BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Large Language Models in Speech Recognition

Martin Tomašovič*

Abstract

This paper aims to explore the conditions under which the Large Language Model (LLM) improves Automatic Speech Recognition (ASR) transcription. Specifically, the study focuses on *n-best rescoring* with masked and autoregressive language models. I score the n-best hypothesis using LLM and interpolate the score with the scores from ASR. I test this approach across different ASR settings and datasets. Results demonstrate that rescoring hypotheses from Wav2Vec 2.0 and Jasper ASR systems reduces the Word Error Rate (WER). LLM fine-tuning proves to be very beneficial – smaller fine-tuned models can surpass larger non-fine-tuned ones. The findings of this research can help practitioners decide which LLM (autoregressive, masked) to use in their own ASR pipeline. And also under what conditions to use it – fine-tuning, normalization and separate scores from a CTC decoder. Additionally, this study provides insights into the expected outcomes of the LLM rescoring in ASR.

*xtomas36@vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

LLMs are nowadays mainly used for various natural language processing tasks. The question is whether bringing LLMs to ASR will improve it.

LLMs are trained on huge amounts of text data. They have learned what correct text should look like. ASR systems output hypothesis with a certain probability. The assumption is LLMs can choose a better hypothesis than the one with the highest ASR score.

Research about using LLM in ASR exists. Some works showed improvements [1, 2], some not [3]. These contrasting findings underscore the need for further investigation.

I chose to explore *n-best* rescoring. I tried various open-source LLMs, with three different ASR systems and three datasets. Among used LLMs, there are masked and autoregressive language models of similar sizes to compare their performance. I also tested the influence of LLM in-domain fine-tuning.

The results of the research contribute to knowledge about using LLM in ASR. The findings provide information about the effectiveness of certain ASR and LLM setups. The information can help implement LLM rescoring into other systems or further research.

2. ASR pipeline

An ASR pipeline is depicted in (Figure 1). Input to the pipeline is an audio (WAV file) sampled to 16kHz. The audio in experiments is taken from English datasets: LibriSpeech [4] (dev other 5.3h), GigaSpeech[1] (dev 12h, XL 10,000h) and TED-LIUM [5, 6](dev 3.7h). They cover read and spontaneous speech from different topics.

For transcribing the datasets to hypotheses, I used End2end ASR systems: `Wav2Vec2 Base 960h` [7], `Whisper Medium` [8] and `STT En Jasper10x5dr` [9]. They all utilize beam search to output multiple hypotheses. Wav2vec and Jasper work with CTC decoders and a small n-gram language model. I used a KenLM 4-gram trained on LibriSpeech during decoding.

I modified the Wav2Vec model's CTC decoder to separately output scores for the acoustic model, KenLM, and word count. I adjusted Whisper's decoder to output multiple hypotheses. The Jasper model and decoder were used without changes.

[1] https://huggingface.co/datasets/speechcolab/gigaspeech

3. Large Language Models

I scored the ASR hypotheses using masked and autoregressive LLMs by summing the probabilities of each token in a hypothesis. There is a difference in the usage of masked (Figure 2) and autoregressive models (Figure 3). The masked model takes the whole text as input and predicts a single token under the `[MASK]` token. To score a text, the same text needs to be fed x times, where x is the number of tokens in the text. Autoregressive models process text from left to right. The text is fed to them only once. One disadvantage is, that they can not access future tokens.

I used these masked language models: BERT [10] uncased with sizes `Base` 110M and `Large` 340M, RoBERTa [11] with sizes `Base` 125M and `Large` 355M. And these autoregressive models: GPT-2 [12] with sizes `Base` 137M and `Medium` 380M, TinyLlama [13] 1.1B, Falcon [14] 7B, Mistral [15] 7B, MPT [16] 7B, Llama2[2] with sizes 7B and 13B.

4. Fine-tuning

To further improve rescoring, I fine-tuned `BERT Base uncased` (Figure 4) and `GPT-2` models on LibriSpeech train clean 100h dataset's text. I tested two GPT-2 fine-tuned models: setting 1 and setting 2, the latter achieves lower validation loss. This is also reflected in (Table 1), where the setting 1 model performs better on TED-LIUM and GigaSpeech and the setting 2 model on LibriSpeech.

Except that, I fine-tuned Llama2 7B using LoRA on GigaSpeech XL text. I used three different rank and α LoRA settings: $r = 8$ $\alpha = 16$, $r = 128$ $\alpha = 256$ and $r = 256$ $\alpha = 128$. These three fine-tuned models manifest consistent improvement across all datasets.

5. Rescoring

In the last part of the pipeline, appropriate weights of an acoustic model, an LLM, an insertion bonus and in some cases, an n-gram's scores should be found, to obtain the best results. The insertion bonus is the number of words in a text. The rescoring weights are in (Figure 5).

For each LLM and ASR setting, I found the best weights by searching a certain range of values, (Figure 6) shows the relationship between WER and the weights of Llama 2 7B fine-tuned on LoRA with $r = 128 \alpha = 256$ on LibriSpeech, hypotheses are from Jasper.

[2] https://huggingface.co/docs/transformers/model_doc/llama2

6. Results

The WER values after rescoring are depicted in (Tables 1-2). The results of Wav2Vec 2.0 rescoring are in (Table 1). Jasper output is decoded using KenLM, and the WER values are in (Table 2). The best WER improvement is 4% absolutely. It is no big surprise, that the more parameters a model has, the better the results of rescoring. There are some exceptions though – some measurements with models GPT-2 and GPT-2 medium.

An interesting finding is, that after LLM in-domain fine-tuning, a smaller model can outperform a model twice its size. The improvement relies on the fine-tuning data, I observed improvement in one dataset and impairment in other datasets – GPT-2 trained on LibriSpeech data. Another finding is that the larger the model, the bigger weight the LLM score should be multiplied with, to get the best WER.

The rescoring depends on the ASR model, rescoring does not improve the WER of Whisper (Figure 7) and (Table 3). To rule out the fact the Whisper outputs non-standard text, I tried different setups. I normalized the output before, and after the LLM scoring, and I tried to pass the text in lowercase to LLM. The fact LLM rescoring does not work may be caused by the Whisper hypotheses scores being very accurate. Also, the Whisper model does not produce diverse hypotheses – many differ only in punctuation.

When it comes to masked vs autoregressive models of the same size, WER after rescoring differs only slightly. However autoregressive (GPT-2 variants) are better for spontaneous speech (GigaSpeech, TED-LIUM) and masked (BERT and RoBERTa variants) are in most cases better for read speech (LibriSpeech). This fact is influenced by the training data. The masked models I used were trained mainly on book data. On the other side, GPT-2 models were trained on internet data. An important thing to consider when deciding between autoregressive and masked models is that the autoregressive model's rescoring is faster.

All 7B models visibly improved WER after rescoring. Surprisingly, a relatively small TinyLlama 1.1B improved WER competitively with bigger models.

Acknowledgements

I would like to thank my supervisor Ing. Karel Beneš for his mentoring and help.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] Liyan Xu, Yile Gu, Jari Kolehmainen, Haidar Khan, Ankur Gandhe, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko. Rescorebert: Discriminative speech recognition rescoring with bert, May 2022.

[2] Hongzhao Huang and Fuchun Peng. An empirical study of efficient asr rescoring with transformers. 2019.

[3] Zeping Min and Jinbo Wang. Exploring the integration of large language models into automatic speech recognition systems: An empirical study. 2023.

[4] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[5] Anthony Rousseau, Paul Deléglise, and Yannick Estève. TED-LIUM: an automatic speech recognition dedicated corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 125–129, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[6] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation, 2018.

[7] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. 2020.

[8] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. 2022.

[9] Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M. Cohen, Huyen Nguyen, and Ravi Teja Gadde. Jasper: An end-to-end convolutional neural acoustic model. 2019.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. 2019.

[12] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[13] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. 2024.

[14] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The falcon series of open language models. 2023.

[15] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. 2023.

[16] MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. Accessed: 2023-05-05.