Review of Paper: End-to-End Spoken Language Understanding Using Joint CTC Loss and Self-Supervised, Pretrained Acoustic Encoders

April 12, 2025

1 Summary of the Paper

The paper discusses improvements in Spoken Language Understanding (SLU) models using self-supervised acoustic encoders fine-tuned with Connectionist Temporal Classification (CTC). Instead of using traditional auto-regressive decoding, which is slower than other methods in this category, the authors put forward a joint training scheme using both CTC and SLU losses. Their experimental results showed significant performance enhancements, achieving a 4% absolute improvement in dialogue act classification on the DSTC2 dataset and a 1.3% accuracy improvement on the SLURP dataset.

2 Main Architecture

The main architecture of the proposed method is shown in Figure 1.

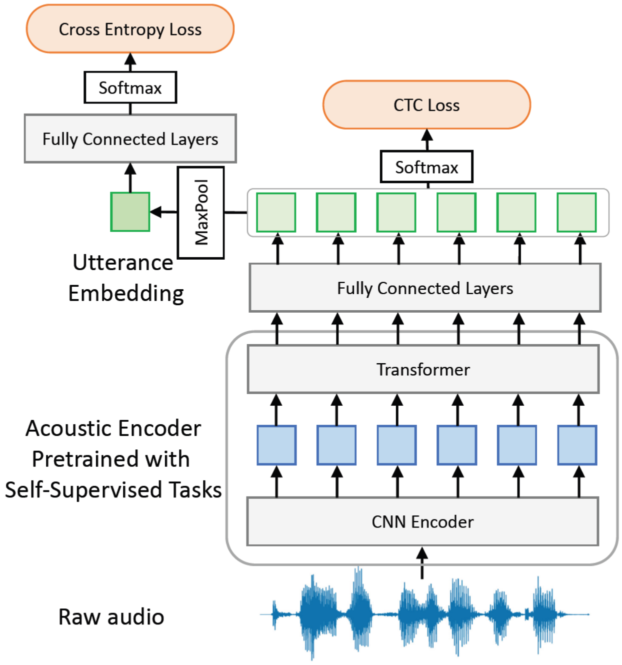

Figure 1: Main architecture of the proposed approach (source: original paper).

3 Technical Strengths

- Joint training of CTC and SLU losses highly enhances both efficiency and accuracy.
- Effective utilization of self-supervised acoustic encoders (Wav2Vec2.0, HuBERT), resulting in state-of-the-art performance.
- Showed that using logits rather than probabilities as input highly improves robustness in SLU tasks.
- Largely outperforms previous architectures on standard benchmarks such as DSTC2, Speech Commands, and SLURP.

4 Technical Weaknesses

- The approach is currently aimed solely at utterance-level classification and does not include sequence-labeling tasks, e.g., slot filling.
- Dependence on highly pre-trained acoustic models, which could restrict adaptability to other languages or specialized domains without extensive fine-tuning.

5 Minor Questions/Weaknesses

- How effectively does the model perform under severe real-life noise conditions?
- Can the method maintain high accuracy when processing longer utterances?

6 Suggestions as a Reviewer

To increase practical applicability, the authors should extend their approach to include sequence-labeling tasks like slot filling. Testing the model performance under different noisy and acoustically diverse environments will help validate how good it is. Additionally, testing scalability regarding computational efficiency across various dataset sizes will strengthen efficiency claims. Experimenting with less extensively pre-trained acoustic models may also demonstrate the method's accessibility and general applicability.

7 Rating and Justification

Rating: 8/10

The paper presents a very effective and state-of-the-art method, clearly shown by experimental results. However, extending its scope to sequence-labeling tasks and evaluating its performance in noisy, real-world conditions would highly enhance its practical value.