MRNet model based on VGG16 for multi-view Knee MRI classification

Muhammad Salah Mahmoud Osman
Alexandria University
Computer and Systems Engineering Dept.
eng-muhammad.salah1621@alexu.edu.eg

Abstract

The most efficient known method for knee diagnoses is the magnetic resonance imaging (MRI) which is characterized by a huge number of scans from several axes. this makes the MRI interpretation time consuming and subject to human error and variability. Using computer vision for MRI interpretation can assist clinical diagnoses to be done in much less time and with a higher accuracy. The state-of-the-art deep learning system for Knee MRI interpretation called MRNet [1] is based on a deep learning model for image classification, AlexNet [2]. this work investigates improving the MRNet model by using a deeper image classification model, VGG16 [3].

1. Introduction

The Magnetic Resonance Imaging for Knee diagnoses provides a detailed set of images from different angles describing the structures within a knee joint, including bones, cartilage, tendons, ligaments, muscles and blood vessels. MRI has repeatedly demonstrated high accuracy for the diagnosis of meniscal and cruciate ligament pathology, and is routinely used to identify those who would benefit from surgery. Due to the size of the MRI exam which include a huge set of detailed images and might be taken from different angles, the interpretation of MRI exams becomes time consuming and prone to error.automating the process of interpreting the knee MRI exams has a number of potential applications, such as assisting the clinical decision and spoting light on high risk patients. However, the variability of the MRI exams; having multiple planes and variable number of images in each exam has limited the use of tradition image analysis techniques and traditional deep learning models which expect fixed input dimensions.

MRNet Model

MRNet can be considered as a building block in the process of interpreting a complete MRI exam from 3 different planes. each MRNet block is used to predict a probability of one of the 3 classifications (Abnormality, ACL tear, Meniscus) for one of the 3 planes (Axial, Coronal, Sagittal) which means a complete interpretation of an MRI exam consists of 9 MRNet blocks each predicts a probability; 3 blocks for abnormality, 3 blocks for ACL tear, and 3 blocks for meniscus. a Logistic regression model for each classification label (Abnormal, ACL, Meniscus) collects the 3 probabilities from the 3 different views to output a single probability to end up with only 3 probabilities for each label.

2. Dataset

We used the origianl MRNet dataset provided by Stanford University Medical Center which contains Reports for knee MRI exams performed between January 1, 2001, and December 31, 2012. the dataset consists of 1370 exams manually reviewed by the original authors of the MRNet paper [1]. The original dataset has 1104 abnormal exams (80.6%), 319 exams with ACL tear (23.2%), and 508 Meniscus exams (37.1%). The original data was split into 1130 training exams, 120 validation exams, and 120 test exams, the later set is a hidden set used for the evaluation of models and we had no access to the set so we split the training set into 1017 training exams (90% or the training set) and 113 validation exams (10% of the training set) then we used the original validation set for our personal testing and evaluation.

3. Methods

3.1. Data loading

Due to the large data size we couldn't load the complete Dataset in the memory; we designed a Data Generator to load the data on the fly and perform random augmentation on the images on the exams.

3.2. Data Pre-Processing

Offline data augmentation could have a good impact on reproducing results and weights but it wasn't efficient because of the large data size. We performed our data augmentation and processing on the fly during the loading step

which result in different augmentation each time an exam was loaded. Our augmentation consisted of random rotation between -25, 25 degrees, random shifting between -25, 25 pixels, and random horizontal flips with probability 50%. The augmentation routine was applied on each image in the exam independently from other images.

3.3. Data Balancing

each of the classification labels (Abnormal, ACL, Meniscus) has different positive/negative ratios which means we cannot balance the data by offline augmentation because it would require building 3 different datasets, one for each label. Our first attempt was to balance the data within the data generator by adding redundant exams form the minority class to each training epoch, this technique wasn't good considering the probability of over-fitting was high due to redundant exams, although each exam was augmented differently this didn't give a satisfying results. However the previous attempt to balance the data didn't give the best results, we then applied class-weighting in the loss function giving the minority class higher weight that the majority class to encourage predicting the minority class. this gave a better results than the previous attempt.

3.4. The Model

We followed the structure of MRNet, however, we replaced the AlexNet block with a VGG16 block. We used the convolution layers of the VGG16 to obtain a feature map for each image in the exam with the shape (7, 7, 512) then applying average pooling and dimensions reduction to obtain a (512) vector for each image in the exam, we then applied max pooling over the exam images to obtain a single (512) vector for the whole exam, this vector was then feed to a Fully Connected layer with a single output probability. This was implemented in a vectoraized fashion to improve the performance of the training on the GPU by treating the number if images in the exam (s) as the batch size of the VGG16 block figure 1.

3.5. Training

First, for experimental purposes we trained the model with randomly initialized weights with the Glorot Uniform initializer. then the training took the following steps:

- training a single model 10 epochs on the axial exams with abnormal labels.
- transfer the trained weights from the previous step to all the 9 models as a weight initialization step.
- train the 9 models on parallel on 2 training cycles each cycle around 15 epochs.
- the validation after each epoch helped us determining when we should stop the training.
- the models was then tested and evaluated against the test set.
- each 3 models predicting a specific class (ex: abnormal) were combined to feed a logistic regression layer for training.
- the combined models with the logistic regression model were then tested against the test set again.

Second, we applied transfer learning as a weight initialization step, using the weights for ImageNet which is a large dataset of natural images. this helped the training to complete in much less time (10-15 epochs instead of the 40 epochs used before). Although the weights were trained on natural images and not medical images, the results were more satisfying than the random weight initialization.

4. Evaluation

during the training we used several metrics to monitor the training and validation, this metrics included:

- True Positives
- False Positives
- True Negatives
- False Negatives
- Binary Accuracy
- Precision
- Recall
- AUC

Howver, for evaluation we used only 3 important metrics:

- Acuraccy
- AUC
- F1 Score

This Metrics helped watching out for skewed classes and data imbalance.

5. Results

Figure 2 describes the detailed metrics' results we obtained.

Although we cannot compare the performance with the performance of the original model because we don't have access to the hidden test set, we will however, share our expectations and the results we obtained from the evaluation.

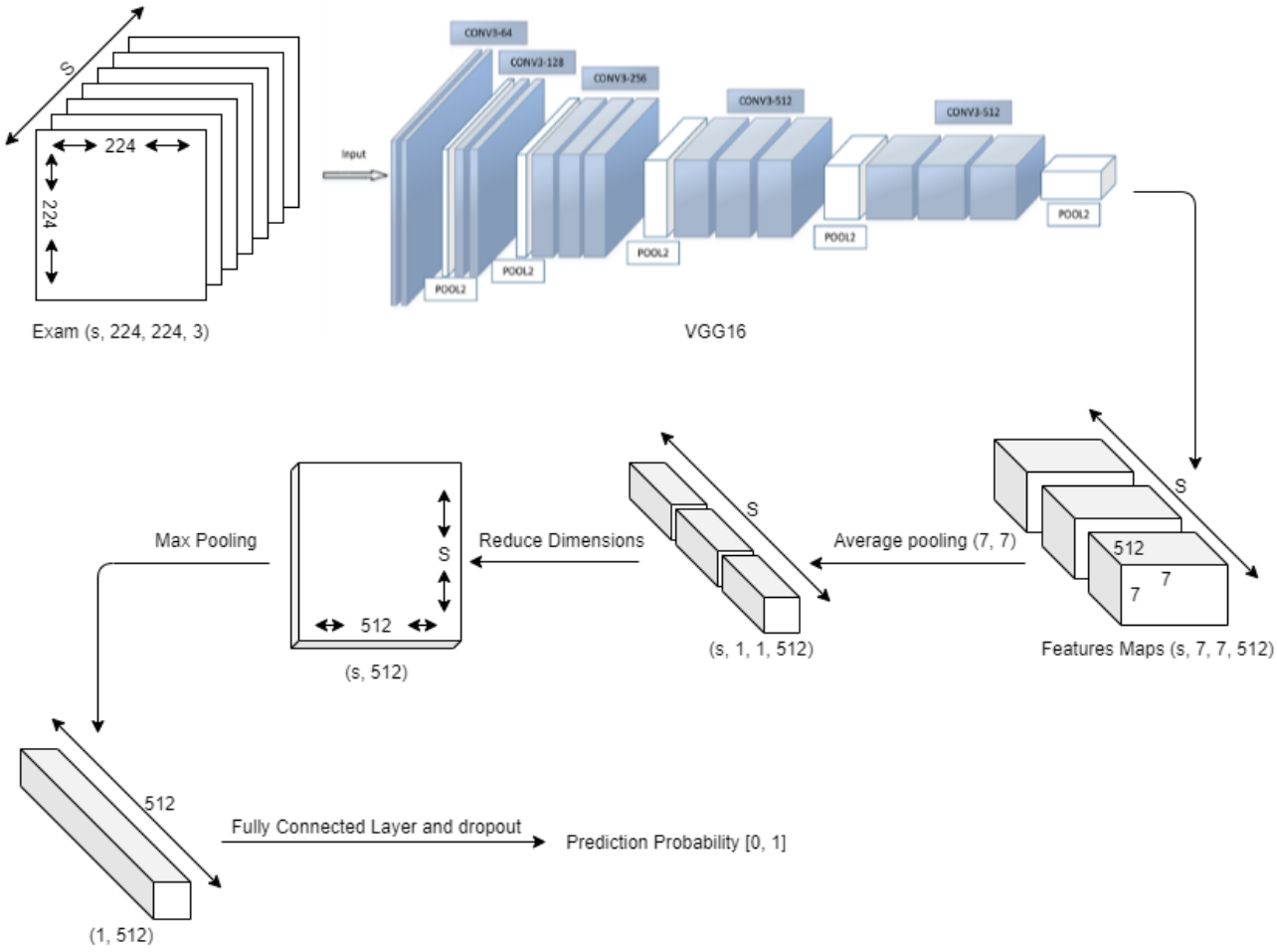

Figure 1. MRNet based on VGG16 model visualization

5.1. Abnormality

As can be observed the model predicts the abnormality the best of other classes with 90% accuracy, F1 Score of 94%, and AUC of 94.5% which is expected to perform better than the original MRNet model on the hidden test set.

5.2. ACL Tear

Although the performance cannot be compared to the abnormality detection the Accuracy is slightly higher than the original model.

5.3. Meniscus

in the meniscus detection our model performs with a slightly lower accuracy than the original model

6. Conclusion and future work

As the results indicates, the original model has a potential to improve, which can have huge impacts on the clinical automated assistance. in this section we list several methods that can be investigated to improve the performance.

- Obtaining more data: Deep Neural Networks are data hungry, they need to be feed with a large enough dataset to be able to generalize and reduce the chances of over-fitting.
- Using weight initialization from models trained on medical images.
- Augmentation, we slightly augmented the data, however, more image transformation can be applied to prevent over-fitting, and better generalize the model predictions.

References

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

MRNet based on VGG16 with ImageNet weights						
	Metrics	Axial	Coronal	Sagittal	1 View Average	3 Views model
Abnormality	Accuracy	85.00%	87.50%	92.50%	88.33%	90.00%
	AUC	87.71%	90.69%	94.65%	91.02%	94.53%
	F1 score	90.00%	92.31%	95.24%	92.52%	94.00%
ACL	Accuracy	81.67%	85.00%	83.33%	83.33%	88.33%
	AUC	91.08%	90.49%	88.33%	89.96%	94.61%
	F1 score	80.70%	82.69%	80.77%	81.39%	86.79%
Meniscus	Accuracy	69.17%	80.00%	73.33%	74.17%	71.67%
	AUC	72.88%	84.36%	76.92%	78.05%	80.01%
	F1 score	69.42%	76.00%	71.43%	72.28%	72.13%
MRNet based on VGG16 with Glorot Uniform Initializer						
	Metrics	Axial	Coronal	Sagittal	1 View Average	3 Views model
Abnormality	Accuracy	81.67%	82.50%	85.00%	83.06%	
	AUC	86.65%	81.39%	85.14%	84.39%	
	F1 score	89.00%	89.55%	91.26%	89.94%	
ACL	Accuracy	76.67%	70.83%	66.67%	71.39%	
	AUC	82.46%	80.13%	71.35%	77.98%	
	F1 score	68.89%	62.37%	51.22%	60.82%	
Meniscus	Accuracy	64.17%	70.83%	64.17%	66.39%	
	AUC	69.26%	74.18%	66.71%	70.05%	
	F1 score	65.60%	68.47%	46.91%	60.33%	

Figure 2. Results: Accuracy, F1 Score and AUC

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Curran Associates Inc.*, 1(25):1097–1105, 2012.

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2014.