Select On-device Spoken Language Understanding Topics

Jimmy Kunzmann
(Credits: Ariya Rastrow & Björn Hoffmeister & Daniel Willett)
Alexa Speech, Amazon
Alexa ASR Science

We do In-Cloud, On-Device and In-Car ASR for
• Human-Machine Interactions (e.g., Alexa)
• Human Speech Transcription (e.g., Voice Search)
• Human-Human-Machine Conversations (e.g., Alexa Conversations)

Where we are:

- Björn Hoffmeister
- Ariya Rastrow
- Mat Hans
- Thanas Mouchtaris
- Daniel Wittern
- Thanasis Mouchtaris
- Sri Garimella
- Thanasis Mouchtaris

Locations:
- Seattle
- Sunnyvale
- Cambridge
- Aachen
- Bangalore
- Pittsburgh
Alexa enabled Products

We build ASR for …

• Headless devices
• Multi-modal devices
• Smart remotes
• Mobile
• Auto
• Wearables
• Robots
Select On-device Spoken Language Understanding Topics

Agenda

• Birds eye view: Finite State Transducer to Neural Transducer ASR
• Dynamic Adaptation and Personalization
• E2E Speech To Understanding
• Edge Processing – Small Footprint ASR
Finite State Transducer (FST) Based ASR

HMM with Neural Acoustic Models (since 2012)

Acoustic Model

Deep Neural Network

Phonetic Decision Tree

Pronunciation Lexicon

Language Model (N-gram)

Finite State Transducer

phone posteriors

Alexa, what’s ..
Neural Transducer Based ASR

Neural Transducer -> Softmax -> Joint Network -> Audio Encoder -> Prediction Network

Speech Signal

30ms

Al__exa, what's ..
Neural Transducer Based ASR – Pros/Cons

Pros
• End-to-end optimizable
• Representation Learning
• Multi-Task Learning
• (Theoretically) Open Vocabulary
• Accuracy wins

Cons
• Not easy to train
• Expensive to train (4-5 weeks on 96 GPUs)
• Rare words are challenging
• Personalization is challenging
• Hotfixing is challenging

H. Tulsiani et al., “Improved training strategies for end-to-end speech recognition in digital voice assistants”, Interspeech 2020
E. Lakomkin et al., “Subword regularization: an analysis of scalability and generalization for end-to-end automatic speech recognition”, Interspeech 2022
Dynamic Adaptation and Personalization

- Difficulty Recognizing uncommon/rare words & phrases (All neural models thrive from data)
  
  *When is movie “X” coming to the theatres?*
  *Call “Y” on his/her cellphone.*
  *Play my “Z” playlist from Spotify.*

- Boost personalized entities and catalogs (ContactNames, PlayList, etc.)

- Domain adaptation
  - Usage shifts overtime
  - Need to support new domains and use cases (cold-start problem) (text-only adaptation)
Dynamic Adaptation and Personalization

- Shallow Fusion
- Attention-based Neural Biasing
- Core RNN-T (1ˢᵗ pass)
- Lattice Encoder
- Generate Lattice
- Neural LM Rescoring (2ⁿᵈ pass)

Audio Encoder

N-best Hypotheses
- who is brad pitt 6.8 ✓
- who is brand bid 24.6 ✗
- who is frank did 43.8 ✗
- ...... ✗
Dynamic Adaptation and Personalization

Attention-based Neural Biasing

Neural Transducer

Softmax

Joint Network

Audio Encoder

Prediction Network

On top of audio, prediction per entity

alexa turn on lily's room
Dynamic Adaptation and Personalization

Attention-based Neural Biasing

Neural Transducer

Softmax

Joint Network

Audio Encoder

Prediction Network

Encoded representation per entity
Dynamic Adaptation and Personalization

Attention-based Neural Biasing

- Neural Transducer
- Softmax
- Joint Network
- Audio Encoder
- Prediction Network

Encoded representation per entity
Dynamic Adaptation and Personalization

Attention-based Neural Biasing

Neural Transducer

Softmax

Joint Network

Audio Encoder

Prediction Network

Encoded representation per entity
Dynamic Adaptation and Personalization

Attention-based Neural Biasing

- Neural Transducer
- Softmax
- Joint Network
- Audio Encoder
- Prediction Network
- alexa

Encoded representation per entity
Dynamic Adaptation and Personalization

Attention-based Neural Biasing

Neural Transducer

Softmax

Joint Network

Audio Encoder

Prediction Network

Alexa turn on lily's room

basement light
kitchen tv
lily's room
ceiling fan
ben's room

Personalized Device Names
Dynamic Adaptation and Personalization

Attention-based Neural Biasing

Pre-trained (frozen)

Neural Transducer

Audio Encoder

Context Encoder

Prediction Network

Joint Network

Softmax

attention weights

alexa

basement light
kitchen tv
lily's room
ceiling fan
ben's room
Dynamic Adaptation and Personalization

Attention-based Neural Biasing

Pre-trained (frozen)

Neural Transducer

Softmax

Joint Network

Attention Module

Audio Encoder

Context Encoder

Prediction Network

alexa turn on lily's room

basement light

kitchen tv

lily's room

ceiling fan

ben's room
Dynamic Adaptation and Personalization

Attention-based Neural Biasing

- Pre-trained (frozen)
- Neural Transducer
- Softmax
- Joint Network
- Attention Module
- Audio Encoder
- Context Encoder
- Prediction Network

- 40% WER Reduction on proper names
**E2E Speech To Understanding**

Conventional Spoken Language Understanding (SLU) System

Drawbacks of a Modular SLU System with Independent ASR & NLU Models

- **Independent Training**
- **Training Errors Are Propagated**
- **Each Error Treated Equally**

- ❌ "turn on the light"
- ✅ "turn on the light"
- ✗ "turn on the light"
E2E Speech To Understanding

Tighter integration for

- Produce an SLU output directly from the speech signal input
- Either trained with a single optimization objective or jointly optimized end-to-end
- “Error-Robust” as well as “Resource Efficient”
E2E Speech To Understanding

E2E Speech To Understanding

RNN-T (ASR) → Neural-Interface → NLU

Backprop NLU loss & improve ASR

Single Stage Streamable SLU
ASR + NLU
Beam Search on (wp, slots)

<table>
<thead>
<tr>
<th>Model</th>
<th>Loss Type</th>
<th>WERR</th>
<th>SemERR</th>
<th>IRERR</th>
<th>ICERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-stage SLU</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Multi-task Semantic RNN-T</td>
<td>$L_{\text{rnnt}}(wp) + L_{\text{ce}}(slot) + L_{\text{ce}}(slot)$</td>
<td>1.41</td>
<td>9.49</td>
<td>14.38</td>
<td>5.13</td>
</tr>
<tr>
<td></td>
<td>$L_{\text{rnnt}}(wp) + L_{\text{rnnt,align}}(slot) + L_{\text{ce}}(slot)$</td>
<td>-0.99</td>
<td>7.43</td>
<td>12.04</td>
<td>-1.26</td>
</tr>
</tbody>
</table>

X. Fu, F. Chang, M. Radfar, K. Wei, J. Liu, G. Strimel, K. M. Sathyendra, “Multitask RNN-T with Semantic Decoder for Streamable Spoken Language Understanding,” ICASSP 2022
E2E SLU - Dialog Context Carry-Over

Transformer-based SLU w/ Context Carry-Over
• BERT embedding for transcription
• Multi-Head Attention with Gating for combining context
• Industrial Voice Assistant (IVA) Data Set

<table>
<thead>
<tr>
<th></th>
<th>WERR</th>
<th>ICERR</th>
<th>SemERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2E T-T SLU</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>+ dialog act</td>
<td>5.4%</td>
<td>4.6%</td>
<td>1.5%</td>
</tr>
<tr>
<td>+ prev. utterance</td>
<td>12.4%</td>
<td>8.9%</td>
<td>6.3%</td>
</tr>
<tr>
<td>+ both</td>
<td>13.8%</td>
<td>11.1%</td>
<td>10.6%</td>
</tr>
</tbody>
</table>

K. Wei et al., “Attentive contextual carryover for multi-turn end-to-end spoken language understanding”, ASRU 2021
Edge Processing – Small Footprint ASR & SLU

Legacy factored HMM

- N-grams are memory inefficient
- Sub-optimal accuracy-vs-footprint curve (disjoint models)

End-to-end all-neural

- Far better accuracy-vs-footprint curve
- Uniform application of compression, quantization and sparsification methods
  - 8-bit (and even 5-bit) quantization-aware training
- Architecture variation and choices
  - LSTM -> LSTM-P
Quantize-Aware Training via Regularization
Achieve 8-bit (and sub 8-bit)

Best weights = \min [\mathcal{L}(weights) + \mathcal{L}_{\text{quantization}}(weights)]

\[ ACosR(x) = -\alpha |\cos(x)| \]

Hieu Nguyen et al., "Quantization aware training with absolute-cosine regularization for automatic speech recognition,” Interspeech 2020
Hey Alexa, turn on the light.
Edge Processing – Small Footprint ASR

Neural Transducer (RNN-T) → Generate N-best list/Lattice → Neural Rescorer (text + lattice) → Generate N-best list/Lattice → Neural LM Rescorer (text)

On-device (using Neural Edge processor)

In-Cloud (Unified Rescoring Pipeline with Cloud-based Devices)
Conclusions

• What we have briefly touched
  • Dynamic Adaptation and Personalization
  • Attention-based Neural Biasing
  • E2E Speech To Understanding
    • Backpropagate NLU loss & improve ASR
    • Semantic decoder & fusion network
    • Dialog Context Carry-Over
  • Small Footprint ASR
    • Quantization aware training
    • Bi-focal RNN-T

• What we haven’t covered
  • Representation Learning
  • Multi-Lingual Modeling
  • Multi-Speaker Modeling
  • Multi-Modal Modeling
  • Closed-loop self-learning, Semi-/weakly-supervised learning
  • Life-long learning
  • Learning on device
  • …
It is still Day One!

A good time to be a speech researcher!

Thank you!

Jimmy Kunzmann
kunzman@amazon.com

Ariya Rastrow
arastrow@amazon.com

Bjorn Hoffmeister
bjornh@amazon.com

Daniel Willett
dawillet@amazon.de