SpineNet: Automatically Pinpointing Classification Evidence in Spinal MRIs
Amir Jamaludin¹, Timor Kadri², Andrew Zisserman¹
¹Visual Geometry Group, Department of Engineering Science, University of Oxford
²Mirada Medical

• Spinal MRIs diagnosis → multiple gradings of discs and vertebrae.
• We:
  i. jointly learn multiple grading systems,
  ii. produce evidence for predictions i.e. “evidence hotspots”.
• Near-human performances across multiple radiological gradings.

Multi-task CNN
We propose a convolutional neural network training scheme which takes in disc volumes, detected using a modified framework proposed in [1], and the corresponding multiple labels per disc, where each label comes from a unique grading system, as inputs during training.

Evidence Hotspots
We show 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.

We follow the saliency method proposed in [2] that correlate pixels in the input image to the predicted output classification by back-propagating the score, at test time, back to the input space.

We extend it to volumetric inputs to produce what we call ‘evidence hotspots’. Multiple ‘evidence hotspots’ heatmaps are produced per disc; each unique to a specific radiological grading. We believe this would help with the integration of automated quantitative scores into clinical practice.

Pfirrmann Grading

Disc Narrowing

Upper Endplate Defect

Lower Endplate Defect

Upper Marrow Change

Lower Marrow Change

Dataset
2009 patients with T2 sagittal MRIs sourced from multiple centres. 12018 individual discs, six per patient, annotated with various radiological gradings by an expert radiologist.

We focused on six main radiological gradings: i) Pfirrmann grading, ii) disc narrowing, iii-iv) upper/lower endplate defects, and v-vi) upper/lower marrow changes.

Pfirrmann Grading
Disc Narrowing
Upper Endplate Defects
Lower Endplate Defects
Upper Marrow Changes
Lower Marrow Changes

Results

The performance (mean ± std %) with various different branch points on the test set. Branch point labelled “None” refers to six independent networks each trained for the six individual tasks.

It can be seen that multi-task training results in a better performance across the multiple tasks and that there is a sweet spot for choosing a branch point at Conv5 in our case. “Intra-rater” is the intra-rater agreement.

References

Acknowledgements
The data was obtained during the EC FP7 project (HEALTH-F2-2008-201620).