# Fast Inference on FPGAs and TPUs

## Classifying Gravitational Lensing Images Efficiently

Callista Christ, University of Chicago
LSST Data Science Program at Fermilab

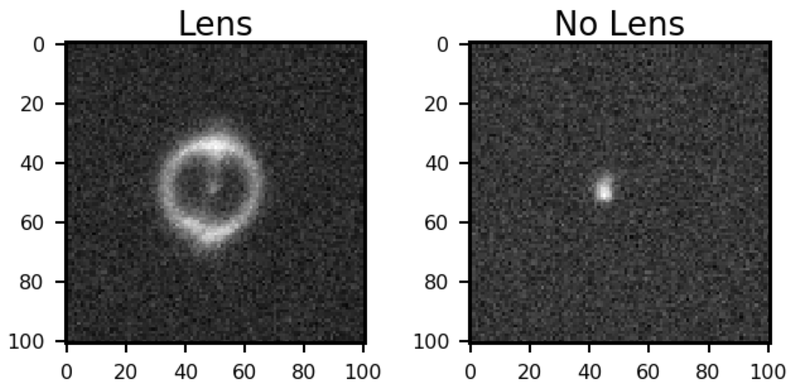
FERMILAB-POSTER-19-105-CD

## Summary

Our goal for this summer is to use various neural network models to classify gravitational lensing in simulated LSST images. We focus on maximizing the efficiency and accuracy of our models for training and testing using fast inference techniques on FPGA and TPU hardware.

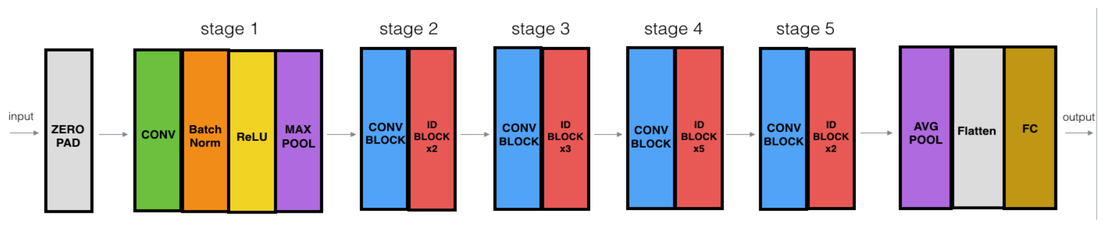
## 1. Data

Our data consists of simulated lensing and non-lensing images in multiple bands (as expected to be observed by LSST) from a Kaggle competition. An example of our data is shown below.




## 2. Architecture and Hardware

**Architecture:** CNNs (Convolutional Neural Networks) are deep learning algorithms that can efficiently analyze image data. We use a type of CNN called ResNet50 (architecture is depicted below) on a cloud computing platform called Microsoft Azure.




**Hardware:** For inference (testing) our model, we use a FPGA (Field Programmable Gate Array). FPGAs are able to be programmed by the user and allow for extremely fast inference which is ideal for LSST data.

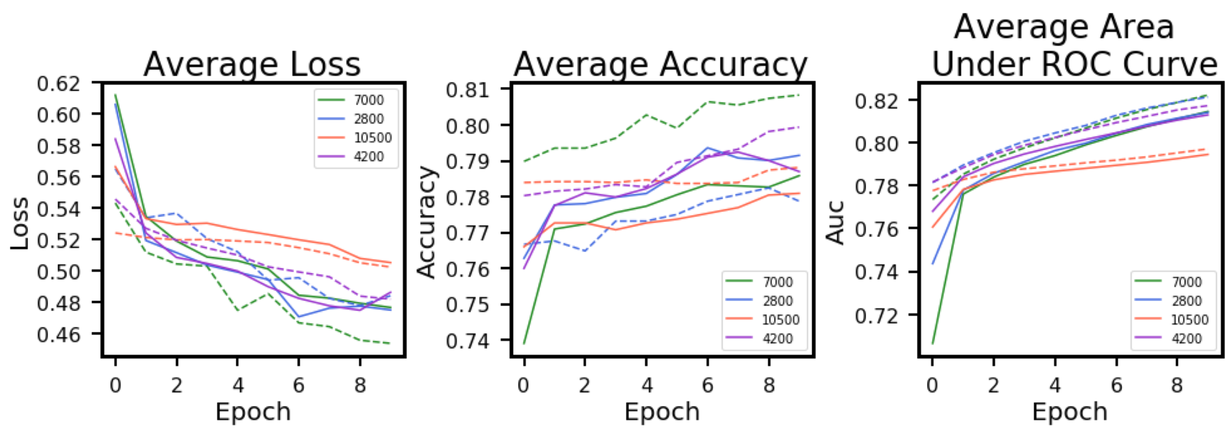
## 3. Method

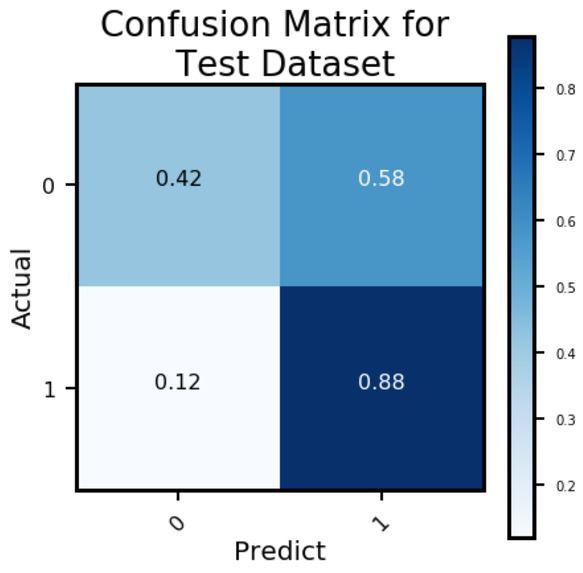
We train our ResNet50 model for 5 epochs, with each epoch taking ~1 minute to run on a GPU. We preprocess the images by upscaling them from 101x101 pixels to 224x224 pixels. We then feed the images into our model in batches of 20 images and classify the image as a lens if the number of pixels of the object is above a certain threshold.

**Acknowledgement:** I would like to thank my advisor Brian Nord and the LSST Data Science Program at Fermilab for their generous support throughout this summer internship.

## 4. Preliminary Results

We trained our neural network as aforementioned but instead of using the whole dataset (2 million images) we use a subset of 20,000 images, with 15,000 used for training and 5,000 used for testing**. The resulting accuracy for the training set is ~81% and for the testing set is ~74%**. How the model performs as a function of the training set size is displayed in the figure below (where the solid lines are for the training data and dashed lines for validation data).




The leftmost figure is a confusion matrix relating our predicted labels to the actual labels for testing (5,000 images). Our model performs well at classifying the lensed images (labeled as “1”, 88%), but struggles to classify the non-lensing images (labeled as “0”, 42%).

By increasing the size of the training/testing datasets, running the model for more epochs, and pre-processing the images by scaling/clipping them, we hope to improve the overall accuracy of our model to over 90%.

## Future Work

**Individual Goal:** My next step will be to train my model using all of the available images for at least 100 epochs. Once my model is trained well, I will write code to access the FPGA for inference and test the overall accuracy and speed of my model.

**Group Goal:** We will test various model/hardware combinations (i.e. MobileNet model on a TPU) with the intention of performing fast inference at or near LSST with our most accurate/efficient combination.

This document was prepared by Deep Skies Collaboration (deepskieslab.com) using the resources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359.