Cervix Type Detection Kaggle Challenge for Cervical Cancer Screening

By Jack Payette, Jake Rachleff, and Cameron Van de Graaf

Problem

The problem that we set out to solve is that of cervix type classification. Intel and MobileODT have teamed up to create a Kaggle competition for classifying cervixes into three classes. This problem is important because healthcare providers are unable to screen for and treat potentially life-threatening cervical cancers if they are unable to classify the cervix type. While healthcare providers in the developed world are skilled at this classification, those in the developing world often lack the necessary time and expertise. We set out to build a classifier from the Kaggle Dataset that would help healthcare providers in low resource areas better classify cervix types, and in turn help them better administer health care services to women in need.

Data Set

Our dataset, which was provided by Kaggle, consists of 6113 training images and 512 test images. The training set consists of 1438 images of Type 1, 2339 images of Type 2, and 2336 images of Type 3. Because submissions go to Kaggle, we do not know the underlying distribution of the test data, but we assume it to be an even distribution. Because our dataset is small and not uniformly distributed, we used a number of data augmentation techniques to prevent our model from predicting the training distribution. Further, because images were of many different sizes, we used extensive resizing in order for our models to fit into memory. The small size of the dataset has limited the possible depth of our model, and caused us to struggle with overfitting even with the data augmentation.

Note: Due to the graphic nature of the images in the dataset, which some audience members may find offensive. we've chosen not to include them on this poster. We have used alternative images to demonstrate augmentation techniques.

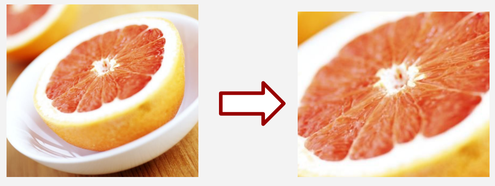
Data Processing

Per Class Equalization

Our dataset initially had imbalanced numbers of examples in each class. Therefore, we sampled disproportionately from less represented classes to stop models from just learning the underlying distribution of training classes

Image Auto Cropping

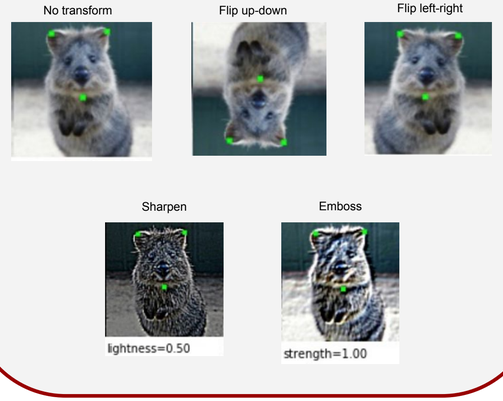
We used a previously developed auto-crop algorithm for identifying cervix features. The general process is demonstrated below, as the algorithm identifies the area of interest, and crops it to be featured. This helps increase consistency in trained features



In the example above, the picture of a grapefruit on a plate is cropped to only feature the pulp of the grapefruit

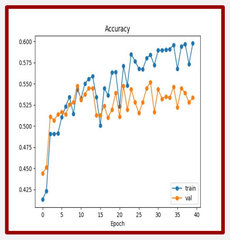
Image Augmentation

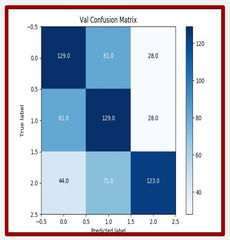
For image augmentation, we took each of our training images and with random probability performed transformations seen below:



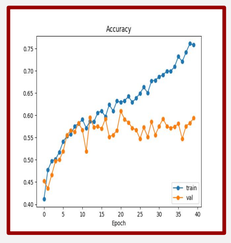
Results

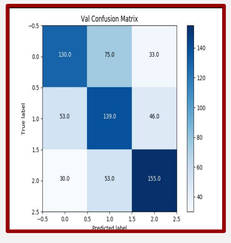
Softmax Classifier



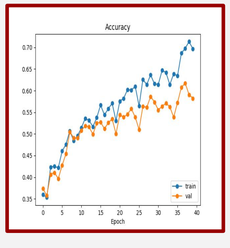


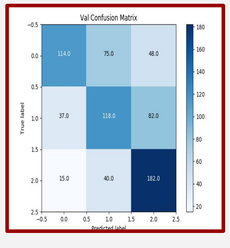
6 Layer Vanilla CNN



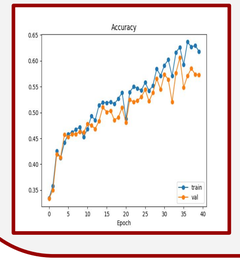


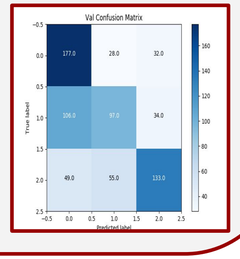
32 Layer Residual CNN





101 Layer Residual CNN



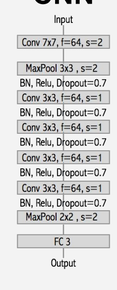


Models

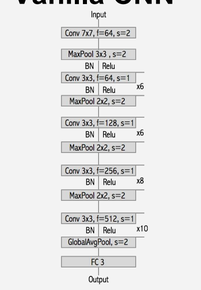
Softmax Classifier

Wx+ b

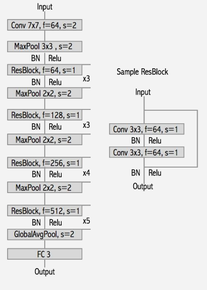
6 Layer Vanilla CNN



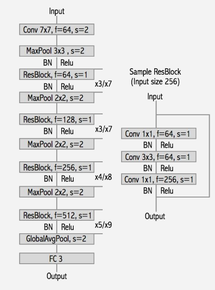
32 Layer Vanilla CNN



32 Layer Residual CNN



53/101 Layer Residual CNN



Next Steps

While our current results are definitely promising, we know that we can get a boost for our final submission by going through rigorous hyperparameter tuning. Specifically, we want to continue to tune learning rate, data augmentation, and data normalization techniques.