-T teachers

The paradox of CTC teachers

- CTC models have a higher WER than RNN-T teachers
- CTC transcripts suffer from linguistic issues
- On long-form test sets, **RNN-T students trained on CTC models outperform their counterparts** trained on RNN-T teachers.

Key findings from ablation studies

- Using **at least 1 CTC teacher** leads to lower student WER
- **Combining** CTC and RNN-T teachers give best results
- RNN-T student models outperform their CTC teachers

Improvement over previous study

- CTC teacher may not outperform RNN-T teacher
- But the resulted student from CTC is always stronger!

Table 5: *Comparison of the WER of streaming models in this paper compared with streaming baselines [1] trained on similar data.*

	Spanish	Portuguese	French
Streaming RNN-T on *Confisland* [1]	35.9	30.8	34.5
Baseline streaming student [1]	28.0	28.3	25.0
Our streaming student	16.4	20.5	16.7
Relative improvement relative to the baseline streaming student	41%	27%	13%

Conclusion

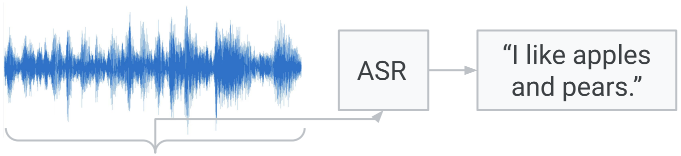
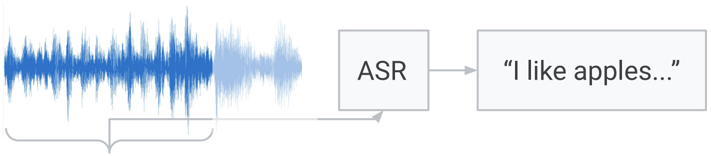
- This talk has introduced RNN-T model together with our team's 1.5+ years of efforts of reducing long-form errors.

- Model upgradation is usually not easy, but imposing interesting problems for researchers.

- Effective collaboration between researchers and engineers are important.

Backup slides

Streaming vs non-streaming ASR

	Non-streaming models	Streaming models
Context	 full context	 limited context
Considerations	• Have access to full context before processing the audio. • Performs better than streaming models. • Less user-friendly.	• Must produce words on-the-fly. • Does not have access to future context.
Use cases	• Offline transcription. • Voice queries.	• Close captions.

Teacher ensemble via ROVER method

Summary diagram

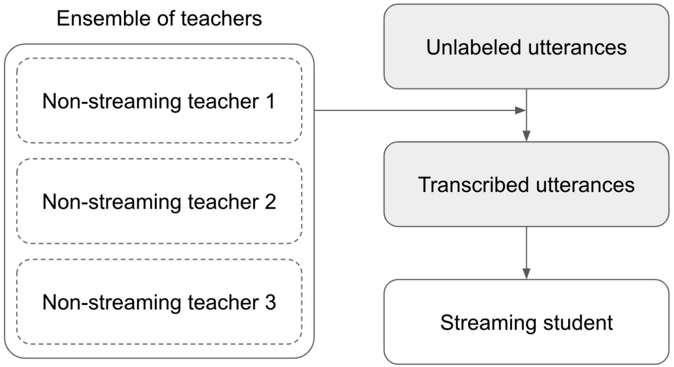


Figure 2: *Teacher-student framework. The student model is trained on an arbitrarily large set of utterances, transcribed by an ensemble of pre-trained teacher models.*

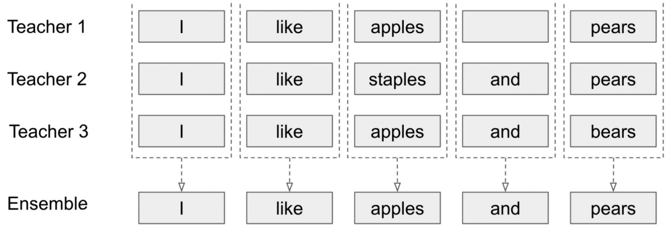
Description of the method

1. Gather **unlabeled audio** from YouTube videos.
2. **Segment** audio, randomly between 5 and 15 seconds.
3. Label the resulting utterances using an ensemble of **teacher models**.
4. Train a **streaming student** on these semi-supervised data.

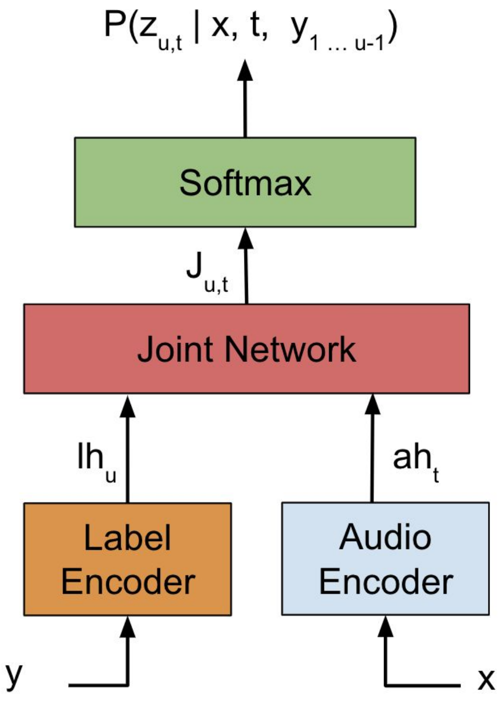
The final model must to be streaming to satisfy deployment constraints.

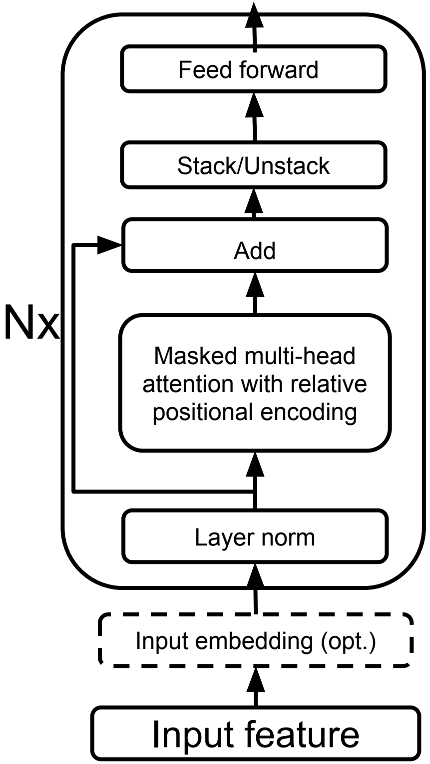
Teacher ensemble

Predictions of multiple teacher models are ensemble using **Recognizer Output Voting Error Reduction (ROVER)**.



Transformer-Transducer





Zhang et al, Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss, ICASSP 2020

Conformer: convolution-augmented transformer

Gulati et al, Conformer: Convolution-augmented Transformer for Speech Recognition, Interspeech 2020