Vol.2018-SLP-125 No.3
2018/12/10

Feature Transfer Learning for Wav2Text Sequence-to-Sequence ASR

Andros Tjandra[1,2,a)] Sakriani Sakti[1,2,b)] Satoshi Nakamura[1,2,c)]

Abstract: In this paper, we construct the first end-to-end attention-based encoder-decoder model to process directly from raw speech waveform to the text transcription. We called the model as "Attention-based Wav2Text". To assist the training process of the end-to-end model, we propose to utilize a feature transfer learning. Experimental results also reveal that the proposed Attention-based Wav2Text model directly with raw waveform could achieve a better result in comparison with the attentional encoder-decoder model trained on standard front-end filterbank features.

1. Introduction

Large-vocabulary continuous speech recognition (LVCSR) systems typically perform multi-level pattern recognition tasks that map the acoustic speech waveform into a hierarchy of speech units such as sub-words (phonemes), words, and strings of words (sentences). Such systems basically consist of several sub-components (feature extractor, acoustic model, pronunciation lexicon, language model) that are trained and tuned separately [1]. First, the speech signal is processed into a set of observation features based on a carefully hand-crafted feature extractor, such as Mel frequency cepstral coefficients (MFCC) or Mel-scale spectrogram. Then the acoustic model classifies the observation features into sub-unit or phoneme classes. Finally, the search algorithm finds the most probable word sequence based on the evidence of the acoustic model, the lexicon, and the language model. But, it is widely known that information loss in the earlier stage can propagate through the later stages. In this paper, we take a step forward to construct an end-to-end ASR using an attentional-based encoder-decoder model for processing raw speech waveform, naming it as "Attention-based Wav2Text". We investigate the performance of our proposed models on standard ASR datasets. However, optimizing an encoder-decoder framework is more difficult than a standard neural network architecture [2]. Therefore, we propose a feature transfer learning method to assist the training process for our attention-based ASR model.

2. Feature Transfer Learning on Encoder-Decoder ASR

Deep learning is well known for its ability to learn directly from low-level feature representation such as raw speech. However, in most cases such models are already conditioned on a fixed input size and a single target output (i.e., predicting one phoneme class for each input frame). In the attention-based encoder-decoder model, the training process is not as easy as in a standard neural network model [2] because the attention-based model needs to jointly optimize three different modules simultaneously: (1) an encoder module for producing representative information from a source sequence; (2) an attention module for calculating the correct alignment; and (3) a decoder module for generating correct transcriptions. If one of these modules has difficulty fulfilling its own tasks, then the model will fail to produce good results.

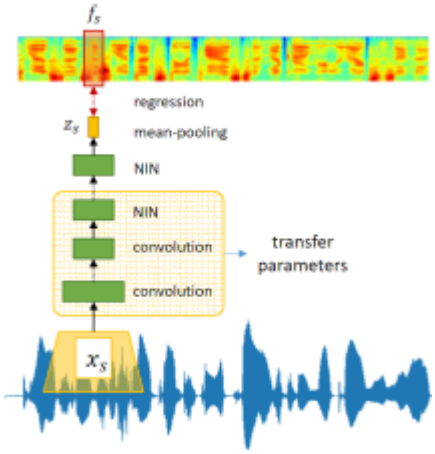

Fig. 1 Feature transfer learning: train lower layers of the encoder (convolutional and NIN layers) to predict spectral features given corresponding raw waveform; then transfer the trained layers and parameters (marked by orange square) into attention-based encoder decoder model.

To ease the burden on training the whole encoder-decoder architecture directly to predict the text transcription given the raw speech waveform, we utilize a transfer learning method on the encoder part. Specifically, we only train the encoder's lower layers consisting of the convolutional and NIN [3] layers to predict the spectral features given the corresponding raw waveform. In this work, we utilize two widely used spectral

[1] Nara Institute of Science and Technology, Japan
[2] RIKEN, Center for Advanced Intelligence Project AIP, Japan
[a)] andros.tjandra.ai6@is.naist.jp
[b)] ssakti@is.naist.jp
[c)] s-nakamura@is.naist.jp

features: MFCC and log Mel-scale spectrogram as the transfer learning target. Figure 1 shows our feature transfer learning architecture. First, given segmented raw speech waveform $\mathbf{x} = [x_1, ..., x_S]$, we extract corresponding D-dimensional spectral features $\mathbf{f} = [f_1, .., f_S]$, $\forall s, f_s \in \mathbb{R}^D$. Then we process raw speech x_s with several convolutions, followed by NIN layers in the encoder part. In the last NIN-layer, we set a fixed number of channels as D channels and apply mean-pooling across time. Finally, we get predictions for spectral features $z_s \in \mathbb{R}^D$ and optimize all of the parameters by minimizing the mean squared error between predicted $\mathbf{z}$ and target spectral features $\mathbf{f}$:

$$\mathcal{L}_{tf} = \frac{1}{S}\sum_{s=1}^{S}\sum_{d=1}^{D}(f_s(d) - z_s(d))^2.$$

In this paper, we also explore multi target feature transfer using a similar structure as in Figure 1 but with two parallel NIN layers, followed by mean-polling at the end. One of the output layers is used to predicts log Mel-scale spectrogram and another predicts MFCC features. We modify the single target loss function from Eq. 1 into the following:

$$\mathcal{L}_{tf} = \frac{1}{S}\sum_{s=1}^{S}\left(\sum_{d=1}^{D_a}(f_s^a(d) - z_s^a(d))^2 + \sum_{d=1}^{D_b}(f_s^b(d) - z_s^b(d))^2\right).$$

where z_s^a, z_s^b are the predicted Mel-scale spectrogram and the MFCC values, and f_s^a, f_s^b are the real Mel-scale spectrogram and MFCC features for frame s. After optimizing all the convolutional and NIN layer parameters, we transfer the trained layers and parameters and integrate them with the Bi-LSTM encoder. Finally, we jointly optimize the whole structure together.

3. Experimental Setup and Results

3.1 Speech Data

In this study, we investigate the performance of our proposed models on WSJ. We follow the training, development and test set as the Kaldi s5 recipe. The raw speech waveforms were segmented into multiple frames with a 25ms window size and a 10ms step size. We normalized the raw speech waveform into the range -1 to 1. For spectral based features such as MFCC and log Mel-spectrogram, we normalized the features for each dimension into zero mean and unit variance. Our training set is WSJ-SI284. We used dev_93 for our validation set and eval_92 for our test set. We used the character sequence as our decoder target where the text from all the utterances was mapped into a 32-character set: 26 (a-z) alphabet, apostrophe, period, dash, space, noise, and "eos".

3.2 Model Architectures

Our attention-based Wav2Text architecture uses four convolutional layers, followed by two NIN layers at the lower part of the encoder module. For all the convolutional layers, we used a leaky rectifier unit (LReLU) activation. Inside the first NIN layers, we stacked three consecutive filters with LReLU activation function. For the second NIN layers, we stacked two consecutive filters with tanh and identity activation function. In details, our convolution layers settings: `Conv(ch=128, k=80, s=4)▷Conv(ch=128, k=25, s=2)▷Conv(ch=128, k=10, s=1)▷Conv(ch=128, k=5, s=1)▷NIN(ch=[128,128])`.

On the top layers of the encoder after the transferred convolutional and NIN layers, we put three bidirectional LSTMs (Bi-LSTM) with 256 hidden units. On the decoder side, we use 128-dimensional for character embedding, followed by an LSTM with 512 hidden units and softmax layer. For the end-to-end training phase, we froze the parameter values from the transferred layers from epoch 0 to epoch 10, and after epoch 10 we jointly optimized all the parameters together until the end of training (a total 40 epochs). For comparison, we also evaluated the standard attention-based encoder decoder with Mel-scale spectrogram input as the baseline.

3.3 Result

Table 1 Character error rate (CER) result from baseline and proposed models on WSJ1 dataset. Word error rate (WER) for Att Wav2Text + transfer multi-target is 17.04%.

Models	Features	Results
Baseline		
Att Enc-Dec (ours)	fbank	7.69%
Proposed		
Att Wav2Text (direct)	raw speech	(not converged)
Att Wav2Text (transfer from fbank)	raw speech	6.78 %
Att Wav2Text (transfer from MFCC)	raw speech	6.58%
Att Wav2Text (transfer from multi target)	raw speech	6.54%

Our proposed Wav2Text models without any transfer learning failed to converge. In contrast, with transfer learning, they significantly surpassed the performance encoder-decoder from Mel-scale spectrogram features. This suggests that by using transfer learning for initializing the lower part of the encoder parameters, our model also performed better then their original features.

4. Conclusion

We described the first attempt to build an end-to-end attention-based encoder-decoder speech recognition that directly predicts the text transcription given raw speech input. We also proposed feature transfer learning to assist the encoder-decoder model training process and presented a novel architecture that combined convolutional, NIN and Bi-LSTM layers into a single encoder part for raw speech recognition. Our results suggest that transfer learning is a very helpful method for constructing an end-to-end system from such low-level features as raw speech signals. With transferred parameters, our proposed attention-based Wav2Text models converged and matched the performance with the attention-based model trained on spectral-based features. The best performance was achieved by Wav2Text models with transfer learning from multi target scheme.

5. Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

References

[1] Gales, Mark and Young, Steve: *The application of hidden Markov models in speech recognition*, Foundations and Trends in Signal Processing (2008).

[2] Chan, William and Jaitly, Navdeep and Le, Quoc and Vinyals, Oriol: *Listen, attend and spell: A neural network for large vocabulary conversational speech recognition*, IEEE ICASSP (2016).

[3] Lin, Min and Chen, Qiang and Yan, Shuicheng: *Listen, attend and spell: A neural network for large vocabulary conversational speech recognition*, arXiv preprint arXiv:1312.4400 (2013).