https://doi.org/10.2352/EI.2022.34.14.COIMG-227

Fully RNN for Knee Ligament Tear Classification and Localization in MRI Scans

Kaiyue Zhu, Ying Chen, Xu Ouyang; Illinois Institute of Technology; Chicago, IL
Gregory White; Rush University; Chicago, IL
Gady Agam; Illinois Institute of Technology; Chicago, IL

Abstract

Diagnosing ligament injuries using MRI scans is a labor-intensive task that requires an expert. In this paper, we propose a fully Recurrent Neural Network (RNN) for detecting Anterior Cruciate Ligament (ACL) tears using MRI scans. The proposed network localizes the ACL and classifies it into several categories: ACL tear, normal tear, and healthy. Existing detection methods use deep learning networks based on single MRI sections, and in this way lose 3D spatial context. To address this, we propose a fully Recurrent Neural Network (RNN) model that processes a sequence of 3D sections and so captures 3D spatial context. The proposed network is based on a YOLOv3 backbone and can produce a sequence of decisions which are then combined by majority voting. Experimental results show improvement over state-of-the-art methods.

Introduction

Ligament injuries may lead to unsatisfactory knee function, reduce the quality of life, and result in an increased risk of osteoarthritis of the knee [16]. Surgery following knee injuries requires months of rehabilitation and may result in permanent disability [5]. Consequently, a fast and reliable diagnosis of ligament injuries may assist in improving patient outcomes. As reported by [25], the Anterior Cruciate Ligament (ACL) is the most commonly injured ligament in the human body. Radiologists diagnose ligament injuries based on Magnetic Resonance Imaging (MRI) scans which provide detailed information about the knee joint, bones, cartilage, tendons, muscles, and blood vessels from all angles. However, inspecting MRI scans is time-consuming and labor-intensive.

Computer-aided diagnosis (CAD) systems can be used to assist physicians in the interpretation of medical images and increase their productivity. The performance of CAD systems need not be comparable or better than that of a physician as the goal of such systems is to be complementary to a physician[6]. In the context of knee MRI, CAD systems help physicians to locate meniscus at the knee [12] and segment bone and cartilage [29]. Examples of CAD systems in other medical fields include lung nodule classification[34], chest pathology detection [3], skin lesion segmentation [10], and brain tumor detection[6].

In this paper, we present a deep learning approach to assist radiologists in localizing the ACL area and detecting the presence of an injury there. Our approach uses a fully Recurrent Neural Network (RNN) to detect a bounding box for the ACL and classifies it into one of several categories: ACL tear, normal tear, and healthy. Existing detection methods only use a single frame to make decisions, while our proposed method uses both spatial (x,y) and frame (z) relationships to detect ligament injuries throughout the processing pipeline, including the feature extraction, regression, and classification phases.

The main contributions of this paper are as follows: 1) We propose a fully RNN model that takes into account relations between frames (sections) to improve object detection. Compared with standard networks, our proposed model shares features and weights across the section dimension. In addition, information from prior sections is used to determine the current output. 2) A modified channel-spatial attention module is implemented to improve detection accuracy. Experimental results show that the channel-spatial attention module helps to increase the $mAP_{.50:.05:.95}$ of our proposed model from 34% to 37.7%. 3) Our proposed model is a fully RNN and thus is capable of producing a sequence of decisions. Experimental evaluation shows that compared with YOLOv3 which uses single sections to make decisions, our proposed model increases the $mAP_{.50:.05:.95}$ from 36.6% to 37.7%. To the best of our knowledge, this is the first work that uses a fully RNN model to perform object detection with three-dimensional medical data.

Related Work

Existing work on knee MRI diagnosis focus mostly on classification tasks [4, 27] where an entire scan is classified without detecting an image region corresponding to the classification. This is in contrast to the approach we describe in this paper where we detect and show the region corresponding to the condition and so justify the decision to the user. Other work, such as [15, 1], can locate the injured area, but require the extraction of a Region of Interest (ROI) first. The prediction is made only based on the ROI. This two-stage approach is somewhat unstable as an inaccurately extracted ROI will degrade the overall performance. In contrast to this, our proposed model is trained in an end-to-end manner. Furthermore, unlike most existing methods where prediction is made based on single MRI sections, our proposed method processes a sequence of 3D sections and fully utilizes 3D spatial context to make decisions.

Our proposed method can be classified as an object detection approach. Object detection aims to detect instances of a semantic object class with a bounding box in an image. Convolutional Neural Network (CNN) has been shown to be an effective tool for this task. The YOLO model [20] is a successful one-stage approach for object detection. It splits an image into grid cells and within each grid cell tests detection in m bounding boxes. For each of the bounding boxes, the network outputs a class probability and a set of bounding box offset values (box center, height, and width) for the final decision. Being a one-stage approach makes the YOLO framework simple and fast. YOLOv2 [18] improved accuracy and made detection faster by replacing VGG16 [24] with Darknet [19]. YOLOv3 [21] improved performance further by using a feature pyramid networks (FPN) [14] to output extracted features at different scales. Our

proposed approach is based on the YOLOv3 framework.

To take into account information at multiple time steps, RNN uses an internal state to save past information. The output of the RNN model is calculated based on not only the current input but also past information stored in internal state variables. RNNs are commonly used to deal with sequential data such as text, speech, and video sequences. The Long short-term memory (LSTM) model [9] is a widely used RNN and has been successfully applied to various tasks such as machine translation [28], text classification [13], feature representation [26], and image captioning[30]. The core idea in LSTM is to use an additional internal state, called cell state, to remove or add information to the cell memory. Most of the above approaches use a fully connected LSTM (FC-LSTM) layer and do not consider spatial correlation.

Xingjian *et al*.[33], proposed a Convolutional LSTM (ConvLSTM) network for precipitation nowcasting to predict the future rainfall intensity in a local region over a relatively short period of time. To model spatiotemporal relationships, they extended the idea of FC-LSTM to ConvLSTM which has convolutional structures in both the input-to-state and state-to-state transitions. Their study shows that the ConvLSTM model consistently outperforms the FC-LSTM.

The ROLO network, a recurrent YOLO network, was proposed in[17]. The ROLO network aims to solve object tracking. In their work, YOLO is used for basic detection and LSTM is used in regression aiming to restrict the location prediction to a spatial range. This approach works with 2D data. Given 3D or time-series data, 2D convolution treats sections individually without considering the relationship between them. To address this problem, our proposed model fully uses the relationship between frames not only at the detection phase but also at the feature extraction phase.

Attention mechanism [2] was first introduced for machine translation and now is widely implemented in computer vision networks. Several attention mechanisms have been successfully applied on various deep learning tasks [31, 32, 11, 22]. Our proposed attention module is based on [35]. Similar to their work, we use two branches to generate channel-wise and spatial-wise attention and then fuse the branches. The difference in our work is that instead of implementing attention using CNN, we apply it in our proposed recurrent network so that the attention area on a frame can be propagated to the following frame.

Proposed Approach

Given 3D MRI data, sections near to each other have a relationship between them. Our proposed approach aims to utilize this additional relationship to improve the performance of ligament injury detection. We treat the relationship between sections as temporal relationships and use our proposed fully RNN model to analyze 3D data. Our model is based on the YOLOv3 framework but replaces the convolutional layers with the ConvLSTM layers. We further simplified the original YOLOv3 framework by replacing the prediction on three different scales with two scales. As a consequence, our proposed model has fewer layers and filters in comparison with YOLOv3. The proposed model architecture is shown in Fig.1. It is based on multiple ConvLSTM and channel-spatial attention blocks.

ConvLSTM Block

The structure of the ConvLSTM layer is shown in Fig.2. The ConvLSTM layer accepts 4D features (time step, channel, height, width) and uses 2D convolution operations. We use the cell structure of FC-LSTM, instead of the one used by Xingjian *et al*.[33]. The cell in Xingjian *et al*.implements a "peephole connection"[7] which allows gates to access the cell state. We removed the "peephole connection" to reduce the number of parameters. The core equations of ConvLSTM are can be formulated as Eq. 1

$$\begin{aligned} i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + b_f) \\ o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + b_o) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \quad , \qquad (1)$$

where in these equations, i_t, f_t, and o_t are the input, forget, and output gates, respectively. The cell and hidden state at time t are denoted as C_t and h_t. The operations $*$ and $\odot$ represent convolution and element-wise multiplication. The Sigmoid function is denoted as σ.

ConvLSTM layers are arranged in blocks based on the residual connection of Darknet [21]. In each block, the first of stack ConvLSTM layers uses a 1×1 kernel to reduce the dimension of the input. This is then followed by a 3×3 kernel designed to increase the number of output channels and extract features at the same time. Then, the block outputs and inputs are added element-wise to form the final output of the block. In comparison with the original residual block of He *et al*.[8] that uses a stack of 3 layers, experiments in [21] suggest that this block produces better results.

Channel-spatial attention module

Our channel-spatial module, as shown in Fig.3, is a modified version of the module proposed in [35] and can analyze 4D features. We switch all dense layers in the original module to LSTM layers and replace 2D convolutional layers with ConvLSTM layers. Our attention module has two branches: channel and spatial attention branches. Given 4D input features Q with shape (t,c,h,w) where h and w are feature dimension, c is the number of channels and t is the number of frames, the module output, Q_2, has the same shape as the input.

The channel attention block re-weights features so that important channels receive higher weights compared with non-essential channels. The output of this block is a channel attention mask Q'. The spatial attention block, similar to channel attention block, re-weights features in the spatial domain and generates spatial attention mask Q''. Channel attention and spatial attention mask re-weight the original features separately by element-wise multiplication. The fused feature map, Q_1, is produced by element-wise addition of the channel and spatial attention outputs. A residual connection is applied to form the final output Q_2.

Experiments

Datasets

Two knee MRI datasets are used in this work: MRNet [4] and kneeMRI [27]. The MRNet dataset consists of 1,370 knee MRI exams performed at Stanford University Medical Center. The dataset contains 1,104 (80.6%) abnormal exams, with 319 (23.3%) ACL tears and 508 (37.1%) meniscal tears. Exam labels were obtained through manual extraction from clinical reports[4]. Spatial annotations were done by a radiologist. Since labeling the entire dataset

Figure 1. *YOLO_ConvLSTM model architecture*

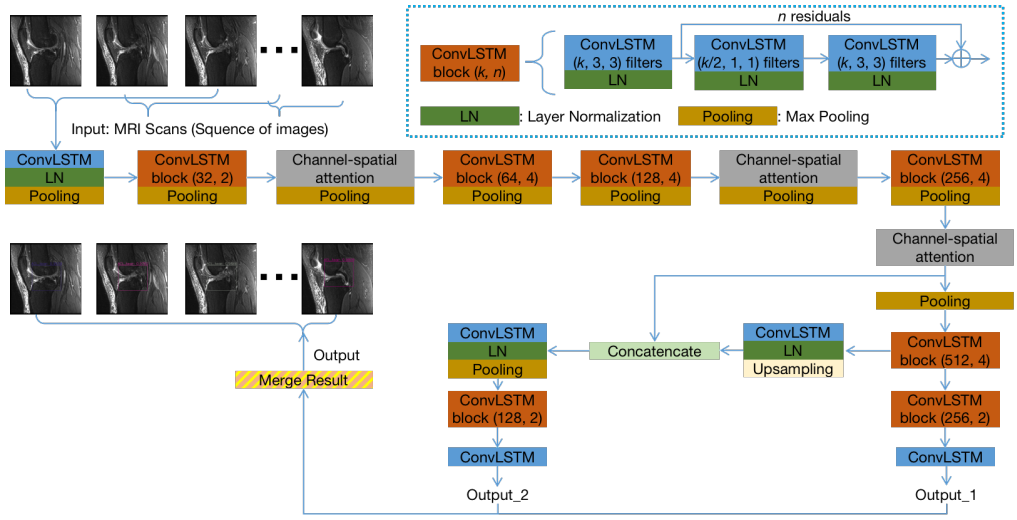


Figure 2. *ConvLSTM cell*

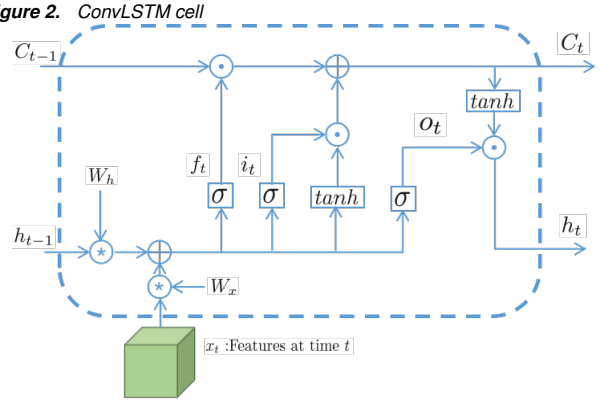
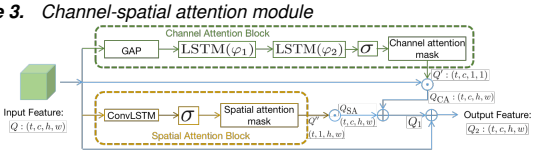


Figure 3. *Channel-spatial attention module*



is time-consuming, the radiologist annotated only 2 to 4 continuous frames with a bounding box and assigned a label (ACL tear or normal tear) if a ligament injury was identified. In all the radiologist labeled 228 exams in this manner.

The kneeMRI Dataset has annotations of healthy and injured ligaments. It contains sagittal views with three different ACL diagnoses: not injured (690 cases), partially injured (172 cases), and completely ruptured (55 cases). We use the non-injured cases as healthy labels in this work. To balance the injured and non-injured (healthy) exams, we randomly selected 263 healthy exams from the kneeMRI dataset and formed a combined dataset with the MRNet dataset. The combined dataset was split into the training and testing subsets as shown in Table 1. Note that the table shows the number of labeled frames, not the number of exams. Each exam normally has multiple (4 to 6) continuous labeled frames.

Table 1. Summary of the combined dataset. There are total 1419 and 306 frames in the training and test sets, respectively.

	Train Set	Test Set	Total
Normal Tear (frames)	224	45	269
ACL Tear (frames)	480	104	584
Healthy (frames)	715	157	872
Total number of frames	1419	306	1725

Evaluation Metrics

To measure the performance of models, we use several metrics. We evaluate object detection as two tasks: classification and localization. To evaluate classification, we build the class confusion matrix and compute based on it the recall, precision, and F_1 score per class type. While in general detection problems, images may contain more than one object, in our specific problem we can assume that each frame contains only one object. To evaluate localization, we use Average Precision (AP) for each class type. We calculate AP_{50} and AP_{75} for each class under certain Intersection over Union (IoU) thresholds (0.5 and 0.75). We then compute the mean AP of all classes under certain IoU threshold: mAP_{50} and mAP_{75}. The final measurement $mAP_{.50:.05:.95}$ is the mean AP for IoU from 0.5 to 0.95 with a step size of 0.05 of all classes.

To our knowledge, while there is some existing work on the classification of ACL conditions without localization, there is no existing work targeting ACL detection. Thus, to form a baseline for comparison we use the original YOLOv3 model which we trained with the labeled knee MRI data. We trained two YOLOv3 models: one trained from scratch and the other starting with a pretrained ILSVRC [23] dataset weights. We also have two implementations of the proposed model: with and without the attention module so that we can measure the contribution of the channel-spatial attention module.

Results

The classification and localization results are shown in Table 2 and Table 3, respectively. The best value is marked in red. As can be observed in the tables, while that the proposed approach does

Table 2. Classification metric results

	Recall			Precision			F1 Score		
Class Type	**ACL Tear**	**Normal Tear**	**Healthy**	**ACL Tear**	**Normal Tear**	**Healthy**	**ACL Tear**	**Normal Tear**	**Healthy**
YOLOv3	0.403	0.444	0.891	0.875	**0.869**	**1.0**	0.552	0.588	0.942
YOLOv3 + pretraining	0.798	0.733	**1.0**	**0.965**	0.825	**1.0**	0.873	**0.776**	**1.0**
Proposed model + without attention	0.875	**0.822**	0.898	0.866	0.685	0.992	0.871	0.747	0.943
Proposed model	**0.894**	0.711	0.987	0.885	0.744	0.981	**0.889**	0.727	0.984

Table 3. Localization metric results

	AP_{50}				AP_{75}				$mAP_{.50:.05:.95}$
Class Type	**ACL Tear**	**Normal Tear**	**Healthy**	**mAP_{50}**	**ACL Tear**	**Normal Tear**	**Healthy**	**mAP_{75}**	**All classes**
YOLOv3	0.432	0.314	0.818	0.521	0.188	0.136	0.259	0.194	0.260
YOLOv3 + pretraining	0.813	0.442	0.903	0.719	0.346	**0.160**	**0.438**	**0.315**	0.366
Proposed model + without attention	0.780	**0.627**	0.818	0.742	0.276	0.036	0.319	0.210	0.340
Proposed model	**0.815**	0.512	**0.905**	**0.744**	**0.373**	0.061	0.357	0.264	**0.377**

not improve the classification F1 score when compared to YOLOv3 with pre-training, it does improve the mAP score thus demonstrating better detection. The tables also show that our attention module contributes to improved detection results. Note that the proposed approach is substantially less complex than YOLOv3.

From the classification results in Table 2, we can observe that the channel-spatial attention module helps to improve the recall of the ACL tear and healthy classes by about 1.9% and 8.9%, respectively, while the recall of the normal tear class reduces by 11.1%. For precision, the ACL tear and normal tear scores increase by 1.9% and 5.9%, respectively, while healthy label scores decrease by 1.1%. The F_1 score for the ACL tear and healthy classes improve by 1.8% and 4.1%, respectively, while scores for the normal tear class decrease by 2%. The overall performance is improved after introducing channel-spatial attention. When compared with the YOLOv3 framework, our proposed model is better than the YOLOv3 without pretraining in terms of recall and F_1 scores, while precision is a little lower. The pretrained YOLOv3 results are comparable to the ones produced by our proposed model. Our proposed model has a higher recall and F_1 scores for the ACL tear class, while the precision of the ACL tear class of our proposed model is lower by 8%. Our proposed model does not perform as well as the pretrained YOLOv3 in the normal tear class, while for the healthy class there is no big difference between the two models, where both models perform well.

From the localization metric results in Table 3, we see that the channel-spatial attention module helps to improve the $mAP_{.50:.05:.95}$ scores from 34.0% to 37.7%. The mAP_{50} and mAP_{75} scores are increased by 0.2% and 5.4%, respectively. For the $mAP_{.50:.05:.95}$ score, the proposed model achieves a score of 37.7% which is about 3.7% higher than the score without attention module. Compared with the YOLOv3 framework, our proposed model is better than the model without pretraining. At AP_{50}, our proposed model is about twice better than YOLOv3 without pretraining in ACL tear cases, 19.8% higher in normal tear cases, and 8.7% higher in healthy cases. The mAP_{50} score is 22.3% higher than the YOLOv3 without pretraining model. However, at AP_{75} the proposed model is 7.5% lower than YOLOv3 without pretraining for the normal tear class, even though the detection of ACL and healthy classes are much better. The mAP_{75} score of our proposed model is 7% higher than YOLOv3 without pretraining. Our model increases the $mAP_{.50:.05:.95}$ score by 11.7% from YOLOv3 without pretraining. We can conclude our proposed model has better performance compared with a YOLOv3 model without pretraining. Compared with pretrained YOLOv3, our model shows advantages at AP_{50}. All three classes are improved by 0.2%, 7%, and 0.2%, respectively. The mAP_{50} of our model is 2.5% higher. However, at mAP_{75}, the pretrained YOLOv3 is about 5.1% higher than our proposed model. At the final measurement $mAP_{.50:.05:.95}$, our model is 1.1% higher than the pretrained YOLOv3 model. In summary, we conclude that at the lower IoU threshold, our proposed model has advantages over the pretrained YOLOv3 and that with increasing threshold values, this advantage becomes smaller. The overall measurement, $mAP_{.50:.05:.95}$, shows that our proposed model is better than the pretrained YOLOv3 and improves the metric from 36.6% to 37.7%.

Fig. 4 shows examples of predicted results produced by our proposed method. The top row (a) shows the ground truth annotation, whereas the bottom row (b) shows our predicted result. As can be observed, the proposed model is able to classify the injury type and detect the injured region at the same time.

Conclusion

We propose a fully RNN deep learning model with a channel-spatial attention module for knee MRI diagnosis using supervised learning. Experimental results show that our proposed approach is better than YOLOv3 and validate our assumption that there is benefit in maintaining section context using RNN to analyze 3D data. From a model complexity aspect, our model has fewer layers and kernels than YOLOv3. The YOLOv3 has about 81 convolutional layers, while our model has 62 ConvLSTM layers (23% reduction). Further, the number of filters is reduced from 1024 to 512. The experimental results also validate that the channel-spatial attention module helps in improving the performance of our proposed model.

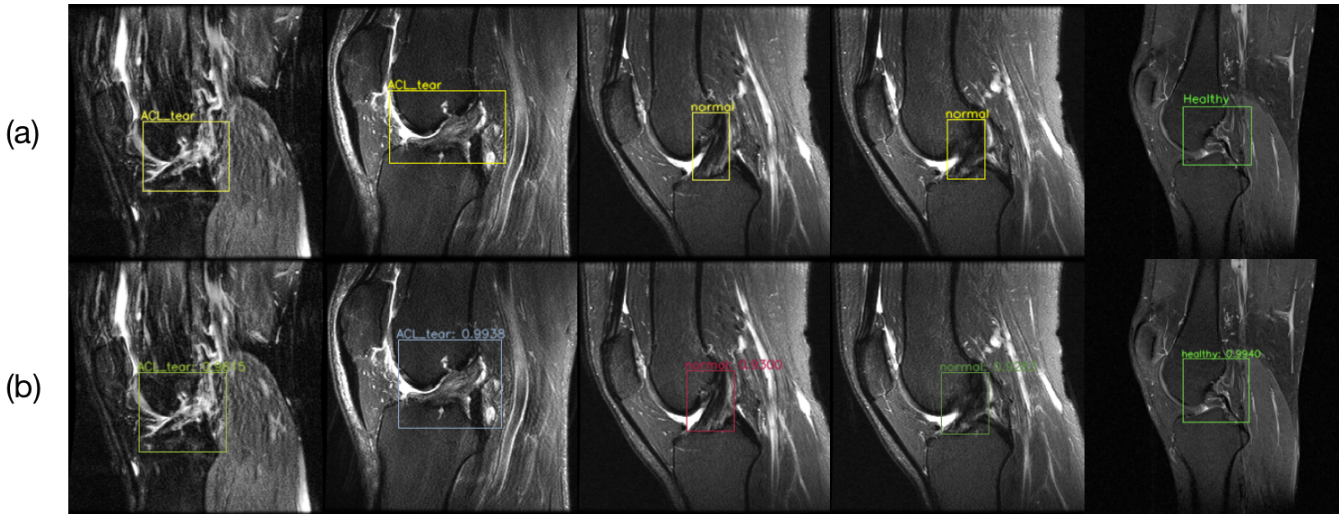


Figure 4. *Final output visualization: (a) shows the groundtruth and (b) is our predicted output.*

In comparison with the YOLOv3 framework, our proposed model has a better performance than a YOLOv3 model with or without pretraining (smaller improvement when compared to pretrained model). The overall performance measure $\text{mAP}_{.50:.05:.95}$, validates that our proposed model is better than the YOLOv3 framework.

References

[1] Mazhar Javed Awan, Mohd Shafry Mohd Rahim, Naomie Salim, Mazin Abed Mohammed, Begonya Garcia-Zapirain, and Karrar Hameed Abdulkareem. Efficient detection of knee anterior cruciate ligament from magnetic resonance imaging using deep learning approach. *Diagnostics*, 11(1), 2021.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Yaniv Bar, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen, and Hayit Greenspan. Chest pathology detection using deep learning with non-medical training. In *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, pages 294–297. IEEE, 2015.

[4] 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.

[5] Steve Bollen. Epidemiology of knee injuries: diagnosis and triage. *British journal of sports medicine*, 34(3):227–228, 2000.

[6] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007.

[7] Felix A Gers and Jürgen Schmidhuber. Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 3, pages 189–194. IEEE, 2000.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[10] Mohammad H Jafari, Nader Karimi, Ebrahim Nasr-Esfahani, Shadrokh Samavi, S Mohamad R Soroushmehr, K Ward, and Kayvan Najarian. Skin lesion segmentation in clinical images using deep learning. In *2016 23rd International conference on pattern recognition (ICPR)*, pages 337–342. IEEE, 2016.

[11] Yuzhu Ji, Haijun Zhang, and QM Jonathan Wu. Salient object detection via multi-scale attention cnn. *Neurocomputing*, 322:130–140, 2018.

[12] Cemal Köse, Okyay Gençalioğlu, and Uğur Şevik. An automatic diagnosis method for the knee meniscus tears in mr images. *Expert systems with applications*, 36(2):1208–1216, 2009.

[13] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[15] Fang Liu, Zhaoye Zhou, Alexey Samsonov, Donna Blankenbaker, Will Larison, Andrew Kanarek, Kevin Lian, Shivkumar Kambhampati, and Richard Kijowski. Deep learning approach for evaluating knee mr images: achieving high diagnostic performance for cartilage lesion detection. *Radiology*, 289(1):160–169, 2018.

[16] L Stefan Lohmander, P Martin Englund, Ludvig L Dahl, and Ewa M Roos. The long-term consequence of anterior cruciate ligament and meniscus injuries: osteoarthritis. *The American journal of sports medicine*, 35(10):1756–1769, 2007.

[17] Guanghan Ning, Zhi Zhang, Chen Huang, Xiaobo Ren, Haohong Wang, Canhui Cai, and Zhihai He. Spatially supervised recurrent convolutional neural networks for visual object tracking. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4. IEEE, 2017.

[18] J Redmon and A Farhadi. Yolo9000: Better, faster, stronger. arxiv 2016. *arXiv preprint arXiv:1612.08242*, 2016.

[19] Joseph Redmon. Darknet: Open source neural networks in c. `http://pjreddie.com/darknet/`, 2013–2016.

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You

only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[21] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[22] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6656–6664, 2017.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] Kurt P Spindler and Rick W Wright. Anterior cruciate ligament tear. *New England Journal of Medicine*, 359(20):2135–2142, 2008.

[26] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

[27] 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.

[28] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[29] Graham Vincent, Chris Wolstenholme, Ian Scott, and Mike Bowes. Fully automatic segmentation of the knee joint using active appearance models. *Medical Image Analysis for the Clinic: A Grand Challenge*, 1:224, 2010.

[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[31] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

[32] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1448–1457, 2019.

[33] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[34] Fan Zhang, Yang Song, Weidong Cai, Min-Zhao Lee, Yun Zhou, Heng Huang, Shimin Shan, Michael J Fulham, and Dagan D Feng. Lung nodule classification with multilevel patch-based context analysis. *IEEE Transactions on Biomedical Engineering*, 61(4):1155–1166, 2013.

[35] Yan Zhang, Min Fang, and Nian Wang. Channel-spatial attention network for fewshot classification. *Plos one*, 14(12):e0225426, 2019.

Author Biography

Kaiyue Zhu *received a BSc degree in Electrical Engineering from the North China Electric Power University and the Illinois Institute of Technology (2013) and two MSc degrees in Electrical Engineering (2014) and Computer Science (2017) from the Illinois Institute of Technology. He is currently a Ph.D. student in the Computer Science department at the Illinois Institute of Technology. His main research areas are in computer vision, deep learning, and machine learning.*

Ying Chen *received her BSc in Computer Science from Fuzhou University (2004) and her PhD in Computer Science from Illinois Institute of Technology (2021). She is currently a researcher at HERE technologies in Cambridge, MA, where she works on the development of 2D perception.*

Xu Ouyang *is a PhD student in the Computer Science department at the Illinois Institute of Technology. He research focuses on image synthesis with deep learning.*

Gregory White*, MD, is an Assistant Professor at the Diagnostic Radiology and Nuclear Medicine department at Rush University Medical Center.*

Gady Agam*, PhD, is an Associate Professor of Computer Science and the director of the Visual Computing Lab, at the Illinois Institute of Technology.*