Searchable Hidden Intermediates for End-to-End Models of Decomposable Sequence Tasks

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, Shinji Watanabe
What is Compositionality?

- Compositionality is the principle behind building complex systems by composing together simpler sub-systems.
What is Compositionality?

• Compositionality is the principle behind building complex systems by composing together simpler sub-systems.
Traditional Cascaded Models

- Traditional Cascaded Models exploited the task compositionality to give many interesting properties that facilitate practicality of these models.

  1. The **strong search capabilities** to compose the final task output from individual system predictions.

  2. The ability to **incorporate external models to re-score** each individual system.

  3. The ability to easily **adapt individual components** towards out-of-domain data

  4. The ability to **monitor performance of the individual systems** towards the decomposed sub-task.
Can we bring these properties into End-to-End Models?
Searchable Hidden Intermediates Framework

• General end-to-end framework to exploit natural decomposition in sequence tasks.

• A sequence task, $A \rightarrow C$ is decomposable, if there is an intermediate sequence $B$ for which $A \rightarrow B$ sequence transduction followed by $B \rightarrow C$ prediction achieves the original task.

  • For instance, Speech Translation using ASR intermediates

• Learn $P(C \mid A)$ through $\max_B (P(C \mid A, B)P(B \mid A))$, approximated using Viterbi search.
Multi-Decoder Model with Searchable Intermediates

(Completed Work)
Multi-Decoder Model with Searchable Intermediates

(Completed Work)

Pass Decoder Hidden Representations:
- ASR Sub-Net maps input to sequence of decoder hidden representations $\mathbf{h}^{D_B}$
- MT Sub-Net maps $\mathbf{h}^{D_B}$ to final ST output
- During inference, approximate $\mathbf{h}^{D_B}$ with $\mathbf{h}_\text{Beam}^{D_B}$
Comparison with Encoder-Decoder

Outperforms both Encoder-Decoder and Cascaded Models -

- +7 BLEU and +3 BLEU on Fisher and CallHome (Es-En)
- +3 BLEU and +4 BLEU on MuST-C En-De and En-Fr

Higher (↑) is better
Performance Monitoring:

ASR Sub-Net: 22.4 WER

MT Sub-Net: 66.6 BLEU

ST: 54.6 BLEU
Retrieval with Beam Search

As ASR quality improves with larger beam, overall ST performance goes up!

Search and Retrieval:
Our framework can use beam search at ASR intermediates to improve the overall ST performance.
Retrieval with External Models

Search and Retrieval:
Our framework has the ability to retrieve better hidden intermediates by -
• Re-scoring using external models at intermediate stages of the network during inference.
• On Fisher Es-En improves by +1 BLEU using CTC and LM re-scoring.
Adapting Individual Components

Search and Retrieval:
Our framework has the ability to adapt individual components of the E2E model towards out-of-domain data.

- We can re-score ASR sub-net with in-domain LM.
- Improves ASR by 10% lower WER, improving the overall ST by +2.4 BLEU

| Model                          | Overall ST(↑) | Sub-Net ASR(↓) |
|-------------------------------|--------------|----------------|
| **IN-DOMAIN ST MODEL**        |              |                |
| Baseline (Wang et al., 2020b) | 12.0         | -              |
| **OUT-OF-DOMAIN ST MODEL**    |              |                |
| Multi-Decoder                 | 12.6         | 46.5           |
| +ASR Re-scoring w/ in-domain LM | **15.0**     | **36.7**       |
Decomposing Speech Transcripts

Lower (↓) is better

|                      | Fisher ASR | CallHome ASR |
|----------------------|------------|--------------|
| Baseline Encoder-Decoder | 23.2       | 45.3         |
| Multi-Decoder (Phoneme) | 20.7       | 40.0         |
| Multi-Decoder (Character) | 20.4       | 39.9         |
| Multi-Decoder (BPE100)   | 19.7       | 38.9         |
Thank you