Universal lymph node detection in T2 MRI using neural networks

Tejas Sudharshan Mathai · Sungwon Lee · Thomas C. Shen · Zhiyong Lu · Ronald M. Summers

Original Article

Abstract

Purpose Identification of lymph nodes (LNs) that are suspicious for metastasis in T2 Magnetic Resonance Imaging (MRI) is critical for assessment of lymphadenopathy. Prior work on LN detection has been limited to specific anatomical regions of the body (pelvis, rectum). Therefore, an approach to universally detect both benign and metastatic nodes in T2 MRI studies of the body is highly desirable.

Methods We developed a Computer Aided Detection (CAD) pipeline to universally identify LN in T2 MRI. First, we trained various neural networks for detecting LN: Faster RCNN with and without Hard Negative Example Mining (HNEM), FCOS, FoveaBox, VFNet, and Detection Transformer (DETR). Next, we show that VFNet with Adaptive Training Sample Selection (ATSS) outperformed Faster RCNN with HNEM. Finally, we ensembled models that surpassed a 45% mAP threshold.

Results Experiments on 122 test studies revealed that VFNet achieved a 51.1% mAP and 78.7% recall at 4 false positives (FP) per volume, while the one-stage model ensemble achieved a mAP of 52.3% and sensitivity of 78.7% at 4FP. We found that VFNet and the one-stage model ensemble can be interchangeably used in the CAD pipeline.

Conclusion Our CAD pipeline universally detected both benign and metastatic nodes in T2 MRI studies, resulting in a sensitivity improvement of \( \approx 14\% \) over the current LN detection approaches (sensitivity of 78.7% at 4 FP vs. 64.6% at 5 FP per volume).

Keywords MRI · T2 · Lymph node · Detection · Deep learning

Introduction

The lymphatic system consists of lymph nodes (LNs) that fight infection in the body by ridding it of foreign matter. Patients with lymphadenopathy may present with enlarged and metastatic nodes, which need to be distinguished from non-metastatic LNs. Furthermore, cancerous LNs appearing at sites in the body that do not correlate with the first site of lymphatic spread indicate distant metastasis [1]. Radiologists follow the AJCC guidelines [1] to measure the nodal size as it defines cancer management and therapy. Sizing a node’s long and short axis diameters (LAD and SAD) enables them to find those with a SAD \( \geq 10 \) mm, which are considered suspicious for cancer [2]. Multi-parametric MRI (mpMRI) imaging is typically ordered to visualize LNs, and among other sequences, it contains T2-weighted MRI, Diffusion Weighted Imaging (DWI), Apparent Diffusion Coefficient (ADC) maps. The inconsistent shapes and myriad appearances of LNs at various anatomical locations in these sequences makes LN measurement challenging. Moreover, the use of multiple scanners with different exam protocols in a clinical setting is an additional confounding factor that can cause radiologists to sometimes miss suspicious LN. Due to these circumstances, an automated CAD pipeline is needed for identification of LN in T2 MRI scans, such that LN sizing can be completed.

There has been a moderate amount of prior work [3–8] on MRI-based LN detection, but they focus on specific anatomical regions. A Mask-RCNN network was used in [3] to identify and segment LN in pelvic MRI images with 62.6% recall at 8.2 FP per volume. In [4], a GentleBoost classifier followed by a convolutional neural network (CNN) was used to achieve a 85% LN detection sensitivity at 5-10 FP per image. In [5], Faster-RCNN identified pelvic LN with an AUC of 91.2%. Universal LN detection in the abdomen has only been researched recently with [6] describing a 64.6%
sensitivity at 5 FP per volume, while [7] posted a 91.6% sensitivity at 4 FP per image on 2D slices only.

In this study, we design a CAD pipeline that detects LN in T2-weighted MRI volumes. We evaluate recently proposed detection models, such as VFNet [9], DETR transformer [10], FoveaBox [11], Faster RCNN [12], etc., on the LN detection task. Next, we describe the superiority of the VFNet model with Adaptive Training Sample Selection (ATSS) [9] over a Faster RCNN model trained with HNEM [13]. Finally, models that exceed a 45% mAP threshold are assembled together, and we observe that the one-stage model ensemble and the VFNet model are similarly proficient at detecting LN in T2 MRI scans: a ≥51% mean average precision (mAP) and 78.7% sensitivity at 4FP per volume. We do not use DWI scans in this work in order to compare our results against prior work (although the addition of DWI would enable reliable LN detection). Despite this, an improvement in sensitivity by 14% (78.7% recall at 4FP vs. 64.6% recall at 5P per volume) is achieved over a recently published LN detection method [6].

Methods

Object detectors

Recently proposed object detection models that were used in this paper are: (1) Faster R-CNN [12], (2) VFNet [9], (3) FCOS [14], (4) FoveaBox [11], and (5) DETR [10]. Faster RCNN is a two-stage anchor-based detector where the first stage generates region proposals of objects of interest, while the second stage classifies these proposals and regresses the object’s bounding box coordinates. As Faster RCNN can generate numerous false positives (FP), Hard Negative Example Mining (HNEM) [13] is a strategy that can be employed to reduce the number and thereby boost precision and recall values. This method involves sampling negative bounding boxes, which do not overlap with the ground-truth and have prediction probability scores that are larger than the true positives, allowing us to iteratively train the model to reduce the incidence of FP.

FCOS [14] extracts features from multi-levels of the Feature Pyramid Network (FPN) [15] using shared heads at different levels to regress the bounding box coordinates, and a centerness score to eliminate false predictions that are far away from the center of the target object location. FoveaBox [11] consists of a backbone and a fovea head network. The backbone is a simple FPN yielding feature maps at multi-levels. The fovea head has two branches: the classification branch performs per-pixel classification on the backbone features, and the box regression branch predicts the coordinates of the box at each location that may be potentially covered by an object.

VFNet combines FCOS (without the centerness branch) with the Adaptive Training Sample Selection (ATSS) module [9] to accurately select the high quality detections from the pool of candidate predictions. By replacing the classification score with the IoU between the ground-truth and the prediction, an IoU-aware classification score (IACS) is obtained that is merged in a novel IoU-aware Varifocal loss, which down-weights the contribution of negative detections to the loss, but up-weights the contribution of positive samples. Inside this setup, a star-shaped bounding box provides context cues to reduce the misalignment between ground-truth and the predicted boxes.

The DEtection Transformer (DETR) [10] uses a transformer encoder–decoder architecture, wherein the self-attention provided by the transformer removes duplicate predictions. Spatial position encoding is added to the feature maps obtained from the ResNet50 backbone inside the encoder–decoder architecture. Prediction feed-forward network heads decode the output embeddings of object queries independently into the bounding box coordinates and labels. A bipartite set matching loss is used to train the model; however, in our implementation, we replaced the original cross entropy classification loss with the focal loss [16] to overcome the class imbalance problem between positive and negative samples, and weighted it with $\lambda_c = 2$.

Weighted Boxes Fusion. At test time, utilizing multiple object detection models or multiple epochs of a single model can yield significantly higher mAP and sensitivity metrics at the cost of a large number of predictions. The predictions consist of the bounding box coordinates of an object and the classification probability. Many strategies, such as Non-Maximal Suppression (NMS) [17,18], exist to reduce these duplicates, but we use Weighted Boxes Fusion (WBF) [18] to reduce the number of FP during test time.

Experiments and results

Data

Between January 2015 and September 2019, patients who underwent MRI imaging of the abdomen were identified. Initially, 584 T2-weighted abdominal MRI studies were downloaded from the National Institutes of Health (NIH) Picture Archiving and Communication System (PACS). Radiology reports for these studies were also downloaded, and natural language processing [19] was used to pull the nodal extent and size measurements in these studies. Patient diagnosis was not extracted from the report using NLP. A radiologist performed a quality check on the collected data, and removed incorrect annotations and scans containing only one LN annotation (either LAD or SAD measures). This process
yielded 421 T2 studies from 383 patients (224 male, 159 female), which were divided into training (~60%, 254 studies), validation (~10%, 45 studies), and test (~30%, 122 studies) splits with unique patients per split. Patients were between the ages of 6 years and 85 years with a median age of 48 years. The train and validation set consisted of only 2D key slices with 1–2 LN annotated with both the LAD and SAD, while the 3D extent of all LN in the test set were fully labeled. N4 bias normalization [20] was first applied to the studies, followed by normalization to [1, 99%] of the voxel intensity range [21], such that the contrast between bright and dark structures in the slices was boosted. The resulting studies had dimensions in the range from (256 ~ 640) × (192 ~ 640) × (18 ~ 60) voxels.

**Implementation**

As radiologists read MRI scans by scrolling through the slices, we replicated their workflow by using 3-slice T2 MRI images where the center slice contained the LN annotated by the radiologist. These 3-channel images were passed to each detection model implemented with the mmDetection framework [22]. Data augmentation was performed: random flips, crops, shifts and rotations in the range of [0, 32] pixels, and [0, 10] degrees, etc. ResNet-50 was the backbone for all models used in this paper (pre-trained with MS COCO weights). A grid search was run across the batch size and learning rate parameters to obtain the optimal values; batch size was kept constant at 2 for all models, while the learning rate was 1e-3 for all models (for 12 epochs) except DETR whose learning rate was 1e-5 (for 150 epochs). Each model was executed 5 times, and the best checkpoint from each run with the lowest validation loss was chosen for testing. The ensemble comprised of the lowest validation loss checkpoint for each model across all runs. All experiments were run on a workstation running Ubuntu 18.04LTS with 4 NVIDIA Tesla V100 GPUs. Evaluation was always performed at an IoU threshold of 25% to be consistent with prior work [6].

**Results**

From prior work, a clinically acceptable result for LN detection is a sensitivity of 65% at 4–6 FP per volume [3,6]. Higher detection sensitivity is required as missing any cancerous LN can be troublesome from a screening and triage perspective; however adequate precision is also necessary as too many FP can overwhelm the interpretation of a scan. Our quantitative results are presented in Table 1 in terms of mean average precision (mAP) and sensitivity, while qualitative results are shown in Fig. 1. Our first experiment was to assess the performance of Faster RCNN with and without HNEM, respectively. Faster RCNN with HNEM improved LN detection mAP and sensitivity by ~2.2% (48.6% vs. 46.4%) and ~3% (76.9% vs. 74.1%) at 4FP, respectively, over Faster RCNN without HNEM. These results are in accordance with prior work on the utility of Faster RCNN with HNEM in reducing FP [13].

In our second experiment, we commissioned the VFNet model to detect LN. VFNet with the ATSS module adequately increased the detection mAP and sensitivity by ~4.7% (51.1% vs. 46.4%) and ~4.6% (78.7% vs. 74.1%) at 4FP, respectively, over Faster RCNN without HNEM. There was a marginal improvement of ~2.5% mAP and ~2% recall by VFNet over Faster RCNN with HNEM. For our third experiment, only the FCOS model in VFNet was used without the ATSS module, and the detection mAP dropped by

| Method                          | mAP | S@0.5 | S@1  | S@2  | S@4  | S@6  | S@8  | S@16 |
|---------------------------------|-----|-------|------|------|------|------|------|------|
| Faster RCNN (without HNEM) [12] | 46.4| 42.3  | 54.3 | 65.1 | 74.1 | 77.8 | 78.1 | 78.1 |
| Faster RCNN (with HNEM) [12]   | 48.6| 39.4  | 53.2 | 66.5 | 76.9 | 79.5 | 79.5 | 79.5 |
| VFNet [9]                      | 51.1| 45.7  | 56.8 | 67.9 | 78.7 | 82.6 | 84.9 | 86.2 |
| FCOS [14]                      | 39.6| 35.1  | 47.1 | 60.1 | 72   | 76.9 | 79.5 | 81.3 |
| FoveaBox [11]                  | 50.2| 46    | 57.2 | 68   | 77.4 | 82.1 | 84.1 | 85.6 |
| DETR [10]                      | 45.5| 39.2  | 51   | 61.3 | 74.8 | 79.8 | 81.8 | 82.5 |
| Ensemble (All)                 | 49.7| 44.6  | 58.4 | 69   | 78.5 | 82.3 | 83.8 | 84.1 |
| Ensemble (One-Stage)           | 52.3| 46.5  | 58   | 68.9 | 78.7 | 82.7 | 85.2 | 86.4 |
| Wang 2022 [6] (3D)             | –   | –     | –    | –    | 64.6 | –    | –    | –    |
| Zhao 2020 [3] (3D)             | 64.5| –     | –    | –    | –    | –    | –    | –    |
| Mathai 2021 [7] (2D only)      | 71.7| 73.8  | 79.7 | 85.7 | 91.6 | 91.6 | 91.6 | 91.6 |
| Mathai 2021 [8] (2D only)      | 65.4| 65.4  | 76.1 | 88.1 | 91.7 | 91.7 | 91.7 | 91.7 |
| Debats 2019 [4] (2D only)      | –   | –     | –    | –    | –    | –    | 80   | –    |

*S* stands for Sensitivity @[0.5, 1, 2, 4, 6, 8, 16] FP. – indicates unavailable metric values.

The bold value that indicates the highest values for each metric across different comparative methods are denoted.
Fig. 1 LN detection results by different neural networks on a slice from a T2 MRI volume. Green boxes: ground-truth, yellow: true positives, and red: false positives. (a) and (b) show the Faster RCNN model without and with HNEM, respectively. Notice the reduced FP count in (b). In (c) and (d), LN detection results by VFNet with ATSS and FCOS are shown with more FP detected by FCOS. (e) and (f) show the results from other models such as FoveaBox and DETR. Finally, (g) shows the ensemble detection result using all models (Faster RCNN + HNEM, VFNet, FoveaBox and DETR), while (h) shows the result of using only the one-stage (OS) model ensemble (VFNet, FoveaBox and DETR). Note the missed LN in (g) captured in (h).

∼11% (51.1% vs. 39.6%), while the sensitivity dropped by ∼7% (78.7% vs. 72%). This experiment attested to the utility of the ATSS module within the VFNet. Note however that these results already showed a significant improvement over a recently proposed LN detection approach [6]; the worst performing FCOS model posted a ∼7.4% increase in sensitivity at 4FP per volume.

The fourth experiment incorporated other one-stage models for LN detection. In recent works [7,8], FoveaBox and DETR have been used to detect LN in T2 MRI scans on a slice level, and the detection metrics for these models were calculated. FoveaBox provided results that were on par with those of VFNet with a ∼50.2% mAP and ∼77.4% recall at 4FP. The DETR transformer model, however, clocked in a lower detection mAP of ∼45.5% and ∼74.8% sensitivity. For our final experiment, we took all the detection models that had a mAP over 45% and assembled them (Faster RCNN with HNEM, VFNet, FoveaBox, and DETR) to be consistent with prior literature [7] in obtaining results for comparison. Initially, we-ensembled all the models together (VFNet, FoveaBox, DETR, and Faster RCNN with HNEM) and noticed that the detection performance did not change significantly; the mAP dropped to ∼49.7% while the sensitivity was ∼78.5%. Then, we-ensembled only the one-stage models (VFNet, FoveaBox, DETR) together, and observed the best mAP of ∼52.3% and sensitivity of ∼78.7% at 4FP per volume.

In prior work specifically focused on LN detection [3,4,6–8], the models were tested on 2D slices where the LN were annotated with the LAD and SAD, and the detection metrics were computed on only these 2D slices resulting in higher mAP and recall as shown in Table 1. However, this is problematic because it does not reflect the true detection performance, i.e., true positives that have not been marked by the radiologist in the current 2D slice or adjoining slices will be counted as FP when in fact they were detected correctly. Since a desirable automated method should process the whole T2 MRI scan, precision and recall should be associated with the FP per volume and not per slice. Therefore, in this work, these models were tested on 122 full 3D scans, where all the LN had been annotated in the test set, and the presented results are more representative of describing the LN existence. From past literature, only two approaches [3,6] have attempted a full 3D volumetric detection of LN in T2 MRI. In [3], a precision of 64.5% was achieved with a combination of T2 and DWI MRI images, whereas we obtained 52.3% precision with T2 MRI only. Our one-stage ensemble and VFNet models improved upon the sensitivity of [3] by ∼16% (78.7% sensitivity at 4FP vs. 62.6% sensitivity at 8FP) and that of [6] by ∼14% (78.7% sensitivity at 4FP vs. 64.6% sensitivity at 5FP).
Discussion

Among the various models and ensembles that were presented above, VFNet and the one-stage ensemble (consisting of VFNet, FoveaBox, DETR) met our pre-clinical usage goals; they achieved 78.7% sensitivity at 4 FP surpassing the required 65% sensitivity threshold at 4-6 FP [3,6]. These models generated detection results that greatly improved upon prior 3D LN detection work as evidenced by Table 1. From our pipeline, WBF was used as a post-processing step, and this was distinct from the original VFNet model that did not use any post-processing step. As we used multiple models (and checkpoints), we found WBF useful in merging the distinct predictions of various models (or epochs) with the mAP and sensitivity either being maintained or improved. In this paper, DWI was not utilized and its exclusion is a limitation of our work as radiologists in current clinical practice refer to these DWI scans to verify their findings. Adding these DWI scans into the detection pipeline will provide a distinct feature representation in the model layers that can improve the detection results. Furthermore, segmentation supervision was not provided in this pipeline and is another limitation of our work as segmentation supervision has been shown to improve detection results [13]. These are avenues that can be explored in future work.

Conclusion

Detection of lymph nodes in T2 MRI studies helps differentiate metastatic nodes from non-metastatic nodes, and aids in identifying cancer that spreads to lymph nodes. In this paper, a CAD pipeline is proposed to identify LN in T2 MRI studies. VFNet and the one-stage ensemble provide the best detection results with ≥51% mAP and 78.7% sensitivity at 4FP per volume. These results significantly improve upon the state-of-the-art methods [3,6] in LN detection with an increase in sensitivity of 16% and 14%, respectively. Our results indicated that VFNet and the one-stage ensemble could be used interchangeably in the proposed CAD pipeline for LN detection.

Funding  This work was supported by the Intramural Research Programs of the NIH National Library of Medicine and NIH Clinical Center (project number 1Z01 CL040004). We also thank Jaclyn Burge for the helpful comments and suggestions.

Declarations

Conflict of interest  RMS receives royalties from iCAD, Philips, PingAn, ScanMed, and Translation Holdings. His lab received research support from PingAn. The authors have no additional conflicts of interest to declare.

Ethical approval  All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and the 1964 Helsinki declaration and its later amendments or comparable ethical standards. For this study, informed consent was not required.

References

1. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, Meyer L, Gress DM, Byrd DR, Winchester DP (2017) The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging. CA Cancer J Clin 67(2):93–99
2. Taupitz M (2007) Imaging of lymph nodes - mri and ct. Springer, pp 321–329
3. Xingyu Z, Peiyi X, Mengmeng W, Pickhardt Perry J, Wei X, Fei X, Rui Z, Yao X, Junming J (2020) Deep learning based fully automated detection and segmentation of lymph nodes on multiparametric mri for rectal cancer: A multicentre study. EBioMedicine 56:102780
4. Debats OA, Litjens GJS, Huisman HJ (2019) Lymph node detection in mr lymphography: false positive reduction using multi-view convolutional neural networks. Peer J 7:e8052
5. Yun L, Qiyue Yu, Gao Y, Zhou Y, Liu G, Dong Q, Ma J, Ding L, Yao H, Zhang Z, Xiao G, An Q, Wang G, Xi J, Yuan W-T, Lian Y, Zhang D, Zhao C-G, Yao Q, Liu W, Zhou X, Liu S, Qingsyao W, Wenjian X, Zhang J, Wang D, Sun Z, Gao Y, Zhang X, Jilin H, Zhang M, Wang G, Zheng X, Wang L, Zhao J, Yang S (2018) Identification of metastatic lymph nodes in mr imaging with faster region-based convolutional neural networks. Cancer Res 78 17:5135–5143
6. Wang S, Zhu Y, Lee S, Elton DC, Shen TC, Tang Y, Peng Y, Lu Z, Summers RM (2022) Global-Local attention network with multi-task uncertainty loss for abnormal lymph node detection in MR images. Med Image Anal 77:102345
7. Mathai TS, Lee S, Elton DC, Shen TC, Peng Y, Zhiyong L, Summers RM (2021) Detection of lymph nodes in t2 mri using neural network ensembles. In: Lian C, Cao X, Rekik I, Xuanang X, Yan P (eds) Machine Learning in Medical Imaging. Springer International Publishing, Cham
8. Mathai TS, Lee S, Elton DC, Shen TC, Peng Y, Lu Z, Summers RM (2021) Lymph node detection in t2 mri with transformers. arXiv
9. Zhang H, Wang Y, Dayoub F, Sünderrahm T (2021) Vafiscaldnet: An IoU-aware Dense Object Detector, in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021 pp. 8510–8519
10. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: Vedaldi A, Bischof H, Brox T, Frahm J-M (eds) Computer vision - ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12346 Springer International Publishing, Cham
11. Kong T, Sun F, Liu H, Jiang Y, Li L, Shi J (2020) Foveabox: beyond anchor-based object detection. IEEE Trans Image Process 29:7389–7398
12. Shaoqing R, Kaiming H, Ross G, Jian S (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R (eds) Advances in Neural Information Processing Systems. Curran Associates Inc, Red Hook
13. Tang Y-B, Yan K, Tang Y-X, Liu J, Xiao J, Summers RM (2019) Ul Dor: a universal lesion detector for ct scans with pseudo masks and hard negative example mining. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp 833–836

© Springer
14. Tian Z, Shen C, Chen H, He T (2019) Fcos: fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
15. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 936–944
16. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2020) Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 42(2):318–327
17. Bodla N, Singh B, Chellappa R, Davis LS (2017) Soft-NMS—improving object detection with one line of code. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)
18. Solovyev R, Wang W, Gabruseva T (2021) Weighted boxes fusion: ensembling boxes from different object detection models. Image Vis Comput 107:104117
19. Peng Y, Lee S, Elton DC, Shen T, Tang Y-X, Chen Q, Wang S, Zhu Y, Summers R, Lu (2020) Automatic recognition of abdominal lymph nodes from clinical text. In: Proceedings of the 3rd clinical natural language processing workshop, pp 101–110, Online, November 2020. Association for Computational Linguistics
20. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC (2010) N4itk: improved n3 bias correction. IEEE Trans Med Imaging 29(6):1310–1320
21. Kociolek M, Strzelecki M, Obuchowicz R (2020) Does image normalization and intensity resolution impact texture classification? Comput Med Imaging Graph 81:101716
22. Chen K, Wang J, Pang J, Yuhang Cao Y, Xiong Xiaoxiao L, Sun S, Feng W, Liu Z, Jiarui X, Zhang Z, Cheng D, Zhu C, Cheng T, Zhao Q, Li B, Xin L, Zhu R, Yue W, Dai J, Wang J, Shi J, Ouyang W, Loy Chen C, Lin D (2019) Mmdetection: open mmocr detection toolbox and benchmark

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.