Comparison of Open Neural Network Exchange (ONNX) and TensorFlow based inferences for the B-scan of interest algorithm

Hugang Ren, PhD[1]; Stefan Duca, MS[2]; Neil D'Souza, MS[1]

[1]Carl Zeiss Meditec, Inc., Dublin, CA, USA; [2]Carl Zeiss Meditec AG, Munich, Germany

Poster #213 - C0050

PURPOSE

- B-scan of interest is a deep learning-based algorithm that aims to improve workflow efficiency for doctors during OCT data review. [1, 2, 3]
- Improving the inference performance can further enhance the user experience and reduce the computation cost.
- ONNX and TensorFlow are two deep learning inference engines. In this study, we compared the inference performance of ONNX and TensorFlow for the B-scan of interest algorithm.

METHODS

- A ResNet-50 neural network was trained using 76,544 OCT B-scans extracted from 598 macular cubes (512x128) acquired from 598 subjects with CIRRUS™ HD-OCT 5000 (ZEISS, Dublin, CA). The trained neural network was then frozen and saved as a protobuf (pb) file for TensorFlow and an onnx file for ONNX.
- TensorFlow 1.13.1 with CUDA 10.0 and cuDNN 7.4 were used for TensorFlow inference. ONNX 1.12 with CUDA 11.4 and cuDNN 8.2.2 were used for ONNX inference. Intel(R) Xeon(R) CPU E5-1620 v3 with 32GB memory was used for CPU based inference and NVIDIA P5000 GPU with 16GB memory was used for GPU based inference.
- The inference performance was assessed using a .NET (C#) based application for both ONNX and TensorFlow.
- To test the inference performance, 25,600 independent B-scans based on 200 macular cubes acquired from 200 subjects at 3 different clinical sites were used as the test set.
- Table 1 shows the detailed information of data used for training and testing and Figure 1 shows the methods on how the comparison was performed.

Inference Performance of ONNX vs TensorFlow for the B-scan of Interest Algorithm

Parameters	Subjects used for training	Volume scans	B-scans for training & validation	B-scans for training	B-scans for validation	Subjects for testing	B-scans for testing
Size/Number	598	598	76,544	61,058	15,338	200	25,600

Table 1. Data used for model training and testing.

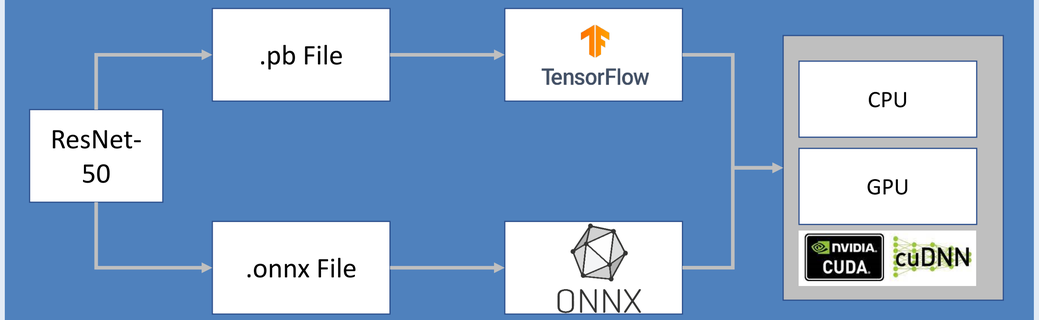

Figure 1. Flowchart comparing ONNX and Tensorflow inference performance running on both CPU and GPU

Mode	TensorFlow (seconds)	ONNX (seconds)	Difference (seconds)	Improvement Percentage
CPU	8.99±0.09	3.57±0.15	5.42±0.17	60.29%
GPU	0.55±0.03	0.35±0.03	0.20±0.03	36.36%

Table 2. Performance comparison of ONNX and TensorFlow based inferences for the B-scan of interest algorithm.

Email: hugang.ren@zeiss.com

Disclosures: HR (E), ND (E) – Carl Zeiss Meditec, Inc.; SD (E) – Carl Zeiss Meditec AG

RESULTS

- In both CPU and GPU modes, the binary (0-normal, 1-B-scan of interest) prediction results for all 25,600 OCT B-scans were identical between TensorFlow and ONNX inferences.
- In CPU mode, the average inference execution times of one macular cube for TensorFlow and ONNX were 8.99±0.09 and 3.57±0.15 seconds respectively. The average difference was 5.42±0.17 seconds and the inference execution time was improved by 60.29%.
- In GPU mode, the average inference execution times of one macular cube for TensorFlow and ONNX were 0.55±0.03 and 0.35±0.03 seconds respectively. The average difference was 0.20±0.03 seconds and the inference execution time was improved by 36.36%.
- Table 2 shows the comparison of the results.

CONCLUSIONS

- ONNX is an open format built to represent machine learning models and support machine learning interoperability.
- In this study, we demonstrated that ONNX can improve the inference execution time of the B-scan of interest algorithm while maintaining the same accuracy in both CPU and GPU modes.
- Future study using latest TensorFlow version can be performed to further compare the inference performance of the two inference engines.

REFERENCES

[1] Ren et al., *IOVS* 2020; 61(7):1635.

[2] Yu et al., *IOVS* 2020; 61(9): PB0085.

[3] Darvishzadeh-Varcheie et al., *IOVS* 2021; 62(8):2450.