AIBench Inference

Chuanxin Lan
Institute of Computing Technology, Chinese Academy of Sciences

Tutorial on ISCA 2021
Diversity of AI models, AI frameworks, data types
- State-of-the-art trained models

Benchmark Affordability
- Comprehensive workloads are not a burden!
  - Inference time is much shorter

Comprehensive workload behaviors
- A new workload characterization methodology - WPC
- Algorithm, System, Architecture

Representative Inference Mode
Content

- AI Models & Datasets
- Benchmark Affordability
- Workload Characterization
- Inference Mode
AIBench Inference Benchmarks

- Coverage of diverse network architectures（CNN, ResNet, LSTM, GRU, Attention, etc.）
  - Text processing (7)
    - Text-to-Text, Text summarization, Learning to Rank, Recommendation, Neural Architecture Search, Advertising, Nature Language Processing (NLP)
  - Image processing (8)
    - Image Classification, Image Generation, Image-to-Text, Image-to-Image, Face Embedding, Object Detection, Image Compression, Spatial Transformer
  - Audio processing (1)
    - Speech Recognition
  - Video processing (1)
    - Video Prediction
  - 3D data processing (2)
    - 3D Face Recognition, 3D Object Reconstruction
# Models & Datasets

<table>
<thead>
<tr>
<th>Inference Benchmark</th>
<th>Algorithm</th>
<th>Dataset</th>
<th>State-of-the-art Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face Embedding</td>
<td>Facenet</td>
<td>LFW</td>
<td>96 MB</td>
</tr>
<tr>
<td>Object Detection</td>
<td>Faster R-CNN</td>
<td>VOC2007</td>
<td>362 MB</td>
</tr>
<tr>
<td>Image Classification</td>
<td>ResNet50</td>
<td>ImageNet</td>
<td>196 MB</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>DeepSpeech2</td>
<td>Librispeech</td>
<td>315 MB</td>
</tr>
<tr>
<td>Recommendation</td>
<td>Neural collaborative filtering</td>
<td>MovieLens</td>
<td>122 MB</td>
</tr>
<tr>
<td>Spatial Transformer</td>
<td>Spatial transformer networks</td>
<td>MNIST</td>
<td>110 KB</td>
</tr>
<tr>
<td>Video Prediction</td>
<td>Motion-Focused predictive models</td>
<td>MNIST</td>
<td>2.8 MB</td>
</tr>
<tr>
<td>Image to Text</td>
<td>Neural Image Caption Model</td>
<td>Microsoft COCO</td>
<td>132 MB</td>
</tr>
<tr>
<td>Learning to Rank</td>
<td>Ranking distillation</td>
<td>Gowalla</td>
<td>64 MB</td>
</tr>
<tr>
<td>Text-to-Text Translation</td>
<td>Transformer</td>
<td>WMT English-German</td>
<td>183 MB</td>
</tr>
<tr>
<td>Image Compression</td>
<td>Recurrent neural network</td>
<td>ImageNet</td>
<td>136 MB</td>
</tr>
<tr>
<td>Natural Language Processing</td>
<td>BERT</td>
<td>Wikipedia</td>
<td>3.8 GB</td>
</tr>
<tr>
<td>Text Summarization</td>
<td>Sequence-to-sequence model</td>
<td>Gigaword</td>
<td>4.5 MB</td>
</tr>
<tr>
<td>3D Face Recognition</td>
<td>3D face models</td>
<td>77,715 samples from 253 face IDs</td>
<td>100 MB</td>
</tr>
<tr>
<td>Image Generation</td>
<td>WassersteinGAN</td>
<td>LSUN</td>
<td>25 MB</td>
</tr>
<tr>
<td>Image-to-Image Translation</td>
<td>CycleGAN</td>
<td>Cityscapes</td>
<td>110 MB</td>
</tr>
<tr>
<td>Neural Architecture Search</td>
<td>Reinforcement learning</td>
<td>PTB</td>
<td>558 MB</td>
</tr>
<tr>
<td>3D Object Reconstruction</td>
<td>Convolutional encoder-decoder network</td>
<td>ShapeNet dataset</td>
<td>172 MB</td>
</tr>
<tr>
<td>Advertising</td>
<td>DLRM</td>
<td>Kaggle Display Advertising</td>
<td>2.1 GB</td>
</tr>
</tbody>
</table>
Content

- AI Models & Datasets
- Benchmark Affordability
- Workload Characterization
- Inference Mode
<table>
<thead>
<tr>
<th>Inference Benchmark</th>
<th>Accuracy</th>
<th>Inference Time (Second)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face Embedding</td>
<td>0.898</td>
<td>29</td>
</tr>
<tr>
<td>Object Detection</td>
<td>0.753</td>
<td>500</td>
</tr>
<tr>
<td>Image Classification</td>
<td>0.747</td>
<td>504</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>21.8</td>
<td>103</td>
</tr>
<tr>
<td>Recommendation</td>
<td>0.601</td>
<td>6</td>
</tr>
<tr>
<td>Spatial Transformer</td>
<td>0.99</td>
<td>2.1</td>
</tr>
<tr>
<td>Video Prediction</td>
<td>70</td>
<td>698</td>
</tr>
<tr>
<td>Image to Text</td>
<td>4.17</td>
<td>948</td>
</tr>
<tr>
<td>Learning to Rank</td>
<td>0.14</td>
<td>763.8</td>
</tr>
<tr>
<td>Text-to-Text Translation</td>
<td>0.566</td>
<td>2.28</td>
</tr>
<tr>
<td>Advertising</td>
<td>78.9</td>
<td>147</td>
</tr>
<tr>
<td>Natural Language Processing (NLP)</td>
<td>71.3</td>
<td>450</td>
</tr>
<tr>
<td>Text Summarization</td>
<td>0.38</td>
<td>84</td>
</tr>
<tr>
<td>3D Face Recognition</td>
<td>94.7</td>
<td>1047</td>
</tr>
</tbody>
</table>
Content

- AI Models & Datasets
- Benchmark Affordability
- Workload Characterization
- Inference Mode
Workload Characterization

- Traditional Methodology
  - Microarchitecture-dependent Workload Characterization
  - Microarchitecture-independent Workload Characterization
  - ISA-independent Workload Characterization
The higher frontend stall is the inherent characteristic of the scale-out workloads

Microarchitecture-independent workload characterization of 118 benchmarks from six benchmark suites

ISA-independent Workload Characterization

ISA-independent workload characterization of SPECCPU 2006 benchmark suites

The Limitations

- **Microarchitecture-independent/dependent**
  - A higher front-end stall is not the inherent characteristic of the scale-out workloads

- **ISA-independent**
  - Instruction characteristics similarity between two workload implementations at the ISA-independent level does not guarantee their instruction behavior characteristics will be similar at the other two levels
AIBench Inference adopts a new workload characterization methodology to achieve the diversity of workload behaviors

- Whole-picture Workload Characterization (WPC) Methodology
Whole-picture Workload Characterization (WPC) Methodology

The framework of WPC methodology

IR Level
- IR Instrument (e.g., LLVM)
- Intermediate Representation Stream
- Logical Processor (Simulator)
  - No Cache and Pipeline

ISA Level
- Source Code
- Binary Instrument (e.g., Pin)
- Binary Stream Trace
- Perfect Processor
- ISA Level Analysis

Microarchitecture Level
- Runtime + OS
- PMU Tool (e.g., Perf)
- Instruction Stream
- Realistic Processor
  - Cache
  - Pipeline
- Microarchitecture Level Analysis

Whole-picture Analysis
IR Level
- ISA-independent analysis on an IR stream that preserves the source code’s computational logic.
  - an execution-driven simulator to analyze computation, control, and data transfer behaviors through instrumentation-based profiling

ISA Level
- ISA-specific analysis of the binary stream with a run-time binary instrument tool.
  - an execution-driven simulator as a perfect processor

Microarchitecture Level
- Analyzing an instruction stream on a specific processor microarchitectural
  - a performance monitoring unit (PMU) tool
WPC: Is having a higher front-end stall an intrinsic characteristic of scale-out workloads?

The instruction locality of Sort relative to Matmul-C at the LLVM IR, ISA, and Microarchitecture levels.
WPC: What is the implication of WPC for architecture metrics

The symmetry matrices measure the cosine similarity of instruction mixes among different implementations of three workloads at three levels.
WPC for AI-Inference workloads

- WPC-based workload characterization
  - Workload Characterization Across Intermediate Representation, ISA, and Microarchitecture

- Subset
  - The previous benchmark subsetting work that aims to remove redundant workloads collect metrics at only one level
Content

- AI Models & Datasets
- Benchmark Affordability
- Workload Characterization
- Inference Mode
Inference Modes

- **Server Mode**
  - Online server mode: query processing capacity of an inference benchmark
  - Metrics: Latency and tail latency

- **Serverless Mode**
  - Serverless architecture: allocating machine resources on demand
  - Metrics: Service start time, memory consumption, price, throughput, etc

- **Offline Mode**
  - Batch processing: a batch inference process of an inference benchmark
  - Metrics: Inference time and inference quality
Conclusion

- AIBench Inference
  - 19 representative AI tasks and models
  - A new workload characterization methodology
    - Whole-picture Workload Characterization (WPC) Methodology
  - Realistic inference modes
    - Server mode
    - Serverless mode
    - Batch mode
Thank you!