Seminar

==========

Institute for Plasma Research

==========

Title: Comparative Analysis of Attention Mechanisms for Improved Hindi Speech Recognition

Speaker: Mr. Gaurav Garg
Institute for Plasma Research, Gandhinagar

Date: 12th April 2024 (Friday)

Time: 03:30 PM

Venue: Committee Room No. 4, IPR

Abstract

Automatic Speech Recognition (ASR), also known as Speech to Text (STT), leverages Machine Learning or Artificial Intelligence (AI) technology to transcribe spoken language into readable text.

Over the last decade, the field has experienced remarkable growth, witnessing the integration of ASR systems into ubiquitous applications like Spotify for podcast transcriptions and Zoom for meeting transcriptions.

Presently, two primary approaches dominate Automatic Speech Recognition: the traditional hybrid methodology and the end-to-end (E2E) Deep Learning methodology.

The traditional hybrid methodology, a stalwart in Speech Recognition for the past fifteen years, combines an acoustic model, lexicon (pronunciation) model, and a language model to generate transcription predictions. The lexicon model defines word pronunciations phonetically, requiring a tailored phoneme set crafted by linguistic experts for each language. The acoustic model (AM) predicts sound or phoneme occurrences within speech segments, often adopting a Hidden Markov Model (HMM) or Gaussian Mixture Model (GMM) variant. Simultaneously, the language model (LM) gauges the likelihood of word sequences, predicting subsequent words and their probabilities. Decoding involves using the lexicon, acoustic, and language model to produce a transcript.

Despite its widespread use, the traditional hybrid approach has drawbacks, including the need for separate and labor-intensive training for each model, reliance on scarce forced-aligned data, and the requirement for expertly crafted phonetic sets for enhanced accuracy.

E2E seeks to address these issues by training models end-to-end, directly mapping speech to transcripts. In an E2E system, sequence-to-sequence learning frameworks, such as Listen-Attend-Spell (LAS), jointly train acoustic and language model components. The LAS system utilizes two key components: an encoder recurrent neural network (RNN) known as the listener, and a decoder RNN referred to as the speller. The listener, designed as a pyramidal RNN, converts raw speech signals into higher-level abstract representations, while the speller converts

these features into output utterances through an attention mechanism. LAS training involves joint training of the listener and the speller.

The attention mechanism determines the importance of each part of the source side in generating a target side word.

In this paper, we present a comparative analysis of various attention mechanisms for the Hindi language. Notably, the end-to-end ASR model was trained without an additional language model.

References :

[1] Rabiner, L., Juang,B.,1986.An introduction to hidden Markov models. IEEEASSPMag.3(1),4–16.
[2] Gales,M.,Young, S.,2007.The application of hidden Markov models in speech recognition. Found.Trends Signal Process. 1(3),195–304.
[3] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results. In Neural Information Processing Systems: Workshop Deep Learning and Representation Learning Workshop, 2014.
[4] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-Based Models for Speech Recognition. In http://arxiv.org/abs/1506.07503, 2015.
[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In International Conference on Learning Representations, 2015
[6] Ilya Sutskever, Oriol Vinyals, and Quoc Le. Sequence to Sequence Learning with Neural Networks. In Neural Information Processing Systems, 2014.
[7] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwen, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Conference on Empirical Methods in Natural Language Processing, 2014.
[8] Chan,W., Jaitly, N., Le, Q. V., and Vinyals, O. (2015). Listen, attend and spell. arXiv:1508.01211 [cs, stat].
[9] Nithya R, Malavika S, Jordan F, Arjun Gangwar, Metilda N J, S Umesh, Rithik Sarab, Akhilesh Kumar Dubey, Govind Divakaran, Samudra Vijaya K, Suryakanth V Gangashetty. (2023). "SPRING-INX: A Multilingual Indian Language Speech Corpus by SPRING Lab, IIT Madras."
[10] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin
Chen, Adithya Renduchintala, and Tsubasa Ochiai, “ESPnet: End-to-end speech processing toolkit,” in Interspeech, 2018, pp. 2207–2211.