MRNet with Alternate Pretrained Features

Shayne Miel
Stanford University
Stanford, CA, USA
smiel@stanford.edu

1. Introduction

Reading and interpreting MRIs is a time-consuming process. Even with trained professionals, it is easy for a clinician to misdiagnose an injury based on an MRI reading. Improving the automated identification of abnormalities in knee MRIs could help prioritize which MRIs to examine first, as well as provide better early results for patients whose scans appear normal. Model predictions could also provide a "second opinion" which would reduce the possibility of missed abnormalities. This could represent a large cost savings for hospitals and an increased level of care for patients.

2. Problem statement

Given a set of three MRI series (axial, coronal, and sagittal) of a patient's knee, we wish to predict the presence of injuries that will require surgery. In particular, we wish to predict whether the knee is healthy, has a meniscal tear, has an ACL tear, or has any other abnormality. Since these injuries can co-occur, we wish to predict three independent binary values: abnormal, acl tear, and meniscus tear.

3. Related work

The main purpose of this study is to replicate the results achieved by Bien, et al's MRNet model [1]. They use a pretrained AlexNet [6] model to extract features from each 2D slice of the 3D MRI volume, followed by a global average pooling per slice to flatten the image, and a max pooling across the volume. Finally, a fully connected layer and sigmoid activation are used to predict a binary label for each series. Those three predictions (axial, coronal, and sagittal) are then used as features in a simple logistic regression to predict the final label in question. This process is repeated for each of the three independent labels.

An important aspect of the MRNet model is the data augmentation done during training. Every volume is randomly flipped horizontally, shifted horizontally by -25 to 25 pixels, and rotated by -25 to 25 degrees each time it is seen during training. This helps prevent the model from overfitting the small data set.

4. Technical approach

My goals for this project are fairly simple:

1. Reproduce the results obtained by Bien, et al. in [1].
2. Experiment with using different pretrained networks to extract features, specifically replacing AlexNet with GoogLeNet [8] as Chi, et al. did for ultrasound images [2], ResNet [3] as done in [5], and Inception-v3 [9] as done in [4].
3. If time allows, try replacing the logistic regression ensemble with a direct concatenation of the features from each of the three series before going through the fully connected layer.

Unfortunately, while Bien, et al. did release code for the architecture of their MRNet model, they did not release the data loading, hyperparameters, data augmentation, ensembling, or training, or evaluation code. This means that more of the effort for this project will be focused on simply reproducing their results.[1]

However, they did release source code for training MRNet on a set of external validation data from the 2017 Štajduhar et al. study [7]. This provides a good starting place for trying replicate their pipeline on the MRNet dataset.

5. Dataset

The MRI data provided in the MRNet challenge contains scans from 3 MRI types (sagittal plane T2, coronal plane T1, and axial plane PD) with 3 labels per MRI (abnormality, ACL tear, and meniscal tear) for 1,250 examinations.

They have designated a training/validation split and have withheld the test set for leaderboard evaluation. In order

[1] I wrote to the authors hoping that they would share their code privately so that I could spend more time on enhancements, but they said that the code is not avaialble.

Diagnosis	Label	Train	Validation	Test
Abnormal	Positive	817	96	95
	Negative	193	24	25
	Total	1010	120	120
ACL	Positive	193	15	54
	Negative	817	105	66
	Total	1010	120	120
Meniscus	Positive	357	40	52
	Negative	653	80	68
	Total	1010	120	120

Table 1. MRNet data splits and label counts.

to assess the changes I will be making, I am calling their validation set the test set, and splitting their training set into a training and validation set. Counts of cases and labels for each set can be seen in Table 1.

Note that this data has already been preprocessed as described in [1]:

> Images were extracted from Digital Imaging and Communications in Medicine (DICOM) files, scaled to 256×256 pixels, and converted to Portable Network Graphics (PNG) format using the Python programming language (version 2.7) and the pydicom library (version 0.9.9).
>
> To account for variable pixel intensity scales within the MRI series, a histogram-based intensity standardization algorithm was applied to the images. For each series, a representative intensity distribution was learned from the training set exams. Then, the parameters of this distribution were used to adjust the pixel intensities of exams in all datasets (training, tuning, and validation). Under this transformation, pixels with similar values correspond to similar tissue types. After intensity standardization, pixel values were clipped between 0 and 255, the standard range for PNG images.

6. Preliminary results

I have been able to successfully recreate the data loading and training pipeline, as well as the data augmentation steps, ensembling, and evaluation code. The hyperparameters and training process are the same for every model. It is unclear whether the current hyperparameters are the optimal ones, but the results look promising. Figure 1 shows the loss curves from the baseline MRNet model for training and validation on each series and diagnosis.

The AUC values for training and validation at each epoch can be seen in Figure 2. The models appear to be learning and generalizing well. The validation scores are noisy though, which is most likely due to the extremely small data set sizes and imbalanced classes.

Table 2 shows the AUC on the test set as reported in [1] and the test set (their validation set) with my reproduction of the MRNet model. I do not have access to the test set that they used for reporting results, so I do not expect to get exactly the same results. The AUC is fairly close for all three diagnoses. My scores are slightly lower on ACL and Meniscus tears, which is most likely due to the smaller data set size and non-optimal hyperparameters.

References

[1] N. Bien, P. Rajpurkar, R. L. Ball, J. Irvin, A. Park, E. Jones, M. Bereket, B. N. Patel, K. W. Yeom, K. Shpanskaya, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of mrnet. *PLoS medicine*, 15(11):e1002699, 2018.

[2] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *Journal of digital imaging*, 30(4):477–486, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] D. Kim and T. MacKinnon. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clinical radiology*, 73(5):439–445, 2018.

[5] P. Korfiatis, T. L. Kline, D. H. Lachance, I. F. Parney, J. C. Buckner, and B. J. Erickson. Residual deep convolutional neural network predicts mgmt methylation status. *Journal of digital imaging*, 30(5):622–628, 2017.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] I. Štajduhar, M. Mamula, D. Miletić, and G. Ünal. Semi-automated detection of anterior cruciate ligament injury from mri. *Computer methods and programs in biomedicine*, 140:151–164, 2017.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

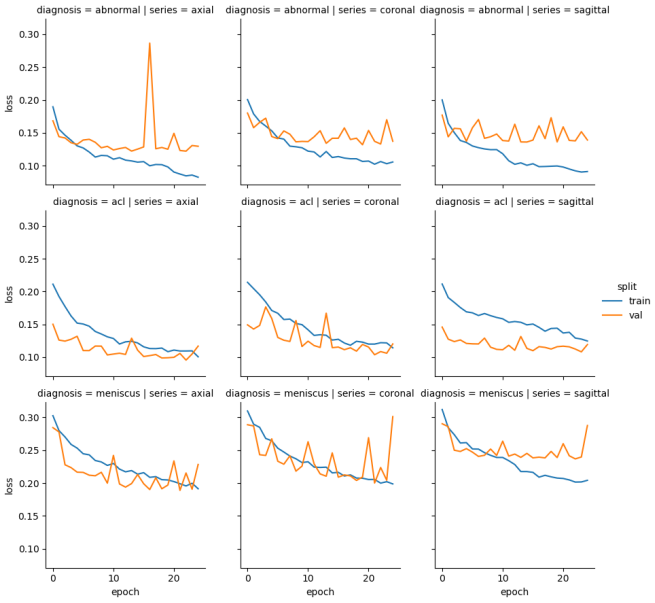

Figure 1. Baseline loss for each diagnosis and series.

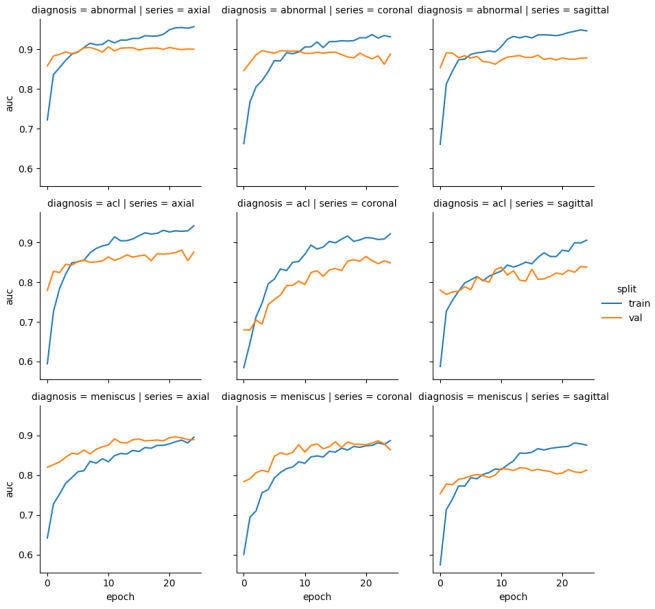

Figure 2. Baseline AUC for each diagnosis and series.

Model	Abnormal	ACL	Meniscus
MRNet (reported)	0.937	0.965	0.847
MRNet (milestone)	0.951	0.949	0.839

Table 2. AUC on the test set