Recurrent and Attention-Based Approaches to Multiple Instance Learning for Musculoskeletal Abnormality Detection

George Horrell
Stanford University
ghorrell@stanford.edu

Vasco Portilheiro
Stanford University
vascop@stanford.edu

Bora Uyumazturk
Stanford University
yuyumaz@stanford.edu

Abstract

Deep convolutional neural networks are applied to the problem of abnormality detection in musculoskeletal X-rays of the lower extremities. Methods build on prior medical image classification approaches by proposing more sophisticated solutions to the problem of Multiple Instance Learning. Two approaches in particular, one based on recurrent neural networks and the other on Attention models, are shown to outperform static pooling methods. General performance approaches radiologist-level abnormality detection on certain body parts.

1. Introduction

Musculoskeletal diseases are extremely common, affecting more than 1.7 billion people worldwide, and accounting for more than 30 million emergency room visits annually ([12]). With this in mind, we have chosen to apply computer vision techniques to the task of detecting abnormalities in musculoskeletal X-rays of the lower extremities. Manual classification is labor intensive, requiring the attention of a trained radiologist. In developing countries this can be a major roadblock to delivering urgent care (the country of Liberia has just two radiologists, for example). A reliable automated system would reduce the workload of radiologists and hospital staff and improve medical efficiency. Recent results in papers such as [9] and [10] by Rajpurkar et al. which attempt classification on upper extremities and chest X-rays, respectively, suggest that accuracy comparable to that of trained radiologists is within reach for such tasks.

In this paper we attempt to replicate such results but on a different dataset, this time consisting of musculoskeletal X-rays of the lower extremities (ankle, hip, foot, knee). Specifically, a single entry of this dataset consists of a study of single patient which is labeled normal or abnormal, where a single study includes multiple images of the body part in question. The models developed in this paper attempt to correctly label the entire study. Formally, given a study $X_i = \{x_1, ..., x_K\}$ consisting of K images, and a label Y_i corresponding to that study, we seek to create a model which given X_i will output $P(Y_i = 1 \mid X_i)$, the probability that the case is abnormal. Our prediction, then, is 1 (abnormal) if $P(Y_i = 1 \mid X_i)$ is greater than 0.5, and 0 (normal) otherwise. Following the lead of [9], we use transfer learning to train a model to classify single images, then pool these scores together into a prediction for the entire study. We use the model architecture in [9] as a baseline against which to compare our new models.

The task of pooling individual image scores together into a single study prediction is a special case of a more general machine learning problem known as Multiple Instance Learning. Our main contribution is implementing and testing dynamic approaches to Multiple Instance Learning, and comparing them to traditional methods, such as taking the arithmetic mean. To this end we developed two models, one based on attention, a concept familiar from the image captioning task, and the other on recurrent neural network architectures.

We found that the attention-based model outperformed the baseline on all extremities except for ankle in terms of the F1 metric, while the recurrent model outperformed it on all extremities except for in AUROC. Ensembling the two together achieved better performance than the baseline in both metrics on average.

2. Dataset

Our dataset has been provided by the Artificial Intelligence for Medicine and Imaging (AIMI) Center. The data consists of 2,000 studies, each of which contains X-rays of a single patient, with each bone seen from multiple views. In total, this amounts to over 22,000 images, each labeled as either normal or abnormal, with an overall normal-abnormal split of 48%-52% .

The dataset is split by body parts into ankle, hip, knee, and foot cases. For each body part, we split the dataset into a training set (64% of the total cases), validation set (16% of the total), and a test set (the remaining 20%), such that there

Table 1: Lower Extremities Dataset

Extremity	Set	Images	Cases	Median
Ankle	Train	3883	320	12
	Val	952	80	12
	Test	1152	100	12
	Total	**5987**	**500**	**12**
Foot	Train	3724	320	12
	Val	948	80	12
	Test	1152	100	12
	Total	**5824**	**500**	**12**
Hip	Train	2804	320	8
	Val	593	80	8
	Test	756	100	8
	Total	**4153**	**500**	**8**
Knee	Train	3881	320	12
	Val	999	80	12
	Test	1242	100	12
	Total	**6122**	**500**	**12**
Total		**22086**	**2000**	**12**

Figure 1: Example view of a normal case from the dataset.

was no overlap in patients between the three sets. Overall, this resulted in 14,242 distinct training images, 3,492 validation images, and 4,302 testing images. Table 1 shows the size of the dataset for each body part, as well as the median number of images per case for each dataset.

3. Background and Related Work

3.1. Classification Through Transfer Learning

Previous attempts at medical imaging classification typically train models through transfer learning. In this approach, the first n layers of a pretrained image classification model, often developed for larger datasets such as CIFAR-10 or ImageNet, are adapted for a more specialized class appending a final classification layer (usually a fully connected layer with a softmax or sigmoid nonlinearity). The final layer is then trained on the new dataset, while the weights of the feature extractor either remain fixed, or are fine-tuned using gradient descent.

The surprising efficacy of transfer learning is documented in [11], in which the authors conduct image classification on various datasets using features extracted from various levels of the OverFeat model. With regard to the amount of layers of the original network one should use, they report better performance the more layers they extract, although this might not be applicable to deeper base network such as DenseNet-169, since their final layers may have specialized more to the original task. The authors conclude that the combination of generic features from a pretrained network and linear classification using those features provides a strong baseline for many classification tasks.

This approach is particularly common in medical classification tasks. Recently it has successfully applied to X-rays of the chest, as well as musculoskeletal X-rays of the upper extremities ([10], [9]). In the latter, they build their model on the Densenet-169 model described in [5], originally trained on ImageNet. They then train the model end-to-end, finally achieving accuracy which exceeds that of radiologists on certain body parts, while remaining competitive on others.

3.2. Dynamic Multiple Instance Learning

The Multiple Instance Learning (MIL) setting is common in medical applications, and in fact was first introduced in [2] for drug activity prediction. Most research surrounding MIL explores methods of pooling outputs for individual objects into a label for a bag of multiple objects.

The most basic approaches involve taking the average of all the prediction of instances in the bag, and this is employed in [9]. Other MIL algorithms involve using Expectation-Maximization to learn a probabilistic graphical model given the probabilities that a given instance is positive or negative ([16]).

In [6], the authors propose a pooling mechanism inspired by the attention mechanism widely used in image captioning and text analysis ([13]). Instead of taking a simple average of the instance predictions, the Attention model uses a weighted average, with the weight of each instance assigned by a two layer neural network. This method has the advantage that it can be trained using back propagation, since the weighted average function is differentiable. Testing on classical MIL datasets, they find that this attention-based approach was comparable with the best performing classical MIL algorithms (such as MI-SVM, EM-DD, etc).

3.3. RNNs and Multiple Instance Learning

Though there are multiple examples of deep learning being applied to MIL, RNNs (Recurrent Neural Networks) have thus far been used sparingly. This is likely due to the fact the MIL problem assumes that the order of the instances in a bag is irrelevant, while RNNs are typically used to model sequentially dependent behavior in tasks such as video classification [8] and image captioning [7]. RNNs have the advantage of being capable of modelling the relationship between entries in a case, but classic RNNs struggle to represent long term dependencies between inputs. Newer variants of the RNN have been developed to alleviate this problem and the LSTM network is most widely used [4]. LSTMs maintain a continuous flow of cell state that allows information from early inputs to propagate throughout the sequence, which is appropriate for inputs of longer, more meaningfully interconnected sequences [4]. Some work has been done to investigate the usage of RNNs in MIL [1], and has demonstrated that they can be comparably effective to classical MIL methods.

4. Methods

4.1. Baseline

Our baseline model is a 169-layer CNN that predicts the probability of a given image containing an abnormality. The network weights are initialized to those of a Densenet-169 model pretrained on ImageNet, and we replace the final classification layer with one that has a single output, to which we apply a sigmoid nonlinearity. We chose this baseline to keep our results comparable with the work performed in [9], the model being architecturally identical to theirs. Note, however, that while [9] performed end-to-end training upon their network, because we were restricted in our computational resources and training set size we only trained the final classification layer.

During training we optimized the weighted cross entropy loss for a given image x_i^t and label y_i^t from body part t (the label for a given image is simply the label assigned to the study it is found in):

$$\begin{aligned} L(x_i^t, y_i^t) = &- w_1^t y_i^t \log P(y_i^t = 1 \mid x_i^t) \\ &- w_0^t (1 - y_i^t) \log(1 - P(y_i^t = 1 \mid x_i^t)), \end{aligned} \tag{1}$$

where

$$w_l^t = \frac{\sum_{i=1}^{N^t} \mathbf{1}\{y_i^t = l\}}{N^t} \tag{2}$$

for $l \in \{0, 1\}$, where N^t is the number of training examples in the dataset for body type t, and $P(y_i^t = 1 \mid x_i^t)$ is the output of the model. This choice is motivated by the fact that the data-set is label-skewed (in total the data contains a 48%-52% normal-abnormal split).

4.2. Multiple Instance Learning

The baseline model described above outputs a probability that a given image is abnormal. However, our task is to assign a label to cases containing many such images. This can be viewed as a Multiple Instance Learning problem. In MIL, one is given a *bag* of instances $X = \{x_1, ..., x_K\}$. K may vary depending on the bag, and each bag is assigned a binary label Y. Furthermore, each instance has a binary label, y_j, which is unknown. The MIL problem assumes the following relationship between the bag label and instance labels:

$$Y = \begin{cases} 1 \text{ if } \sum_{j=1}^{K} y_k > 0 \\ 0 \text{ otherwise} \end{cases}, \tag{3}$$

that is, the bag is positive if and only if at least one of the instances is positive.

MIL also assumes that probability that a bag is positive, $P(Y = 1 \mid X)$, should be permutation invariant. In this paper, we explore three distinct methods of pooling single instance label prediction into prediction for an entire bag. For the following let $P(y = 1 \mid x)$ denote the output of the baseline model given an X-ray image x.

4.2.1 Arithmetic Mean

In this approach, also used in [9], the output probability for a bag X is simply the average of the output probabilities for the instances in that bag. Precisely, if $X = \{x_1, .., x_K\}$ is a bag of images, then

$$P(Y = 1 \mid X) = \sum_{i=1}^{K} \frac{1}{K} P(y_i = 1 \mid x_i). \tag{4}$$

4.2.2 Attention

The arithmetic mean approach has some clear disadvantages. For example, consider a case X with two images, x_1 and x_2 of the same body part, where x_1 obscures the abnormality, and in x_2 it is completely clear. We would expect $P(y_1 = 1 \mid x_1)$ to be close to 0, and $P(y_2 = 1 \mid x_2)$ to be close to 1. With averaging, $P(Y = 1 \mid X)$ will be approximately 0.5, when it should actually be very high, since x_2 clearly shows the abnormality.

To remedy this, we consider taking a weighted average of our baseline outputs, where the weights are given by a two layer neural network, the intuition being that the network may learn to recognize views that contain important signal. Given a bag X of images $x_1, ..., x_K$, this network takes as input a feature embedding of a single image x_i, and outputs a score, s_i. Let $\mathbf{s}$ be the vector of scores. Then the weights, $\mathbf{w} = [w_1, ..., w_K] = \text{softmax}(\mathbf{s})$ (this is to ensure that the weights add up to 1). Finally, the output of the

entire model is

$$P(Y = 1 \mid X) = \sum_{i=1}^{K} w_i P(y_i = 1 \mid x_i). \quad (5)$$

Here the output probabilities $P(y_i = 1 \mid x_i)$ are given by the classification layer of the baseline model trained initially trained on single images. An advantage of this architecture is that it is still permutation invariant, and can handle any bag size without alteration.

We implemented the Attention model (Figure 2a) using a two layer network, with a ReLU nonlinearity preceded by batch normalization, and dropout layers before each fully connected layer to prevent overfitting. This was then trained to optimize weighted binary cross entropy on a per-case basis, keeping both the feature extraction and classification layer from the original baseline network fixed.

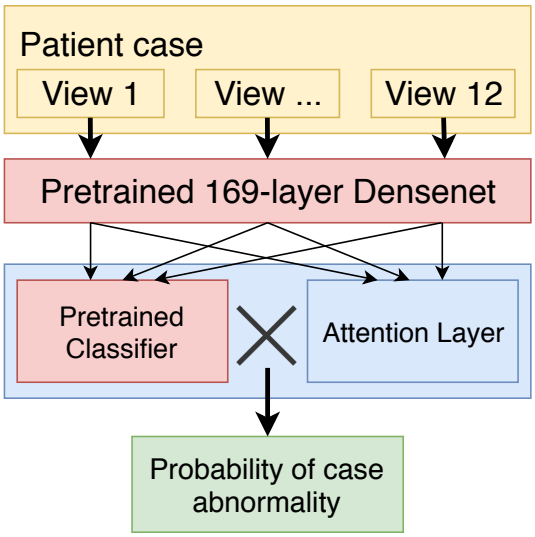
4.2.3 LSTM Pooling

Arithmetic and attention methods in MIL have the advantage of being easily interpretable. Unfortunately they cannot incorporate information about the relationship between views within a case. For instance, consider two views that suggest an subtle abnormality in the same region — neither of the methods proposed so far could adequately share this information.

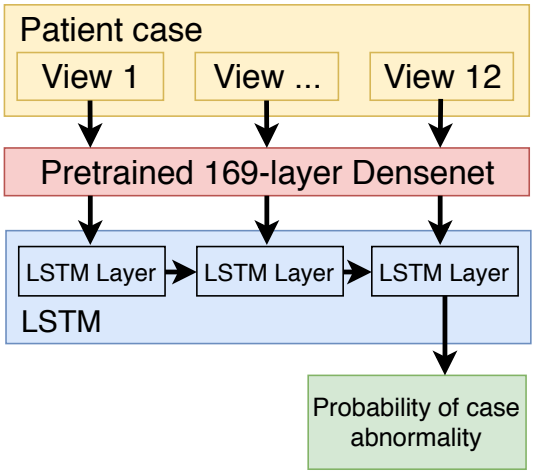
To alleviate this problem, we propose using a recurrent network as our pooling layer. As discussed above, there has been limited research into the efficacy of recurrent networks for permutation invariant MIL. Recurrent models are typically not permutation invariant. However, we hypothesized that a recurrent network could be effective in discerning between obscured or irrelevant views and "remember" features of important views that clearly show or refute abnormalities.

Since relationships between immediately sequential views are equally relevant as separated views, we encounter the issue of long term dependencies. To mitigate this issue, we used an LSTM network. The system of gates that the LSTM network uses to protect cell state neatly represents how we wish for our prediction to be updated by each view ([4]) - updating our predicted outcome if new information passes a threshold.

In an attempt to encourage permutation invariance, we randomly permute the views within each case every time they are loaded. We continue to use the pretrained Densenet-169 as a feature extractor, which we do not update. Another possible concern is that the network might learn the length of cases, as these can vary and there is a slight correlation between case size and abnormality (0.07). To account for this, we fixed our case length to the median case length for the dataset (twelve images). For any cases with fewer than twelve images, we repeated previous images within the case to make up to twelve. For any cases with more than twelve images, we randomly sample twelve views. While we anticipated that this would negatively affect predictions for these cases, we found that only 14% of cases were above size twelve. Standardizing the length of a case also provides training benefits, allowing full vectorization of cases and larger batch sizes.



(a) Attention



(b) LSTM Pooling

Figure 2: a). The Attention model outputs a weight for each embedded input, then applies a softmax so that they collectively add up to one. These are then dot-producted with the classifier scores to produce the output. b). In the LSTM pooling model, features for each view are extracted using Densenet and then sequentially fed into LSTM model. The hidden representation is then classified once the last view has been inputted.

5. Experiments

After initially attempting to train a model that could classify abnormal X-rays without discriminating between the different extremities, we found that this approach yielded poor results. To improve performance, we split our dataset and trained individual models for each of the four extremities (as suggested in [9]).

We ran experiments on three classes of models, training each one individually on each of the four extremities: ankle, foot, hip, and knee. In particular, our baselines were the DenseNet-169 models pretrained on ImageNet with the final classification layer then trained on each of the extremities; these are evaluated with arithmetic mean pooling — the equivalent of running the model used in [9] on our dataset, allowing for comparable results. Our two experimental models were the Attention model described in §4.2.2, and the LSTM model described in §4.2.3.

5.1. Training

For training, we normalize the data to have the same mean and variance as the images in ImageNet. We also resize the images to have the same 224×224 size as those in ImageNet. As previously mentioned, we augment the dataset using random lateral inversions as well as rotations of up to 30 degrees.

Using grid-search on a validation set of 80 studies, we found that the best performing learning rate for the baseline was 10^{-4}, 10^{-5} for Attention, and 10^{-3} for the LSTM. Due to limited time and computing resources, we identified that a good limit on the batch-size for training was 32 views. For the baseline and Attention models, this allowed for fast-enough training when data loading was parallelized through a worker pool. For the LSTM model, we reduced batch-size to 6, finding that larger batch sizes could not be loaded. This was due to the fact that each batch was composed of multiple cases, and each case contained multiple views, in particular, twelve views, as described in §4.2.3.

For all models, we used the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, as in [9], and decayed the learning rate by a factor of 10 each time the validation loss has plateaued for 10 epochs. We save the best model weights with respect to F1 evaluated on the validation set during training as a form of early-stopping.

5.2. Testing

In testing, we grouped each case of views to perform the analog of MIL for prediction. Namely, for each model, we fed in all the views for a particular study, and predicted $P(Y = 1|X)$ for the bag (for Y whether the bag X is abnormal, as in equation 3). For the baseline model, we used the arithmetic mean as pooling described in §4.2.1. As described above, the Attention and LSTM models in turn are designed to pool the results of different views and output one prediction per case. Additionally, we tested an ensemble of the Attention and LSTM models.

5.3. Metrics

We use AUROC (area under the ROC curve) and F1 as our primary evaluation metrics, the former being considered a standard for the evaluation of medical diagnostic test ([3]), and the latter to provide a comparable metric to [9], which lists F1 scores. The ROC curve plots the true-positive rate against the false-positive rate of a binary classifier, which are found as functions of the classification threshold α, $TPR(\alpha)$ and f $FPR(\alpha)$. The ROC curve for an uninformative model will be a straight diagonal line between these two points $TPR(0) = FPR(0) = 0$ and $TPR(1) = FPR(1) = 1$. An ROC curve above this line shows positive predictive power, as well as the extent of tradeoff between true and false positives. Thus, the area under the curve, which can range from 0 (always predict incorrectly) to 1 (always correct), provides a measure of classifier performance. One benefit of using AUROC is that it is independent of the classification threshold α. Another benefit AUROC has over more traditional measures on vision tasks (such as accuracy) is that it takes into account both true positive rate (sensitivity) and true negative rate (specificity, which is just $1 - FPR(\alpha)$), which ensures it is unaffected by the prevalence of positive training examples ([3]).

This latter benefit extends to the F1 score, which is the harmonic mean of the precision, i.e. positive predictive power

$$\frac{TPR(\alpha)}{FPR(\alpha) + TPR(\alpha)}, \tag{6}$$

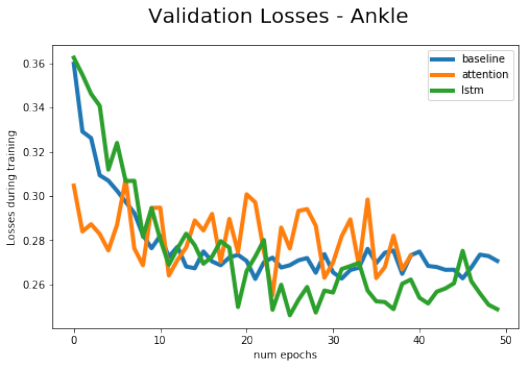
and true positive rate, i.e.

$$F_1(\alpha) = 2\frac{\frac{TPR(\alpha)^2}{FPR(\alpha)+TPR(\alpha)}}{\frac{TPR(\alpha)}{FPR(\alpha)+TPR(\alpha)} + TPR(\alpha)}. \tag{7}$$

Thus, the F1 score can be seen as measuring the performance at a particular point on the ROC curve. Since both our models and those in [9] round $P(Y = 1|X)$ to predict labels, we report $F_1(0.5)$.

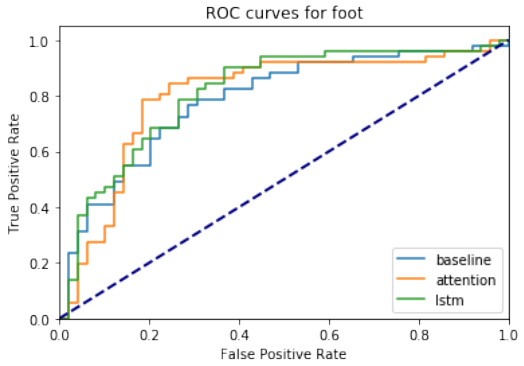
5.4. Results

The results of testing the model are reported in Table 2, which contains the AUROC and F1 scores for three models evaluated on the various test sets. Comparison to the baseline is reasonable in this case since its architecture matches that of previous muskuloskeletal classification attempts ([9], [10]). We found that the model did not grossly overfit during training, which suggests that our choice training hyperparameters were reasonable.



(a) Validation Losses for Baseline Model

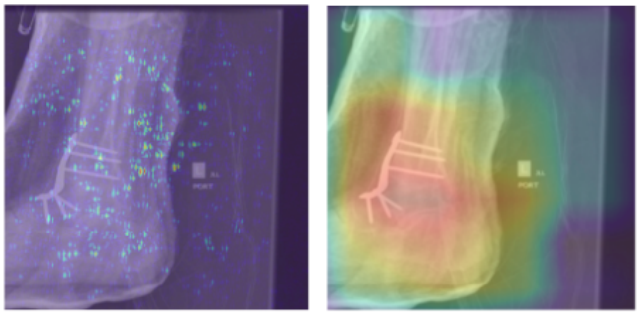
Figure 3: The validation loss per epoch of the various models trained on ankle X-rays.



(a) Model ROC Curves for Foot Studies

Figure 4: The ROC curves show the tradeoff each model makes between false and true positives. Both the Attention and LSTM models outperform the baseline on area under the curve and in F1, as can be seen by the shapes of the curves. While LSTM has a better AUROC, the Attention model has a substantially better F1, which may be explained by the region in ROC space where its curve is higher than that of the LSTM, which is likely where the decision threshold boundary lies.

We found that there was some variation in performance of the three classifiers between the four extremities. The Attention model out-performed the baseline architecture from [9] when evaluated on lower extremities in F1 on average, and achieved comparable AUROC scores. On the other hand, the LSTM model registered higher AUROC scores on average than the Attention model, but did not improve on the baseline F1 scores. Finally, ensembling the Attention and LSTM models resulted in a model which out-performed the baseline in both AUROC and F1 scores.



(a) Saliency map and Class Activation Map for abnormal ankle X-ray. The baseline model outputs an abnormality probability of 0.75, correctly classifying the image.

Figure 5: Saliency maps and Class Activation Maps (CAMs) show the region of the input image which was important in classificaiton.

6. Discussion

6.1. Model Interpretation

6.1.1 Saliency Maps and Class Activation Mappings

To better understand the performance of baseline single image classification layer, we create both saliency and class activation mappings. The saliency map is computed by taking the absolute value of each pixel of the input image with respect to the score of the correct class. Colloquially, one might say that the saliency map shows which part of the image the model output is sensitive to, that is, where small changes in the pixel values might have large effects on the model output. Class activation mappings, on the other hand, might be said to highlight the areas which were highly activated during the classification of an abnormality. These are computing by taking an weighted average of the activations of an image before the final average pooling layer of the feature extraction, where the weights come from the final classification layer. Formally, if we have an image x, denote the i-th activation mapping by $f_i(x)$ and the i-th fully connected weight by c_i. Then $M(x)$, the class activation mapping, is given by the formula

$$M(x) = \sum_i c_i f_i(x). \tag{8}$$

$M(x)$ is then upscaled to the dimensions of the input image to highlight which regions were important in classification.

Note that the example given in Figure 5 demonstrates any writing or labeling on the images was not valuable for classification. This is often a concern with medical datasets, as labels or writing on the views can often give a hint as to whether the case was normal or abnormal, distorting model results.

Table 2: Testing Results (with winners in each row bolded)

Extremity	Metric	DN-169	Attention	LSTM	Ensemble
Ankle	AUROC	0.782	0.767	**0.797**	**0.800**
	F1	**0.659**	0.564	0.634	0.617
Foot	AUROC	0.785	0.800	**0.812**	**0.812**
	F1	0.629	**0.763**	0.681	0.716
Hip	AUROC	0.722	0.697	**0.756**	0.744
	F1	0.581	**0.633**	0.622	0.630
Knee	AUROC	**0.831**	0.798	0.725	0.786
	F1	0.588	**0.699**	0.506	0.621
Average	AUROC	**0.780**	0.766	0.773	**0.786**
	F1	0.614	**0.665**	0.611	0.646

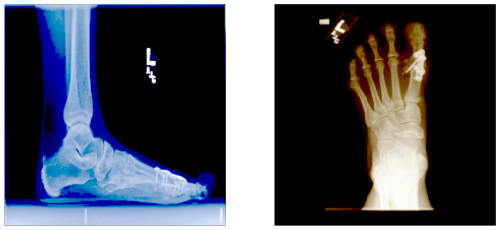
6.1.2 Attention

The point of the Attention model is to be able to differentiate between important and unimportant views of a case. Figure 6 demonstrates the effect the Attention model can have on the output of a specific study. In this case, it attributes a much higher weight to the correctly classified image, changing the final prediction. One disadvantage of attention, however, is that it generally makes predictions more confident, and therefore is highly dependent on predictive capability of the baseline image classifier. This can be observed in the histograms in Figure 7. This may explain the lower AUROC performance of Attention on knee studies, since changing the decision threshold has less of an effect on an extremely confident predictor. Note, however, that the Attention model may effectively trade AUROC performance for better F1, as shown in Figure 4 for the case of foot studies.

Despite good performance by the Attention model, its loss curve in Figure 3 suggests it may have had trouble generalizing, as its validation loss curve remains relatively constant throughout training. Solutions may involve increasing dropout, adding explicit regularizations, or increasing the size and quality of the dataset.

6.1.3 LSTM

We see from the validation loss curves that the LSTM was able to generalize, yet it remains the most difficult model to interpret. The saliency map in Figure 8 shows that images which are input later generally exhibit more gradient activity, suggesting that although the LSTM does retain some memory of previous views, it still is order dependent, which is undesirable in the MIL setting. The dispersed nature of the saliency map also suggests that it might have trouble targeting the area of the image most relevant to abnormality detection, perhaps due to the variance in the alignment of the input images.



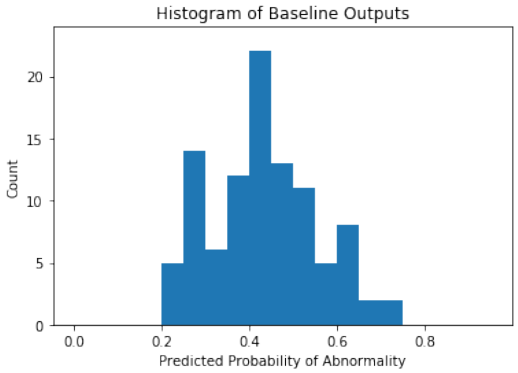
$p_1 = .54, w_1 = 0.02$ $p_2 = 0.21, w_2 \approx 0.00$

$p_3 = 0.65, w_3 = 0.97$

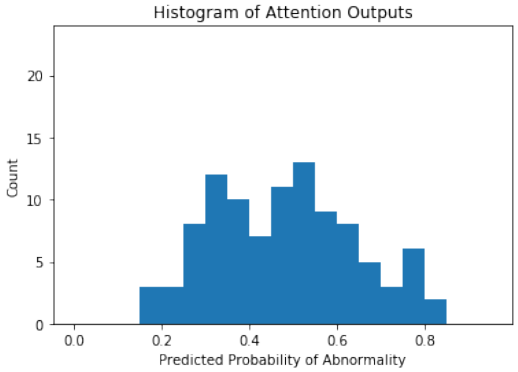
Figure 6: Model output probability and attention weights for each view in a single abnormal foot study. Due to the weighting mechanism, which allows it to assign more importance to the third image, the Attention model outputs an abnormality probability of 0.65, correctly classifying the study, while a naive average would give 0.47, resulting in a misclassification by the baseline.

6.2. Model Performance

Metrics for radiologist performance upon lower extremity abnormality classification were unavailable, but the F1 score on our best performing model (Attention model on foot) is comparable to radiologist F1 score on more challenging upper body extremities, such as the finger. Our dataset was quite small, with only 320 training points per



(a) Baseline Predictions for Knee Studies



(b) Attention Predictions for Knee Studies

Figure 7: a). The baseline predicts probabilities of abnormality in a distribution with mode of around 0.5. b). The Attention model substantially spreads out predicted probabilities, indicating that it makes more confident predictions.

extremity after removing training and validation sets. The added difficulty of MIL, where a case might contain only 2 salient views out of twelve or more, compounded this issue. As a result, our less successful models performed worse than previous attempts at X-ray abnormality classification. Furthermore, the F1 scores for our baseline underperformed those achieved by the same model architecture [9] (low of 0.792, high of 0.968), likely due to a combination of our substantially smaller data set and potentially different scheduling and/or other optimizations.

These models show the possibility of strong abnormality classification performance on lower body extremities. With more data per extremity and more compute, we believe that comparable results to the MURA paper would be possible. There is also an issue of data cleanliness — in the MURA paper, the median case contains fewer than five views, whereas our dataset median case contained more than ten views. This dilution of the signal as discussed above constitutes a significant challenge in MIL.

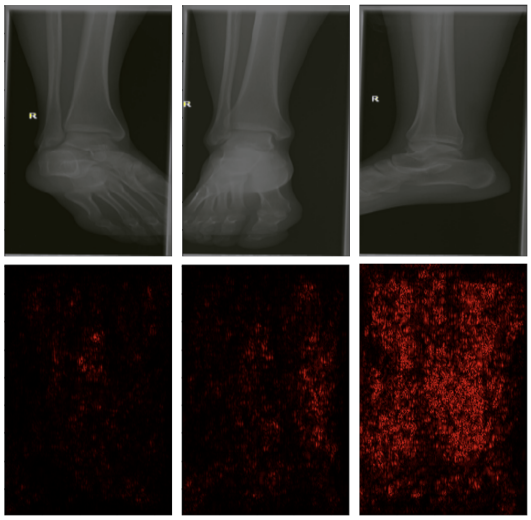


Figure 8: LSTM saliency maps for images 2, 7 and 12 within one case. Decreasing intensity of salient pixels stepping backwards through case reveals highly sequential processing of cases.

7. Conclusion

In this paper, we presented three distinct pooling methods that could be applied to the problem of MIL in medical imaging. The baseline model, when compared with performance of similar models on more complete datasets, illustrates the relative difficulty of this lower extremity classification task. Comparing performance of the two novel models — an Attention based model and a recurrent LSTM model against the baseline, we find that Attention outperforms the baseline in F1 (except for on the knee dataset), while LSTM outperforms the baseline in AUROC in all extremities except for the knee. Future directions might include improving the interpretability of models like the LSTM, and tuning models like Attention to better generalize.

As medical datasets become widely available, we believe further inquiries into MIL techniques to be worth pursuing, especially if models such as these are to be deployed in the clinical setting.

Contributions and Acknowledgements

We would like to thank the Stanford AIMI for the use of their lower extremity X-ray dataset. We would also like to further thank Matt Lungren, Bhavik Patel, Katie Shpanskaya and Kristen Yeom of the AIMI for their help with understanding the dataset, and Mu Zhou for helping to set up our usage of their GPUs for training. We made use of the PyTorch libraries for our modelling, including their pretrained Densenet-169 model. We built upon a basic training loop from CS231n assignment two.

All team members collaborated equally on the milestone report, final project and poster.

References

[1] T. G. Dietterich, R. H. Lathrop, and L. Tomz. Multi-instance Learning Using Recurrent Neural Networks. *WCCI 2012 IEEE World Congress on Computational Intelligence*, 89(2):31–71, 1997.

[2] T. G. Dietterich, R. H. Lathrop, and L. Tomz. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(2):31–71, 1997.

[3] K. Hajian-Tilaki. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, (2):627635, 2013.

[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. *ArXiv e-prints*, Aug. 2016.

[6] M. Ilse, J. M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. *CoRR*, abs/1802.04712, 2018.

[7] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[9] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Ng. MURA Dataset: Towards Radiologist-Level Abnormality Detection in Musculoskeletal Radiographs. *ArXiv e-prints*, Dec. 2017.

[10] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *ArXiv e-prints*, Nov. 2017.

[11] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition. *ArXiv e-prints*, Mar. 2014.

[12] A. D. Woolf and B. Pfleger. Burden of major musculoskeletal conditions. *Bulletin of the World Health Organization*, 81(9):646–656, 2003.

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *ArXiv e-prints*, Feb. 2015.

[14] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *ArXiv e-prints*, Nov. 2014.

[15] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola. Deep Sets. *ArXiv e-prints*, Mar. 2017.

[16] Q. Zhang and S. A. Goldman. Em-dd: An improved multiple-instance learning technique. In *NIPS*, 2001.