Multitask Learning with Low-Level Auxiliary Tasks for Encoder-Decoder Based Speech Recognition

Shubham Toshniwal, Hao Tang, Liang Lu, and Karen Livescu

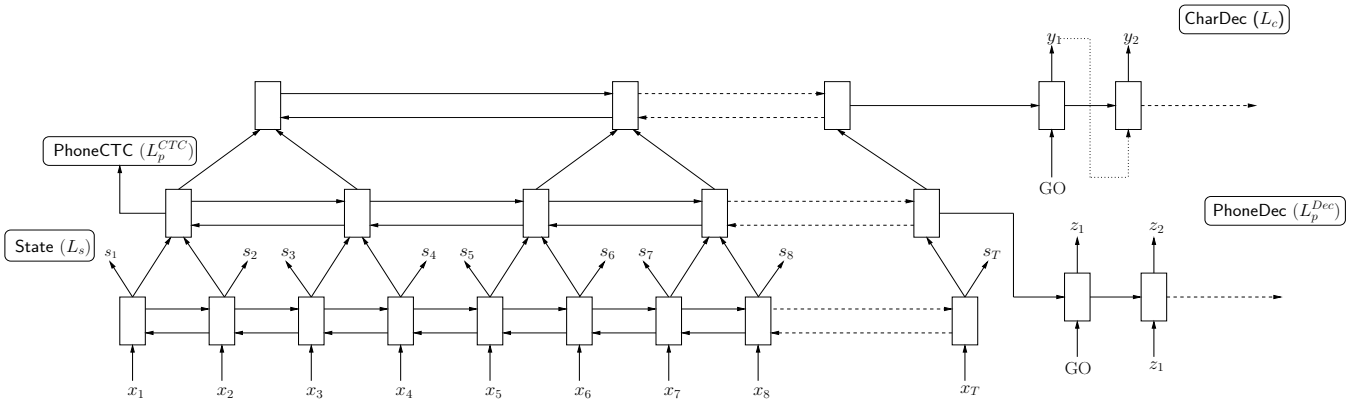

Figure 1: Sketch of our training-time model with multiple losses applied at different layers. Different hidden state layers of this encoder are used for predicting HMM state label s_i, phone sequence $\boldsymbol{z}$, and character sequence $\boldsymbol{y}$ via a LSTM decoder. At test time, only the character decoder is used for transcription.

Introduction

Automatic speech recognition (ASR) has historically been addressed with modular approaches, in which multiple parts of the system are trained separately. This breaking up into modules makes the training simpler. Recently, end-to-end training approaches for ASR have become viable and popular. End-to-end training is conceptually simple and allows all parameters to be learned based on final task loss. However, such an approach ignores the domain knowledge available as explicit intermediate-level supervision. We propose a multitasking approach for deep neural ASR that aims to maintain the advantages of end-to-end approaches, while utilizing the intermediate supervision through auxiliary task losses. We find that applying an auxiliary loss at an appropriate intermediate layer is key to getting performance gains [1].

Models

Our baseline model is based on attention-enabled encoder-decoder RNNs. The *speech encoder* is a 4-layer deep pyramidal bidirectional LSTM that reads in acoustic features $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ and outputs a sequence of high-level features (hidden states) $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{\lceil T/8 \rceil})$ which the *character decoder* attends to while generating the output character sequence $\boldsymbol{y} = (y_1, \ldots, y_K)$, as shown in Figure 1.

We explore two types of auxiliary labels for multitask learning: phonemes $\boldsymbol{z}$ and sub-phonetic states $\boldsymbol{s}$ which are applied to lower layer of encoder, as shown in Figure 1.

[1] A full paper on this work is currently under review in Interspeech 2017

Experiments

We use the Switchboard corpus, which contains roughly 300 hours of conversational telephone speech, as our training set and Eval2000, consisting of two subsets Switchboard (SWB) and CallHome (CHE), as our test set.

Table 1 presents our results on this test set. Our baseline model has better performance than the most similar previous encoder-decoder result (Lu, Zhang, and Renals 2016), which we further improve via the addition of low-level auxiliary tasks. It should be noted that adding these tasks at lower level is key to getting these gains.

Table 1: Word Error Rate (WER, %) on Eval2000 for different encoder-decoder models. We refer to the baseline model as "Enc-Dec" and the models with multitask training as "Enc-Dec + [auxiliary task]-[layer]".

Model	SWB	CHE	Full
Our models			
Enc-Dec (baseline)	25.0	42.4	33.7
Enc-Dec + PhoneDec-3	24.5	40.6	32.6
Enc-Dec + PhoneDec-4	25.4	41.9	33.7
Enc-Dec + PhoneCTC-3	24.6	41.3	33.0
Enc-Dec + PhoneDec-3 + State-2	**23.1**	**40.8**	**32.0**
Lu et al.(Lu, Zhang, and Renals 2016)			
Enc-Dec	27.3	48.2	37.8
Enc-Dec (word) + 3-gram	25.8	46.0	36.0

References

Lu, L.; Zhang, X.; and Renals, S. 2016. On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition. In *ICASSP*.