Athens University of Economics and Business
Department of Informatics
Master of Science in Computer Science

Deep Learning
Instructor: P. Malakasiotis

MURA Dataset (Binary Classification)

Panagiotis-Christos Kyrmpatsos (f3322204)
Nikos Nikolaidis (ds3322212)

July 18, 2023

1 Dataset Exploration

This project is part of the Deep Learning course. We will explore different techniques to acquire a high score in the task of abnormality detection based on the MURA Dataset. The MURA (Musculoskeletal Radiographs) dataset is a popular benchmark dataset in the field of deep learning for musculoskeletal imaging. While it provides a valuable resource for researchers, it also presents certain challenges that need to be carefully addressed. We mention three key difficulties encountered when working with the MURA dataset and discusses their implications for our project. Our code can be found on Kaggle and Github.

1.1 Limitations of the Dataset

One of the primary challenges in the MURA dataset is its limited size and imbalanced distribution of images across different classes. The dataset consists of approximately 40,000 radiographs from diverse musculoskeletal exams, making it relatively small compared to other medical imaging datasets. Additionally, the distribution of images is highly skewed, with a significant class imbalance observed in some categories. This scarcity of data and class imbalance pose challenges for training deep learning models effectively. It can lead to biased predictions, poor generalization, and difficulty in capturing rare abnormalities or conditions. Figure :1 illustrates said imbalance.

Another challenge in the MURA dataset is the inherent annotation variability and subjectivity associated with musculoskeletal radiographs. The dataset is labeled by expert radiologists, and while their expertise is invaluable, there can still be inconsistencies and discrepancies in labeling due to the complexity and interpretive nature of musculoskeletal imaging. Variations in positioning, image quality, and the subjective nature of identifying abnormalities contribute to annotation variability and ultimately to the task difficulty.

Lastly, the challenge of processing images in the MURA dataset lies in reconciling the variations in size, orientation, color representation, and other factors, to ensure a standardized and homogeneous dataset that can be effectively utilized for deep learning tasks. Take for example the two very different examples pictured in Figures 3 and 4.

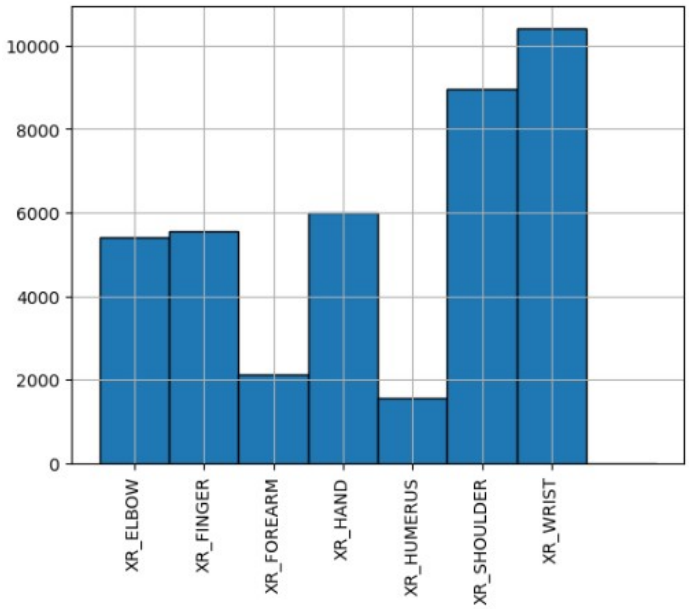

Figure 1: Differences in occurrence of each body part

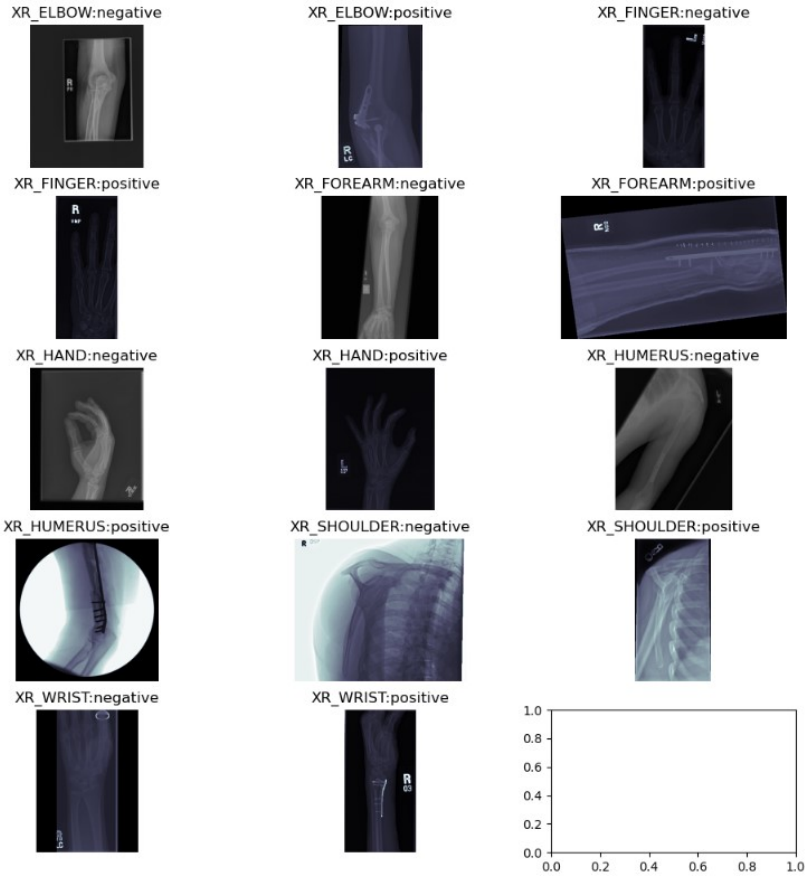

Figure 2: A sample of positive and negative items for each body part

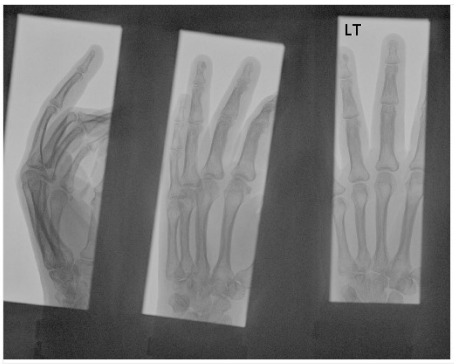

Figure 3: Three Fingers

Figure 4: One Arm

2 Preprocessing

2.1 Image Representation

Harmonizing the color representation across the dataset ensures consistent and meaningful feature extraction. Images in the dataset came in both RGB and Grayscale format. To ensure homogeneity of our samples we converted all images to RGB. We also resized all images to 256 x 256 size. Also, the values are first rescaled to [0.0, 1.0] and then normalized using mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225]. These values were chosen to fit out samples to our pretrained model. Furthermore, we used random auto contrast and Random horizontal flip transformations on the training dataset. Both transformations were applied with a %50 probability.

2.2 Transfer Learning

We chose ResNet as a pre-trained model to further fine-tune for the MURA dataset offers several distinct advantages. Firstly, ResNet has demonstrated exceptional performance in various computer vision tasks, including image classification, object detection, and segmentation. Its deep architecture, featuring residual connections, enables effective learning of intricate features and helps alleviate the vanishing gradient problem during training. Secondly, ResNet has been pretrained on large-scale datasets like ImageNet, which enables it to capture a wide range of generalizable visual features. By leveraging this pretraining, we can effectively transfer knowledge from the domain of natural images to the musculoskeletal radiographs in the MURA dataset. This initialization with learned parameters provides a head start for training, reducing the overall training time and resource requirements.

3 Experiments

3.1 Pretrained frozen Resnet18 and Linear Layer

The experiment involving fine-tuning only a linear layer on top of ResNet18 and training for 10 epochs resulted in an accuracy of 69.51%. This outcome suggests that the fine-tuned model was able to capture certain features relevant to the MURA dataset. This initial result 5 highlights the potential effectiveness of leveraging transfer learning with ResNet18.

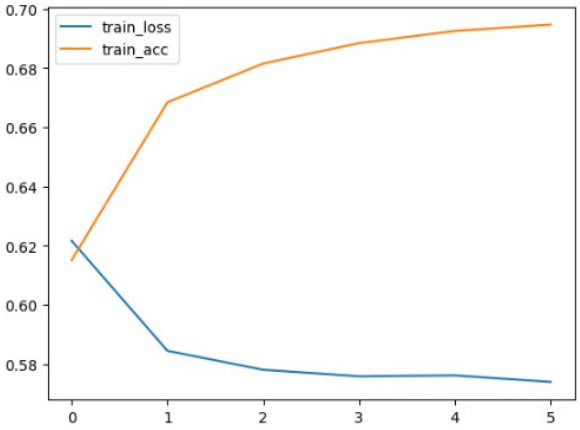

Figure 5: ResNet18+Linear Layer

3.2 ResNet18 and two Linear Layers

Since the loss in our previous model quickly plateaued we tried to augment the capacity of our model by inserting a second linear layer of dimensions 512 x 512 on top of the frozen ResNet18 model. By doing this , our model now is ResNet18+Linear(512,512)+Linear(512,1). The increased capacity proved effective in raising the test accuracy to 72,17%. Figure 9 illustrates the progress of our model during training for 10 epochs using Adam and learning rate=0.001.

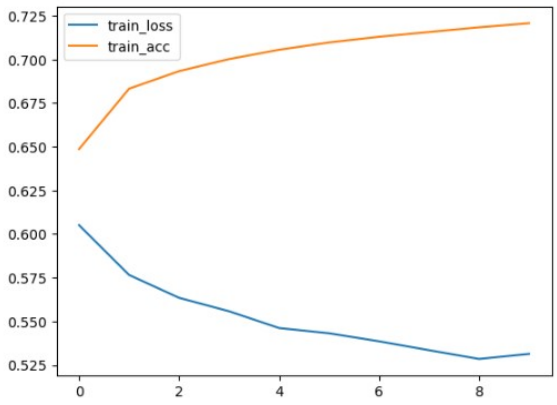

Figure 6: ResNet18+Linear+Linear for 10 epochs

3.3 Unfrozen ResNet18 with intermediate task and two Linear Layers

Since the last two cases ignored the additional information stored in the "BONE" column, indicating the type of Bone each image contains, we tried to test an approach that includes an intermediate training task.

We unfreeze the *RESNET* weights and train using the BONE attribute as a label, in a 6-label multi-class task. Here we use Macro F1 score as or accuracy metric. Then, we freeze the *RESNET* weights and we throw away the last linear layer, and we train a binary classification problem on the *"Label"* attribute.

The purpose of this approach is to force the model to create internal representations of latent features present in our dataset before training in our ultimate task. Also, this can serve as a way to introduce auxiliary information about the current image that could potentially assist classification.

3.4 ResNet18+Linear+Linear only on humerus X-Rays

Since our previous approaches did not come to fruition we tried to train separately our model for only one bone category, the humerus. For our training we used Adam with learning rate=0.01 for training of 1000 steps after which we achieved an accuracy of 79.36% on our test data. We followed by further pretraining the same model using a learning rate of 0.001 by which we finally achieved an accuracy of 88.92% on the humerus test data. In the following figure we observe the loss and accuracy curves for the training and development datasets.

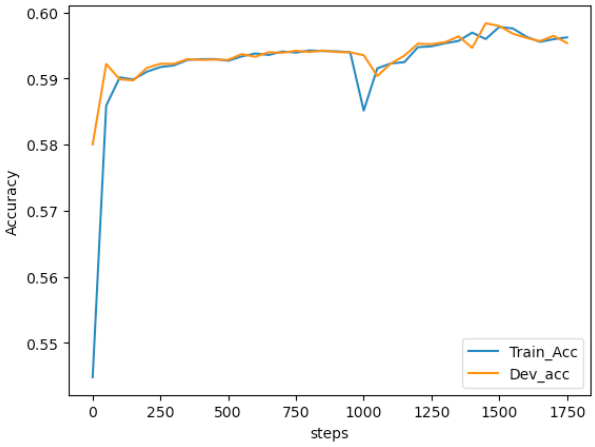

Figure 7: Unfrozen RESNET with intermediate-task

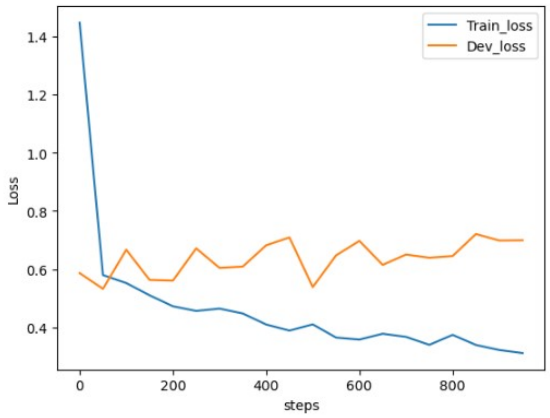

Figure 8

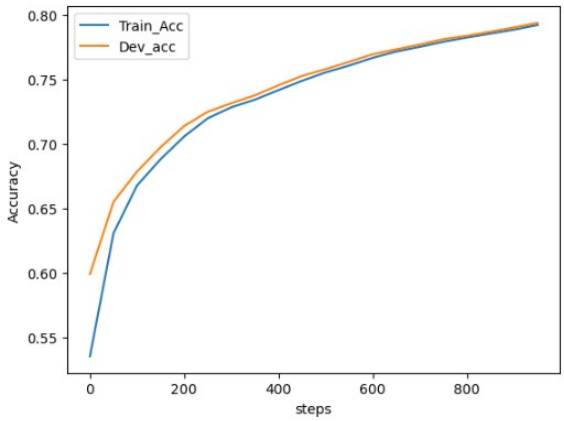

Figure 9: CCC

4 Difficulties and solutions

The first object that we had to overcome came in the form of finding a suitable platform to run our experiments. After unsuccessful experimentation with Google's colab we resorted to using Kaggle for experiments mainly due to the ability to save our data independently of the kernel.

Another problem we faced was the difference in the number of samples regarding each bodypart. To compensate for this obstacle, we performed stratified sampling on our training data. Moreover, since development data were unavailable (dev data were used as test), we created them using 10% of our training dataset.

We also took advantage of early stopping techniques to gauge when our training became counterproductive. Most of the time we used the value

of patience=4, ie wait for 4 epochs to check whether validation Accuracy improves.

5 Conclusion

We conducted three different approaches in our classification task. Initially, we attempted to fine-tune a pretrained model using our entire dataset. Although this approach provided satisfactory results, it did not achieve state-of-the-art performance. To improve our approach, we focused on enhancing the feature representation by splitting our task into two parts: a multiclass classification for identifying body parts and a binary classification task. During training, the loss consistently decreased for the unfrozen RESNET model in the multiclass classification task. However, when it came to the binary classification task with the frozen RESNET layer, the training started at 40% accuracy and did not exhibit a significant improvement.

The training process for this experiment concluded with an accuracy score of 59.1%. Evidently, the representations and additional information embedded into our pre-trained model for the multiclass classification did not effectively capture relevant and useful information for our binary task.

Lastly, we explored an alternative approach by separately training each bone class. This strategy yielded the most promising results and achieved performance on par with state-of-the-art methods for the same task. Taking everything into consideration, we assert that a system that first identifies the bone and then independently trains on images from each bone class is the most effective approach for our task at hand.