Medical Image Analysis 41 (2017) 63–73

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

SpineNet: Automated classification and evidence visualization in spinal MRIs

Amir Jamaludin[a,*], Timor Kadir[b], Andrew Zisserman[a]

[a] VGG, Department of Engineering Science, University of Oxford, Oxford, UK
[b] Optellum, Oxford, UK

ARTICLE INFO

Article history:
Received 3 March 2017
Revised 19 July 2017
Accepted 20 July 2017
Available online 21 July 2017

Keywords:
MRI analysis
Radiological classification
Spinal MRI

ABSTRACT

The objective of this work is to automatically produce radiological gradings of spinal lumbar MRIs and also localize the predicted pathologies. We show that this can be achieved via a Convolutional Neural Network (CNN) framework that takes intervertebral disc volumes as inputs and is trained only on disc-specific class labels. Our contributions are: (i) a CNN architecture that predicts multiple gradings at once, and we propose variants of the architecture including using 3D convolutions; (ii) showing that this architecture can be trained using a multi-task loss function without requiring segmentation level annotation; and (iii) a localization method that clearly shows pathological regions in the disc volumes. We compare three visualization methods for the localization.

The network is applied to a large corpus of MRI T2 sagittal spinal MRIs (using a standard clinical scan protocol) acquired from multiple machines, and is used to automatically compute disk and vertebra gradings for each MRI. These are: Pfirrmann grading, disc narrowing, upper/lower endplate defects, upper/lower marrow changes, spondylolisthesis, and central canal stenosis. We report near human performances across the eight gradings, and also visualize the evidence for these gradings localized on the original scans.

1. Introduction

Automated detection, grading and localization of abnormalities in medical images is an important task to support clinical decision making across many medical specialties. As the utilization and interpretation of medical images continues to expand beyond the radiology department, as it has in cardiology and neurology, it is becoming increasingly important to supplement the typically qualitative radiological read with objective quantitative methods.

However, to date, many so-called Computer Aided Detection (CADe) and Computer Aided Diagnosis (CADx) techniques have relied on strong supervision for their development. For example, conventional CADe algorithms for oncology typically require pixel level segmentation or at least the placement of bounded regions of interest around the lesion as training data. Yet obtaining sufficiently large, accurately delineated datasets in medical imaging can be prohibitively expensive due to limits in clinical time. In contrast, scores or even indicators relating to the presence or absence of pathology can be readily harvested from clinical radiological reports. It is therefore desirable to develop techniques that can learn to predict radiological scores and localize them from such weakly labelled training data.

In this paper, we propose a CNN-based framework to classify and qualitatively localize multiple abnormalities in T2 weighted sagittal lumbar MRIs. The method is trained, validated and tested on a dataset of 2009 patients (12018 individual spine disc volumes). The localizations are achieved implicitly through training the network for the classification tasks, and no other labels are needed apart from the disc-level classification labels. Given an input scan, the trained model: (i) predicts eight separate labels corresponding to eight different radiological gradings, and (ii) produces eight heatmaps, that localizes possible pathological evidence unique to each grading. A simplified view of the overall pipeline is given in Fig. 3.

Why multi-task? Despite their advantages and leading performance, one of the challenges of utilizing CNNs in medical imaging remains: their need for large training datasets. We address this in our work partly through aggressive dataset augmentation but also through the use of multi-tasking where one architecture serves to address multiple problems and hence provides multiple supervisory signals for the common trunk convolutional layers. We show that multi-tasking can improve classification performance, demonstrating that the features and spatial relationships learnt by the convolutional layers generalize to other related tasks. Since clini-

[*] Corresponding author.
E-mail address: amirj@robots.ox.ac.uk (A. Jamaludin).

http://dx.doi.org/10.1016/j.media.2017.07.002
1361-8415/© 2017 Elsevier B.V. All rights reserved.

cians are typically required to assess the state of different anatomical regions within the medical image, either because they are relevant to the clinical question or because they are required to assess visible anatomy for so-called incidental findings, the development of techniques that can predict multiple gradings simultaneously is desirable. In our application of interest, each disc and vertebra can have grades to describe their state of normality or degradation.

Why qualitative localization? We believe that the integration of automated quantitative scores into clinical practice can be aided if the system can highlight the regions within the image that lead to the prediction. However, as mentioned above, the cost of obtaining such image mark-up can be prohibitive. Moreover, it is often difficult for a clinician to identify precisely the voxels that resulted in their opinion. Often it is the overall appearance of an area that leads to the conclusion. To this end, we leverage the ability of CNNs to learn to localize important contributing image regions when trained for classification tasks. We evaluate several previously proposed techniques and adapt them to our task.

1.1. Related work

CNNs have started to be used for medical vision problems in two main ways: either (1) using networks pre-trained on a larger normally non-medical dataset of natural images as features, e.g. to detect skin cancer (Esteva et al., 2017); or (2) training a network from scratch, e.g. to detect sclerotic metastases in spinal Computed Tomography (CT) images (Roth et al., 2014). However, most research on spinal diagnosis classification has been primarily conducted with handcrafted features, (what is now referred to as “shallow” learning), e.g. in works concerning radiological scoring of the intervertebral discs (Ghosh et al., 2011; Jamaludin et al., 2015; Lootus, 2015; Roberts et al., 2010). One recent successful example of using CNNs on medical images is a segmentation framework called U-Net proposed by Ronneberger et al. (2015) which overcame the problem of a small amount of data through the use of elastic augmentation, though requiring strong (pixel-level) supervision. Visualizing disease regions via networks has been investigated in other works in medical imaging (Esteva et al., 2017; Shin et al., 2016), while multi-tasking is prevalent in both medical and non-medical task e.g. a multi-task segmentation CNN by Moeskops et al. (2016) and the UberNet network proposed by Kokkinos (2016).

Over the years several methods have been developed for visualizing saliency maps of predictions from CNN models trained on classification tasks. In this work, we focus on three methods: (i) error backpropagation by Simonyan et al. (2014), which computes the gradient of the class score prediction with respect to the input image via backpropagation; (ii) guided backpropagation by Springenberg et al. (2015), which changes the backward pass of ReLU in a similar manner to the ‘deconvnet’ method in Zeiler and Fergus (2014); and, (iii) excitation backpropagation by Zhang et al. (2016), which changes the gradients of the convolutional and average pooling layers. Another method that works extremely well that is not covered in this work is the saliency via class activation maps (CAM) by Zhou et al. (2016) which is applicable principally to later layers in the CNNs (it is usually applied to the global average pooling prior to the final fully-connected layer). Discussions of what the visualizations mean and how they can be changed by architectural choices in the CNN models can be seen in the work by Mahendran and Vedaldi (2016).

On the content:. This paper is an extended version of our MICCAI conference paper (Jamaludin et al., 2016). The extensions include: (i) the addition of two new classification gradings: spondylolisthesis and central canal stenosis; (ii) comparison of multiple CNN architectures: the original from Jamaludin et al. (2016), against a version with 3D kernels; (iii) alternative backpropagation techniques to visualize a saliency map or heatmap of the predictions; and (iv) quantitative evaluations of the evidence hotspots.

2. The Genodisc Dataset

In this paper we use T2 sagittal MRI scans obtained from the Genodisc Dataset which consists of 12018 individual disc volumes, from 2009 patients, and their corresponding radiological gradings. Each scan contains up to six lumbar discs per patient, with the majority having a complete field-of-view of the lumbar region and including six discs.

The dataset was collected by the Genodisc Project (http://www.physiol.ox.ac.uk/genodisc) in 2009. It recruited from “patients who seek secondary care for their back pain or spinal problem”, and recorded their MR scans and clinical information. The scans were sourced from multiple centres in the UK, Hungary, Slovenia and Italy. They came from different machines, and were acquired with a variety of acquisition protocols following clinical standards. Because of this, the slice thickness of the sagittal scans also vary substantially from one centre to the other, ranging from 2.6 mm to 6.0 mm, with a median of 4mm. It contains multiple type of scans: T1 axial, T1 sagittal, T2 axial and T2 sagittal. For the purposes of this research, only T2 sagittal scans are considered, as most radiological gradings for the spine can be assessed via T2 sagittal scans, and these often emphasize pathologies compared to T1. The scans were assessed by a single expert spinal radiologist who produced a range of global, i.e. the whole spine, and local, i.e. per disc, gradings.

2.1. Radiological gradings

We consider the following six gradings from the dataset (and the SpineNet is trained to automatically classify these):

1. **Pfirrmann grading.** A grading system of disc degeneration using criteria of disc signal heterogeneity, brightness of the nucleus and disc height; 5 grades.
2. **Disc narrowing.** Defined as a multi-class measurement of the disc heights; 4 grades.
3. **Spondylolisthesis.** A binary measure of the vertebral slip; i.e. is the pair of vertebrae above and below a disc in-line or has it slipped?
4. **Central canal stenosis.** The constriction of the central canal in the region adjacent to each intervertebral disc. The grading is based on assessment of both sagittal and axial images. We have only studied a binary ‘presence’ or ‘absence’ of stenosis here.
5. **Endplate defects.** Deformities of the endplate regions, both upper and lower, with respect to the intervertebral disc.
6. **Marrow changes.** Visible signal variations of the vertebral endplate region.

Endplate defects and **marrow changes** are further divided into a classification of lower and upper endplate gradings such that the two gradings can be viewed as four separate binary classification tasks. In all, we deal with eight unique classification tasks. Fig. 1 gives the grade distribution for each radiological grading in the dataset, and Fig. 2 shows example MRI slices for each radiological grading.

3. Classification overview & input volumes

A standard lumbar spinal MR scan typically includes multiple inter-vertebral discs in its field-of-view. The number of discs varies depending on the pathology of the patient, the clinical question of interest and the protocol standards of the clinical centre, but typically contain the six lumbar discs from the T12-L1 disc to the

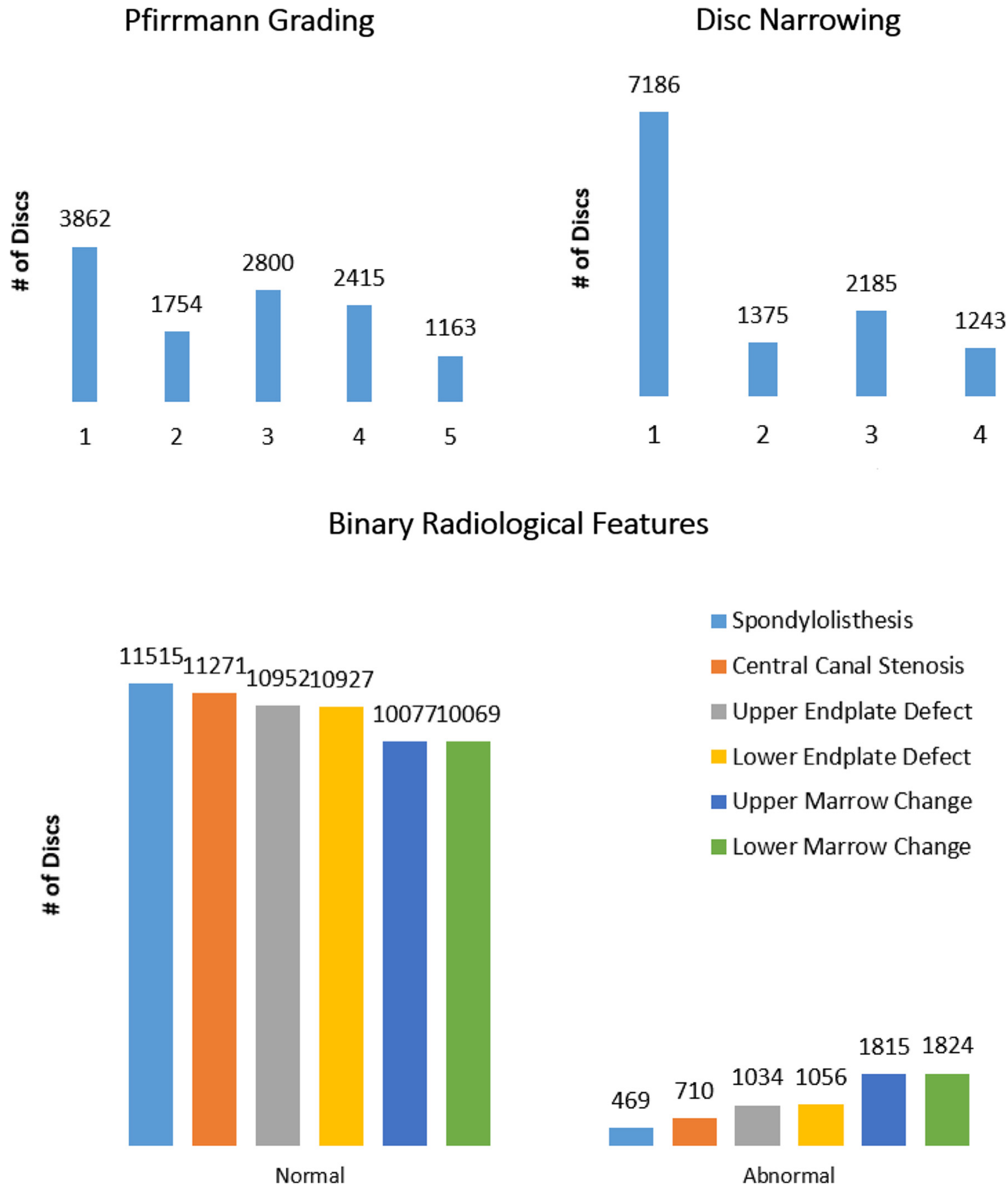

Fig. 1. Score label distribution for all the gradings. Both **endplate defects** and **marrow changes** have two separate gradings, one each for both the upper and lower endplate regions. Note, there is a total of 12,018 discs but since there are missing labels (independent of grading), the totals of labelled discs shown in the table for each grading are different. **Pfirrmann grading** and **disc narrowing** are multi-class while the other gradings are binary. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

L5-S1 disc. Our goal is to predict the radiological gradings for the discs and assocated vertebrae, and we choose as input to the CNN a volume around each disc that includes the disc and part of the vertebral bodies above and below the disc. The CNN predicts all of the gradings for the disc volumes simultaneously, as well as the heatmap of evidence hotspots. We have experimented with whole spine images as input and found that there is a significant decrease in terms of classification performance. The network is trained simultaneously to predict all the radiological gradings via a multi-task loss function. Our overall pipeline can be seen in Figure 3.

Disc volume extraction: First a tight bounding volume is obtained for each of the seven vertebral bodies adjacent to the six radiologically labelled discs via the detection and regression steps based on the vertebral bodies detection framework of Jamaludin et al. (2015) and Lootus (2015). Then, from each pair of vertebral bodies, rough estimates of disc bounding volumes are obtained. Finally, the disc volumes for classification are defined as follows: the region is rotated within the sagittal slice so that the disc is horizontal and centred. The regions are resized, while maintaining aspect ratio, to be the same dimension 112 × 224 per slice; this 1:2 ratio is to ensure that the disc region would not include the up-

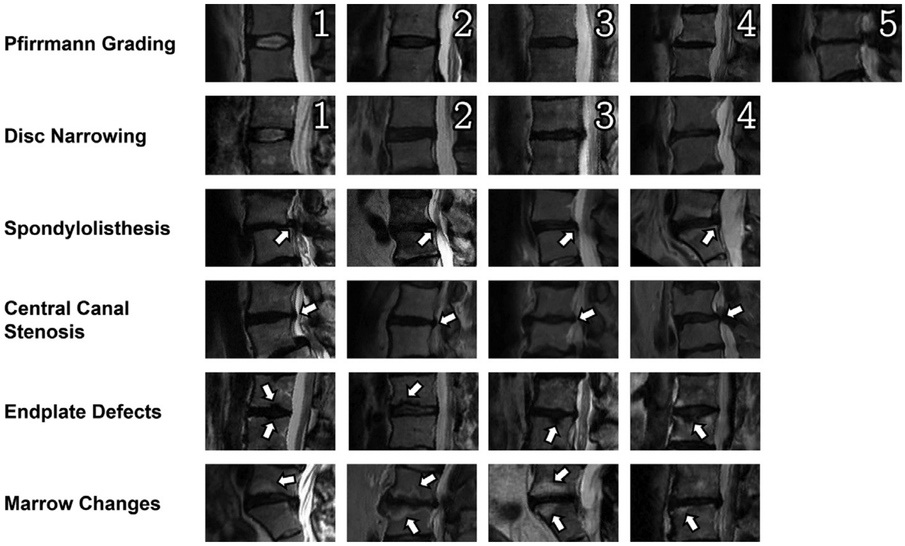

Fig. 2. **Pfirrmann grading** grades and **disc narrowing** are multi-class, 5 classes for **Pfirrmann grading** and 4 classes for **disc narrowing**, while the other gradings are binary.

per and lower endplates of the adjacent vertebral bodies. Roughly 40% of each vertebrae, upper and lower, appear in each region of interest. The discs are aligned according to their mid-sagittal slices and include up to 9 sagittal slices. Narrow discs with less than 9 slices are zero-padded slice-wise. Each slice of the disc is normalized such that the median intensity of its pair of vertebral bodies is 0.5 to mitigate against bias field effects. The range of the intensity inside the disc volume is set to be between 0 and 1.

4. Loss functions & CNN architectures

In this section we first describe the multi-task loss functions that are used to train the network, followed by the CNN architectures that we compare. Training and implementation details are then given in Section 4.3.

4.1. Loss functions

Multi-task loss: Each of the eight radiological gradings of Section 2 are classifications, and we follow the conventional practice of training a network by minimising the softmax log-losses (cross-entropy) for the grading. Each grading is considered as a separate task, and a separate head is defined on the network for each task (with a common trunk shared between the tasks, see below). For each task, t, where $t \in \{1 \dots T\}$, and input volume, x, the network outputs a vector y of size C_t which corresponds to the number of classes in task t. The loss, $\mathcal{L}_t$, of each task over N training volumes can be defined as:

$$\mathcal{L}_t = -\sum_{n=1}^{N} \left(y_c(x_n) - \log \sum_{j=1}^{C_t} e^{y_j(x_n)} \right) \tag{1}$$

where y_j is the jth component of the **FC8** output, and c is the true class of x_n. Solving the multi-task problem in an end-to-end fashion translates to minimizing the summation of all the losses i.e. all t in Eq. (1):

$$\mathcal{L} = \sum_{t} \omega_t \mathcal{L}_t \tag{2}$$

where ω_t is the weight of task t. We find setting $\omega_t = 1$ for every t works well on our tasks, but it might be beneficial to fine tune ω_t for different problems. Setting one weight to 1 and the rest to 0 results in a standard training of a single task. At training time, the loss of task t is only calculated for valid labels i.e. missing labels of task t are ignored. This is extremely beneficial as inputs can possess missing labels in one task but not others.

Class-balanced loss: Class imbalance refers to the different number of training examples for each possible outcome of a multi-way classification. Since most of the classification tasks we deal with are unbalanced, a common problem in medical classification tasks, we use a class-balanced loss during training. For each task, we reweight the loss such that the combined losses are balanced. To achieve balance in training, class-specific weights are introduced. These weights are determined to be:

$$\alpha_c = freq(c)^{-1} \tag{3}$$

where $freq(c)$ is the class label frequency in the training sets for each task e.g. $freq(1)$ is the frequency of the first class ($c = 1$) in the training set. The loss for each task can then be expressed as:

$$\mathcal{L}_t = \sum_{c} \alpha_c \ell_t(c) \tag{4}$$

where $\ell_t(c)$ is the component of the loss for class c. This is equivalent to oversampling the minority class but, since our data is multi-labelled, this balance cannot be achieved by a trivial solution such as simply oversampling or undersampling each disc.

4.2. CNN architectures

The base network architecture is a modified version of the 1024 variant of the VGG-M network introduced in Chatfield et al. (2014). The input dimension of the disc volume is 112 × 224 × 9 where 9 refers to the number of slices in an extracted disc volume (see Section 3 for disc volume details). The **Conv1** kernels of VGG-M are changed to accept 9 input channels instead of the standard 3 for RGB images. We also omit the use of local response normalization after the **Conv1** and **Conv2** layers and the stride of the **Conv2** layer is changed to 1 to ensure no information is lost after the **Conv5** pooling layer. The main configuration we use for parameter sharing is a branch point after **Conv5**. To predict the multiple tasks, each task branches out after the specified branch point i.e. unique **FC6, FC7**, and **FC8** layers for each task. The layers succeeding the branch point are identical in terms of the number of weights for each task except for the final **FC8** layer, which has specific output dimensionality according to the number of classes in each task e.g.

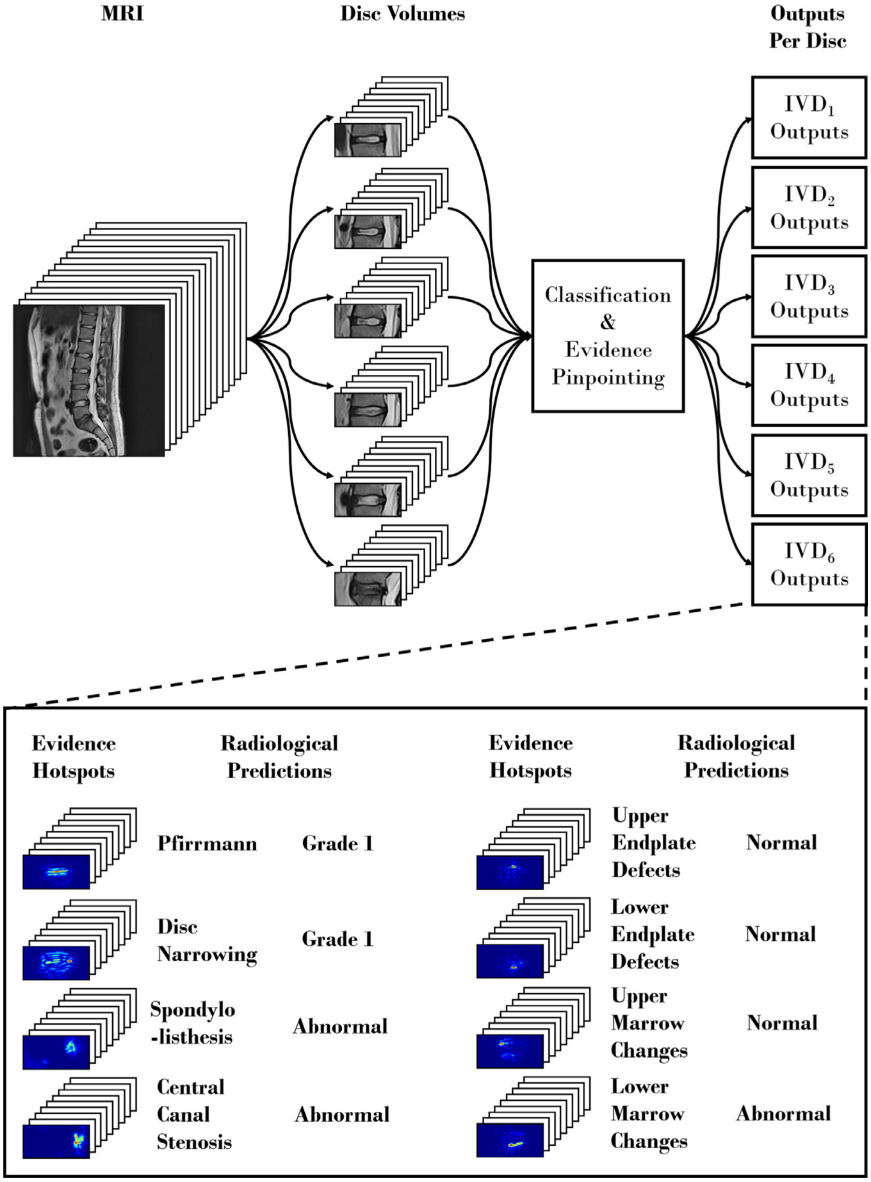

Fig. 3. Overview: Each of the six intervertebral disc volumes is assessed as an input to the CNN. There are eight outputs for each disk consisting of a radiological grading prediction and an associated hotspot mapping. The CNN architecture is given in Fig. 4. The dimension of each input disc volume is 112 × 224 × 9, essentially a stack of 9 image slices of the disc, which is also the dimension of the evidence hotspot heatmap.

5-way softmax for Pfirrmann, and 4-way softmax for disc narrowing classifications. Fig. 4 shows the network with a branch point after **Conv5**.

3D kernels: Since MR scans comprise a stack of 2D images forming a 3D volume, we can either use a 2D CNN which at the **Conv1** layer treats individual slice of the whole stack of 2D images as separate feature channels or a 3D CNN which uses 3D convolution kernels. Both ideas have merits: using a 2D CNN makes sense since spinal MRIs, and in fact many other MRIs in clinical practice, are often assessed by radiologists on a per slice basis and as such is closer to the condition at which the radiological gradings were annotated; while using a 3D CNN means keeping the integrity of the volumetric information we are assessing. In the 3D case every convolutional kernel except for **Conv4** is extended to have an extra dimension of size 3, going from right-to-left slice-wise in the volume, with no pooling and the stride set to be 1 e.g. in **Conv1** the standard 7 × 7 kernel is extended to 7 × 7 × 3. The smallest odd kernel size of 3 was used for the last dimension since the input disc is limited to just 9 slices at maximum. To keep the number of parameters comparable, the **Conv4** layer is set to be 3 × 3 × 9 with no padding slice-wise which results in a reduction of the **Conv4** activation dimension. Subsequent layers proceeding **Conv4** are thus exactly the same as those in a 2D CNN. Overall, the number of parameters for the 3D CNN is roughly 2% more than the 2D CNN (434M vs 426M).

4.3. Implementation details

Training: Training of all the tasks is done end-to-end simultaneously via stochastic gradient descent with momentum from scratch without any pre-training. The inputs are normalized with per-channel mean subtraction as per common convention for the 2D models but per-volume mean for the 3D models. The hyper-parameters are: mini-batch size 256 (2D) & 128 (3D); momentum 0.9; weight decay 0.0005; initial learning rate 0.001, which is lowered by a factor of 10 as the error plateaus. The weights are initialized according to the method in Glorot and Bengio (2010) and normally reach convergence in about 1000 epochs. The networks

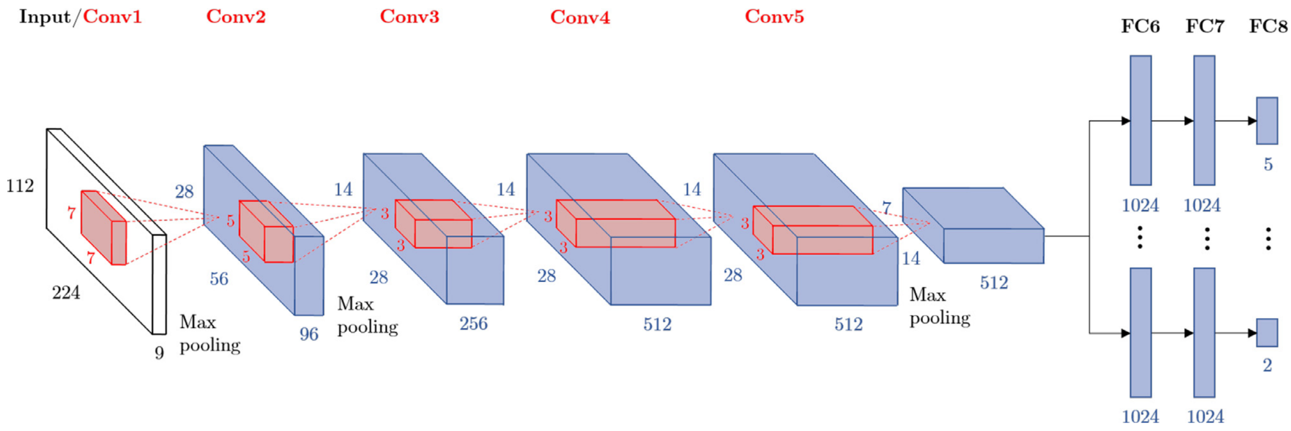

Fig. 4. Multi-task version of the VGG-M architecture (Chatfield et al., 2014) with a branch point after the **Conv5** layer resulting in a multi-way classification tasks. The numbers in black refer to the input, the numbers in blue are for the activations after an operation (e.g. convolution) while the numbers in red (for example 7 in **Conv1**) refer to the size of the kernels. The max-pooling window is set to be 2 × 2 with a stride of 2. For all the convolutional kernels, the stride is set to be 1 except for **Conv1** which is set to be 2. For the 3D network, all the convolutional kernels are extended to have an extra dimension of size 3 e.g. for **Conv1** the kernel would be 7 × 7 × 3. In the figure, the output of the upper **FC8** layer is set to size 5 to predict Pfirrmann grading, while the lower one illustrates a binary output with a size 2 output.

were trained via the MatConvNet toolbox (Vedaldi and Lenc, 2014) using an NVIDIA Titan X GPU.

Data augmentation: We employ several aggressive random on-the-fly augmentation strategies during training. We include the methods suggested by Krizhevsky et al. (2012) for natural images, and also define additional augmentations that are suited to 3D MR scans. The training augmentation strategies are:

1. Rotation with $\theta = -15^\circ$ to 15°
2. Translation of ± 32 pixels in the x-axis, ± 24 pixels in the y-axis, ± 2 slices in the z-axis
3. Rescaling with a scaling factor between 90% to 110%
4. Intensity variation between −0.1 to 0.1
5. Random slice-wise flip i.e. reflection of the slices across the mid-sagittal

At test time, the final prediction is calculated from the average of 54 predictions: (i) 8 patches, ± 16 pixels from the origin, alongside the centre patch, (ii) their slice-wise reflections, and (iii) sliding the slice window ± 1 across the volume.

5. Radiological grading classification experiments & results

The dataset is split into a 80:10:10 train:validation:test sets on a per patient basis (not per disc). This results in 1806 patients (10,836 discs) train/val samples, and 203 patients (1,224 discs) test samples. All the results reported here are from two models trained for each experiment by swapping the validation and test sets.

Evaluation protocols: To evaluate classification performance, we use average per-class accuracy which is suitable for highly unbalanced classifications. For comparison, we provide the average per-class intra-rater agreement calculated from two separate sets of labels by the same radiologist. These labels were obtained for a subset of the dataset (121 patients, 726 discs) assessed by the radiologist at different times (in comparison the full test set consists of 203 patients). The intra-rater agreement serves as a good benchmark of performance since we are essentially limited to the quality of the label i.e. we can only be as good as the radiologist. Note that inter-rater agreement, i.e. agreement between two different radiologists, tends to be consistently worse than the intra-rater agreement e.g. Pfirrmann et al. (2001) reported that **Pfirrmann grading** has an inter-rater kappa range of **0.69 – 0.81** and an intra-rater range of **0.84 – 0.90**; while the radiologist in our dataset has an above average intra-rater kappa value of ***0.91***. To obtain a standard deviation over the results, two models are trained for each experiment by swapping the validation and test sets.

Multi-task experiment: We investigate the relationship between the addition of tasks and its impact on the performance of the model. For this experiment we used the 2D network that share layers up till **Conv5**, and each individual task afterwards has its own **FC6, FC7**, and **FC8** layers. The accuracy for each task and the intra-rater agreement is given in Table 1. It can be seen that going from a CNN trained on a single task, in this case **lower endplate defects**, to one trained on multiple tasks improves the performance considerably **(79.5% → 86.4%)**. This is also true one other tasks, although not shown in Table 1, i.e. **Pfirrmann grading (69.8% → 71.2%), disc narrowing (72.3% → 73.9%), upper endplate defects (79.0% → 85.7%), upper marrow changes (88.1% → 89.2%)**, and **lower marrow changes (87.3% → 88.2%)**.

Overall, all the learned tasks have better results if the tasks are jointly learned and the inclusion of more tasks generally results in a better performance.

2D vs 3D architectures: The classification performance of the 2D and 3D models is compared in Table 2. It appears that most gradings, except for **Pfirrmann grading (71.9% → 71.5%)** and **disc narrowing (75.9% → 75.0%)**, are better or equal in performance with the 3D CNN. These exceptions might be due to the fact that both **Pfirrmann grading** and **disc narrowing** are normally graded by radiologists purely on only the mid-sagittal slice of each disc. The most significant difference in terms of performance going to 3D can be seen in the prediction of **spondylolisthesis (92.9% → 95.2%)**, slippage of the vertebral bodies, which is the most global grading (in terms of the disc volume fed into the CNN) and benefits most with the addition of the 3D kernels. Overall, the 3D network is on average 0.6% **(85.7% → 86.3%)** better in terms of per task accuracy.

We investigated alternative 3D networks i.e. those without the reduction of slice-wise dimension at **Conv4** but these proved to be only marginally better than our baseline 2D network. We suspect with the addition of more training data the 3D model may improve further, as in Tran et al. (2015) where the presence of a large-scale dataset made 3D CNNs shine. We also experimented with deeper networks (VGG-16, ResNet, etc.) using 2D kernels, but obtained little to no improvements in performance. Again, we suspect that this lack of improvement is due to insufficient training data so that we are not able to take advantage of the additional capacity of the deeper networks.

Adding disc level supervision: Since all the disc volumes extracted are used as independent samples without any information regarding the disc level, we also experimented on adding more disc level supervision into the network. The intuition was that the

Table 1
Performance under variations of the number of tasks for the 2D network. The network was first trained on the **lower endplate defects** and from left to right the network is reinitialized and trained again from scratch with the addition of more classification tasks. The rightmost column shows the results with a network trained with six of the eight tasks. The tasks are added after the **Conv5** branch point. Overall, there is a positive correlation between the amount of shared tasks and the performance of the network.

Task	Intrarater	**Lower endplate defects**	+ **Upper endplate defects**	+ **Lower marrow changes**	+ **Upper marrow changes**	+ **Pfirrmann**	+ **Disc narrowing**
Lower endplate defects	83.3	79.5 ± 1.2	84.3 ± 0.7	85.4 ± 0.9	85.2 ± 0.1	86.0 ± 2.3	86.4 ± 2.0
Upper endplate defects	80.7	–	84.8 ± 1.9	84.7 ± 3.0	85.0 ± 1.9	85.6 ± 0.5	85.7 ± 0.1
Lower marrow changes	91.4	–	–	88.6 ± 0.2	88.4 ± 2.4	87.9 ± 1.2	88.2 ± 2.0
Upper marrow changes	92.5	–	–	–	89.5 ± 0.1	89.7 ± 0.2	89.2 ± 0.5
Pfirrmann	70.4	–	–	–	–	68.8 ± 0.7	71.2 ± 0.4
Disc narrowing	72.0	–	–	–	–	–	73.9 ± 0.7

Table 2
The performance (%) of various models on the test set. "**Intra-rater**" is the intra-rater agreement. "**2D**" and "**3D**" are the results from the 2D and 3D CNNs, two models each with swapped validation and training sets.

Tasks	Intra-rater	Models	
		2D	3D
Pfirrmann	70.4	71.9 ± 1.5	71.5 ± 1.0
Disc narrowing	72.0	75.9 ± 0.4	75.0 ± 2.3
Upper endplate defects	80.7	83.7 ± 1.8	85.2 ± 2.1
Lower endplate defects	83.3	86.9 ± 1.3	87.5 ± 0.4
Upper marrow changes	92.5	89.9 ± 2.2	91.0 ± 1.3
Lower marrow changes	91.4	90.1 ± 1.7	90.3 ± 2.1
Spondylolisthesis	89.6	92.9 ± 0.7	95.2 ± 0.0
Central canal stenosis	79.7	94.3 ± 0.0	94.3 ± 0.9
Average	82.5	85.7 ± 0.9	86.3 ± 0.3

disc level might improve performance as discs of differing levels have different grading distributions: it has been observed that pathological cases appear significantly more on lower level discs (L4-L5 or L5-S1) when compared to upper level discs. We specified the lumbar disc level (which one of six) as a one-hot encoding vector, and added intermediate fully connected layers prior to concatenation with the **FC6** activations of each of the eight gradings. However, we did not see any significant improvement on adding this supervision.

Comparison to other methods: We evaluate our performance on **Pfirrman grading** and **disc narrowing** classifications using the test set and evaluation protocol of Lootus (2015). It is important to note that in Lootus (2015): Pfirrmann grading is measured in terms of accuracy to ± 1 of the radiologist grade, and disc narrowing grading is simplified to a binary classification of normal/abnormal discs where grade 1 is considered to be normal and grades 2 to 4 are abnormal. For the comparison, we adapt our gradings to match those of Lootus (2015). We surpass their performance by **+8.7/+8.2%** (**87.4%** → **96.1**/**95.6%**) for Pfirrmann grading, and **+4.1/+6.3%** (**83.7%** → **87.8**/**90.0%**) for disc narrowing; compared against our 2D/3D models.

We also compare our **central canal stenosis** performance against Zhang et al. (2017) at different disc levels where we obtain a comparable performance when comparing their results on their data against our results on our test set: **+7.5%** (**87.2%** → **94.7**) for L3-L4, **+0.8%** (**85.1%** → **85.9**) for L4-L5, and **+6.2%** (**87.5%** → **93.7**) for L5-S1). Note that the method described in Zhang et al. (2017) uses axial images, the standard for stenosis classification, while we use sagittal.

Limitations: We only use T2-weighted sagittal scans in our experiments while certain gradings are normally assessed with the addition of other modalities e.g. axial scans are commonly used to assess **central canal stenosis**, and Modic changes grading typically requires both T1-weighted and T2-weighted scans. We compensate for the classification of Modic changes by only looking at Modic changes specific to T2. This is why we call our grading, **marrow changes**. We are able to achieve reasonable performance in **central canal stenosis** with only sagittal scans.

6. Visualizing evidence hotspots

We show here that a network trained for a fine-grained classification task can produce a heatmap which pinpoints the region in the image space responsible for the prediction. This map lights up pathological areas, 'hotspots' of the prediction, specific to the trained task in the image. Only the class labels attached to the disc volume need be used in training without any other extra supervisory information. We term the heatmap produced 'evidence hotspots'. In the following, we consider three different saliency extraction methods to produce the evidence hotspots:

Saliency by backpropagation: Simonyan et al. (2014) proposed a method that ranks the influence of an image, x, according to its influence on a specific class score. The method proceeds by linearizing the relationship between a specific output class score and the input x as prediction of a specific class, y. For CNNs, we can approximate the highly non-linear function of y to be $y \approx w^{\mathbf{T}}x + b$ where w and b are the weight and bias of the model. So ranking the influence of the input, x, can be posed as ranking the magnitude of the weight, w, that influences the output y. The weight can be obtained as $w = \frac{\partial y}{\partial x}$, which can be found by backpropagation. However, unlike Simonyan et al. (2014), where the input x is a 3 channel RGB image ($z = 3$), our input consists of 9 channels ($z = 9$), each a greyscale image of an MRI slice in the disc volume. Furthermore, instead of producing a 1 channel saliency map calculated from the maximum magnitude of w, $\mathcal{M} = max_z|w|$ as in Simonyan et al. (2014), the saliency map is also 9 dimensional, such that $\mathcal{M}_z = |w_z|$ where $z \in \{1 \ldots 9\}$ since each input channel corresponds to an actual volumetric slice of the disc. The main visual difference from our maps to the maps shown in Simonyan et al. (2014) is that our salient regions are more localized and more specific to the area that is the cause for the classification. We suspect that this might be because our classification tasks are more fine-grained, and because our input images are visually very similar.

Saliency by guided backpropagation: Guided backpropagation proposed by Springenberg et al. (2015) is a slightly modified, compared to Simonyan et al. (2014), saliency extraction method also based on backpropagation. The proposed method changes the gradient, δ, of ReLUs during backpropagation. As we are only interested in finding the derivative of the class score with respect to the input, the gradient, δ, here refers to the derivative with respect to the input (or activations in the case of gradients of intermediate layers), x, and not the weights of the CNN. During standard backpropagation, a ReLU unit, $y = \max(0, x)$, backpropagates the gradient of the layer next to the ReLU but for only the positive portion of the input $\delta_{i-1} = \delta_i \cdot [x > 0]$, where δ is the gradient and i denotes the layer of the gradient i.e. δ_i is the gradient of

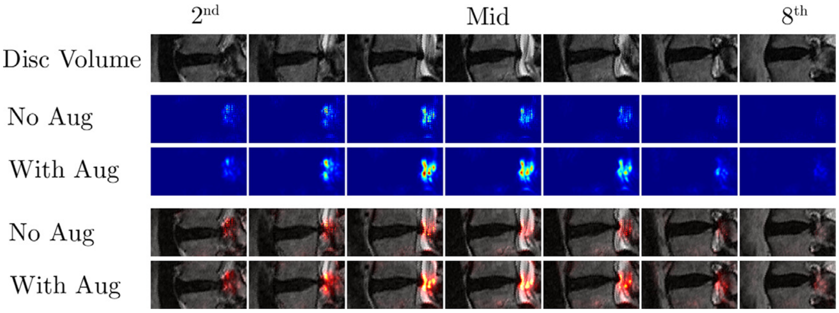

Fig. 5. Test time augmentation: Example of central canal stenosis example taken from the test set alongside its hotspot via excitation backpropagation with and without test time augmentation. The top row shows images of the slices of a disc volume (slice 2 to slice 8) followed by two rows of the heatmaps of the hotspots, and two rows of the heatmaps overlaid on the image slices. Note the improvement in hotspot visibility brought by the augmentation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

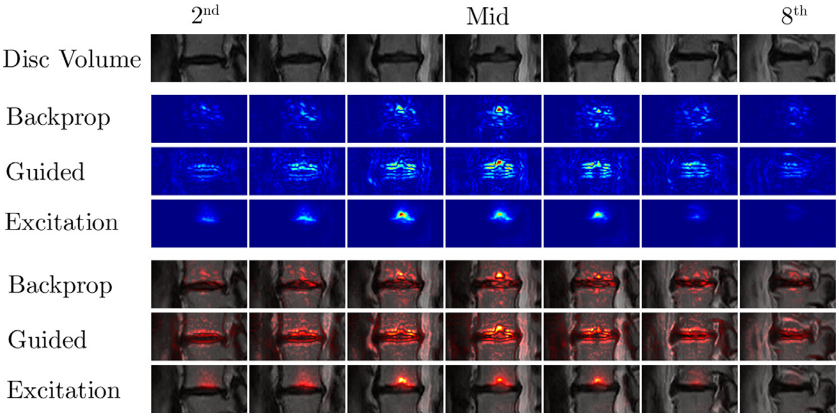

Fig. 6. Hotspot comparison 1: Example of upper endplate defects example taken from the test set. Topmost set of images are slices of a disc volume (slice 2 to slice 8) followed by: (i) the heatmaps of the hotspots, and (ii) the heatmaps overlaid on top of the disc volume. We show the three different methods of producing hotspots with test time augmentation. The excitation method gives the best visibility of the three. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the layer succeeding the ReLU and δ_{i-1} is the gradient after backpropagating through ReLU (gradient after backward pass through ReLU). Springenberg et al. (2015) further restricts the gradient of the ReLU unit by only propagating positive gradients with positive inputs $\delta_{i-1} = \delta_i \cdot [x > 0] \cdot [\delta_i > 0]$. This is in a way a combination of normal backpropagation used by Simonyan et al. (2014) and the 'deconvnet' method proposed by Zeiler and Fergus (2014) which has a ReLU gradient that depended only on the positivity of the gradient of the next layer $\delta_{i-1} = \delta_i \cdot [\delta_i > 0]$.

Saliency by excitation backpropagation: Zhang et al. (2016) proposed a method based on a probabilistic Winner-Take-All process. Similarly to guided backpropagation, this method changes the backpropagation of the gradient by looking at both the gradient δ_i, and the input, x. Unlike guided backpropagation however, only gradients of convolutional and the average pooling layers are changed. The methods starts by thresholding the convolutional weights, if backpropagating in the convolutional layer, $w^+ = \max(0, w)$ and running a forward pass producing an intermediate output y^+. The gradient δ_i is then normalized by a factor of y^+ via element-wise division: $\hat{\delta}_i = \delta_i \oslash y^+$. Then the backpropagated gradient δ_{i-1} is calculated from the normalized $\hat{\delta}_i$ via a backward pass, which is then normalized via element-wise multiplication with the input: $\hat{\delta}_{i-1} = \delta_{i-1} \odot x$. Zhang et al. (2016) also suggested an extension to their method called contrastive excitation backpropagation which was shown to be better and we follow this in our implementation. Contrastive excitation saliency can be found by backpropagating the difference between the gradients of the penultimate layer: $\delta_i^+ - \delta_i^-$ where δ_i^+ is the gradient calculated via positive weights w^+ and δ_i^- is the gradient calculated via the negation of the positive weights $-w^+$. For our implementation, we backpropagate only up till **Conv1** which gives us visually better results. As in Zhang et al. (2016), we find that backpropagating back to the input space does not give good saliency maps.

Test time augmentation: Instead of just one standard run of backpropagation, we find that aggressive augmentations is key to producing more salient localized hotspots. The final heatmap is computed from the average of multiple saliency maps produced from randomly augmented images using our training augmentation scheme; the resulting saliency map is transformed back to the original image space with the reverse of the applied augmentation. We use around 150 augmentations (each requiring a backpropagation pass) which takes around 90 s to complete for a single disc volume. It can be seen in Fig. 5 that test time augmentation helps improve the hotspots essentially by smoothing the heatmaps.

6.1. Qualitative results

We use the 3D CNN trained to predict the eight different gradings to visualize the hotspots. We find no significant difference in terms of hotspots quality between 2D and 3D CNNs for standard and guided backpropagations. Since the 2D CNN essentially merges the slices of the disc volume after **Conv1**, it is necessary to backpropagate all the way back to the input space to get the hotspot for each channel/slice in the disc volume. This in turn results in poorer heatmaps when using a 2D CNN and excitation backpropagation (which works poorly when propagated all the way back to the input layer). This is different in the 3D case, where we can

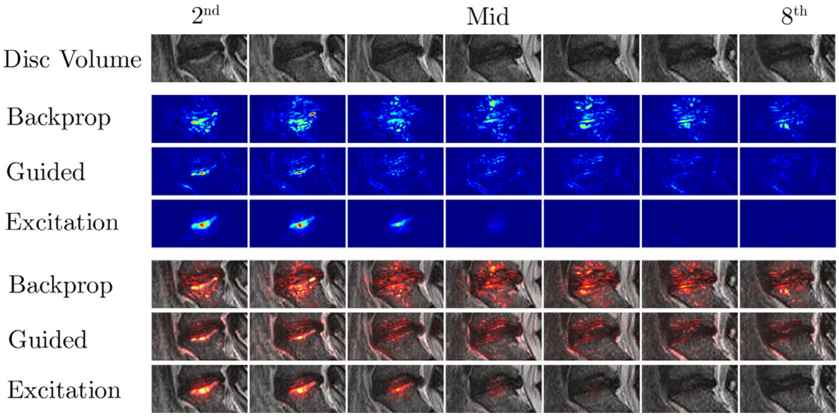

Fig. 7. Hotspot comparison 2: Example of lower marrow change example taken from the test set. Identical configuration to Fig. 6. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

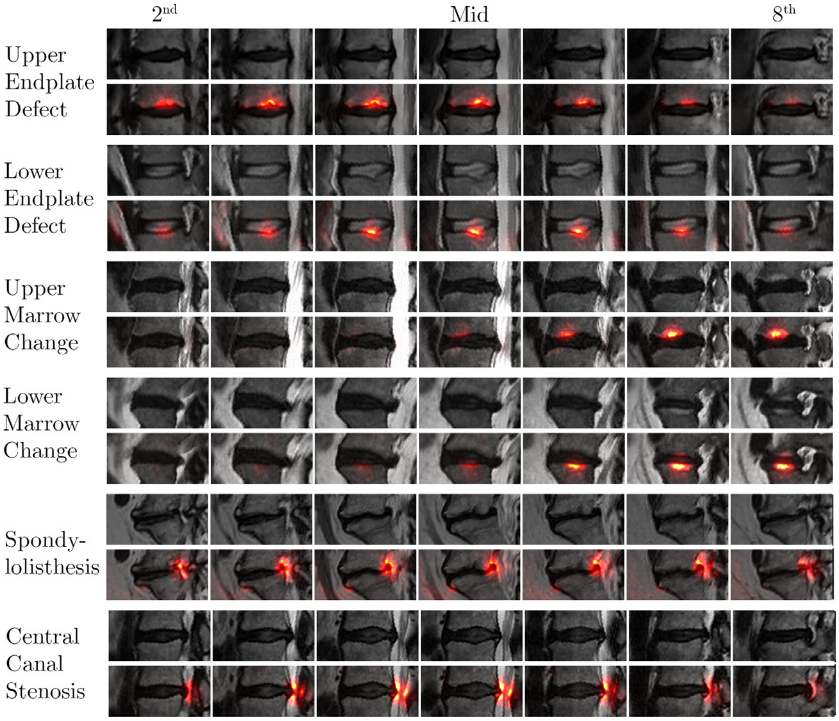

Fig. 8. Evidence hotspots of different gradings: Examples of disc volumes (upper in each pair) and their corresponding evidence hotspots (lower in each pair). The leftmost and rightmost columns of images are the second and eighth slice for each disc, out of the full volume of 9 slices. Going from top to bottom are examples for each of the binary gradings: (i) **upper endplate defects**, (ii) **lower endplate defects**, (iii) **upper marrow changes**, (iv) **lower marrow change**, (v) **spondylolisthesis**, and (vi) **central canal stenosis**. Only pathological/abnormal examples are shown for each radiological grading/classification task, i.e. endplate defects appearing as protrusions of the discs into the vertebral bodies, and marrow changes appearing as localized discolourations of the vertebral bodies near the vertebral endplates. Note that these hotspots localize extremely well to the assigned tasks e.g. in the lower endplate defects example the hotspots appear only in the lower endplate even though there are defects on the upper endplate. These examples are randomly selected on different patients in the test set. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

backpropagate to any layer in the network (**Conv1** is best for our dataset).

Figs. 6 and 7 compare the three methods we experimented on, in each case using test time augmentation. Qualitatively, excitation backpropagation produces the best results in our dataset with cleaner hotspots and better localizations. For example, in Fig. 6, the upper endplate defect was correctly highlighted only by backpropagation and excitation backpropagation. Similarly, in Fig. 7, the lower marrow change was correctly highlighted only by the guided and excitation backpropagations.

Fig. 8 shows hotspots (via excitation backpropagation) on different tasks or gradings on randomly selected discs in the test set. It can be seen that the hotspots of endplate defects and marrow changes appear to highlight the respective abnormalities in the endplate regions of the vertebral bodies. Interestingly, since spondylolisthesis is the measure of the vertebral slip, the hotspots highlight: (i) both vertebral bodies, and (ii) their misalignment near the posterior. Similarly, in the case of central canal stenosis the hotspots highlight the specific region of only the pinched area near the posterior of the disc and not the whole canal.

6.2. Quantitative results

To validate the evidence hotspots produced by the 3D CNN we compare them against ‘ground truth’ bounding boxes annotated by a clinician. These comparisons are done on 312 disc volumes in

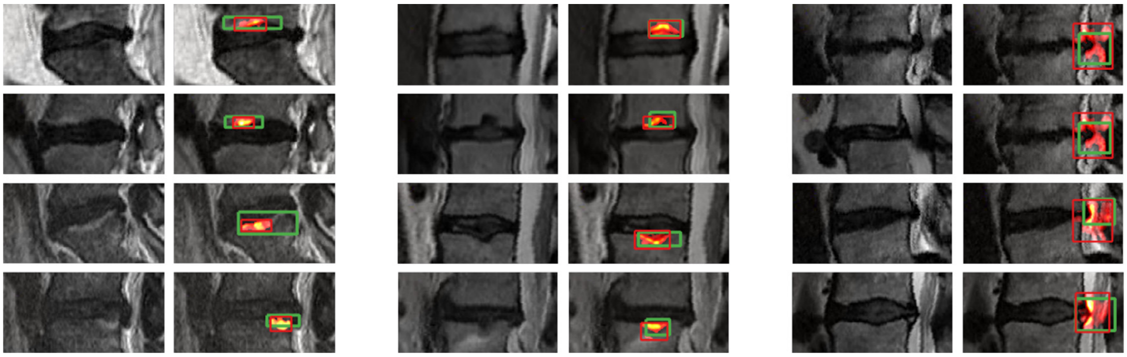

Fig. 9. Bounding boxes examples: Ground truth in green and predicted in red. **Left: marrow changes** examples. **Middle: endplate defects** examples. **Right: central canal stenosis** examples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

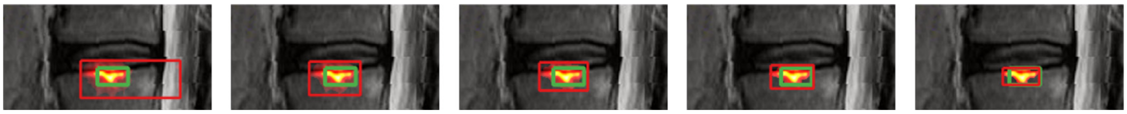

Fig. 10. Effects of thresholding: The threshold increases going from the left to right i.e. from the 95th to the 99th percentile. The ground truth bounding boxes are green in colour and the predicted bounding boxes are red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Jaccard/IOU of the fitted bounding boxes against ground truth bounding boxes calculated on a per volume basis (9 slices per volume). Note that the ground truth annotations are not complete. By only comparing the localization performance when there exists a ground truth bounding box in each slice we get: marrow changes 0.283, endplate defects 0.240, and central canal stenosis 0.501.

Marrow changes	0.242
Endplate defects	0.149
Central canal stenosis	0.405

the test set that have been graded with one or more of: (i) **marrow changes**, (ii) **endplate defects**, or (iii) **central canal stenosis**. These discs were read by a spinal surgeon, who has access to both the disc level information and the radiological gradings reported by the radiologist. The task was to annotate the region that best supported the grading with a tight bounding box in each slice of the scan.

To measure the localization performance of the hotspots, we calculate the Jaccard similarity coefficient, or the Interesection over Union (IoU), between the ground truth bounding boxes and the hotspots. To do so, for a given input disc volume, we produce a binary mask by thresholding based on the percentile of the intensity distribution of the hotspot and fit a minimum bounding box for each mask. Examples of discs with the hotspots, predicted bounding boxes and 'ground truth' bounding boxes can be seen in Fig. 9 while the effects of thresholding on the sizes of the fitted bounding boxes can be seen in Fig. 10.

We find that thresholding around the 90th – 99th percentile consistently works better than other percentiles. Since the magnitude of the hotspots is correlated with the certainty of the grading prediction, it is not surprising to see that higher thresholds have better localization performance. However, since the ground truth is constrained to regular boxes instead of delineated segmentations of the pathological regions, there is a trade-off in performance after a certain threshold unique to each grading. Table 3 shows the Jaccard/IoU of the three gradings.

As a baseline, we also calculate the Jaccard/IOU of predicting the whole of the disc volume as the detected bounding boxes (each box per slice) for each grading. Since sizes of the bounding boxes vary depending on the grading, **endplate defects** tend to be smaller than **central canal stenosis** for example, comparing them to this baseline is a good indicator of the localization performance. Comparing each grading: **marrow changes** (**0.023** → **0.242**), **endplate defects** (**0.013** → **0.149**), and **central canal stenosis** (**0.053** → **0.405**). Overall, we see that the performance of the bounding boxes from the evidence hotspots for each grading, produced using the 3D CNN trained only for classification, is around 10 times better than the naive baseline.

7. Summary & conclusion

We have shown that multi-task training is extremely beneficial especially in medical vision problems where it is common to have multiple-labels for a single modality. Training with multiple tasks in general improves over the performance of training on individual tasks. Furthermore, we have shown that we can also produce high quality visualizations of pathology hotspots that can aid understanding of the predictions of models trained only on weak class labels. Also, we have improved upon results in our original conference publication (Jamaludin et al., 2016) by using 3D CNN and adding more tasks useful for lumbar MR spinal analysis. One clear extension is to experiment with resampling the disc volumes as the effects of slice thickness variations were not assessed which is especially useful for 3D CNNs.

Acknowledgements

We are grateful to Dr. Jill Urban for discussions, to Prof. Jeremy Fairbank for his annotations of grading locations and discussions, to Prof. Iain McCall for the radiological gradings, and to Aravindh Mahendran for discussions on the visualization methods. This work was supported by the RCUK CDT in Healthcare Innovation (EP/G036861/1) and the EPSRC Programme Grant Seebibyte (EP/M013774/1). The data was obtained during the EC FP7 project GENODISC (HEALTH-F2-2008-201626).

References

Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of the British Machine Vision Conference.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115–118. doi:10.1038/nature21056.

Ghosh, S., Alomari, R.S., Chaudhary, V., Dhillon, G., 2011. Computer-aided diagnosis for lumbar mri using heterogeneous classifiers. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1179–1182. doi:10.1109/ISBI.2011.5872612.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the International Conference on Artificial Intelligence and Statistics.

Jamaludin, A., Kadir, T., Zisserman, A., 2015. Automatic modic changes classication in spinal MRI. MICCAI Workshop: CSI.

Jamaludin, A., Kadir, T., Zisserman, A., 2016. SpineNet: automatically pinpointing classification evidence in spinal MRIs. In: International Conference on Medical Image Computing and Computer Assisted Intervention.

Kokkinos, I., 2016. Ubernet: training a 'universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. http://arxiv.org/abs/1609.02132. CoRR abs/1609.02132.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems.

Lootus, M., 2015. Automated radiological analysis of spinal MRI. University of Oxford Ph.D. thesis.

Mahendran, A., Vedaldi, A., 2016. Salient deconvolutional networks. In: European Conference on Computer Vision.

Moeskops, P., Wolterink, J.M., van der Velden, B.H.M., Gilhuijs, K.G.A., Leiner, T., Viergever, M.A., Igum, I., 2016. Deep learning for multi-task medical image segmentation in multiple modalities. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II. Springer International Publishing, Cham, pp. 478–486. doi:10.1007/978-3-319-46723-8_55.

Pfirrmann, C.W.A., Metzdorf, A., Zanetti, M., Hodler, J., Boos, N., 2001. Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine 26 (17). http://journals.lww.com/spinejournal/Fulltext/2001/09010/Magnetic_Resonance_Classification_of_Lumbar.11.aspx.

Roberts, M.G., Pacheco, E.M.B., Mohankumar, R., Cootes, T.F., Adams, J.E., 2010. Detection of vertebral fractures in dxa vfa images using statistical models of appearance and a semi-automatic segmentation. Osteoporosis Int. 21 (12), 2037–2046. doi:10.1007/s00198-009-1169-6.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention.

Roth, H.R., Yao, J., Lu, L., Stieger, J., Burns, J.E., Summers, R.M., 2014. Detection of sclerotic spine metastases via random aggregation of deep convolutionalneural network classifications. MICCAI Workshop: CSI.

Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35 (5), 1285–1298. doi:10.1109/TMI.2016.2528162.

Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep inside convolutional networks: visualising image classification models and saliency maps. In: Workshop at International Conference on Learning Representations.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A., 2015. Striving for simplicity: the all convolutional net. In: Workshop at International Conference on Learning Representations.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15. IEEE Computer Society, Washington, DC, USA, pp. 4489–4497. doi:10.1109/ICCV.2015.510.

Vedaldi, A., Lenc, K., 2014. MatConvNet – Convolutional neural networks for MATLAB.CoRR abs/1412.4564. http://dblp.uni-trier.de/rec/bibtex/journals/corr/VedaldiL14

Zeiler, M.D., Fergus, R., 2014. Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I. Springer International Publishing, Cham, pp. 818–833. doi:10.1007/978-3-319-10590-1_53. Visualizing and Understanding Convolutional Networks.

Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S., 2016. Top-Down Neural Attention by Excitation Backprop. Springer International Publishing, pp. 543–559. doi:10.1007/978-3-319-46493-0_33.

Zhang, Q., Bhalerao, A., Hutchinson, C., 2017. Weakly-Supervised Evidence Pinpointing and Description. Springer International Publishing, Cham, pp. 210–222. doi:10.1007/978-3-319-59050-9_17.

Zhou, B., Khosla, A., A. Oliva, L.A., Torralba, A., 2016. Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition.