Phase 3: Model Training, Part 2

Welcome to Phase 3 of the capstone project. This section will be the second of two parts that concerns the model training process of the model development cycle. You continue to play the role of a bioinformatics professor. The questions will relate to the various challenges faced by the teams working on the two projects introduced in the first section.

You have made recommendations (based on your answers in the prior phase) to both of the research teams. They have taken your suggestions into account, and have since refined their results. Both teams have e-mailed you summaries of recent progress, which are shown below.

Project 1: CXR-based COVID-19 Detector

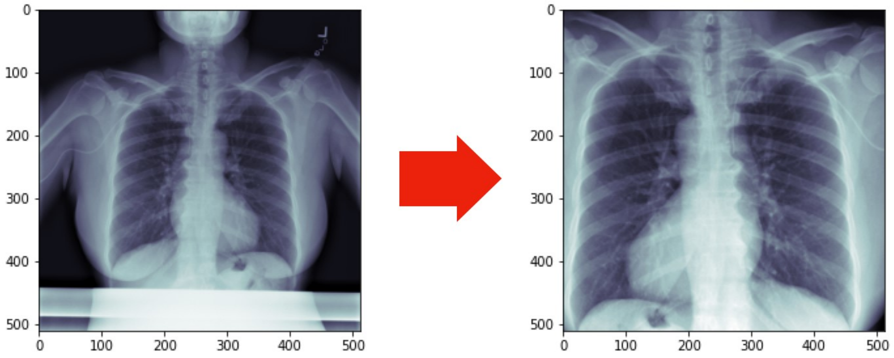
Hi,

Thank you for your excellent feedback. We are now facing the opposite problem- our model is now memorizing the training data and failing to generalize to new, unseen data. As a recap, below are the changes we’ve implemented since our last check-in.

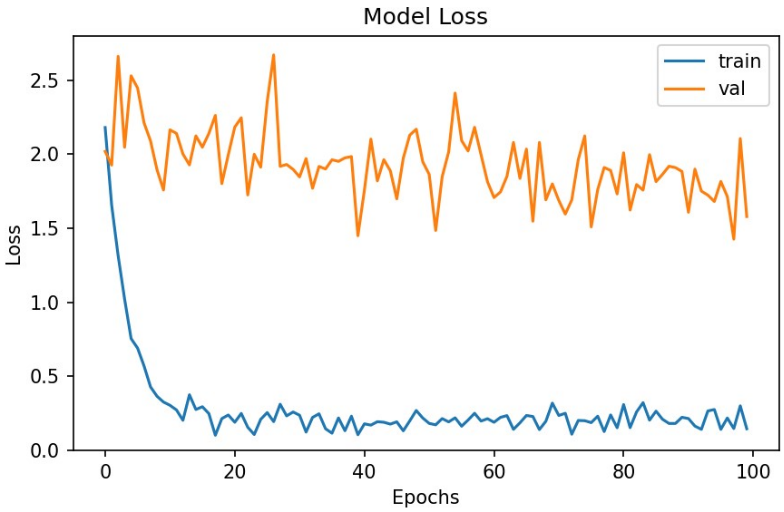
We re-split the data into a training, validation, and test set. We are placing 80% of the data into the training set, 10% of the data into the validation set, and 10% of the data into the test set. We split the data by patient this time, to prevent patient overlap. We are now evaluating the model using the validation set.

We tried out your suggestion to upsize the images from 224 by 224 pixels to 512 by 512 pixels in order to retain some of the fine-grained resolution while keeping the memory requirements manageable. We adapted the first few layers of the model architecture to accommodate for this change.

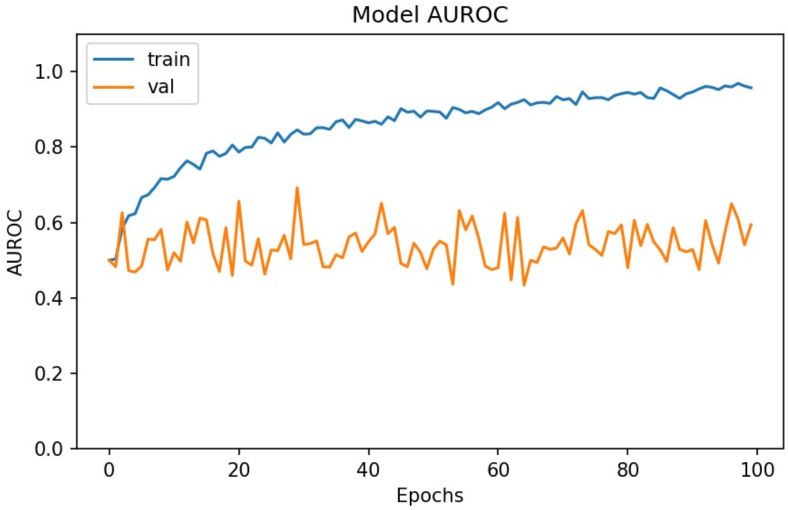
We eased back on the data augmentation. Now we do a simple horizontal flip and incorporate only a slight amount of zoom.



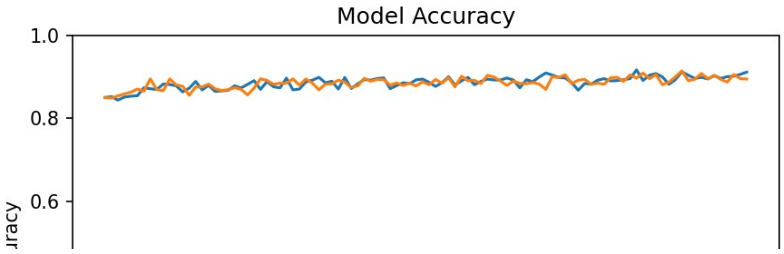
Here are our new training curves from our model. Per your recommendation we’ve become oversampling the COVID-positive exams in the training set. It was helpful, but we’re starting to see some real learning occurring. However, as you can see, the loss for the training set is now far lower than that of the validation set.

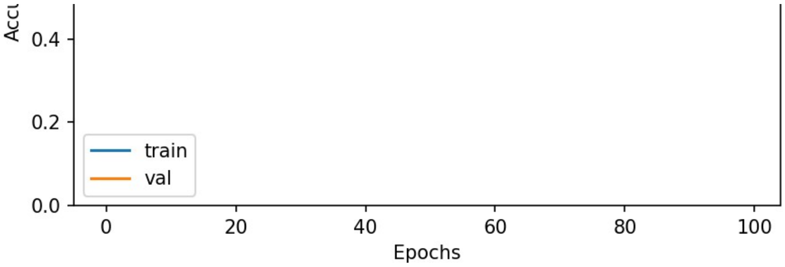


Now that our model is training, we are tracking both the AUROC and accuracy of the model during training on the training and validation sets. Here is the model's AUROC over time:



On the epoch where the model achieves the highest validation set AUROC, we see a 0.846 AUROC on the training set and 0.692 AUROC on the validation set. However, when we visualize the accuracy of the model, we get a very different story:





On the epoch where the model achieves the highest validation set accuracy, the model attains an accuracy of 0.912 on the training set and 0.914 on the validation set. We’re not sure why its accuracy is so high. We double-checked the code and there don’t seem to be any bugs in the program.

The model is certainly performing better than it was before, and I think there are still some bugs to work out. Let me know if you have any suggestions, thanks.

Project 2: EHR-based Intubation Predictor

Hi,

Thank you for your guidance– your suggestions were much needed and have allowed us to make significant progress.

We are now using the 40,000 exams from the “COVID-like” dataset as our training and validation sets. We are using the 3,000 exams from the COVID dataset as our test set.

Specifically, we are splitting the “COVID-like” dataset such that 70% of exams are in the training set and the remaining 30% are in the validation set. We are planning on using 10-fold cross validation on the training set in order to choose the best hyperparameters. Once we have those, we plan on training the model on the full training set with early stopping in order to produce our final model.

We are training both logistic regression models and random forest models. As always, let us know if you have any feedback or questions, thanks!

In [ ]: