AUTOMATIC DEEP LEARNING BASED CERVICAL SPINAL FRACTURE DIAGNOSIS

Sage Hahn* *James Allison*† *Richard Watts*† *Safwan Wshah**

* University of Vermont, Complex Systems Center, Burlington, Vermont 05401, USA
† University of Vermont, Department of Radiology, Burlington, Vermont 05401, USA

ABSTRACT

Fractures of the spine are potentially serious injuries that are associated with significant morbidity and mortality. The second cervical (C2) vertebrae, in particular, has a unique morphology and associated fracture patterns. In this paper, we propose a deep learning based approach for the automatic diagnosis of C2 fractures based on Computed Tomography (CT) volumes. The proposed approach makes use of a two separate 3-dimensional convolutional neural networks (CNN) to first localize the C2 vertebra, and secondly to predict the presence of a fracture. We evaluate and train the stages of our pipeline, along with a comparative 2D CNN approach, on a dataset of CT scans collected from 465 patients. The proposed method shows promising experimental results, obtaining a cross-validated area-under-the-curve of .88 for fracture diagnosis.

Index Terms— C2 fracture, Dens fracture, Deep learning, Cervical spine, 3D convolutional neural network

1. INTRODUCTION

Spinal injury is a major source of morbidity and mortality in the United States. Approximately 7800 new cases of spinal cord injury occur each year [1]. Detection of fracture is essential for the appropriate care of trauma patients who have either suffered or are at risk of developing spinal cord injury as a consequence of trauma. Additionally, certain fracture patterns in the cervical region are associated with other pathology such as blunt cerebrovascular injury and may be the impetus for additional vascular evaluation and treatment [2]. Detection of spinal fractures reflects an opportunity for the application of machine learning in radiographic imaging. Initially, a functional system could "pre-read" examinations and alert clinicians to the presence of fractures which require immediate attention. Within the realm of radiology, this could be helpful to triage patients with injuries for more rapid scan interpretation. Ultimately, one may envision a fully automated system for the interpretation of spinal CT imaging.

As a first step towards developing a comprehensive tool for assessment of spinal fractures on CT images, we focused on developing a system that could first reliably identify the C2 cervical vertebrae and secondly determine the presence of fractures involving the odontoid process.

As far as we can determine there has been no previous work done specifically on the automated detection of C2 fractures. However, there is a large body of work in closely related areas including spinal fracture diagnosis, general fracture diagnosis, and general disease diagnosis. Of particular interest within the scope of automated spinal fracture detection is work done by Roth et al. [3] on the use of ConvNets in a 2.5D approach. They obtain an an area-under-the-curve (AUC) of 0.857 on the detection of posterior-element fractures from spinal CT's. A number of other 2D approaches exist on the task of fracture detection within 2D images obtained from plain radiographs and other similar modalities. Approaches on this arguably less complex data tend to yield more impressive results, with an AUC of .954 on detecting wrist fractures [4] and 95.5% accuracy on detecting intertrochanteric hip fractures [5] among others. In general, due to its vastly greater size, working with 3D volumes tends to be more difficult and memory intensive for diagnostic deep learning approaches. Despite the increased computational difficulty, a number of approaches exist that make use of three dimensional convolutional neural networks (3D CNNs) in order to predict directly from a 3D representation of the data. Examples include work done by Payan et al. [6] with 3D CNNs on predicting Alzheimer's disease and by Korolev et al [7] on brain MRI classification.

On the other hand, our initial sub problem of vertebra localization within a large 3D medical volume has received more attention. Approaches to this task tend to vary greatly, and include a range of older machine learning based methods [8] in addition to more recent deep learning inspired approaches. Yang et al. [9] offers a deep learning approach to the automated labelling of vertebrae within CT volumes over range of different pathologies. Recent efforts have also yielded results on more general anatomical localization through two pass or cascaded methods, for example Zheng et al. [10] makes use of an initial shallow candidate detector followed by a deeper more accurate classification network, ultimately obtaining a mean error of 2.64mm on the task of carotid artery bifurcation detection.

Within this paper we explore a new approach towards region of interest selection, building on existing approaches for both localization and segmentation. Our method presented below evaluates the use of 3D segmentation networks trained on imperfect, and importantly easy to create, segmentation labels as a tool for anatomical localization. We additionally contribute to the growing understanding surrounding the use of deep 3D CNNs on large medical volumes, evaluating their performance on the task of C2 fracture diagnosis.

2. METHODS

In this section we outline the details of the proposed deep learning pipeline for automatic C2 fracture diagnosis.

2.1. Dataset Details and Labels

Our dataset was gathered from the University of Vermont Medical Center, with an IRB waiver of consent, for the purpose of this study. The full dataset consists of 465 cases in total, where all 62 scans with C2 fractures and 403 without were reviewed by a board certified radiologist in addition to the original case report for the presence of a C2 fracture. At this stage of the study, we decided to ignore cases with significant metal artifacts close to the C2 vertebra. Additionally, the radiologist provided for each of the fracture positive cases a range of slices within the sagittal plane reconstruction where evidence of the fracture is visible for use in the 2D diagnostic network. The dimensionality of the gathered CT dataset varies greatly in terms of amount of 512x512 axial slices. Typically scans contain between 300-700 slices, with the C2 vertebra tending to encompass at most a 128x128x128 region within this larger 300-700x512x512 volume. A rough manual segmentation was further preformed on down-sampled 128x128x128 representations of 128 scans (an even mix of fracture and non-fracture) for the purpose of localization. Specifically, the area segmented is from the apex of the Dens down to roughly the border between the Dens and the body of the C2 vertebra, as can be seen in more detail in Fig. 1.

2.2. Localization

In order to achieve localization down to the 128x128x128 region of interest we make use of an established 3D convolutional network designed for segmentation. The novelty of our approach lies in the use of rough segmentation as tool for localization. CT volumes are first re-sampled from their original input size down to 128x128x128, a necessary step due to memory constraints on input size for the network. The network architecture employed at this stage is heavily inspired by the 3D U-Net approach [11], additionally employing residual weights and deep supervision [12]. The network is trained, with a batch size of one along with basic training augmentation in the form of 3D permutations, to segment the portion of the Dens described earlier and as seen in Fig 1.

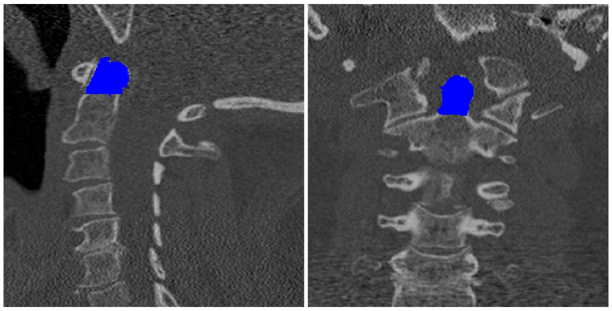

Fig. 1. Sample segmentations of the Dens on a sample cervical spine CT. On the left, the sagittal view, on the right a coronal view.

Given a ground truth or predicted segmentation, a basic post processing scheme is employed in choosing the final 128x128x128 crop. First, the predicted segmentation is re-scaled to fit the original volume. The volume is then processed in all two dimensional reconstructions (axial, coronal and sagittal) removing predicted segmentations with less than 9 pixels on a given slice as a coarse method for outlier removal. Next, a smaller 3D region of interest is calculated to include all of the remaining segmented output, and lastly this region is expanded in all dimensions (increasing the size of the crop) to ultimately represent a 128x128x128 region of the original volume. In degenerative cases where the initial outlier threshold fails to reduce all dimensions of the predicted volume to 128 or less, the threshold is raised by increments of 1 and continually re-evaluated until sufficiently reduced. It should be noted that the ratios for volume expansion are not equal, but were instead chosen based on experimentation. In general though, most ratios examined along with choices of pixel outlier threshold tended to capture the full C2 vertebra, the presence of hand tuned parameters were added only to more accurately center the vertebra within the region interest as well as to help capture outlier cases.

2.3. 3D Diagnosis Network

The 3D networks used at this stage are heavily influenced by the Res-Net architectures [13], in particular we use a ‘vanilla’ 3D version which has been used for brain MRI classification [7], among other medical and non-medical tasks such as 3D action recognition. A number of different configurations for a 3D Res-Net are possible, these include in general different depths from shallower networks with 18 or 50 layers, to deeper and therefore more memory intensive networks with 101 or 151 layers. We sought to evaluate the effect of different depths on performance, and therefore made use of a number of different configurations during testing. Raw input at this stage comes in the form of a 128x128x128 crop on the original scan, which we either use as is or down-sample to 96x96x96 or 64x64x64. Networks are trained with a binary cross entropy loss function and the Adam optimizer in order to predict a probability between 0 (no fracture) and 1 (fracture) for each sample. All network configurations are additionally trained through use of a technique known as snapshot ensembles [14], where a cyclic learning rate is employed and 'snapshots' of the networks weights are taken over a fixed number of total epochs. In our case we took between 2-5 snapshots over typically 100 epochs, corresponding to a saved set of network weights after every 20 epochs, ultimately taking the average score from each set of weights during testing. This technique is a useful tool in relieving noise from various hyper-parameters, especially when working with 3D CNNs which tend to be very sensitive to small parameters changes, thus reducing overall time needed to fine tune performance.

2.4. 2D Diagnosis Network

In comparison with the 3D approach, our 2D network receives as input 128 sagittal slices with dimensions 128x128. Each individual slice here has a corresponding label as to if a fracture is visible on that slice individually, and likewise the network is trained to make slice by slice predictions in that regard. We make use of a slightly modified ResNet-50 [13] architecture as the underlying prediction network. Likewise, the same snapshot ensemble scheme, optimizer and loss function as discussed with regard to the 3D diagnosis network are used. Importantly, the network outputs 128 individual predictions from each scan which must be combined in order to make an overarching prediction on the scan itself. This is done simply by taking the maximum predicted value from the series of slices, and associating that value with the scan itself.

2.5. Implementation Details

All work presented was implemented in python, making extensive use of the Keras library [15]. Further all training and experimental results were obtained on a desktop with a Nvidia GeForce GTX 1080 graphics card.

3. EXPERIMENTAL RESULTS

In order to evaluate our full localization pipeline as described in Section 2.2, we trained a model on 124 labelled scans, using these scans to additionally fine tune the post processing parameters. The remaining 341 unlabelled scans were then utilized as our test set for evaluating overall performance. As these scans are unlabelled, the outputted final 128x128x128 predictions were evaluated by a human to validate if the C2 vertebra was accurately captured within the predicted region of interest. The following methodology yielded only one complete miss out of all 341 evaluated scans, along with two additional partial misses where at least 50% of the body of the C2 vertebra was still captured. While it can be difficult to determine exactly why the full miss occurred, we suspect it is related to metal artifacts present in the lower C-spine. That being said, our presented methodology remained robust over the remaining 338 scans which notably contain a broad range of different pathologies, artifacts and resolutions.

During our evaluation of the diagnosis networks we alternatively made use of stratified 5-fold cross validation (CV), as we have access to only 62 positive fracture cases. The dataset used for training and testing over 5-fold CV at this stage is comprised of all 62 positive cases and a random sample of 62 negative cases. Over the course of a large number of experimental runs we evaluated the performance of different models through use of the following metrics: Area Under the Receiver Operating Characteristic Curve (ROC AUC), Average Precision (AP) defined as the weighted mean of precisions achieved at each threshold of the precision-recall curve, and the highest F1 score achieved over all thresholds (F1).

While we chose to focus on optimizing model performance and parameters for a 3D network approach, it nonetheless bears a brief comparison with its 2D counterpart. To this end, we presents experimental results at the bottom of Table 1 for the 2D Diagnosis network trained and evaluated on the dataset described above. Table 1 further shows the 3D CNN network parameters used to generate our most successful results over 5 fold CV. Our best models converge around a ROC AUC score of .88, notably achieved with variability surrounding most introduced parameters. We found additionally that on the fly training data augmentation in the form of slight random 3D scaling improved accuracy over all models. On the other hand, we were unable to improve results through use of data augmentation techniques during testing, a technique which in some cases has been shown to improve accuracy.

4. CONCLUSIONS AND DISCUSSION

We have introduced a robust technique for the localization of the C2 vertebra from rough easy to generate labels, as well as a subsequent predictive model for fracture diagnosis. It offers accurate region of interest localization over a huge range of different anatomies, achieving an accuracy of 99+% over 341 scans. We further achieve a ROC AUC of 88% and AP of 91% over 5-fold cross validation on the task of C2 fracture diagnosis through use of 3D CNNs. Likewise, we were able to achieve comparable performance across a number of similar 3D and 2D CNN approaches, regardless of additional time spent fine-tuning the 3D network, which suggests we might be able to improve 2D CNN results further. We also believe that additional fracture cases will prove essential towards training a more accurate pipeline. In future work we plan to expand our dataset, both in terms of size and location of fracture within the C-Spine, our eventual goal being a system capable of scanning a full CT volume for fractures.

Layers	Input Size	Snapshots	Epochs	Batch Size	ROC AUC	AP	F1
3D-101	64x64x64	5	100	6	.88 ± .03	.91 ± .05	.85 ± .07
3D-18	64x64x64	3	60	4	.88 ± .07	.90 ± .07	.87 ± .06
3D-18	64x64x64	5	100	8	.88 ± .06	.89 ± .08	.85 ± .06
3D-151	64x64x64	5	100	4	.88 ± .06	.88 ± .09	.84 ± .04
3D-50	64x64x64	6	120	5	.88 ± .07	.90 ± .07	.85 ± .07
3D-101	96x96x96	5	100	3	.87 ± .02	.89 ± .02	.82 ± .02
2D-50	128x128	5	100	64	.89 ± .07	.88 ± .10	.85 ± .05
2D-50	128x128	3	60	32	.84 ± .04	.85 ± .07	.82 ± .03

Table 1. Best 2D and 3D network results from various models and hyper-parameters evaluated over 5-Fold CV.

5. REFERENCES

[1] "National spinal cord injury association resource center.," www.sci-info-pages.com/factsheets.html, Accessed on October 20, 2018.

[2] William J Bromberg, Bryan C Collier, Larry N Diebel, Kevin M Dwyer, et al., "Blunt cerebrovascular injury practice management guidelines: the eastern association for the surgery of trauma," *Journal of Trauma and Acute Care Surgery*, vol. 68, no. 2, pp. 471–477, 2010.

[3] Holger R Roth, Yinong Wang, Jianhua Yao, Le Lu, Joseph E Burns, and Ronald M Summers, "Deep convolutional networks for automated detection of posterior-element fractures on spine ct," in *Medical Imaging 2016: Computer-Aided Diagnosis*. International Society for Optics and Photonics, 2016, vol. 9785, p. 97850P.

[4] DH Kim and T MacKinnon, "Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks," *Clinical radiology*, vol. 73, no. 5, pp. 439–445, 2018.

[5] Takaaki Urakawa, Yuki Tanaka, Shinichi Goto, Hitoshi Matsuzawa, Kei Watanabe, and Naoto Endo, "Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network," *Skeletal radiology*, pp. 1–6, 2018.

[6] Adrien Payan and Giovanni Montana, "Predicting alzheimer's disease: a neuroimaging study with 3d convolutional neural networks," *arXiv preprint arXiv:1502.02506*, 2015.

[7] Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova, "Residual and plain convolutional neural networks for 3d brain mri classification," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 835–838.

[8] Tobias Klinder, Jörn Ostermann, Matthias Ehm, Astrid Franz, Reinhard Kneser, and Cristian Lorenz, "Automated model-based vertebra detection, identification, and segmentation in ct images," *Medical image analysis*, vol. 13, no. 3, pp. 471–482, 2009.

[9] Dong Yang, Tao Xiong, Daguang Xu, Qiangui Huang, David Liu, Zhou, et al., "Automatic vertebra labeling in large-scale 3d ct using deep image-to-image network with message passing and sparsity regularization," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 633–644.

[10] Yefeng Zheng, David Liu, Bogdan Georgescu, Hien Nguyen, and Dorin Comaniciu, "Robust landmark detection in volumetric data with efficient 3d deep learning," in *Deep Learning and Convolutional Neural Networks for Medical Image Computing*, pp. 49–61. Springer, 2017.

[11] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432.

[12] Baris Kayalibay, Grady Jensen, and Patrick van der Smagt, "Cnn-based segmentation of medical imaging data," *arXiv preprint arXiv:1701.03056*, 2017.

[13] 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*, 2016, pp. 770–778.

[14] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger, "Snapshot ensembles: Train 1, get m for free," *arXiv preprint arXiv:1704.00109*, 2017.

[15] François Chollet et al., "Keras," https://keras.io, 2015.