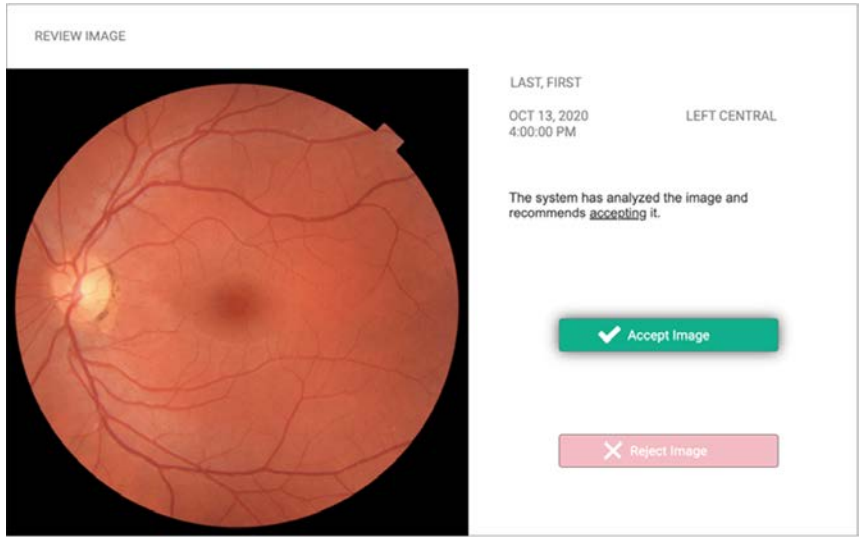
Optimizing inference performance of a fundus image quality neural network model for edge computing using TensorFlow Lite



Hugang Ren, PhD; Simon A. Bello, PhD; Minhaj N. Alam, PhD;
Niranchana Manivannan, PhD; Taylor Shagam, BS; Neil D'Souza, MS

Disclosures

Carl Zeiss Meditec, Inc. (CZMI), Dublin, CA, USA

- Hugang Ren, PhD: CZMI (E)
- Simon A. Bello, PhD: CZMI (E)
- Minhaj N. Alam, PhD: CZMI (C)
- Niranchana Manivannan, PhD: CZMI (E)
- Taylor Shagam, BS: CZMI (E)
- Neil D'Souza, MS: CZMI (E)

Background

- Based on world report on vision[1] from World Health Organization (WHO) published in 2019, globally, at least 2.2 billion people have a vision impairment.

- Among those cases, it is estimated that 3 million people were vision impairment due to diabetic retinopathy (DR) and 146 million people have DR[1]. DR is also the leading cause of vision loss in working-age adults[2].

- Since most of the vision impairment from DR is avoidable through early detection , early screening using tools such as fundus camera has been endorsed[1].

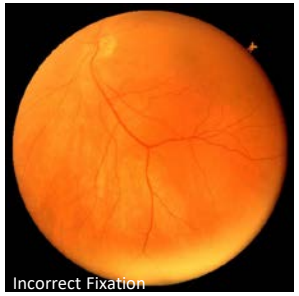
1. https://www.who.int/publications/i/item/9789241516570, ISBN: 9789241516570
2. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. The Lancet 2010;376:124–36. 10.1016/S0140-6736(09)62124-3

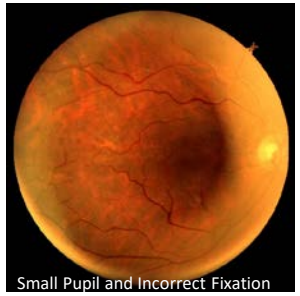
Motivation

- To achieve effective disease screening and diagnosis, acquiring good quality fundus camera images is crucial.

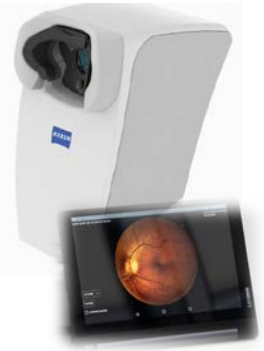

Good Quality


Small Pupil


Incorrect Fixation


Small Pupil and Incorrect Fixation

VELARA 200



- For untrained technicians, a quality indicator after image capture can assist them in the identification of the low-quality images to recapture them while the patient is still available.

- Deep learning-based image quality algorithms have been demonstrated to provide good performance on identifying bad quality fundus images automatically[1].

1. Raj, Aditya, Kumar Tiwari, Anil and Martini, Maria (2019) Fundus image quality assessment : survey, challenges, and future scope. IET Image Processing, 13(8), pp. 1211-1224. ISSN (print) 1751-9659

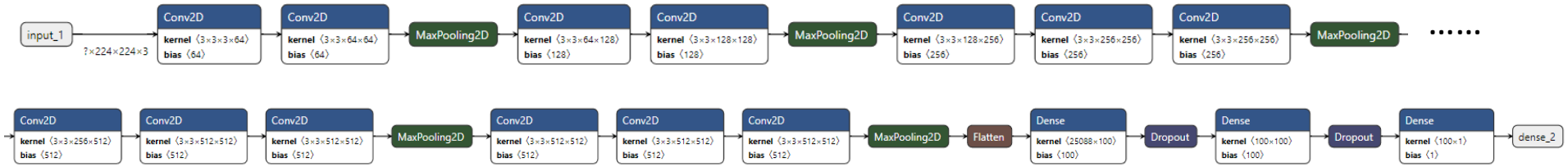
Purpose

- However, running a deep learning image quality algorithm on low-cost fundus cameras can be slow.

- TensorFlow Lite (TFLite) is a deep learning framework designed for inference on the device, also known as edge computing. In this study, we investigated methods to optimize the inference performance using TFLite.

Methods

- A VGG-16 neural network was trained using fundus images captured with **VELARA™ 200** (ZEISS, Dublin, CA), a fully automated non-mydriatic fundus camera with a 45-degree field of view centered around the macula.



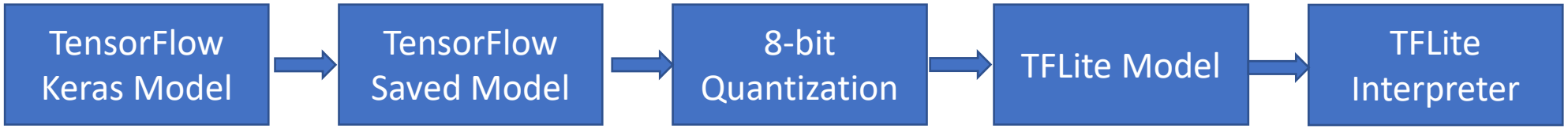
- 4574 images were used for training and 597 images were used for testing. Details can be found in the table below. The grading of the images was performed by a majority vote among 3 subject matter experts.

Category	# of Good Quality	# of Bad Quality	Total # of Images
Training	3158	1416	4574
Testing	353	244	597

VELARA 200

Methods

- First, the floating-point (FP) TensorFlow model was converted and saved as a TFLite file(.tflite). Then, post-training dynamic range quantization was applied using TFLite. The weight of the trained model was quantized into 8-bit integer.



- To test the accuracy, both the FP and quantized TFLite models were evaluated on desktop using the same test set. To test the speed, an Android app was built and installed on the fundus camera tablet that comes with a Qualcomm Snapdragon 439 CPU.

Type	Specification
CPU	Qualcomm® Snapdragon™ 439 octacore
OS	Android 9 Pie™
Memory	4GB
Storage	64GB

Results

- The sensitivity (p=0.5) and specificity (p=0.5) are not statistically different after TFLite optimization.
- Inference speed improved substantially using quantized model on CPU, and the model size reduced by 75% from 67.3MB to 17.8MB.

Model Type	Sensitivity	Specificity	AUC score	CPU speed (ms)		GPU speed (ms)	Model Size (MB)
				1 thread	8 threads		
TensorFlow	85.7% [80.6%, 89.8%]	99.2% [97.5%, 99.8%]	0.976 [0.962, 0.990]	N/A	N/A	N/A	67.3
Floating-point TensorFlow Lite	85.7% [80.6%, 89.8%]	99.2% [97.5%, 99.8%]	0.976 [0.962, 0.990]	4808	1847	1263	67.3
Quantized TensorFlow Lite	85.2% [80.2%, 89.4%]	99.4% [98.0%, 99.9%]	0.977 [0.964, 0.990]	3008	783	1261	17.8

Improvement

- To further improve the performance, a new deep learning model based on SqueezeNet was trained to perform a three-class classification: image quality, small pupil artifact and incorrect fixation artifact[1].

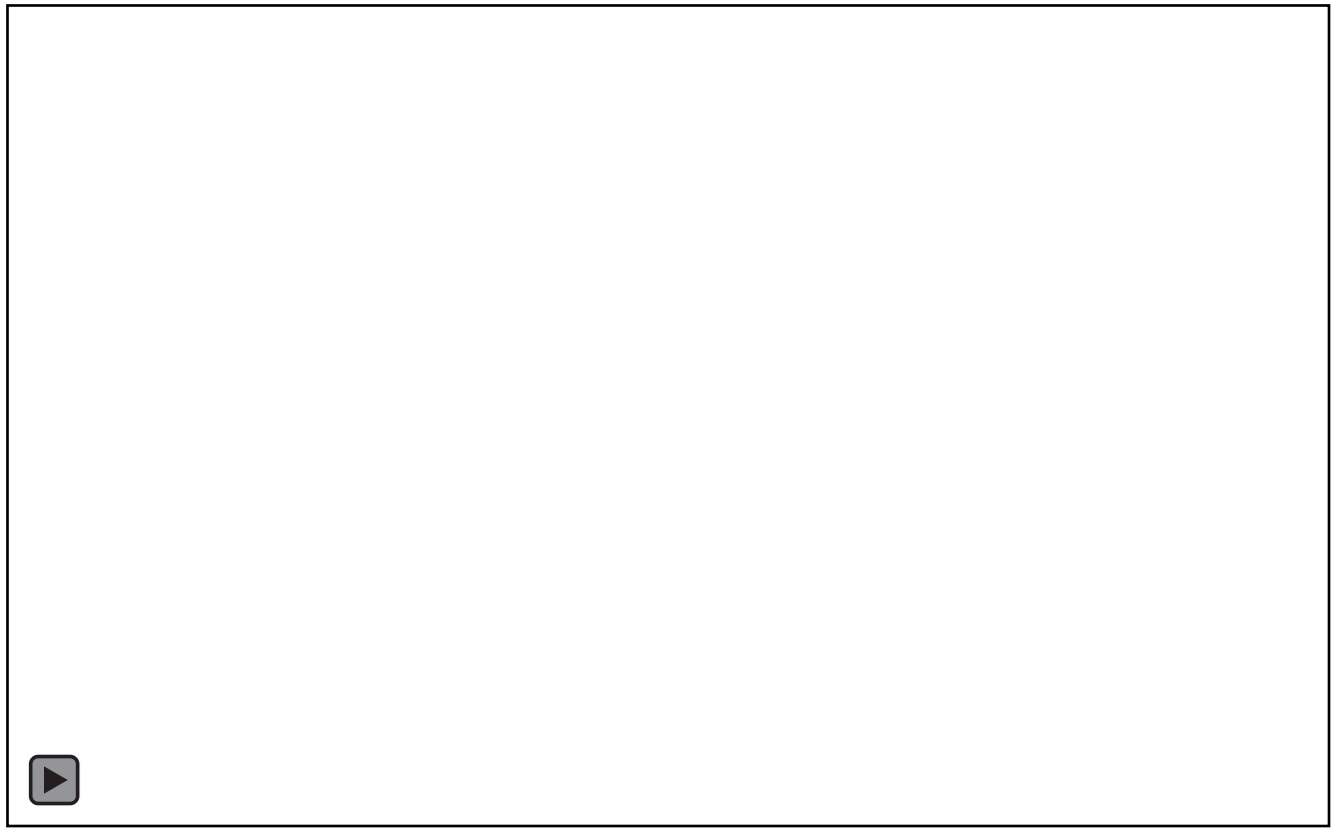
Model Output	Result of 0	Result of 1
Good/Bad Quality	Bad	Good
Small Pupil Artifact	No	Yes
Incorrect Fixation Artifact	No	Yes

- Based on a new development data set with 993 images, by running the new model on the Android tablet, the following performance was achieved.

Model Type	Good/Bad Quality		Small Pupil		Incorrect Fixation		Average Speed (ms)		Size (MB)
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	1-thread	8-thread	
TensorFlow	91.60%	91.43%	93.83%	95.67%	85.71%	93.42%	N/A	N/A	**13.1**
Floating-point TFLite	92.85%	90.57%	93.21%	95.67%	85.71%	93.22%	1462.31	848.93	**13.1**
Quantized TFLite	90.67%	91.43%	93.83%	95.55%	85.71%	93.11%	1438.52	978.31	**4.0**

1. Omlor et. al: Image quality and artifact feedback for fundus imaging using edge device for teleretinal application, 2021 ARVO Imaging in the Eye Conference, 3571361, May 14th, Friday, 10:15-11:15

Recorded Demo App on an Android Tablet



Conclusions

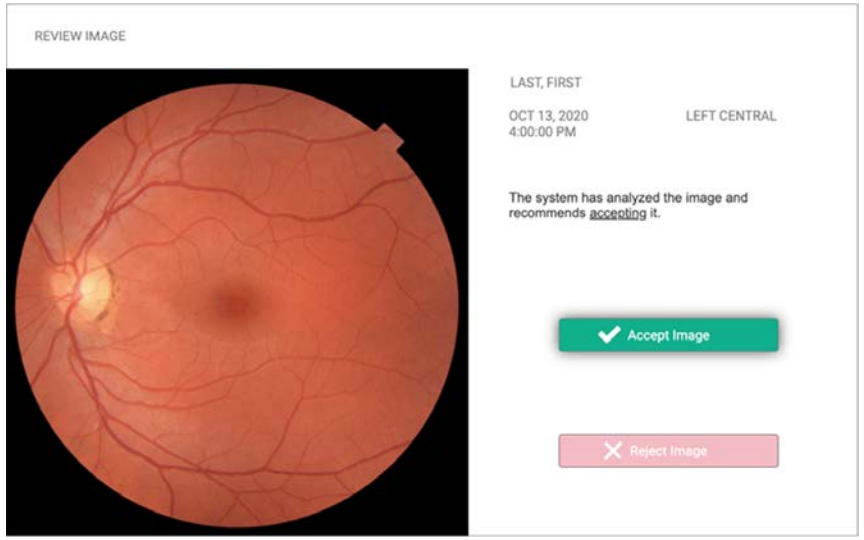
- In this study, we demonstrated a method to optimize the inference speed of a fundus image quality neural network model for edge computing using TFLite.

- The optimized TFLite model was successfully integrated into an Android app running on a low-cost fundus camera tablet.

Acknowledgement

- Dr. Dorothy Hitchmoth
- Dr. Bryan Rogoff
- Dr. Patty Sha

Optimizing inference performance of a fundus image quality neural network model for edge computing using TensorFlow Lite



Contact: hugang.ren@zeiss.com