MONICA2: Mobile Neural Voice Command Assistants towards Smaller and Smarter

Yoonseok Hong, Shounan An, Sunwoo Im, Jaegeon Jo, Insoo Oh

Netmarble AI Center
38, Digital-ro 26-gil
Guro-gu, Seoul, Republic of Korea
{yhong, ethan.an, ism07, jg623, ioh}@netmarble.com

Abstract

In this paper, we propose on-device voice command assistants for mobile games to increase user experiences even in hands-busy situations such as driving and cooking. Since most of the current mobile games cost large memory (e.g. more than 1GB memory), so it is necessary to reduce memory usage further to integrate voice commands systems on mobile clients. Therefore a need to design an on-device automatic speech recognition system that costs minimal memory and CPU resources rises. To this end, we apply cross layer parameter sharing to Conformer, named MONICA2 which results in lower memory usage for on-device speech recognition. MONICA2 reduces the number of parameters of deep neural network by 58%, with minimal recognition accuracy degradation measured in word error rate on Librispeech benchmark. As an on-device voice command user interface, MONICA2 costs only 12.8MB mobile memory and the average inference time for 3-seconds voice command is about 30ms, which is profiled in Samsung Galaxy S9. As far as we know, MONICA2 is the most memory efficient yet accurate on-device speech recognition which could be applied to various applications such as mobile games, IoT devices, etc.

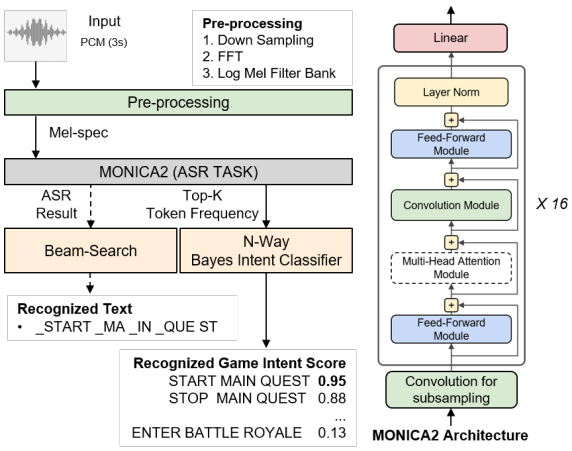

Figure 1: Pipeline of voice commands assistants SDK. The modules with dash lines mean that one module is shared in MONICA2 architecture.

Introduction

Nowadays, transformer architectures (Vaswani et al. 2017) show breakthrough progress on automatic speech recognition (ASR) tasks. (Dong, Xu, and Xu 2018; Karita et al. 2019a) have demonstrated that transformers are suitable networks in terms of accuracy performance and easy to train than RNN based architectures. In addition, there are studies about combining transformer encoders with the connectionist temporal classification (CTC) loss or the transducer loss (Xu, Li, and Zhang 2021), which have shown better performance than RNNs. (Karita et al. 2019b) improves the recognition accuracy by combining CTC based acoustic models (AM) and language models (LM). Recently there have been studies involving convolution modules in transformer blocks, named Conformer (Gulati et al. 2020), which successfully achieved the state-of-the-art (SOTA) recognition accuracy even with fewer parameter sizes.

Although the Conformer has smaller parameters than any other SOTA models, there are difficulties to use Conformer as an on-device voice assistant for mobile games, aspect on restrained memory usage. In addition, most of current mobile massively multiplayer online role-playing games (MMORPG) cost large memory and mobile GPUs are occupied by game engines for graphical rendering. Therefore the goal of this work is to reduce memory footprints and CPU resources for on-device ASR system with minimal recognition accuracy degradation. To this end, we apply cross-layer parameter sharing (CLPS) (Lan et al. 2019; Li et al. 2019) for the encoder blocks of Conformer, and name our model as MONICA2.

MONICA2

We propose MONICA2 to further reduce memory usage with minimum accuracy degradation. The modules of MONICA2 and Conformer are very similar, however, MONICA2 is applied CLPS on multi-head self-attention modules (MHSAM). As Figure 1, the 16 encoder blocks of MONICA2 share one MHSAM. In additional, we replace the RNN decoder which is used in Conformer with transducer loss to linear projection layer with CTC. Also, MONICA2 is designed with the smaller network size,

Model	# Params (M)	Librispeech benchmark (WER) dev-clean / others	test-clean / others	Mobile Factor Weights Size	Memory	RTF
Conformer S (Gulati et al. 2020)	10.3	-	2.7 / 6.3	-	-	-
S-T lite (Lim et al. 2020)	52.9	5.1 / 12.6	5.3 / 12.8	-	199	0.471
MONICA (Lim et al. 2020)	5.4	-	9.2 / 19.9	3.9	27	0.087
Conformer - CTC	5.2	5.4 / 13.5	5.5 / 13.7	4.1	17.1	**0.013**
MONICA2	**3.5**	5.9 / 14.8	6.1 / 15.1	**2.9**	**12.8**	**0.013**

Table 1: Evaluation results on Librispeech datasets. Weights size means occupied tflite model's size on disks (MB) and memory is peak memory footprint during inference (MB). RTF is the seconds to inference 1 second audio. Mobile factor is measured on Galaxy S9.

	Response Time	Success Rate
MONICA2		
+ Cosine Similarity	150 ms	82 %
+ N-way NBIC	130 ms	88 %

Table 2: The results are done with an android application for demo on Samsung Galaxy S9. Response time covers all steps in the SDK pipeline.

(a) Key-word activated (b) Voice command recognized

Figure 2: MONICA2 integrated into *A3: Still Alive* as voice commands assistant. (a) After key-word is activated. (b) The main character take action based the recognized voice command via MONICA2: "start auto-battle".

$\{num_{blocks}$=16, dim_{units}=144, dim_{attn}=144, att_{head}=4, $size_{conv}$=32$\}$. The training models are performed using ESPNet framework (Watanabe et al. 2018).

To evaluate the effectiveness of MONICA2, we perform experiments on the Librispeech corpus(Panayotov et al. 2015) and summarize the recognition accuracy measured with word error rate (WER) and mobile resource profiling on Table 1. S-T lite and MONICA are our previous works which is based on transformer. Conformer-CTC and MONICA2 has similar architecture, the main difference is MONICA2 applied sharing MHSAM to further compress parameters. When sharing MHSAM like MONICA2, the memory usage is reduced about $4.3MB$ with only 0.5% WER drop compare to Conformer-CTC. Also, MONICA2 has improved by 3% WER compared to MONICA and the memory footprint is reduced by $14.2MB$.

Demonstration

As Figure 1, we developed the SDK that has a simple pipeline to integrate MONICA2 into MMORPG, *A3: Still Alive* (Netmarble Corp. & Netmarble N2 Inc. 2020) as prototype demonstration. In pre-processing step, mel-spectrogram is extracted from PCM as 3-seconds waveform using CKFFT (Cricket Technology 2014) which has highly optimized STFT library for mobile CPU. The ASR tasks with MONICA2 is running on Tensorflow lite interpreter and ASR decoding is done with beam search ($size = 1$) which is same as greedy search. To understand the intents of the voice command such as *"start auto-battle", and "show equipment"*, we design a simple N-ways Naive bayesian intent classifier (NBIC) to execute the 16 voice commands actions for games. The NBIC uses a frequency vector which is extracted from top-5 token for each frame as input and consists of term frequency, inverse document frequency and two multi-class naive bayesian classifiers in that order. The one classifier is for actions, *"start", "stop", "show", etc* and the other is for target *"main quest", "map", "battle royale", etc* . With NBIC, we can improve the intent recognition accuracy of MONICA2 without any LMs and the results is on Table 2.

As shown in Figure 2, with the work developed in (An et al. 2019; Xu et al. 2020), on-device key-word spotting is already on the service for *A3: Still Alive* in Korea.

In addition, we build the Android application powered by Unity Engine for mobile environment tests. For an interactive demonstrations, we release our android test app and voice commands chess games in our demo link [1].

Conclusions

In this paper, we have presented memory efficient MONICA2 for on-device ASR, which reduce the number of parameters of DNNs by 58% with minimal recognition accuracy degradation measured in WER (Conformer: 4.3% vs MONICA2: 6.1%) on benchmark Librispeech data sets. MONICA2 costs only 12.8MB mobile memory and running real-time on most of current mobile phones as an voice command user interface. We also integrate MONICA2 into mobile game $A3$: *Still Alive* and MONICA2 will be on the service next year.

[1]Demo Url : https://yshong93.github.io/monica2_demo_public

References

An, S.; Kim, Y.; Xu, H.; Lee, J.; Lee, M.; and Oh, I. 2019. Robust keyword spotting via recycle-pooling for mobile game. *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*.

Cricket Technology. 2014. Cricket FFT. https://github.com/sjmerel/ckfft. Accessed: 2021-12-01.

Dong, L.; Xu, S.; and Xu, B. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5884–5888. IEEE.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; and Pang, R. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, 5036–5040.

Karita, S.; Chen, N.; Hayashi, T.; Hori, T.; Inaguma, H.; Jiang, Z.; Someki, M.; Soplin, N. E. Y.; Yamamoto, R.; Wang, X.; et al. 2019a. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 449–456. IEEE.

Karita, S.; Soplin, N. E. Y.; Watanabe, S.; Delcroix, M.; Ogawa, A.; and Nakatani, T. 2019b. Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration. *Proc. Interspeech 2019*, 1408–1412.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Li, S.; Raj, D.; Lu, X.; Shen, P.; Kawahara, T.; and Kawai, H. 2019. Improving Transformer-Based Speech Recognition Systems with Compressed Structure and Speech Attributes Augmentation. In *INTERSPEECH*, 4400–4404.

Lim, Y.; Hong, Y.; An, S.; Jo, J.; Lee, H.; Jeong, S.; Eum, A.; Im, S.; and Oh, I. 2020. MONICA: MObile Neural voIce Command Assistant for mobile games. *Proceedings of the NeurIPS 2020 Demonstration Track*.

Netmarble Corp. & Netmarble N2 Inc. 2020. A3: Still Alive. https://a3.netmarble.com/. Accessed: 2021-12-01.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Enrique Yalta Soplin, N.; Heymann, J.; Wiesner, M.; Chen, N.; Renduchintala, A.; and Ochiai, T. 2018. ESPnet: End-to-End Speech Processing Toolkit. In *Proceedings of Interspeech*, 2207–2211.

Xu, H.; Lim, Y.; An, S.; Cho, H.; Hong, Y.; and Oh, I. 2020. Robust end-to-end keyword spotting and voice command recognition for mobile game. *Acoustics, Speech and Signal Processing (ICASSP), 2020 IEEE International Conference on*.

Xu, M.; Li, S.; and Zhang, X.-L. 2021. Transformer-Based End-to-End Speech Recognition with Local Dense Synthesizer Attention. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5899–5903.