Applying Attention and Semi-supervised Learning to Knee Injury Diagnosis

Sahil Arora
sarora@gatech.edu

Randy Michnovicz
rmichnovicz@gatech.edu

Chidozie Onyeze
chidozieonyeze@gatech.edu

Samuel Stentz
samuelstentz@gatech.edu

Abstract

This work attempts to improve upon existing models that detect knee injuries and tears by incorporating attention, positional embedding, and semi-supervised learning into MRNet, a deep convolutional network developed by Stanford Machine Learning Group. Multi-headed self-attention sublayers replaced the max pooling layers, and further modifications such as adaptive average pooling were incorporated. Data augmentation was also performed by generating labels for the sagittal views of MRI images from the fastMRI dataset and pretraining the classifier with those. The performance of these models were evaluated with an external validation dataset as a benchmark.

1. Introduction/Background/Motivation

Medical imaging is a field that has gained a lot from the improvements in machine learning techniques and the development of new techniques. This field has specifically benefited from the advent of deep neural networks in their modern form. We consider the domain of MRI analysis. As of today, the analysis of MRI images is a generally slow process, requiring a substantial amount of trained man-hours [3]. Thus, accurate deep learning models can help speed up diagnosis or at least help provide guidance to radiologists analyzing MRI images.

Our ultimate goal was to design a model capable of diagnosing pathologies in a knee from MRI scans of the knee. This project builds on MRNet, a model created by a team from Stanford [1], and attempts to improve on the design of the network by incorporating self-attention [7] between the different layers of the knee MRI images. A similar technique was used by [4] and has been shown to improve similar models.

The dataset used by the original Stanford team is also somewhat limited in size (less than 1200). Thus, we also try to tackle this problem by applying techniques of semi-supervised learning using an alternate but unlabeled dataset, fastMRI [9], to boost the training set size. We hoped that these changes together would ultimately improve the accuracy of the model.

Our primary guiding reference has been MRNet [1]. This model has proven to be effective in the metric of the area under the receiver operating characteristic curve (AUC) that compares false positives and true negatives for different cut-offs. MRNet utilizes a imagenet-pretrained copy of Alexnet to extract features from the MRI images. Each feature is then converted to a single number by averaging the feature. MRNet then takes the maximum of these 1-dimensional features over the different images and performs a logistic regression using the maximal 1-dimensional features. We noticed that in taking this maximum, some of the potentially useful information about the different layers and their features relative to each other may be lost, so we replaced this with a multi-headed self-attention layer to provide the logistic regression more informative input.

We also implement a GradCam [5] on our model. This is of particular use to the radiologists and surgeons being assisted by our model or similar models as it helps provide a visual representation of the regions of interest in determining whether or not a pathology exists. This makes human verification of results a much more efficient task and makes for a much more interpretable model.

This work is of great importance as an aid to radiologists and surgeons in improving the speed and accuracy of diagnosing patients. From the work of MRNet, similar models have been shown to significantly reduce the rate of false positives for certain types of pathologies. Similar models have also been shown to improve the accuracy in the specificity of a diagnosis which leads to more measured treatment recommendations. Overall, our work has value as a demonstration of the possible value of the techniques we apply (attention and semi-supervised learning) to the domain of MRI analysis and medical imaging as a whole.

2. Data

We used three datasets of DICOM-based images of knee scans from 2 datasets: MRNet, fastMRI, and an independent external validation dataset. They all provide DICOMs in the form of PNG images where a single MRI scan is represented by a variable number of sequential image slices and a number of attributes, most notably the plane in which an MRI scan was taken. MRNet contains MRI scans taken from 1370 exams. Each exam has 3 MRI scans, labeled with the plane in which the scan was taken: sagittal, axial or coronal. A scan contains a number of slices in the range [20, 61]. Each exam is labeled with binary values of non-exclusive diagnosis types from exams taken at Stanford University Medical Center in the years 2001-2012.

The specifications of the dataset were originally given in [1] but we restate them here for clarity. The dataset consists of 1,370 exams with 1,104 being abnormal. Of those 1,104 abnormal exams, 319 were ACL tears and 508 were meniscal tears with 194 being both. The data was split into training, tuning and validation sets containing 1,130, 120 and 120 exams respectively. The tuning and validation sets were constructed such that there were at least 50 positive examples of each label (abnormal, ACL tear, and meniscal tear) in each set.Each image in each MRI scan is resolution 224 x 224 and contains 3 color channels.

FastMRI is a dataset collected by NYU school of Medicine and Facebook AI Research intended for reconstructing MRI images from a subset of scans. From this dataset, we use the DICOM-based images of knee MRI scans taken from the sagittal, axial and coronal planes as unlabeled data in our semi-supervised learning setup. The dataset contains X sagittal plane [X, 90].

The external validation dataset consists of 917 sagittal plane knee joint images gathered at the Clinical Hospital Centre Rijeka, Croatia, from 2007 to 2014. Each of these images came with a corresponding binary label for both injury and complete rupture. This dataset first appeared in [6] and was used as a benchmark in the development of the original MRNet. Since this was still an available benchmark, it was used accordingly in comparing the performances of the attention model to the original MRNet.

Consider Figures 1, 2, 3 for examples of the exams in the MRNet dataset.

3. Approach

Our approach involved replicating the original MRNet model [1] as a baseline then making two major modifications to it: replacing a max pool layer with multi-head attention, and using a semi-supervised approach to add training data.

The baseline approach uses an imagenet-pretrained Alexnet to extract features of the dimension $N \times 256 \times 6 \times 6$ where N is the number of slices. Each 6×6 feature is average pooled to get features of the dimension $N \times 256$. The features from each layer are combined using a max pool layer that results in a 1×256 tensor for each set of images. This is then passed into a fully connected layer, resulting in a 1x1 output which is passed into a sigmoid layer to output a binary prediction. Training uses cross-entropy loss and a batch size of 1. This batch size is required due to the incon-

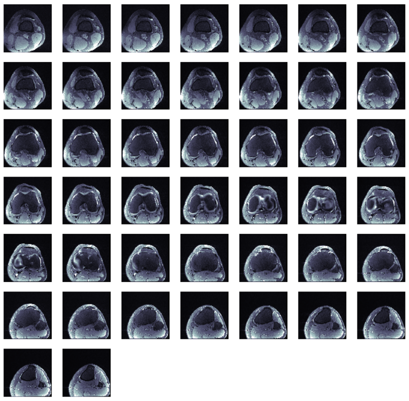


Figure 1. Example of Axial slices from an exam in the MRNet Dataset

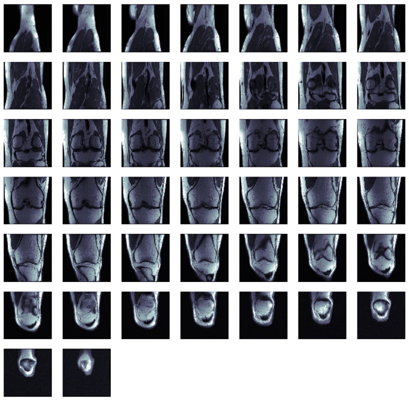


Figure 2. Example of Coronal slices from an exam in the MRNet Dataset

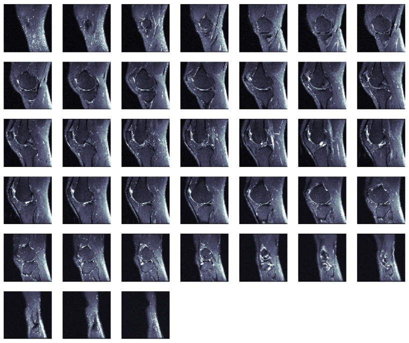
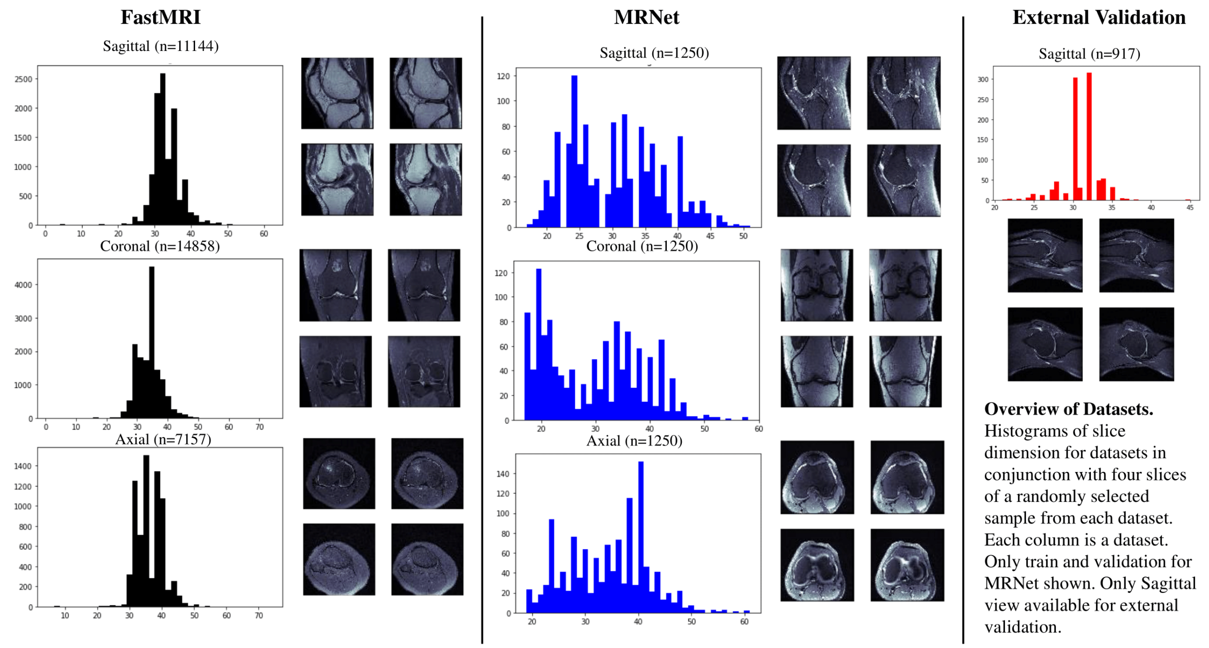


Figure 3. Example of Sagittal slices from an exam in the MRNet Dataset



Overview of Datasets. Histograms of slice dimension for datasets in conjunction with four slices of a randomly selected sample from each dataset. Each column is a dataset. Only train and validation for MRNet shown. Only Sagittal view available for external validation.

Figure 4. Diagram showing the distribution of number of slices for exams in the MRNet and fastMRI datasets

sistent number of layers for each examination. A separate model is trained for each plane, and the model predictions are combined using a logistic regression from the 3 model outputs. For the external validation dataset, all results reported are directly from a single CNN for sagittal view.

The modification we made to the original MRNet model was replacing the max pool layer with multiple layers of multi-headed self-attention followed by a feedforward layer, both incorporating layer normalization. After this, we average pool the results before passing it through the fully connected classifier. An alternative approach we tried is using MRNet with attention and fully connecting the output of the attention without average pooling. This approach results in a model that depends on the maximum number of layers per input and requires that we pad the input to ensure all the inputs have the correct number of layers.

The idea behind using attention is that attention may be a less naive, more trainable, and more interpretable idea than a max pool of features. While max pooling takes the max of feature outputs from many slices, these feature outputs are not easy to interpret, so we instead are able to weight individual layers. The attention layer also takes into account the slice's position (via a sinusoidal positional encoding), which may be useful if, for example, a certain slice provides evidence for a certain condition.

We also augmented the MRNet's external validation data with the scans from fastMRI in a semi-supervised learning technique. Because the fastMRI data has MRI scans of the sagittal plane which are unlabelled, we use a model trained on the MRNet data to predict labels for each fastMRI scan using a threshold of 0.5 for the logits output by the model (teacher model). The teacher model is then trained on the fastMRI data using the artificial labels (student model). Finally, the student model is retrained on the MRNet test set.

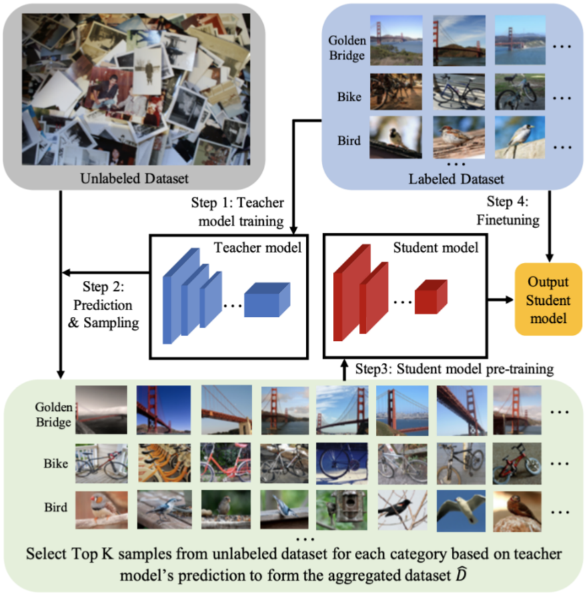

Unlabeled Dataset
Golden Bridge
Bike
Bird
Labeled Dataset
Step 1: Teacher model training
Step 4: Finetuning
Teacher model
Student model
Step 2: Prediction & Sampling
Output Student model
Step3: Student model pre-training
Golden Bridge
Bike
Bird
Select Top K samples from unlabeled dataset for each category based on teacher model's prediction to form the aggregated dataset $\hat{D}$

Figure 5. Pictorial explanation of the Semi-supervised learning approach we used from [8].

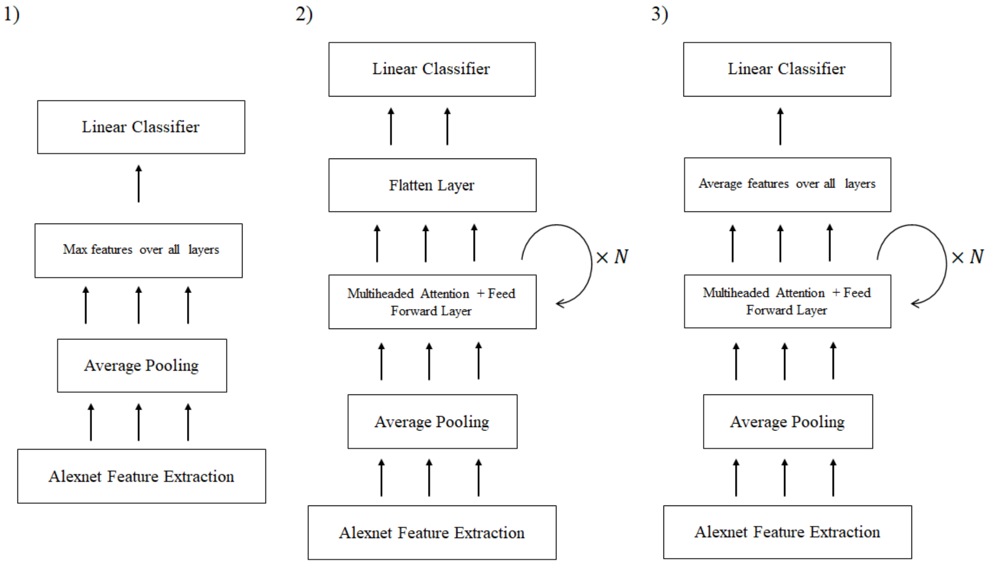

1)
Linear Classifier
Max features over all layers
Average Pooling
Alexnet Feature Extraction
2)
Linear Classifier
Flatten Layer
Multiheaded Attention + Feed Forward Layer
× N
Average Pooling
Alexnet Feature Extraction
3)
Linear Classifier
Average features over all layers
Multiheaded Attention + Feed Forward Layer
× N
Average Pooling
Alexnet Feature Extraction

Figure 6. Diagrams of Models with 1) being the original MRNet, 2) being MRNet with attention and flattened output and 3) being MRNet with attention and average pooling.

This method was heavily inspired by the work done by the Facebook AI Research Team in [8], with two differences. First, all of our unlabelled data is used in pretraining without sampling as we are confident all our unlabelled data belong to one of our classes (i.e. a diagnosis exists), an assumption that [8] cannot make. Second, our teacher and student model have identical architectures and all learned weights are used from previous stages to minimize required time for training for the scope of the project.

The existing infrastructure provided by Stanford [1] allowed only for training and evaluating the MRNet model on the external validation dataset. So in addition to the changes made to the model, changes had to be made to the repository to load, train, and evaluate the original MRNet dataset as well as the fastMRI dataset. All of the models and surrounding infrastructure were developed in python and pytorch.

4. Experiments and Results

The goal of this work was to improve upon the classification tasks of MRNet. The quantitative metric used to measure improvement was performance on the test set in the external validation dataset after training. This was the only benchmark relied upon since it had accurate labels and there was room for improvement, so models could be distinguished. Four models were analyzed by this benchmark: the original MRNet, the initial attention model with a multiheaded attention sublayer, the attention model with average pooling, and the final attention model pretrained with the semi-supervised data.

These models were all trained with an Adam optimizer over the cross-entropy loss, paired with a learning rate scheduler that decreases the learning rate by a certain factor after a set number of epochs (known as *patience*) of no improvement. Between these models, hyperparameter tuning was performed for the number of epochs, learning rate, weight decay, factor, and patience. An exhaustive grid search was not performed due to computational constraints, but multiple values were tested and the best combinations were reported in the results. Despite the tuning, we were not successful by the given metric.

With the original MRNet dataset, the performance of our basic attention model (the original model with a multiheaded attention sublayer and a fully connected layer afterwards) was promising, with the model perfectly classifying the knee MRIs for both abnormalities and tears in the validation set. However, the original model also performs just as well, achieving 1.000 AUC for both tasks. Since both models perform the same and there was no way to perceive any improvements, a new benchmark was required. The basic attention model and further changes to it were therefore analyzed through its performance on an external validation dataset independently collected by a Croatian hospital that also was doing research on automating injury detection in knees.

The results for the models on the external validation dataset were disappointing. For abnormality classification, the basic attention model appeared to overfit, as it outperformed the original model in the training and validation set, but then was 7.5% worse in AUC on the testing set. Im-

Models	Abnormality Performance			Tear Performance		
	Training AUC	Validation AUC	Testing AUC	Training AUC	Validation AUC	Testing AUC
MRNet	0.9997	0.8513	**0.8922**	0.9998	0.9171	**0.8692**
Attention and Fully Connected Layer	0.9999	0.8789	0.8187	0.9918	0.9000	0.7965
Attention and Average Pooling	0.9591	0.8419	0.8864	0.8670	0.7705	0.8081
Pretraining Attention Model with fastMRI	0.9593	0.8416	0.8866	0.8668	0.7686	0.8004

Table 1. Performance on External Validation Dataset

provements to the basic attention model yielded some gains, but not enough to outperform the original MRNet. Replacing the fully connected layer with an adaptive average pooling layer had worse training and validation performance but earned an AUC of 0.8864 on the testing set. Training with the weakly labeled fastMRI data did not do much to help, as it only improved the testing AUC by 0.0002. Both of these improvements did not yield a better AUC than the 0.8922 that MRNet achieved.

For tear classification, the results were worse. The basic attention model similarly did about 7.5% worse in AUC compared to MRNet. Incorporating average pooling and pretraining with the semi-supervised learning were not very successful, as they only yielded AUCs of 0.8081 and 0.8004 respectively. In this case, augmenting the data seemed to make the model worse at classifying tears. Another issue was that in training the accuracy of the models seemed to plateau earlier, only reaching 0.86 AUC on the training set and 0.77 AUC on the validation set, despite the original models reaching 0.99 and 0.90 respectively. While this suggests there was more for the models to learn, the losses were converging much earlier despite tweaking with the learning rate and tuning other hyperparameters.

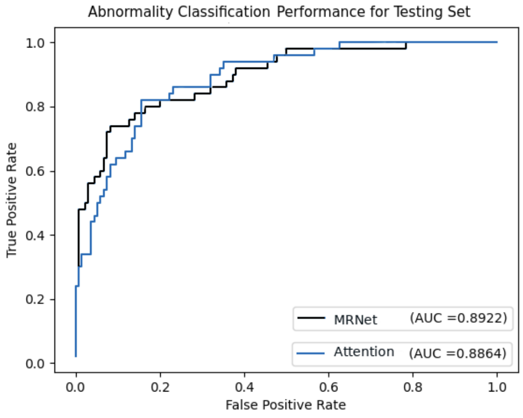


Figure 7. Comparison of model performance

For tear classification, the results were worse. The basic attention model similarly did about 7.5% worse in AUC compared to MRNet. Incorporating average pooling and pretraining with the semi-supervised learning were not very successful, as they only yielded AUCs of 0.8081 and 0.8004 respectively. In this case, augmenting the data seemed to make the model worse at classifying tears. Another issue was that in training the accuracy of the models seemed to plateau earlier, only reaching 0.86 AUC on the training set and 0.77 AUC on the validation set, despite the original models reaching 0.99 and 0.90 respectively. While this suggests there was more for the models to learn, the losses were converging much earlier despite tweaking with the learning rate and tuning other hyperparameters.

We also implemented GradCam [5] for our model that allows us to determine regions of interest and the input images in determining whether or not a pathology exists. In doing this we consider the final convolutional layer within the embedded Alexnet. The result of this implementation of GradCam on the attention model with average pooling can be observed in Figure 8. Of course, as people without medical experience, it is not possible for use to draw conclusion about the efficacy of the GradCam since we cannot make independent verification of its the results. This being said, the existence of the GradCam allows the possibility to further understand the model and could lead to a more reliable model in the future.

5. Discussion and Future Direction

This work was able to show a preliminary application of multiple attention layers directly into a fully connected layer with padding and/or followed by averaging across the variable slice dimension was unable to show significant improvement in accuracy for MRI classification. As with any optimization, learning rate, optimizer choice, and many hyperparameters could be tuned to potentially improve our AUC. Due to the time constraints and scope of this project, little fine tuning was done on model parameters, meaning our gap in performance may simply be an artifact of limited resources and time.

Future investigation could be done into using different semi-supervised methods for utilizing the fastMRI dataset and evaluation of our proposed models on the privately held MRNet test set. Additionally, to account for imbalance in the representation of both classes for all classification tasks, a balanced loss function could be applied. Our work performs no data augmentation to the inputs such as

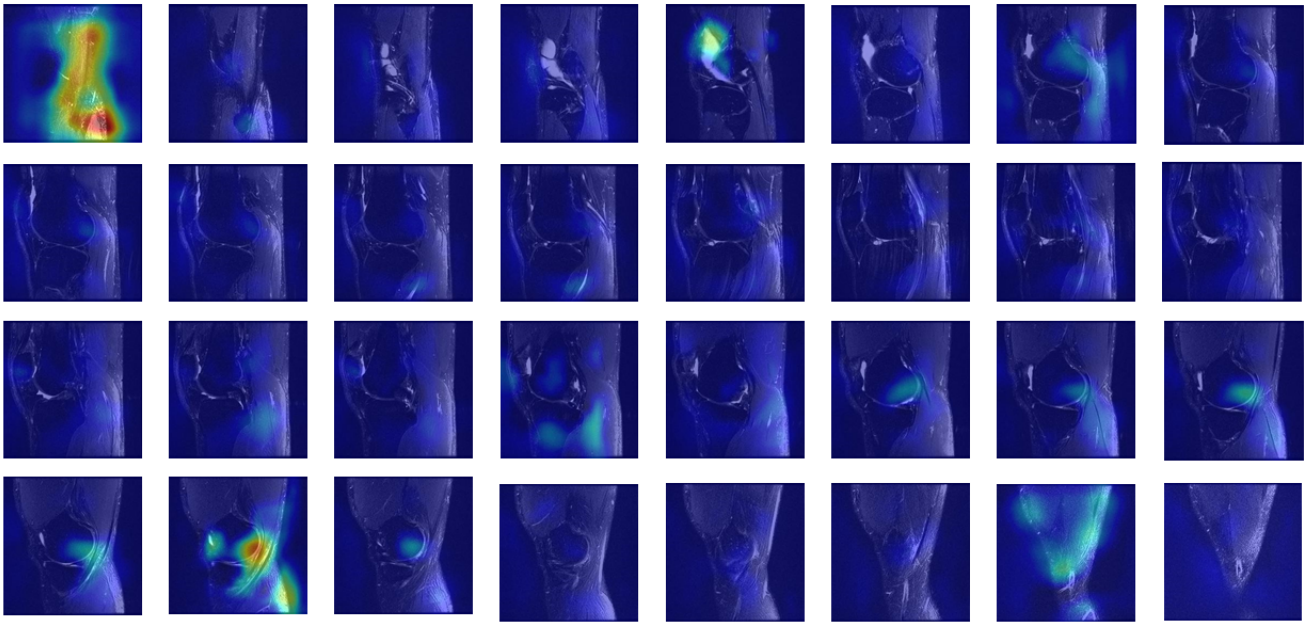


Figure 8. GradCam of images from an Abnormal Sagittal Exam

shear, scale, or rotations. This may be a serious limitation to our semi-supervised approach, as manual inspection of the sagittal view slices shows that the three datasets vary in what direction the sagittal plane is viewed from Figure 4.

Finally, alternative changes to the model architecture could be taken. First, the 6x6x256 output of Alexnet could be flattened instead of averaged as no clear explanation is provided to explain averaging across this dimension. Second, a different CNN with residual layers could be used such as ResNet [2] which has been shown to outperform Alexnet on popular benchmarks. Finally, an alternative to multi-headed self-attention could be applied for the problem. Attention is really a mapping of keys, queries, and values to outputs. It may be possible that using an architecture such as using each slice's vector as the key and value and every other slice as queries followed by averaging may be advantageous. This architecture more naturally models how the model should normalize each individual slice based on the other slices. We offer this potential architecture change because we suspect our model is limited by an inability to effectively account for how important different slices are in the input. For example, an input padded with meaningless slices on either end would suffer from the averaging operation we perform. Another simpler change could be max pooling over the slice dimension to compare results.

In conclusion, it appears addition of attention with average pooling performs similarly to the original MRNet model. Further work remains to be done to improve accuracy by effectively using larger unlabelled datasets such as fastMRI or through the use of alternative architectures.

6. Work Division

The division of work in this project was as fair as was reasonable and is expressed in Table 2. All members were also involved in various miscellaneous tasks such as researching other possible datasets and generating plots and diagrams.

Student Name	Contributed Aspects	Details
Sahil Arora	Implementation and Analysis	Developed the loading, training, and evaluation of the datasets and analyzed the results
Randy Michnovicz	Interpretation and Writing	Developed the code to parse and analyze the datasets and wrote multiple sections of report
Chidozie Onyeze	Implementation	Implemented the different MRNet with Attention models and implemented GradCam
Sam Stentz	Implementaion and Pre-processing	Performed semi-supervised learning on data and pre-processed the different datasets

Table 2. Contributions of team members.

References

[1] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. *PLoS medicine*, 15(11):e1002699, 2018. 1, 2, 4

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 6

[3] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019. 1

[4] Yifan Peng, Ke Yan, Veit Sandfort, Ronald M Summers, and Zhiyong Lu. A self-attention based deep learning method for lesion attribute detection from ct reports. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE, 2019. 1

[5] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 5

[6] Ivan Štajduhar, Mihaela Mamula, Damir Miletić, and Gözde Ünal. Semi-automated detection of anterior cruciate ligament injury from mri. *Computer methods and programs in biomedicine*, 140:151–164, 2017. 2

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[8] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification, 2019. 3, 4

[9] Jure Zbontar, Florian Knoll, Anuroop Sriram, Matthew J. Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, James Pinkerton, Duo Wang, Nafissa Yakubova, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI. 2018. 1