Improving Myanmar Automatic Speech Recognition with End-to-End Technique

Hay Mar Soe Naing , Win Pa Pa
Natural Language Processing Lab.
University of Computer Studies Yangon, Myanmar
haymarsoenaing@ucsy.edu.mm, winpapa@ucsy.edu.mm

***Abstract* — Sequence-to-sequence models have been widely used in the end-to-end speech and natural language processing community. This paper explores an emerging sequence-to-sequence model called Transformer in Myanmar automatic speech recognition. We compared and analyzed Transformer-based end-to-end deep learning models and traditional hybrid approaches. Experiments are conducted on a corpus of nearly 82 hours of broadcast News and recorded conversational speech. According to the experiments, proposed transformer-based model provides the superior results than traditional approaches. When using the transformer model, the best performance in the News domain was a word error rate (WER) of 12.7% and a character error rate (CER) of 8.3%. Furthermore, the best WER of 9.6% and CER of 7.3% were achieved in recorded conversational speech.**

Keywords— automatic speech recognition, DNN-HMM, end-to-end, LSTM-TDNN, transformer

I. INTRODUCTION

Automatic Speech Recognition (ASR) is the processing of human speech into readable text using machine learning or artificial intelligence (AI) technology. The field had more and more rapid growth over the past decade. ASR has made great strides in mainstream languages such as English, Chinese and Japanese. However, other minority languages such as Myanmar still need to promote the development of new models and methodologies due to low attention and limited open-source corpora. Last few years, Myanmar ASR systems have taken traditional hybrid approach. It consists of three main components - pronunciation dictionary, acoustic and language models. Each component trained independently and faced a complex pipeline. Thus, the construction and fine-tuning of these components is difficult.

In recent, end-to-end deep learning approaches are state-of-the-art in speech recognition and synthesis community, leveraging deep neural network capabilities. This end-to-end model replaces the traditional pipeline approach with a single neural network architecture. Thus, end-to-end deep learning models are easier to train and require less human effort than traditional hybrid methods. It is also more accurate than the traditional models currently in used. In this paper, the experiments are carried out with transformer based end-to-end ASR model on Myanmar to improve the performance of traditional hybrid system.

II. TRANSFORMER BASED END-TO-END ASR

An end-to-end system allows the direct mapping of input acoustic feature sequences into word sequences. The data does not necessitate to be force-alignment. Depending on the usage architecture, deep learning systems can be built to produce the accurate transcriptions without using lexicon or language model, although language models can benefit to produce more accurate word sequences.

A. Encoder and Decoder Networks

The ASR transformer maps input frame level acoustic features sequence $X = \{x_1, x_2 \dots, x_T\}$ to the intermediate high-level representation sequence $H = \{h_1, h_2 \dots, h_T\}$ by the encoder. The decoder generates the sequence of word level tokens or sub-word unit $Y = \{y_1, y_2 \dots, y_T\}$ given the intermediate representations. Generally, it relies entirely on the attention and feed forward networks [1]. The ASR transformer architecture is illustrated in Figure 1.

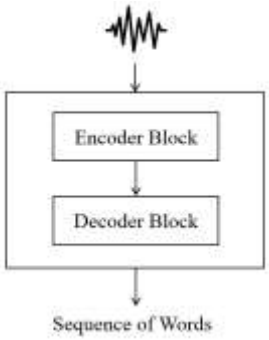

Fig. 1. The Architecture of Transformer-based ASR Model

An encoder block that stacks multiple encoders. All encoders have the same structure but different weights. Each encoder is divided into two sub-layers: self-attention and feed forward neural network. The encoder receives the acoustic feature vectors from speech signal. First encoder converts these data into a set of vectors using self-attention. The outputs of the self-attention layer are fed to a feed forward network and transmits the outputs to the next encoder. The final encoder processes these vectors and hand-overs the data of encoding function to the decoder block. The decoder also has these two layers, but there is an attention layer between them, which helps the decoder focus on the important parts of the input sentence, similar to the usual attention mechanism in seq2seq models.

B. Self-Attention Mechanism

The scaled dot-product attention is commonly used in the self-attention mechanism of transformer model. It has fast computation and shortened paths between words, as well as potential interpretability. This attention consists of three elements: queries (*Q*), keys (*K*) of dimension d_{att} and Values (*V*). The scale dot-product attention is computed as the following equation (1).

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_{att}}}\right) V \qquad (1)$$

Multi-head attention combines several self-attention maps into a general matrix computation. This mechanism can be

utilized as an optimization problem. It can be used to bypass the problem associated with the initialization failures and improve the training speed [2].

$$MHA(Q,K,V) = concat(s_1, \dots, s_h)\, W \quad (2)$$

$$s_h = attention\left(QW_h^Q, KW_h^K, VW_h^V\right) \quad (3)$$

where W is the trained weight matric and h is the number of attention heads.

III. Experimental Setup

The proposed transformer-based speech recognizer was conducted with the open-source, ESPnet toolkit [4]. All experiments were performed on an 82-hour speech corpus, which contains web news and recorded conversational speech from UCSY. This corpus has approximately 81 hours for training, and 1 hour for testing. Totally, 554 speakers are participated in training corpus. The system performance is evaluated on two sets – broadcast news (41 minutes) and recorded conversational speech domain (30 minutes). In the News domain testing set, there are 7 female and 8 male speakers. In the conversational speech testing set, 6 female and 4 male speakers are included. These speakers are not included in training process. The detail statistics of speech corpus in described in Table 1. For speech acoustic features, an 80-dimensional log-Mel filterbank with 3-dimensional pitch features is extracted on every 10 ms frame. Most of the hyper parameters of transformer models observed the default settings of ESPnet toolkit, especially the WSJ recipe.

TABLE I. Detail Description of Speech Corpus

	Domain	***No. of Speakers***	***Duration***
Training	News	261	~ 25 hours
	Conversation	293	~ 57 hours
Eval-News	News	15	41 min 48 sec
Eval-Convs	Conversation	10	30 min 10 sec

For hybrid traditional DNN-HMM and LSTM-TDNN methods, Kaldi toolkit [3] is used for baseline experiment. The SRILM toolkit is applied to build language model. This experiment used a lexicon based on the Myanmar Language Commission (MLC) which include 110 phonemes in training corpus. Default configurations provided by Kaldi; the recipe of WSJ is utilized in this experiment.

In the Transformer architecture, the encoder uses 12 self-attention blocks stacked on 2D convolutional blocks, while the decoder uses 6 self-attention blocks. We employed a 2048-dimensional feed forward network for each transformer layer. Four attention heads with 256 dimensions are used for multi-head attention. The Adam optimizer and learning rate of 0.002 is exploited without early stopping. We also adopted a warmup step of 20,000, gradient clipping, and cumulative gradients for regularization. For decoding, a beam search algorithm with a beam size of 20 and CTC weight of 0.3 was used, respectively. This experiment does not use an external language model. Since Myanmar is a syllabic language, it is expected that syllable level tokens will be used. Thus, this transformer-based model used the byte pair encoding (BPE) to generate the sub-word units. This study examined the usage of sub-word units suitable for Myanmar language. Since there are nearly 2,500 syllables in the transcription corpus, the experiments were performed using 2,500 sub-word units as the BPE vocabulary size.

IV. Experimental Results

In this paper, we analyzed two traditional hybrid methods DNN-HMM, LSTM-TDNN methods and transformer-based ASR recognizer. We reported both WER% and CER% as the evaluation metrics. The recognition performances are listed in Table 2 and 3.

TABLE II. Evaluaton on Broadcast News Domain

Methods	***WER (%)***	***CER (%)***
DNN-HMM	16.8	-
LSTM-TDNN	14.0	-
TRANSFORMER	**12.7**	**8.3**

TABLE III. Evaluation on Conversation Domain

Methods	***WER (%)***	***CER (%)***
DNN-HMM	18.2	-
LSTM-TDNN	17.6	-
TRANSFORMER	**9.6**	**7.3**

According the experimental results, it can be seen that two conventional hybrid DNN-HMM and LSTM-TDNN models obtained lower WER% than the transformer-based end-to-end method. Compared the transformer with DNN-HMM model on News domain, the WER decreases about 4.1%. While comparing with the LSTM-TDNN model, the proposed transformer model descents the WER of 1.3%. When the model is evaluated on recorded conversational test set, it shows that the transformer lessens the WER of 8.6% than the DNN-HMM and WER of 8% than the LSTM-TDNN approaches. Traditional ASR systems need several components and are challenging to fine-tune on each part. Thus, end-to-end model facilitates these components into a single pipeline for better performance.

V. Conclusion

This paper examines the transformer-based end-to-end architecture on Myanmar speech recognizer to enhance the performance. Moreover, the traditional hybrid DNN-HMM and LSTM-TDNN models are empirically compared with the transformer. This work shows that a transformer-based model can achieve the superior results in both News and conversational speech domain. In the future, we will further promote the benchmark system with the state-of-the-art techniques such as conformer, wav2vec, etc.

References

[1] M. Orken, O. Dina, A. Keylan, T. Tolganay, O. Mohamed, "A study of transformer-based end-to-end speech recognition system for Kazakh language", Scientific Reports, 12(1), pp.1-11, 2022.

[2] A. Bie, B. Venkitesh, J. Monteiro, M. A. Haidar, M. Rezagholizadeh, "Fully quantizing a simplified transformer for end-to-end speech recognition", CoRR. 2019.

[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, "The Kaldi speech recognition toolkit", In IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF). IEEE Signal Processing Society, 2011.

[4] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, "Espnet: End-to-end speech processing toolkit", arXiv preprint arXiv:1804.00015, Mar 30, 2018.