Towards Automatic Report Generation in Spine Radiology using Weakly Supervised Framework

Zhongyi Han[1,2], Benzheng Wei[1,2,*], Stephanie Leung[3,4], Jonathan Chung[3,4], and Shuo Li[3,4,*]

[1] College of Science and Technology, Shandong University of Traditional Chinese Medicine, Jinan, Shandong, China
wbz99@sina.com
[2] Computational Medicine Lab (CML), Shandong University of Traditional Chinese Medicine, Jinan, Shandong, China
[3] Department of Medical Imaging, Western Univeristy, London, ON, Canada
slishuo@gmail.com
[4] Digital Imaging Group (DIG), London, ON, Canada

Abstract. The objective of this work is to automatically generate unified reports of lumbar spinal MRIs in the field of radiology, i.e., given an MRI of a lumbar spine, directly generate a radiologist-level report to support clinical decision making. We show that this can be achieved via a weakly supervised framework that combines deep learning and symbolic program synthesis theory to overcome four inevitable tasks: semantic segmentation, radiological classification, positional labeling, and structural captioning. The weakly supervised framework using object level annotations without requiring radiologist-level report annotations to generate unified reports. Each generated report covers almost type lumbar structures comprised of six intervertebral discs, six neural foramina, and five lumbar vertebrae. The contents of each report contain the exact locations and pathological correlations of these lumbar structures as well as their normalities in terms of three type relevant spinal diseases: intervertebral disc degeneration, neural foraminal stenosis, and lumbar vertebrae deformities. This framework is applied to a large corpus of T1/T2-weighted sagittal MRIs of 253 subjects acquired from multiple vendors. Extensive experiments demonstrate that the framework is able to generate unified radiological reports, which reveals its effectiveness and potential as a clinical tool to relieve spinal radiologists from laborious workloads to a certain extent, such that contributes to relevant time savings and expedites the initiation of many specific therapies.

1 Introduction

Automated report generation is a worthwhile work to expedite the initiation of many specific therapies and contribute to relevant time savings in spine radiology. Nowadays, multiple lumbar spinal diseases not only have deteriorated the quality of life but have high morbidity rates worldwide. For instance, Lumbar Neural Foraminal Stenosis (LNFS) has attacked about 80% of the elderly population

[10]. In daily radiological practice, time-consuming report generation leads to the problem of the delay of a patient's stay in the hospital and increases the costs of hospital treatment [14]. Automatic report generation systems would offer the potential for faster and more efficient delivery of radiological reports and thus would accelerate the diagnostic process [12]. However, to date, most so-called Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx) techniques cannot generate radiological reports in the medical image analysis domain, let alone the spinal image analysis. In addition, MRI is widely used in clinical diagnosis of spinal diseases as is better to demonstrate the spinal anatomy [6]. Therefore, this work is devoted to the radiological report generation of lumbar MRIs to support clinical decision making.

Proposed framework. Our proposed weakly supervised framework combines deep learning and symbolic program synthesis theory, achieving fully-automatic radiological report generation through semantic segmentation, radiological classification, positional labeling, and structural captioning. Firstly, we propose a novel Recurrent Generative Adversarial Network (RGAN) for semantic segmentation and radiological classification of intervertebral discs, neural foramina, and lumbar vertebrae. The RGAN is constituted by 1) an atrous convolution autoencoder module for fine-grained feature embedding of spinal structures; 2) a followed spatial long short-term memory based Recurrent Neural Network (RNN) module for spatial dynamic modeling; and 3) an adversarial module for correcting predicted errors and global contiguity. Secondly, we propose a strong prior knowledge based unsupervised symbolic program synthesis approach for positional labeling of multiple spinal structures. Finally, we propose a symbolic template based structural captioning method for generating unified radiological reports. The generated radiological reports contain the exact locations and pathological correlations of three spinal diseases: LNFS, Intervertebral Disc Degeneration (IDD), and Lumbar Vertebrae Deformities (LVD).

Why weak supervision? In this paper weakly supervised learning refers to using object level annotations (i.e., segmentation and class annotations) without requiring radiologist-level report annotations (i.e., whole text reports) to generate unified reports. To date, the weakly supervised learning manner is supposedly the one and only resolution. Because if it was possible to have a large amount of data, like natural image captioning dataset MS COCO [9], Visual Genome [7], we would directly use text report annotations to train end-to-end report generation modules as the natural image captioning technology [8]. Such technology needs a large amount of descriptive annotations (i.e., sentences, paragraphs) to train fully supervised learning model and generate coarse descriptions only. However, in daily radiological practice, different radiologists always write radiological reports in different styles and different structures, which cannot be learned with a little amount dataset. Clinical-radiological reports also contain exactly important clinical concerns, such as locations, normalities, and gradings. Since clinical concerns inside few words decide the correctness of a radiological report, it is also impossible to judge the correctness of computer-made reports compared with radiologist-made reports using Natural Language Processing (NLP) tech-

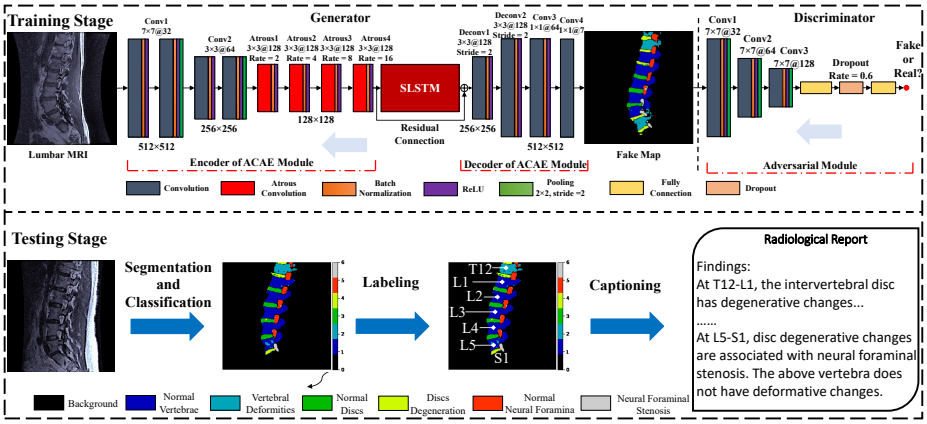


Fig. 1. The workflow of the proposed weakly supervised framework.

nologies. On contrary, it is possible to use weakly supervised learning manner decomposing the task into multiple procedures, i.e., detect clinical concerns using object level annotations first and then fill these concerns in a universal template to generate unified radiological reports.

Related works. To the best of our knowledge, neither CADe nor CADx work has achieved spinal report generation. Existing works include but are limited to detection [2], segmentation [5], labeling [1] or classification [4] of one type spinal structure. In other fields, a few works devoted to automated report generation. [16] uses a large amount of data and NLP based image captioning to study the report generation of pathology bladder cancer images. [15] uses chest X-rays data to study the report generation of thorax diseases.

2 The Proposed Framework for Report Generation

The workflow of the weakly supervised framework is illustrated in Fig. 1. The framework combines two type theories. The first type is our proposed learning-based methods RGAN for segmentation and classification (see Sec. 2.1). The second type is a strong prior knowledge based unsupervised symbolic program synthesis for labeling and captioning (see Sec. 2.2). The framework thus can concentrate on intuitive perceptual thinking while focuses on rule-based thinking.

2.1 Recurrent Generative Adversarial Network

RGAN comprises two sub-networks: a generative network and a discriminative network. The generative network is designed to generate pixel-level predicted maps, i.e., each pixel in a generated map has the possibility of seven classes comprised of normal/abnormal neural foramina, normal/abnormal intervertebral discs, normal/abnormal vertebrae, and background. The discriminative network

is designed to supervise and encourage the generative network. When training, inspired by [3], the generative network aims at fooling the discriminative network, while the discriminative network makes great efforts to discriminate its inputs whether are fake maps generated by the generative network or real maps from ground truth. When a strong confrontation occurs, the discriminative network eagerly prompts the generative network to look out mismatches in a wide range of higher-order statistics. The generative network comprises of a deep Atrous Convolution Autoencoder (ACAE) module and a Spatial Long Short-Term Memory (SLSTM) based RNN module, while the discriminative network comprises of an adversarial module.

ACAE module. The ACAE module comprises of four standard convolution layers, four atrous convolution layers as an encoder, and two deconvolution layers as a decoder. For each location i on the output feature map y and each kernel k on the weight w and bias b, atrous convolution are applied over the input feature map x as $y[i] = f(\sum_k x[i + r \cdot k] * w[k] + b[k])$. The $r \cdot k$ is equivalent to convolving the input x with upsampled kernels, which is produced by inserting zeros between two consecutive values of each kernel along each spatial dimension. Progressive rates r of $\{2, 4, 8, 16\}$ is adopted after cross-validation, which modifies kernel's receptive fields adaptively. The ACAE module practically produces semantic task-aware features using fewer parameters and larger receptive fields. The ACAE module also has little-stacked downsampling operations, so that avoids severely reducing the feature resolution among low-dimensional manifold. The ACAE module thus enables RGAN to not only address the high variability and complexity of spinal appearances in MRI explicitly but also effectively preserve fine-grained differences between normal and abnormal structures.

RLSTM based RNN module. This module is to memorize the spatial pathological correlations between neighboring structures. For instance, current neural foramen has a high probability of being abnormal when neighboring discs or vertebra are abnormal. The module has a spatial top-down structure. Assuming $M \in \mathbb{R}^{n \times n \times c}$ represents a set of deep convolutional feature maps generated by the encoder of ACAE module with widths of n, heights of n, and channels of c. Firstly, the module downsamples its input feature maps to $M^{'} \in \mathbb{R}^{\frac{n}{i} \times \frac{n}{i} \times c}$, where i is the size of 4 according to the receptive fields of spinal structures. Secondly, the module patch-wisely unstacks these downsampled feature maps $M^{'}$ to a set of spatial sequences $M^{''} \in \mathbb{R}^{(\frac{n}{i})^2 \times c}$. Thirdly, the module recurrently memorizes long-period context information between spatial sequences and generates outputs $S \in \mathbb{R}^{(\frac{n}{i})^2 \times c}$. Finally, the module adaptively upsamples the outputs S into $S^{'} \in \mathbb{R}^{n \times n \times c}$ using two deconvolution layers. Accordingly, the module has $(\frac{n}{i})^2$ LSTM units and c-dimensions cell state. The module is capable of selectively memorizing and forgetting semantic information of previous spinal structures when transforming the high-level semantic features into sequential inputs of LSTM units.

Adversarial module. The adversarial module of the discriminative network comprises of three convolutional layers with large kernels, three batch normalizations, three average pooling layers, and two fully connected layers with dropout.

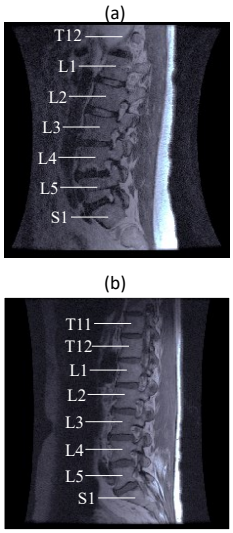
When training, the adversarial module first receives the predicted maps from the generative network and manual maps from ground truth and then outputs a single scalar. The adversarial processes substantially correct predicted errors and break through small dataset limitation, so as to achieve continued gains on global-level contiguity and avoidable over-fitting.

2.2 Prior Knowledge-based Symbolic Program Synthesis

In this paper, unsupervised symbolic program synthesis refers to leveraging prior human knowledge to discover inherent patterns in spinal structures. This study assumed that human knowledge representation is symbolic and that reasoning, language, planning, and vision could be understood in terms of symbolic operations. Therefore, we can design a model-free symbolic programing to realize labeling and captioning.

Unsupervised labeling. The input of the unsupervised labeling process is the predicted maps generated by RGAN, and the output is three dictionaries comprised of locations and normalities of three spinal structures. The keys of each dictionary are the locations of one type structure, while the values of the dictionary are the normality conditions at the locations of one type structure in a lumbar spine. The first step of the unsupervised labeling process is to discover patterns for location assignment of each pixel. According to our observations, locations, surrounding correlations are the inherent patterns inside lumbar spinal structures, i.e., in a lumbar spine, all intervertebral discs are separated by vertebrae that like the blank of a piano. Let intervertebral disc as an example, we first calculate out the minimal height of vertebral in the training set and then let the height divided by four be the margin between pixels of intervertebral discs. We thus get the margined pixels of intervertebral discs into lists. Since generated maps have a few spots, the second step is to decide the true label of margined pixels. For instance, at the L5-S1 intervertebral disc, we compare the pixel amounts between normal and abnormal labels and then choose the one that has the most amount pixels as the final label. We collect the final results into dictionary for the next captioning process.

Template-based captioning. The input of this captioning process is three dictionaries and the output is a fully structural radiological report. Although reports wrote by different radiologists always have different styles and different patterns, the focus is still the clinical concern. After summarizing common patterns inside radiological reports as a decision problem, we can use If-Then symbolic operations to create a unified template. For instance, at L5-S1 if the neural foramen is abnormal, intervertebral disc and vertebra are normal, then the output would be *"At L5-S1, the neural foramen has obvious stenosis, the intervertebral disc does not have obvious degenerative changes, and the above vertebra does not have deformative changes."*. It is noteworthy that this process can significantly promote clinical pathogenesis-based diagnosis. While the SLSTM-based RNN module can memorize the spatial pathological correlations between neighboring structures, LVD and LDD are substantially crucial pathogenic factors and vital predictors of LNFS. Accordingly, for instance, if the neural foramen, disc, and



(a) Radiological Report

At T12-L1, the intervertebral disc has obvious degenerative changes. The neural foramen does not have stenosis.

At L1-L2, the above vertebra has deformative changes. The intervertebral disc does not have obvious degenerative changes. The neural foramen does not have obvious stenosis.

At L2-L3, the neural foramen has obvious stenosis. The intervertebral disc does not have obvious degenerative changes. The above vertebra does not have deformative changes.

At L3-L4, disc degenerative changes are associated with neural foraminal stenosis.

At L4-L5, disc degenerative changes are associated with neural foraminal stenosis.

At L5-S1, the intervertebral disc has obvious degenerative changes. The above vertebra also has deformative changes. They lead to the neural foraminal stenosis.

(b) Radiological Report

At T12-L1, the intervertebral disc has obvious degenerative changes. The neural foramen does not have stenosis.

At L1-L2, the above vertebra does not have deformative changes. The intervertebral disc does not have degenerative changes. The neural foramen also does not have stenosis.

At L2-L3, the above vertebra does not have deformative changes. The intervertebral disc does not have degenerative changes. The neural foramen also does not have stenosis.

At L3-L4, the neural foramen has obvious stenosis. The intervertebral disc does not have obvious degenerative changes. The above vertebra does not have deformative changes.

At L4-L5, disc degenerative changes are associated with neural foraminal stenosis.

At L5-S1, the intervertebral disc has obvious degenerative changes. The above vertebra also has deformative changes. They lead to neural foraminal stenosis to a certain extent.

Fig. 2. The generated radiological reports. The text in purple color represents that our framework is helpful for building comprehensive pathological analysis.

vertebra are abnormal at L3-L4, the captioning process can output *"At L3-L4, the intervertebral disc has obvious degenerative changes. The above vertebra also has deformative changes. They lead to the neural foraminal stenosis to a certain extent."*. If the neural foramen is normal but disc or vertebra is abnormal, one can predict that the neural foramen has a great possibility to be stenosis. Therefore, this captioning process promotes early diagnosis when the pathogenic factor is solely occurring. This process is also helpful for building comprehensive pathological analysis and benefits to clinical surgery planning.

3 Results

Dataset. Our dataset is collected from multicenter and different models of vendors including 253 multicenter clinical patients. Average years of patient age is 53±38 with 147 females and 106 males. Among sequential T1/T2-weighted MRI scans of each patient, one lumbar middle scan was selected to better present neural foramina, discs, and vertebra simultaneously in the sagittal direction. The segmentation ground truth is easily labeled by our lab tool according to the clinical criterion. The ground truth of classification was annotated as normal or abnormal by extracting from clinical reports, which are double-checked by board-certified radiologists. Five-fold cross-validation is employed for the performance evaluation and comparison.

Performance of radiological report generation. The representative radiological reports generated by our proposed framework are illustrated in Fig. 2. The framework can produce pathological correlations between LNFS, LVD, and

Table 1. Performance comparisons between RGAN and other models.

Method	Pixel accuracy	Dice coefficient	Specificity	Sensitivity
FCN [13]	0.917±0.004	0.754±0.033	0.754±0.035	0.712±0.032
U-Net [11]	0.920±0.004	0.797±0.013	0.816±0.027	0.770±0.026
ACAE	0.958±0.002	0.841±0.013	0.862±0.018	0.823±0.024
ACAE+SLSTM	0.959±0.002	0.848±0.009	0.865±0.021	0.837±0.025
ACAE+Adversarial	0.960±0.004	0.863±0.006	0.873±0.015	0.855±0.027
RGAN	**0.962±0.003**	**0.871±0.004**	**0.891±0.017**	**0.860±0.025**

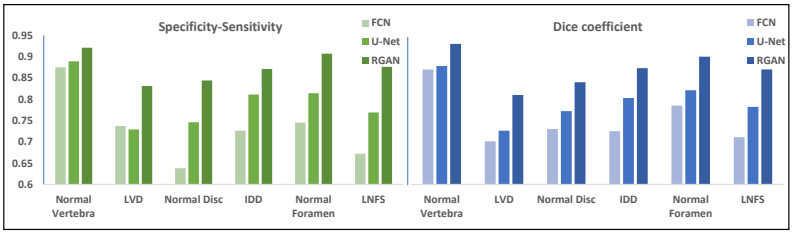


Fig. 3. Specificity-Sensitivity and Dice coefficient of three type models.

IDD as shown in the text in purple color. Representative results demonstrate the advantages of the framework, which efficiently combines deep learning that is robust to noisy data and symbolic program synthesis that is easier to interpret and requires less training data. Generated unified reports also demonstrate that the weakly supervision is robust and efficient, and endows our framework an potential as a clinical tool to relieve spinal radiologists from laborious workloads to a certain extent.

Performance inter- and intra-comparison of RGAN. As illustrated in Table 1 and Fig. 3, RGAN achieves more higher performance than Fully Convolutional Network (FCN) [13] and U-Net [11] in the segmentation and classification of three type spinal structures. FCN and U-Net are implemented strictly upon public resources. After removing the SLSTM based RNN module and the adversarial module, RGAN also achieves higher performance than its ablated versions as shown in the third-sixth rows of Table 1. Since no existing works achieved simultaneous segmentation and classification of multiple spinal structures, we do not conduct extra comparisons.

4 Discussion and Conclusion

We show that using the weakly supervised framework that combines deep learning and symbolic program synthesis is very efficient and flexible to generate spinal radiological reports. The reason for using object segmentation rather than object detection is that segmentation is better to present the spatial correlations between spinal structures. The study just has a try, and further work will focus on 1) considering more uncommon spinal diseases, and 2) collecting more clinical data in order to realize end-to-end report generation.

References

1. Alomari, R.S., Corso, J.J., Chaudhary, V.: Labeling of lumbar discs using both pixel- and object-level features with a two-level probabilistic model. IEEE Transactions on Medical Imaging 30(1), 1–10 (Jan 2011)
2. Cai, Y., Osman, S., Sharma, M., Landis, M., Li, S.: Multi-modality vertebra recognition in arbitrary views using 3d deformable hierarchical model. IEEE Transactions on Medical Imaging 34(8), 1676–1693 (Aug 2015)
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
4. He, X., Yin, Y., Sharma, M., Brahm, G., Mercado, A., Li, S.: Automated diagnosis of neural foraminal stenosis using synchronized superpixels representation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 335–343. Springer (2016)
5. He, X., Zhang, H., Landis, M., Sharma, M., Warrington, J., Li, S.: Unsupervised boundary delineation of spinal neural foramina using a multi-feature and adaptive spectral segmentation. Medical image analysis 36, 22–40 (2017)
6. Kim, S., Lee, J.W., Chai, J.W., Yoo, H.J., Kang, Y., Seo, J., Ahn, J.M., Kang, H.S.: A new mri grading system for cervical foraminal stenosis based on axial t2-weighted images. Korean journal of radiology 16(6), 1294–1302 (2015)
7. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision 123(1), 32–73 (2017)
8. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating image descriptions. In: Proceedings of the 24th CVPR. Citeseer (2011)
9. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
10. Rajaee, S.S., Bae, H.W., Kanim, L.E., Delamarter, R.B.: Spinal fusion in the united states: analysis of trends from 1998 to 2008. Spine 37(1), 67–76 (2012)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
12. Rosenthal, D.F., Bos, J.M., Sokolowski, R.A., Mayo, J.B., Quigley, K.A., Powell, R.A., Teel, M.M.: A voice-enabled, structured medical reporting system. Journal of the american medical informatics association 4(6), 436–441 (1997)
13. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE transactions on pattern analysis and machine intelligence 39(4), 640–651 (2017)
14. Vorbeck, F., Ba-Ssalamah, A., Kettenbach, J., Huebsch, P.: Report generation using digital speech recognition in radiology. European Radiology 10(12), 1976–1982 (Nov 2000), https://doi.org/10.1007/s003300000459
15. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. arXiv preprint arXiv:1801.04334 (2018)
16. Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L.: Mdnet: A semantically and visually interpretable medical image diagnosis network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3549–3557 (July 2017)