Enhancing the OpenAI Whisper model for Czech Language Recognition: A Hybrid Approach of classic fine-tuning, LORA and INT8 Training

Vít Tlustoš (xtlust05), Jiří Vlasák (xvlasa15), Josef Kotoun (xkotou06),

May 13, 2023

Abstract

This paper presents an approach to enhance the Whisper model by OpenAI for Czech language recognition, specifically trained on a large dataset composed of Czech parliamentary hearings and the Czech subset of the Common Voice dataset. We implement a hybrid technique involving classical fine-tuning, Low-Rank Adaptation (LORA) and INT8 training. Our empirical evaluation demonstrates that our model significantly outperforms the original Whisper model on the test sets of these datasets in terms of accuracy and/or computational efficiency when considering INT8 models.

1 Introduction

Automatic Speech Recognition (ASR) systems, essential in a wide range of applications, transform spoken language into written text. OpenAI's Whisper [RKX+22], an ASR system based on the Transformer architecture, has shown impressive performance across various languages and contexts due to extensive training on diverse data.

However, Whisper's performance can be further optimized for specific languages, such as Czech, that are under-represented in its training data. This paper presents our approach to enhance Whisper's performance for the Czech language using classical fine-tuning, Low-Rank Adaptation (LORA), and INT8 training, and utilizing a rich dataset composed of Czech parliamentary hearings and the Czech subset of the Common Voice dataset. The following sections detail our approach, experimental setup, and the significant improvements achieved over the original Whisper model.

2 Related Work

Whisper [RKX+22] and Wav2vec [BZMA20] are two modern advanced ASR systems. Whisper is a Transformer-based model trained on 680,000 hours of multilingual and multitask supervised data, allowing it to perform various tasks with improved robustness.

Wav2vec, on the other hand, uses self-supervised learning of speech representations and a contrastive pretraining objective to fine-tune on very little labeled data for competitive results on various ASR benchmarks.

These models have end-to-end architectures and can leverage large amounts of unlabeled data for pretraining, eliminating the need for separate components or feature engineering which is often needed when using classical ASR systems.

3 Approach

3.1 Model

Whisper is an automatic speech recognition (ASR) system developed by OpenAI. It's based on the standard transformer encoder-decoder architecture. For its training, Whisper utilizes a vast dataset comprising 680,000 hours of multilingual and multitask supervised data collected from the web. This

extensive training allows Whisper to recognize and transcribe/translate speech from a wide variety of languages and contexts, enhancing its overall performance and versatility. The model is scaled to the following scaling variant each having different number of parameters: tiny (39M), base (74M), small (244M), medium (769M) and large (1550M). In our work we use the base, small and large v2 variants.

3.2 Data

3.2.1 Used datasets

We used the Czech Parliament Plenary Hearings dataset created by Kratochvil, Polak, and Bojar. The source data are available at the website of Chamber of Desputies[1] in form of approximately 15 minutes long audio recordings of plenary hearings and corresponding transcripts from stenographers. The dataset is created by processing these recordings into form of short audio snippets and corresponding transcripts. For more information about the original dataset, including its pre-processing and structure, please refer to Kratochvil, Polak, and Bojar [KPB20].

To load the dataset, we used the Hugging Face Datasets repository, where we developed a loading script for the dataset. The script allows for the loading of individual splits using the Hugging Face Datasets library.

In addition to the Czech Parliament Plenary Hearings dataset, we incorporated the Common Voice 13.0 [ABD+19]. This dataset contains 256 hours of Czech audio and transcripts in the form of one sentence per sample.

3.2.2 Preprocessing

The transcripts in the original dataset were in uppercase and without punctuation, which is not ideal for transformer-based models. To address this issue, we fine-tuned the NeMo [KLN+19] Punctuation And Capitalization Model for Czech and applied it to process the transcripts. The NeMo model is based on BERT [DCLT18], which is a transformer-based language model, and has been pre-trained on a large corpus of data to recognize and correct common capitalization and punctuation errors. For fine-tuning the model, we used the UD_Czech-CAC treebank and UD_Czech-PDT treebank, which together contain approximately 120,000 Czech sentences. These datasets were pre-processed to meet the NeMo model's required format using their script from the github example. After fine-tuning the model, we evaluated its performance in adding punctuation and capitalization on test dataset split. The model achieved an acceptable precision of 70% for adding punctuation to sentences and 95% for capitalization.

We then preprocessed the dataset by removing the 'SEL' word from the start or end of some transcription, which likely originated from the parliament dataset's automatic system for splitting the data into shorter snippets. Processed dataset is available at Hugging Face repository.

Finally, we merged the datasets, resulting in a dataset consisting of 700 hours of Czech audio and its corresponding transcripts.

3.3 Training

Our goal was to enhance the Whisper model for better performance on the Czech language recognition. To achieve this, we employed a multiple techniques involving classical fine-tuning, performance efficient tuning (PEFT) and Low-Rank Adaptation (LORA) and training in reduced precision (INT8). The last point worth noting is that all the training was conducted on the Galdor cluster, which features powerful NVIDIA A40 48GB GPUs provided by Metacentrum.

3.3.1 Fine-tuning

Fine-tuning is a common approach where a pre-trained model is further trained (fine-tuned) on a new dataset. In our case, we fine-tuned the Whisper model on the Czech parliamentary hearings dataset and the Czech subset of the Common Voice dataset. It is notable that when employing this approach all model weights of the model are updated. This may result in to the so called *catastrophic forgetting* since all the weights of the model are updated and the dataset is usually task specific. Therefore we employ the PEFT + LORA when training the Large V2 version of the model.

[1] https://www.psp.cz/eknih/2017ps/audio/2017/index.htm

3.3.2 Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) [MGD+22] is a method for fine-tuning deep learning models that optimizes the trade-off between the model's performance and computational efficiency thereby performing the performance efficient tuning. Without the PEFT the training of the large version of the model would not be possible because it would not fit at all or a batch of only few samples would be possible to fit to the GPU.

3.3.3 Low-Rank Adaptation

Low-Rank Adaptation (LORA) [HSW+21] is a method used to efficiently fine-tune pre-trained models. When performing the LORA we have only fine-tuned the the Query and Value matrices of each attention layer. It is worth noting that the original matrices from the pre-trained model are kept frozen to preserve the general knowledge embedded within them. In LORA, instead of directly fine-tuning these original matrices, two new matrices per each Query/Value matrix are introduced: first matrix $M_1^{D\times R}$ and second matrix $M_2^{R\times D}$, where D is the original dimension and R is the rank of the low-rank structure. When multiplied like $M_{LORA} = M_1 \times M_2$ the resulting matrix $M_{LORA}^{D\times D}$ has a same as the original matrix making is suitable for later addition. These matrices are fine-tuned on the new task, significantly reducing computational cost due to their lower dimensionality. Importantly, the outputs of these new matrices do not replace, but rather supplement the outputs of the original matrices. The low-rank outputs are added to the original outputs, integrating the task-specific adaptations with the pre-trained model's general knowledge. Also note that an important hyper-parameter scale S is introduced to the equation $M_{result} = M_{LORA} * S + M_{original}$ giving more importance to the adaptation. This balanced approach allows for efficient and effective adaptation of the model to the new task. In this work we have used a rank $R = 16$ and scale $S = 16$. In the case of the Large V2 model we train only 0.57 % of the total parameters thereby performing the performance efficient tuning.

3.3.4 INT8

Lastly, we implemented INT8 training. INT8 quantization is a method that reduces the numerical precision of the weights in a neural network. This reduction in precision allows for faster computation and less memory usage, significantly improving the efficiency of the model. This approach does not significantly compromise the accuracy of the predictions by performing clever quantization with respect to the dataset, making it an excellent choice for our task. In the case of the Large V2 version of the model we would be able to fit a batch of only 2 samples per one A40 48GB GPU without the PEFT + LORA + INT8. While using all the mentioned techniques we can easily fit a batch of 32 samples to the one A40 48GB GPU.

4 Experiments and evaluation

In this section, we describe the experiments conducted to evaluate the effectiveness of our hybrid approach to enhance the Whisper model for Czech language recognition. We also present the evaluation metrics used, the results obtained from the experiments, and a comparison between our enhanced model and the original Whisper model.

4.1 Experimental setup

We merged the Czech parliamentary hearings dataset and the Czech subset of the Common Voice dataset and trained our models on this dataset. For training we used the train and validation subsets of these datasets and used the test subsets for evaluation.

To evaluate the model's generalization capability, we conducted a zero-shot evaluation on the VoxPopuli dataset. Here we measure the model's ability to generalize to unseen data. We compared our enhanced model with the original Whisper model.

4.2 Evaluation Metrics

To evaluate the performance of our fine-tuned models, we used two metrics: cross-entropy loss and word error rate (WER).

Cross-entropy loss is a standard metric for evaluating the performance of a neural network. It measures the difference between the predicted probabilities of the model and the actual target labels.

In addition, we used Word Error Rate to measure the accuracy of the transcribed text output by the model compared to the ground truth text. WER measures the percentage of words in the transcribed text output that are incorrect compared to the ground truth text. It is calculated by dividing the total number of errors (insertions, deletions, and substitutions) by the total number of words in the ground truth text. A lower WER indicates better accuracy of the transcribed text output.

4.3 Results

4.3.1 Czech parliament and Common Voice 13.0 dataset

The table 1 shows the performance of our models on the merged Czech parliament and Common Voice 13.0 Czech dataset. We compared our models with the OpenAI base model, OpenAI Small model, and OpenAI LargeV2 model. We can see that the OpenAI base model performed the worst with a high cross-entropy loss and WER (as expected). In all cases the training the finetuning yields notable improvements.

Model	**Cross-entropy loss**	**WER**
OpenAI Base	2.37	75.19
Common Voice and parliament Base (ours)	0.25	23.18
OpenAI Small	3.104	49.3
Common Voice and parliament Small (ours)	0.154	13.8
OpenAI LargeV2	2.39	34.2
Common Voice and parliament LargeV2-LORA-Int8 (ours)	0.17	15.53

Table 1: Performance of our models on the merged Czech parliament and Common Voice 13.0 Czech dataset.

4.3.2 Zero-shot evaluation on the Czech subset of the VoxPopuli dataset

In addition to evaluating our models on the merged Czech parliament and Common Voice dataset, we also conducted a zero-shot evaluation on the Czech subset of the VoxPopuli dataset. The results of this evaluation are shown in Table 2. As expected, the models that were trained on the merged Czech parliament and Common Voice dataset outperformed the OpenAI models on this dataset. Our best-performing model achieved a WER of 16.9, which is still better than the OpenAI's equivalent model. However, our model's performance on this dataset was not as strong as on the merged dataset, indicating that there is still room for improvement in our model's generalization to unseen data.

Model	**Cross-entropy loss**	**WER**
OpenAI Base	1.715	64.66
Common Voice and parliament Base (ours)	0.51	35.31
OpenAI Small	1.81	36.2
Common Voice and parliament Small (ours)	0.36	23.7
OpenAI LargeV2	1.32	17.8
Common Voice and parliament LargeV2-LORA-Int8 (ours)	1.21	16.9

Table 2: Performance of our models on the Czech subset of the VoxPopuli dataset.

5 Conclusion

In conclusion, the empirical evidence obtained from this study clearly attests to the superior performance of the fine-tuned model over its original counterpart in handling the Czech language. This superiority is clearly demonstrated by the substantial improvements noted in the zero-shot evaluation results on the VoxPopuli dataset. The base model's Word Error Rate (WER) decreased from 64.66

to 35.31, while the small model experienced a drop from 36.2 WER to 23.7 WER. Furthermore, the LargeV2 model saw a minor yet significant improvement from 17.8 WER to 16.9 WER.

However one might anticipate a more pronounced difference when fine-tuning the LargeV2 model. It's worth noting that the Large V2 model, despite being fine-tuned with the combined application of LORA, PEFT, and INT8 on our dataset, did not outperform the small, traditionally fine-tuned model. In fact, its performance was only marginally superior on the VoxPopuli dataset. This phenomenon may be attributable to the quantization process or potentially a rank deficiency that hinders it from surpassing the performance of the fully fine-tuned small model. These factors offer prospects for further investigation and analysis.

References

[ABD+19] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[BZMA20] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[HSW+21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[KLN+19] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019.

[KPB20] Jonas Kratochvil, Peter Polak, and Ondrej Bojar. Large corpus of Czech parliament plenary hearings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6363–6367, Marseille, France, May 2020. European Language Resources Association.

[MGD+22] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

[RKX+22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.