AIBench: Towards Scalable and Comprehensive Datacenter AI Benchmarking

Wanling Gao[1,2,4], Chunjie Luo[1,2,4], Lei Wang[1,2,4], Xingwang Xiong[1,4], Jianan Chen[1,4], Tianshu Hao[1,4], Zihan Jiang[1,4], Fanda Fan[1,4], Mengjia Du[1,4], Yunyou Huang[1,4], Fan Zhang[1], Xu Wen[1,4], Chen Zheng[1,2,4], Xiwen He[1], Jiahui Dai[2,3], Hainan Ye[2,3], Zheng Cao[5], Zhen Jia[6], Kent Zhan[7], Haoning Tang[8], Daoyi Zheng[9], Biwei Xie[10], Wei Li[11], Xiaoyu Wang[12], and Jianfeng Zhan[1,2,4] ⋆

[1] State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences
{gaowanling,wanglei_2011,luochunjie,xiongxingwang,chenjianan,haotianshu, jiangzihan,fanfanda,dumengjia,huangyunyou,zhangfan,wenxu,zhengchen, zhanjianfeng}@ict.ac.cn
[2] BenchCouncil (International Open Benchmarking Council)
[3] Beijing Academy of Frontier Sciences and Technology, daijiahui,yehainan@mail.bafst.com
[4] University of Chinese Academy of Sciences
[5] Alibaba, zhengzhi.cz@alibaba-inc.com
[6] Princeton University, zhenj@cs.princeton.edu
[7] Wuba, zhankunlin@58.com
[8] Tencent, haoningtang@tencent.com
[9] Baidu, zhengdaoyi@baidu.com
[10] China RISC-V Alliance
[11] Cambricon, liwei1@cambricon.com
[12] Intellifusion, wang.xiaoyu@intellif.com

Abstract. AI benchmarking provides yardsticks for benchmarking, measuring and evaluating innovative AI algorithms, architecture, and systems. Coordinated by BenchCouncil, this paper presents our joint research and engineering efforts with several academic and industrial partners on the datacenter AI benchmarks—AIBench. The benchmarks are publicly available from http://www.benchcouncil.org/AIBench/index.html. Presently, AIBench covers 16 problem domains, including image classification, image generation, text-to-text translation, image-to-text, image-to-image, speech-to-text, face embedding, 3D face recognition, object detection, video prediction, image compression, recommendation, 3D object reconstruction, text summarization, spatial transformer, and learning to rank, and two end-to-end application AI benchmarks. Meanwhile, the AI benchmark suites for high performance computing (HPC), IoT, Edge are also released on the BenchCouncil web site. This is by far the most comprehensive AI benchmarking research and engineering effort.

Keywords: Datacenter · AI · Benchmark.

⋆ Jianfeng Zhan is the corresponding author.

1 Introduction

AIBench provides a scalable and comprehensive datacenter AI benchmark suite. In total, it includes 12 micro benchmarks, 16 component benchmarks, covering 16 AI problem domains: image classification, image generation, text-to-text translation, image-to-text, image-to-image, speech-to-text, face embedding, 3D face recognition, object detection, video prediction, image compression, recommendation, 3D object reconstruction, text summarization, spatial transformer, learning to rank, and two end-to-end application AI benchmarks: DCMix [1]—a datacenter AI application combination mixed with AI workloads, and E-commerce AI—an end-to-end business AI benchmark. The details of AIBench is introduced in our technical report [2].

We provide both training and inference benchmarks. The training metrics are the wall clock time to train the specific epochs, the wall clock time to train a model achieving a target accuracy [3], and the energy consumption to train a model achieving a target accuracy [3]. The inference metrics are the wall clock time, accuracy, and energy consumption. Additionally, the performance numbers are reported on the BenchCouncil web site (http://www.benchcouncil.org/numbers.html), to measure the training and inference speeds of different hardware platforms, including multiple types of NIVDIA GPUs, Intel CPUs, AI accelerator chips, and to measure the performance of different software stacks, including TensorFlow, PyTorch, and etc.

Using the benchmarks from AIBench, BenchCouncil is organizing the 2019 BenchCouncil International AI System and Algorithm Competition, including four tracks: AI System Competitions on RISC-V—an open-source chip, Cambricon—an AI accelerator Chip, and X86 processors, and 3D Face Recognition Algorithm Competition sponsored by Intellifusion.

2 Related Work

Much previous work focuses on datacenter AI benchmarking. Table 1 summarizes the differences between AIBench and the state-of-the-art and state-of-the-practise datacenter AI benchmarks. Previous work like MLPerf [4], Fathom [5], DAWNBench [3], and TBD suite [6] only targets at component benchmarks, while lacking of the micro and application benchmarks. On the contrary, benchmarks like DeepBench [7] and DNNMark [8] only provide several micro benchmarks, while lacking of the component and application benchmarks. Thus, previous work adopts a narrow vision of datacenter AI scenario, and fails to propose a comprehensive AI benchmark suite.

AIBench includes a series of micro, component and application benchmarks to benchmark the AI systems, architectures, and algorithms. Also, a wide variety of data types and data sources are covered, including text, images, street scenes, audios, videos, etc. The workloads are implemented not only based on mainstream deep learning frameworks like TensorFlow and PyTorch, but also based on traditional programming model like Pthreads, to conduct an apple-to-apple

comparison. Meanwhile, the HPC AI benchmarks [9], IoT AI benchmarks [10], Edge AI benchmarks [11], and big data benchmarks [12–14] are also released on the BenchCouncil web site.

Table 1. The Summary of Different AI Benchmarks.

	Micro benchmark	Component benchmark	Application benchmark	Dataset	Software Stacks
AIBench	12	16	2	16	3
MLPerf [4]	N/A	7	N/A	3	2
Fathom [5]	N/A	8	N/A	6	1
DeepBench [7]	4	N/A	N/A	N/A	1
DNNMark [8]	8	N/A	N/A	N/A	1
DAWNBench [3]	N/A	2	N/A	3	2
TBD [6]	N/A	7	N/A	6	4

3 Datacenter AI Benchmark Suite—AIBench

Totally, AIBench covers 16 representative real-world data sets widely used in AI scenario and provides 12 AI micro benchmarks and 16 AI component benchmarks. Among them, each micro benchmark provides a neural network kernel implementation, consisting of a single unit of computation [15]; Each component benchmark provides a full neural network model to solve multiple tasks, each of which is a combination of multiple units of computation; Each application benchmark provides an end-to-end application scenario.

3.1 Datacenter AI Micro Benchmarks

Micro benchmarks in AIBench abstracts units of computation among a majority of AI algorithms, and covers 12 units of computation in total. The micro benchmarks are convolution, fully connected, relu, sigmoid, tanh, maximum pooling, average pooling, cosine normalization, batch normalization, dropout, element-wise operation, and softmax.

3.2 Datacenter AI Component Benchmarks

Component benchmarks in AIBench cover 16 problem domains and contain both training and inference. For both training and inference, TensorFlow and PyTorch implementations are provided.

Image classification uses ResNet neural network [16] and uses ImageNet [17] as data input to solve image classification task.

Image generation uses WGAN [18] algorithms and uses LSUN [19] dataset as data input to generate image data.

Text-to-Text Translation uses recurrent neural networks [20] and takes WMT English-German [21] as data input to translate text data.

Image-to-Text uses Neural Image Caption [22] model and takes Microsoft COCO dataset [23] as input to describe image using text.

Image-to-Image uses the cycleGAN [24] algorithm and takes Cityscapes [25] dataset as input to transform the image to another image.

Speech-to-Text uses the DeepSpeech2 [26] algorithm and takes Librispeech [27] dataset as input to recognize the speech data.

Face embedding uses the FaceNet [28] algorithm and takes the LFW (Labeled Faces in the Wild) dataset [29] or VGGFace2 [30] as input to convert image to an embedding vector.

3D face recognition uses 3D face modes to recognize 3D information within images. The input data includes 77,715 samples from 253 face IDs, which is published on the BenchCouncil web site.

Object detection uses the Faster R-CNN [31] algorithm and takes Microsoft COCO dataset [23] as input to detect objects in images.

Recommendation uses collaborative filtering algorithm and takes MovieLens dataset [32] as input to provide recommendations.

Video prediction uses motion-focused predictive models [33] and takes Robot pushing dataset [33] as input to predict video frames.

Image compression uses recurrent neural networks and takes ImageNet dataset as input to compression images.

3D object reconstruction uses a convolutional encoder-decoder network and takes ShapeNet Dataset [34] as input to reconstruct 3D object.

Text summarization uses sequence-to-sequence model [35] and takes Gigaword dataset [36] as input to generate summary description for text.

Spatial transformer uses spatial transformer networks and takes MNIST dataset [37] as input to make spatial transformations.

Learning to Rank uses ranking distillation algorithm [38] and uses Gowalla dataset [39] to generate ranking scores.

3.3 Application Benchmarks

The suite also provides two end-to-end application benchmarks: DCMix [1]—mixed datacenter workloads, and E-commerce AI—an end-to-end business AI benchmark. Among them, DCMix is to model the datacenter application scenario, and generate mixed workloads with different latencies, including AI workloads (i.e., image recognition, speech recognition), online service (e.g., Online search), etc.

E-commerce AI is to mimic complex modern Internet services workloads, which is a joint work with Alibaba. An AI-based recommendation module is included.

3.4 AI competition

Using the benchmark implementations from AIBench as the baselines, BenchCouncil is organizing the International AI System and Algorithm Competition, advancing the state-of-the-art or state-of-the-practice algorithms on different systems or architecture, like X86, Cambricon, RISC-V, and GPU. This year, there are four tracks, including AI System Competition based on RISC-V, Cambricon, and X86 chips, and Intellifusion 3D Face Recognition Algorithm Competition. The competition information is publicly available from `http://www.benchcouncil.org/competition/index.html`. Any companies and research institutes are welcomed to join and organize a competition track each year.

Among the four tracks., RISC-V and Cambricon-based AI System Competitions are to implement and optimize image classification on RISC-V and Cambricon, respectively. The X86-based AI System Competition is to implement and optimize the recommendation algorithm. The algorithm Competition is to develop innovative algorithms for 3D Face Recognition.

4 Conclusion

This paper proposes a comprehensive datacenter AI benchmarks—AIBench, covering 12 micro benchmarks, 16 component benchmarks, and 2 end-to-end application benchmarks. The benchmark suite is publicly available from `http://www.benchcouncil.org/AIBench/index.html` .

Acknowledgment

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

References

1. X. Xiong, L. Wang, W. Gao, R. Ren, K. Liu, C. Zheng, Y. Wen, and Y. Liang, "Dcmix: Generating mixed workloads for the cloud data center," *BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench18)*, 2018.
2. 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, Gang Lu, Junchao Shao, Zhenyu Wang, Xiaoyu Wang, and Hainan Ye. AIBench: An Industry Standard Internet Service AI Benchmark Suite. Technical Report 2019.
3. C. Coleman, D. Narayanan, D. Kang, T. Zhao, J. Zhang, L. Nardi, P. Bailis, K. Olukotun, C. Ré, and M. Zaharia, "Dawnbench: An end-to-end deep learning benchmark and competition," *Training*, vol. 100, no. 101, p. 102, 2017.

4. "Mlperf," https://mlperf.org.
5. R. Adolf, S. Rama, B. Reagen, G.-Y. Wei, and D. Brooks, "Fathom: reference workloads for modern deep learning methods," in *Workload Characterization (IISWC)*. IEEE, 2016, pp. 1–10.
6. H. Zhu, M. Akrout, B. Zheng, A. Pelegris, A. Phanishayee, B. Schroeder, and G. Pekhimenko, "Tbd: Benchmarking and analyzing deep neural network training," *arXiv preprint arXiv:1803.06905*, 2018.
7. "Deepbench," https://svail.github.io/DeepBench/.
8. S. Dong and D. Kaeli, "Dnnmark: A deep neural network benchmark suite for gpus," in *Proceedings of the General Purpose GPUs*. ACM, 2017, pp. 63–72.
9. Z. Jiang, W. Gao, L. Wang, X. Xiong, Y. Zhang, X. Wen, C. Luo, H. Ye, X. Lu, Y. Zhang, S. Feng, K. Li, W. Xu, and J. Zhan, "HPC AI500: A Benchmark Suite for HPC AI systems," *2018 BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench18)*, 2018.
10. C. Luo, F. Zhang, C. Huang, X. Xiong, J. Chen, L. Wang, W. Gao, H. Ye, T. Wu, R. Zhou, and J. Zhan, "AIoT Bench: Towards Comprehensive Benchmarking Mobile and Embedded Device Intelligence," *2018 BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench18)*, 2018.
11. T. Hao, Y. Huang, X. Wen, W. Gao, F. Zhang, C. Zheng, L. Wang, H. Ye, K. Hwang, Z. Ren, and J. Zhan, "Edge AIBench: Towards Comprehensive End-to-end Edge Computing Benchmarking," *2018 BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench18)*, 2018.
12. W. Gao, J. Zhan, L. Wang, C. Luo, D. Zheng, X. Wen, R. Ren, C. Zheng, X. He, H. Ye *et al.*, "BigDataBench: A scalable and unified big data and ai benchmark suite," *arXiv preprint arXiv:1802.08254*, 2018.
13. 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.
14. Z. Jia, L. Wang, J. Zhan, L. Zhang, and C. Luo, "Characterizing data analysis workloads in data centers," in *2013 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 2013, pp. 66–76.
15. W. Gao, J. Zhan, L. Wang, C. Luo, D. Zheng, F. Tang, B. Xie, C. Zheng, X. Wen, X. He, H. Ye, and R. Ren, "Data Motifs: A lens towards fully understanding big data and ai workloads," *Parallel Architectures and Compilation Techniques (PACT), 2018 27th International Conference on*, 2018.
16. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
17. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.
18. M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.
19. F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.
20. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

21. https://nlp.stanford.edu/projects/nmt/.
22. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 652–663, 2017.
23. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
24. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
25. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.
26. D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.
27. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
28. F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
29. G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.
30. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 67–74.
31. S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
32. F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 5, no. 4, p. 19, 2016.
33. C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Advances in neural information processing systems*, 2016, pp. 64–72.
34. A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.
35. R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang *et al.*, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:1602.06023*, 2016.
36. A. M. Rush, S. Harvard, S. Chopra, and J. Weston, "A neural attention model for sentence summarization," in *ACLWeb. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2017.
37. Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, vol. 2, p. 18, 2010.

38. J. Tang and K. Wang, "Ranking distillation: Learning compact ranking models with high performance for recommender system," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 2289–2298.
39. "Gowalla dataset," https://snap.stanford.edu/data/loc-gowalla.html.