BERT-ASR

Wen-Chin Huang[1,2], *Chia-Hua Wu*[2], *Shang-Bao Luo*[2], *Kuan-Yu Chen*[3], *Hsin-Min Wang*[2], *Tomoki Toda*[1]

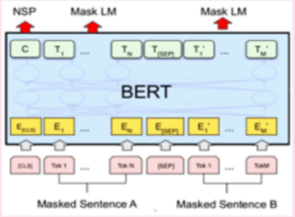
[1]Nagoya University, Japan [2]Academia Sinica, Taiwan [3]National Taiwan University of Science and Technology, Taiwan

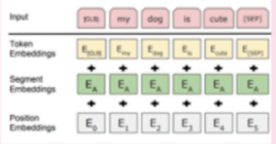
Overview

- A simple method for automatic speech recognition (ASR) by fine-tuning BERT
- **BERT-ASR**, formulates ASR as a classification problem, where the objective is to correctly classify the next word given the acoustic speech signals and the history words.

BERT

- **B**idirectional **E**ncoder **R**epresent ations from **T**ransformers
- A language model (LM) trained on large-scale unlabeled text data and can generate rich contextual representations
- Learns language by using 2 main methods: **MLM** (Masked Language Model) & **NSP** (Next Sentence Prediction)
- Adopts a multi-layer Transformer encoder architecture





BERT-LM

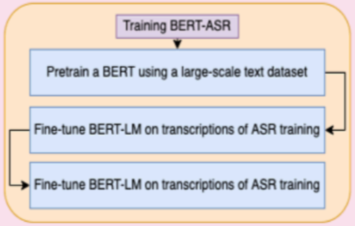
$$(y_1, ..., y_T) \rightarrow \begin{cases} ([\mathtt{CLS}]) \\ ([\mathtt{CLS}], y_1) \\ ([\mathtt{CLS}], y_1, y_2) \\ \dots \\ ([\mathtt{CLS}], y_1, \dots, y_{t-1}) \end{cases}$$

- A probabilistic **Language Model (LM)** using BERT
- Exhaustively enumerate all possible training samples
- Training becomes simply minimizing cross-entropy objective

$$\mathcal{L}_{\mathrm{LM}} = -\sum_{i=1}^{N}\sum_{t=1}^{T} P(y_t^{(i)} | [\mathtt{CLS}], y_1^{(i)}, \dots, y_{t-1}^{(i)}).$$

BERT-ASR Training and Fine-Tuning

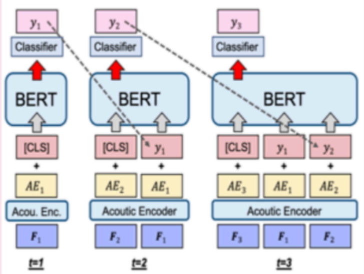
- Acoustic frames can be segmented into T groups (T is # of tokens in transcript)
- Acoustic embeddings concatenated with original BERT inputs fed into model.
- This way we augment the BERT-LM into BERT-ASR


Training BERT-ASR

Pretrain a BERT using a large-scale text dataset

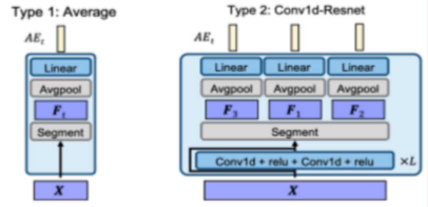
Fine-tune BERT-LM on transcriptions of ASR training

Fine-tune BERT-LM on transcriptions of ASR training

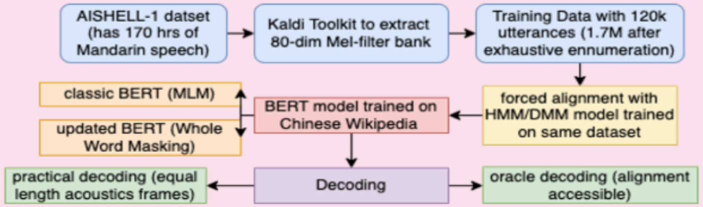


Acoustic Encoder

- Converts raw acoustic feature segments (Fi) into acoustic embeddings
- Authors experimented with the
  - ➢ **Average Encoder** and
  - ➢ **Conv1d-Resnet Encoder** (takes temporal relationship between segments into account)



Experiments


AISHELL-1 datset (has 170 hrs of Mandarin speech)

Kaldi Toolkit to extract 80-dim Mel-filter bank

Training Data with 120k utterances (1.7M after exhaustive ennumeration)

forced alignment with HMM/DMM model trained on same dataset

BERT model trained on Chinese Wikipedia

classic BERT (MLM)

updated BERT (Whole Word Masking)

Decoding

practical decoding (equal length acoustics frames)

oracle decoding (alignment accessible)

Table 1: Results on the AISHELL-1 dataset. "Orac." and "Prac." denote the oracle decoding and practical decoding, respectively. "Conv1d resnet X" denotes the conv1d resnet encoder with X resnet blocks. Best performance of the BERT-ASR are shown in bold.

Model	Acoustic encoder	Perplexity		CER (Orac.)		CER (Prac.)		SER (Orac.)		SER (Prac.)	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Trigram-LM	-	133.32	127.88	-		-		-		-	
LSTM-LM	-	79.97	78.80	-		-		-		-	
BERT-LM	-	39.74	41.72	-		-		-		-	
BERT-ASR	Average	5.88	9.02	65.8	68.9	96.4	105.8	60.3	63.5	91.5	100.3
	Conv1d resnet 1	4.91	7.63	55.8	59.0	89.6	99.6	50.0	53.8	84.4	94.1
	Conv1d resnet 2	**4.77**	**6.94**	**54.6**	**58.8**	89.7	**99.1**	49.5	53.6	84.6	93.5
	Conv1d resnet 3	4.83	7.41	54.8	58.9	89.8	99.4	49.6	53.6	84.6	93.9
	Conv1d-resnet 4	4.78	7.29	**54.6**	59.0	**89.5**	99.3	49.4	53.9	84.4	93.8
GMM-HMM	-	-		-		10.4	12.2	-		-	
DNN-HMM	-	-		-		7.2	8.4	-		-	

References

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019

[5] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre- training," 2018.

[8] A.Vaswani,N.Shazeer,N.Parmar,J.Uszkoreit,L.Jones,A.N Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," *NeruIPS*, 2017.

[17] J. Shin, Y. Lee, and K. Jung, "Effective sentence scoring method using BERT for speech recognition," 2019

[18] J.Salazar,D.Liang,T.Q.Nguyen,andK.Kirchhoff,"Masked language model scoring," 2020

[19] H. Futami, H. Inaguma, S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Distilling the knowledge of bert for sequence- to-sequence asr," 2020.

[20] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," 2017

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," Dec. 2011.

[23] T.Wolf,L.Debut,V.Sanh,J.Chaumond,C.Delangue,A.Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., "Hugging- face's transformers: State-of-the-art natural language processing," 2019.