AIoT Bench: Towards Comprehensive Benchmarking Mobile and Embedded device Intelligence

Chunjie Luo[124], Fan Zhang[1], Cheng Huang[12], Xingwang Xiong[12], Jianan Chen[12], Lei Wang[14], Wanling Gao[14], Hainan Ye[34], Tong Wu[5], Runsong Zhou[6], and Jianfeng Zhan[124] ⋆

[1] State Key Laboratory of Computer Architecture,
Institute of Computing Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Beijing Academy of Frontier Science and Technology
[4] BenchCouncil (International Open Benchmarking Council)
[5] China National Institute of Metrology
[6] China Software Testing Center

Abstract. Due to increasing amounts of data and compute resources, the deep learning achieves many successes in various domains. Recently, researchers and engineers make effort to apply the intelligent algorithms to the mobile or embedded devices. In this paper, we propose a benchmark suite, AIoT Bench, to evaluate the AI ability of mobile and embedded devices. Our benchmark 1) covers different application domains, e.g. image recognition, speech recognition and natural language processing; 2) covers different platforms, including Android and Raspberry Pi; 3) covers different development frameworks, including TensorFlow and Caffe2; 4) offers both end-to-end application workloads and micro workloads.

Keywords: AI · IoT · Benchmark.

1 Introduction

Due to increasing amounts of data and compute resources, the deep learning achieves many successes in various domains. Recently, researchers and engineers make effort to apply the intelligent algorithms to the mobile or embedded devices, e.g. smart phone, self-driving cars, smart home. On one hand, the neural networks are made more lightweight to adapt the mobile or embedded devices by using simpler architecture, or by quantizing, pruning and compressing the networks. On the other hand, the mobile and embedded devices provide additional hardware acceleration using GPUs or NPUs to support the AI applications. Since AI applications on mobile and embedded devices get more and more attention,

⋆ Jianfeng Zhan is the corresponding author.

the benchmarking of the AI ability of those devices becomes an urgent problem to be solved.

Benchmarking the AI ability of mobile and embedded devices is non-trivial. We consider that the benchmark should meet the following requirements. 1) It should cover typical AI application domains. Currently, the AI application mainly focuses on the image, speech, and text domain. The workloads should satisfy the diversity of the AI application domain. 2) It should cover the typical platforms of the IoT devices. Android devices and Raspberry Pi are the widely used in IoT environments. 3) It should consider the different development frameworks of the AI applications on the mobile and embed devices. 4) Beyond the end-to-end application workloads which can reflect the performance of the system comprehensively, we also need micro workloads to obtain the fine-grained analysis of the performance.

Recently, there are several AI related benchmarks have been proposed. For example, ETH Zurich AI benchmark [10] aim to benchmark the Android smartphone using different vision tasks implemented with TensorFlow Lite. Other AI related benchmarks have Fathom [2], DAWNBench [4]. The existing AI related benchmarks do not satisfy the requirements mentioned above.

In this paper, we propose a benchmark suite, AIoT Bench, to evaluate the AI ability of mobile and embedded devices. Our benchmark 1) covers different application domains, e.g. image recognition, speech recognition and natural language processing; 2) covers different platforms, including Android devices and Raspberry Pi; 3) covers different development tools, including TensorFlow and Caffe2; 4) offers both end-to-end application workloads and micro workloads. Coordinated by BenchCouncil, we also release AIBench [5, 6], HPC AI500 [12], Edge AIBench [8] and BigDataBench [14, 15].

2 Benchmarking Requirements

Here we will discuss the requirements of benchmarking mobile and embedded devices intelligence.

- **Domain diversity**. Computer vision is the most active research area for AI applications, typical vision tasks include image classification, face recognition, and object detection. Speech recognition, to map an acoustic signal into the corresponding sequence of words, is another active area of AI research and application. Natural language processing (NLP) is one of the main AI areas. Natural language processing includes applications such as language model, machine translation, sentiment analysis and so on [7]. There are different features in different areas. The AI benchmarks should cover those typical application areas.
- **Platform diversity**. Android is designed primarily for touchscreen mobile devices such as smartphones and tablets. In addition, Google has developed Android TV for televisions, Android Auto for cars, and Wear OS for wrist watches. Because of its openness, Android becomes the world's most popular mobile platform. The Raspberry Pi is a series of small single-board

computers. The Raspberry Pi is a powerful platform when it comes to AI. Because of its strong processing capability, the small form factor, and low power requirement, the Raspberry Pi is very popular for smart robotics and embedded projects.

- **Framework diversity**. There are a number of popular deep learning frameworks. The benchmarks should cover the main frameworks, which are widely used on mobile and embedded devices. TensorFlow [1] is an open-source machine learning library, released by Google in 2015. TensorFlow Lite, designed for mobile and embedded devices, is presented in 2017. Caffe [11] is another popular open-source deep learning framework, developed at UC Berkeley. Facebooks releases Caffe2 in 2017, the mobile version for iOS and Android platforms.
- **Testing Hierarchy**. The end-to-end application benchmark can reflect the performance of the system comprehensively, while micro benchmark can get the fine-grained analysis of the performance. Both of them are useful for evaluating the mobile and embedded devices.

3 AIoT Bench

We propose a benchmark suite, AIoT Bench, to evaluate the AI ability of mobile and embedded devices. Our benchmark 1) covers different application domains, e.g. image recognition, speech recognition and natural language processing; 2) covers different platforms, including Android and Raspberry Pi; 3) covers different development frameworks, including TensorFlow and Caffe2; 4) offers both end-to-end application workloads and micro workloads.

Image classification workload. This is an end-to-end application workload of vision domain, which takes an image as input and outputs the image label. The model we use for image classification is MobileNet [9], which is a light weight convolutional network designed for mobile and embedded devices.

Speech recognition workload. This is an end-to-end application workload of speech domain, which takes words and phrases in a spoken language as input and converts them to the text format. The model we use is the DeepSpeech 2 [3], which consists of 2 convolutional layers, 5 bidirectional RNN layers, and a fully connected layer.

Transformer translation workload. This is an end-to-end application workload of NLP domain, which takes the text of one language as input and translates into another language. The model we use is transformer translation model [13], which solves sequence to sequence problems using attention mechanisms without recurrent connections used in traditional neural seq2seq models.

Micro workloads. In our benchmarks, we provide the micro workloads, which are the basic operations to compose different networks. In detail, the micro workloads include convolutional operation, pointwise convolution, depthwise convolution, matrix multiply, pointwise add, ReLU activation, sigmoid activation, max pooling, average pooling.

The workloads in AIoT Bench are implemented using both TensorFlow Lite and Caffe 2 on the platform of Android as well as Raspberry Pi. We only include the prediction procedure since the training are usually carried out on datacenters.

4 Related Work

ETH Zurich AI benchmark [10] contains workloads covering the tasks of object recognition, face recognition, playing atari games, image deblurring, image super-resolution, bokeh simulation, semantic segmentation, photo enhancement. Those tasks are mainly focus on the vision application. The benchmark suite is implemented only using TensorFlow Lite and aims to evaluate the Android smartphones.

Table 1. The comparison of AIoT Bench against ETH Zurich AI benchmark.

		AIoT Bench	ETH Zurich AI benchmark
Domain diversity	Vision	Yes	Yes
	Speech	Yes	No
	NLP	Yes	No
Platform diversity	Android	Yes	Yes
	Raspberry Pi	Yes	No
Framework diversity	Tensorflow Lite	Yes	Yes
	Caffe2	Yes	No
Testing Hierarchy	End-to-end	Yes	Yes
	Micro	Yes	No

5 Conclusion

In this paper, we analyze the requirements of benchmarking IoT devices intelligence. And to meet the requirements, we propose a benchmark suite, AIoT Bench, to evaluate the AI ability of mobile and embedded devices. Our benchmark covers different application domains, different platforms, different development frameworks. Moreover, we offer both end-to-end application workloads and micro workloads in our benchmark.

Acknowledgment

This work is supported by the Standardization Research Project of Chinese Academy of Sciences No.BZ201800001.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). pp. 265–283 (2016)
2. Adolf, R., Rama, S., Reagen, B., Wei, G.Y., Brooks, D.: Fathom: Reference workloads for modern deep learning methods. In: 2016 IEEE International Symposium on Workload Characterization (IISWC). pp. 1–10. IEEE (2016)
3. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: End-to-end speech recognition in english and mandarin. In: International conference on machine learning. pp. 173–182 (2016)
4. Coleman, C., Narayanan, D., Kang, D., Zhao, T., Zhang, J., Nardi, L., Bailis, P., Olukotun, K., Ré, C., Zaharia, M.: Dawnbench: An end-to-end deep learning benchmark and competition. Training **100**(101), 102 (2017)
5. Wanling Gao, Fei Tang, Lei Wang, Jianfeng Zhan, Chunxin Lan, Chunjie Luo, Yunyou Huang, Chen Zheng, Jiahui Dai, Zheng Cao, Daoyi Zheng, Haoning Tang, Kunlin Zhan, Biao Wang, Defei Kong, Tong Wu, Minghe Yu, Chongkang Tan, Huan Li, Xinhui Tian, Yatao Li, Junchao Shao, Zhenyu Wang, Xiaoyu Wang, and Hainan Ye.: AIBench: An Industry Standard Internet Service AI Benchmark Suite. Technical Report 2019.
6. Gao, W., Luo, C., Wang, L., Xiong, X., Chen, J., Hao, T., Jiang, Z., Fan, F., Du, M., Huang, Y., Zhang, F., Wen, X., Zheng, C., He, X., Dai, J., Ye, H., Cao, Z., Jia, Z., Zhan, K., Tang, H., Zheng, D., Xie, B., Li, W., Wang, X., Zhan, J.: The report of datacenter ai benchmarks in bigdatabench: Towards scalable and comprehensive ai and big data benchmarking. 2018 BenchCouncil Symposium on Benchmarking, Measuring and Optimizing (Bench18) (2018)
7. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
8. Hao, T., Huang, Y., Wen, X., Gao, W., Zhang, F., Wang, L., Ye, H., Hwang, K., Ren, Z., Zhan, J.: Edge aibench: Towards comprehensive end-to-end edge computing benchmarking. 2018 BenchCouncil Symposium on Benchmarking, Measuring and Optimizing (Bench18) (2018)
9. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
10. Ignatov, A., Timofte, R., Chou, W., Wang, K., Wu, M., Hartley, T., Van Gool, L.: Ai benchmark: Running deep neural networks on android smartphones. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
11. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 675–678. ACM (2014)
12. Jiang, Z., Gao, W., Wang, L., Xiong, X., Zhang, Y., Wen, X., Luo, C., Ye, H., Lu, X., Xu, W., Zhang, Y., Feng, S., Li, K., Zhan, J.: Hpc ai500: A benchmark suite for hpc ai systems. 2018 BenchCouncil Symposium on Benchmarking, Measuring and Optimizing (Bench18) (2018)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)

14. L. Wang, J. Zhan, C. Luo, Y. Zhu, Q. Yang, Y. He, W. Gao, Z. Jia, Y. Shi, S. Zhang *et al.*, "BigDataBench: A big data benchmark suite from internet services," *IEEE International Symposium On High Performance Computer Architecture (HPCA)*, 2014.
15. Jia Z, Wang L, Zhan J, Zhang L, Luo C.: Characterizing data analysis workloads in data centers. In 2013 IEEE International Symposium on Workload Characterization (IISWC) 2013 Sep 22 (pp. 66-76). IEEE.