Accelerating Alibaba Transformer model performance with 3rd Gen Intel® Xeon® Scalable Processors (Ice Lake) and Intel® Deep Learning Boost

Wanchen Sui
Minmin Sun
Alibaba Group

Feng Tian
Penghui Cheng
Changqing Li
Pujiang He
Haihao Shen
Intel Corporation

Introduction

3rd Gen Intel® Xeon® Scalable Processors are based on Intel's 10nm+ process technology. They offer more CPU cores, higher memory capacity, and frequency than previous generation. Technologists from Alibaba Group and Intel worked together to explore what these capabilities mean for AI applications, particularly when used with Intel® Deep Learning Boost (Intel® DL Boost). We also explored the Intel® Neural Compressor (formerly known as Intel® Low Precision Optimization Tool), which helps customers rapidly develop and deploy their AI INT8 models on Intel® Xeon® Scalable processor-based platforms. We optimized the Alibaba Transformer model on 3rd Gen Intel® Xeon® Scalable Processors and demonstrated 1.36x and 1.42x performance improvement in FP32 and INT8 inference over Intel's previous generation processors.

Technology Overview

Transformer is a key model used in Alibaba's end-to-end Machine Learning Platform for AI (PAI). It is widely used in real-world natural language processing (NLP) tasks, serving millions of users through Alibaba's online service. Low latency and high throughput are keys to Transformer's success, and 8-bit low precision is a promising technique to meet such requirements.

Intel® DL Boost offers powerful capabilities for 8-bit low precision inference on AI workloads. With the support of Intel® Neural Compressor, we can optimize 8-bit inference performance while significantly reducing accuracy loss. These capabilities demonstrate leadership in AI inference and shows the power of Intel® DL Boost and 3rd Gen Intel® Xeon® Scalable Processors.

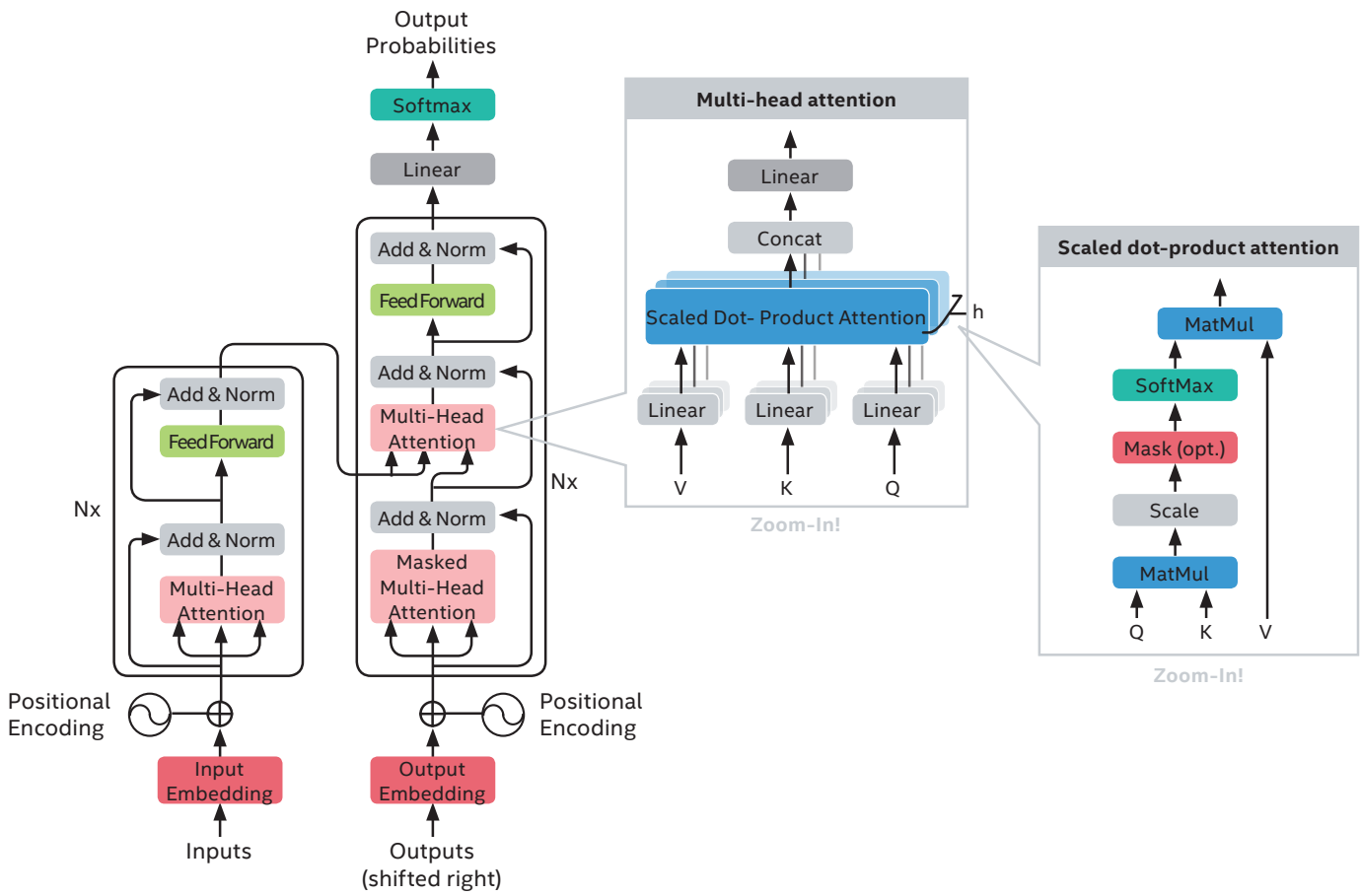

Figure 1. Subgraph of Transformer building block. (Image source: Vaswani, et al., 2017)

Model Analysis

Mode structure

Figure 1 shows the subgraph of the Transformer building block.

According to the graph, some ops are applicable for INT8 quantization to leverage the power of the Intel® Advanced Vector Extensions 512 with Intel® DL Boost Vector Neural Network Instructions (AVX512_VNNI). We utilize Intel® Neural Compressor to automatically generate an INT8 model that meets a predefined accuracy loss goal. Currently, Intel® Neural Compressor supports quantization tuning on the PyTorch imperative path. We rely on the Intel® Neural Compressor to traverse all possible quantization schema tuning space, such as per-tensor, per-channel, asymmetric, and symmetric setting for each quantizable operations. This approach produces an optimized quantized model. The following is a code snippet of using Intel® Neural Compressor to enable the Transformer model.

```
import neural_compressor as nc
from nc.metric import BaseMetric

class Dataset(object):
  def __init__(self):
    …
  def __getitem__(self, index):
    …
  def __len__(self):
    …

# Define a customized Metric function
class MyMetric(BaseMetric):
  def __init__(self, *args):
    …
  def update(self, predict, label):
    …
  def reset(self):
    …
  def result(self):
    …

# Quantize with customized dataloader and metric
from neural_compressor.experimental import Quantization, common
quantizer = Quantization('./conf.yaml')
dataset = Dataset()
quantizer.metric = common.Metric(MyMetric)
quantizer.calib_dataloader = common.DataLoader(dataset, batch_size=1)
quantizer.eval_dataloader = common.DataLoader(dataset, batch_size=1)
quantizer.model = common.Model(model)
q_model = quantizer()
```

For more information about how to use Intel® Neural Compressor to enable the new quantized model, please refer to the Intel® Neural Compressor page on GitHub.

Model pro iling

Alibaba's Transformer model is a PyTorch model. We adopted profiling methodology to analyze the model performance. From the FP32 model profiling log below, we see that model is computation intensive. We see further that 70 percent of total time occupied by computation-intensive ops, such as conv and matmul. This information tells us that the AVX512_VNNI should bring a significant performance speed on the Transformer model. Meanwhile, increased memory bandwidth and frequency of the 3rd Gen Intel® Xeon® Scalable Processors should also deliver benefits for memory-intensive operations.

Name	Self CPU %	Self CPU	...	# of Calls
aten::mm	37.67%	76.644ms	...	331
aten::mkldnn_convolution	29.32%	59.650ms	...	2
aten::copy_	3.84%	7.821ms	...	664
aten::bmm	2.84%	5.779ms	...	144
aten::add_	2.35%	4.791ms	...	331
forward	2.10%	4.277ms	...	1
aten::threshold	1.98%	4.038ms	...	44
aten::_softmax	1.93%	3.922ms	...	72

The INT8 model profiling log below shows if all matmul ops get quantized, the computation speedup would be 76.644/(20.296 + 6.632) = 2.84x. NOTE: Here quantized conv ops get 59.65/11.65 = 5.12x performance speedup, which beyond 4x theoretical peak performance speedup. It is because the fp32 convolution op runs on oneDNN path which in fact includes two extra reorders before and after real convolution computation comparing with int8 convolution FBGEMM op.

Name	Self CPU %	Self CPU	...	# of Calls
quantized::linear	20.27%	20.296ms	...	289
quantized::conv2d_relu	11.64%	11.653ms	...	2
aten::copy_	7.63%	7.636ms	...	664
quantized::linear_relu	6.62%	6.632ms	...	42
aten::bmm	5.91%	5.913ms	...	144
forward	4.20%	4.208ms	...	1
aten::_softmax	3.92%	3.926ms	...	72

Performance and Validation

After measuring the Transformer model on the 2nd Gen and 3rd Gen Intel® Xeon® Scalable Processors we saw impressive performance gains. Tables 1 and 2 show the end-to-end performance improvement for FP32 (Table 1) and INT8 (Table 2).

Test case			2nd Gen Intel® Xeon® Scalable Processors Precision: FP32 Framework: PyTorch 1.7.1 No. of instances: 26	3rd Gen Intel® Xeon® Scalable Processors Precision: FP32 Framework: PyTorch 1.7.1 No. of instances: 32	Performance Gain
Model	Dataset	Batch size	Throughput (Sentences Per Second)	Throughput (Sentences Per Second)	percent
Transformer	Customer Dataset	1	51.32	69.99	36%

Table 1. Transformer Model Performance Gain for FP32

Test case			2nd Gen Intel® Xeon® Scalable Processors Precision: INT8 Framework: PyTorch 1.7.1 No. of instances: 26	3rd Gen Intel® Xeon® Scalable Processors Precision: INT8 Framework: PyTorch 1.7.1 No. of instances: 32	Performance Gain
Model	Dataset	Batch size	Throughput (Sentences Per Second)	Throughput (Sentences Per Second)	percent
Transformer	Customer Dataset	1	155.39	219.96	42%

Table 2. Transformer Model Performance Gain for INT8

Figure 2 shows these results in chart form.

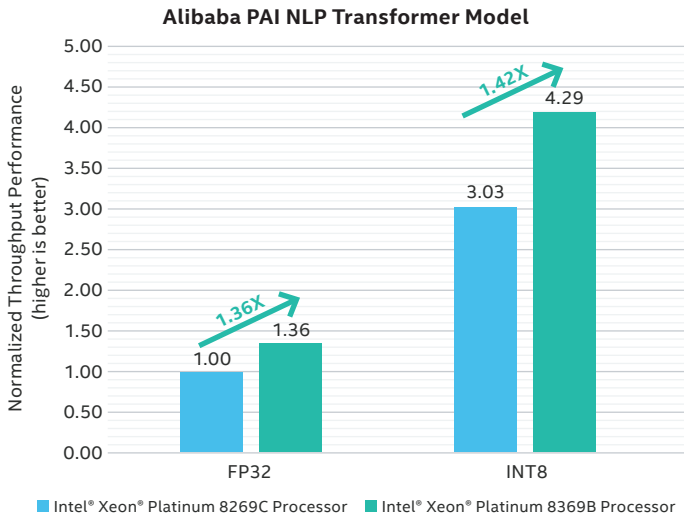

Figure 2. Generational speedups for FP32 and INT8 data types

By leveraging the latest Intel® DL Boost (INT8) Technology, we gained a significant performance increase observing a speedup of approximately 3.1x compared to the FP32 solution. As Alibaba customized the 3rd Gen Intel® Xeon® Scalable platform, the total

throughput performance improved by approximately 42 percent compared to the 2nd Gen Intel® Xeon® Scalable platform.

For accuracy, the INT8 transformer model is validated on customer data and found the accuracy loss is 0.4% which can meet customer needs.

Test case			Cascade Lake Precision: FP32	Ice Lake Precision: INT8	Accuracy loss
Model	Dataset	Batch size	Score (Reported by customer's evaluation program)		
Transformer	Customer Data	1	91.45%	91.05%	0.4%

Conclusion

The 3rd Gen Intel® Xeon® Scalable Processors family increases CPU core number, frequency and memory bandwidth compared to the 2nd Gen Intel® Xeon® Scalable Processors family. These changes brought a 1.42x performance improvement on the PyTorch Transformer INT8 model and a 1.36x performance improvement on the PyTorch Transformer FP32 model. Alibaba partner with Intel's latest CPU and INT8 quantization tool could bring up to 3.1x performance boost on Alibaba PAI blade inference toolkit. Alibaba Cloud expects the processors will help speed up Transformer tasks and provide more efficient services to Alibaba's million+ customers.

Configuration Details

Alibaba PAI NLP Transformer Model on PyTorch 1.7.1 Throughput Performance on 3rd Gen Intel® Xeon® Scalable Processors Family

Baseline Configuration: Test by Intel as of 03/19/2021. 2-node, 2x Intel® Xeon® Platinum 8269C Processor, 26 cores, HT On, Turbo ON, Total Memory 192 GB (12 slots/ 16 GB/ 2933 MHz), BIOS: SE5C620.86B.02.01.0013.121520200651(0x4003003), CentOS 8.3, 4.18.0-240.1.1.el8_3.x86_64, gcc 8.3.1 compiler, Transformer Model, Deep Learning Framework: PyTorch 1.7.1, https://download.pytorch.org/whl/cpu/torch-1.7.1%2Bcpu-cp36-cp36m-linux_x86_64.whl, BS=1, Customer Data, 26 instances/2 sockets, Datatype: FP32/INT8

New Configuration: Test by Intel as of 03/19/2021. 2-node, 2x Intel® Xeon® Platinum 8369B Processor, 32 cores, HT On, Turbo ON, Total Memory 512 GB (16 slots/ 32GB/ 3200 MHz), BIOS: WLYDCRB1.SYS.0020.P92.2103170501 (0xd000260), CentOS 8.3, 4.18.0-240.1.1.el8_3.x86_64, gcc 8.3.1 compiler, Transformer Model, Deep Learning Framework: PyTorch 1.7.1, https://download.pytorch.org/whl/cpu/torch-1.7.1%2Bcpu-cp36-cp36m-linux_x86_64.whl, BS=1, Customer Data, 32 instances/2 sockets, Datatype: FP32/INT8

All performance data is tested in lab environment.

Learn More

- 3rd Gen Intel® Xeon® Scalable Processors
- Alibaba Machine Learning for AI Platform
- Intel® DL Boost
- Intel® Neural Compressor