Deep Learning with MURA

Shiye Cao

Dataset: The MURA (musculoskeletal radiographs) dataset provided by Stanford University School of Medicine consists of 14,863 studies from 12,173 patients, with a total of 40,561 multi-view radiographic images (each manually labeled as normal or abnormal by radiologists from the Stanford Hospital and separated into 7 groups based on the body part: elbow, finger, forearm, hand, humerus, shoulder, and wrist).

Due to the limited time constraint, I chose to work with only 2 datasets (finger-4708 images and wrist-6974 images). I pre-processed the data by picking out the images that showed more than one x-ray at a time. I also randomly picked out 75 normal images and 75 abnormal images from the MURA dataset to use as the validation dataset.

Clinical Significance: Today, musculoskeletal conditions affect more than 1.7 billion people worldwide, and are the most common cause of severe, long-term pain and disability, with 30 million emergency department visits annually and increasing.[1] I hope that through deep learning, my model can make diagnosis beyond the level of current expert radiologist. That way, in the future, the model can be used to help radiologist make better judgements on whether an X-ray study is normal or abnormal.

Below is a chart with the accuracy rate of 3 radiologists from the Stanford Hospital[2]:

	Radiologist 1	Radiologist 2	Radiologist 3
Wrist	0.791	0.931	0.931
Finger	0.304	0.403	0.410
Overall (all 7 parts)	0.731	0.763	0.778

Experiments: To find the best deep learning model, I experimented with 3 different models including: ResNet18, DenseNet161, and VGG16. I have also experimented with various other factors including: Convolution Neural Net, the number of epochs, and Pre-training. I chose to not test the effects of data augmentation on the dataset since there is enough variation in the dataset with images from all kinds of different view point, although images with more than one X-ray in the frame was taken out of the dataset during pre-processing.

1. Wrist: With ConvNet vs. Without ConvNet, Epoch = 25

Model	Training	Validation
ResNet18 with ConvNet	Train loss: 0.3509 Acc: 0.8508	Val loss: 0.1135 Acc: 0.9600
ResNet18 without ConvNet	Train loss: 0.0754 Acc: 0.9762	Val loss: 0.9762 Acc: 1.0000

Better Model: Without Convolutional Neural Network

---

[1] https://stanfordmlgroup.github.io/competitions/mura/

[2] https://stanfordmlgroup.github.io/competitions/mura/

2. Wrist: With 25 epochs vs. with 50 epochs

Model	Training	Validation
ResNet18 with 25 epochs without ConvNet	Train loss: 0.0754 Acc: 0.9762	Val loss: 0.9762 Acc: 1.0000
ResNet18 with 50 epochs without ConvNet	Train loss: 0.0709 Acc: 0.0235	Val loss: 0.9787 Acc: 0.9933

Better Model: With 25 epochs

3. Finger: Pre-trained vs. no Pre-training, Epoch 25

Model	Training	Validation
Pre-trained ResNet18	Train loss: 0.5063 Acc: 0.7421	Val loss: 0.6131 Acc: 0.7267
Not Pre-trained ResNet18	Train loss: 0.6156 Acc: 0.6521	Val loss: 0.6873 Acc: 0.6133

Better Model: With Pre-training

4. Wrist: Pre-trained with ImageNet, Without ConvNet, Epoch = 25

Model	Training	Validation
ResNet18	Train loss: 0.0754 Acc: 0.9762	Val loss: 0.9762 Acc: 1.0000
DenseNet161	Train loss: 0.0625 Acc: 0.9805	Val loss: 0.0130 Acc: 1.0000
VGG16	Train loss: 0.0659 Acc: 0.9808	Val loss: 0.0105 Acc: 1.0000

Better Model: VGG16

5. Finger: Pre-trained with ImageNet, Without ConvNet, Epoch = 25

Model	Training	Validation
ResNet18	Train loss: 0.5063 Acc: 0.7421	Val loss: 0.6131 Acc: 0.7267
DenseNet161	Train loss: 0.4734 Acc: 0.7669	Val loss: 0.5640 Acc: 0.7467
VGG16	Train loss: 0.4593 Acc: 0.7777	Val loss: 0.5460 Acc: 0.7667

Better Model: VGG16

Final Model: VGG16 Model without ConvNet and with pre-training and 25 epochs.

Conclusion: The deep learning model performed better than the radiologists on both the wrist and the finger dataset. So, I see a value in generalizing the model to all 7 datasets and together it will likely become a computer radiologist that performs better than the current human radiologists.