Clinical Radiology 73 (2018) 439–445

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net

Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks

D.H. Kim*, T. MacKinnon

Medical Imaging Department, Royal Devon and Exeter Hospital, Barrack Road, Exeter EX2 5DW, UK

ARTICLE INFORMATION

Article history:
Received 28 May 2017
Accepted 14 November 2017

AIM: To identify the extent to which transfer learning from deep convolutional neural networks (CNNs), pre-trained on non-medical images, can be used for automated fracture detection on plain radiographs.

MATERIALS AND METHODS: The top layer of the Inception v3 network was re-trained using lateral wrist radiographs to produce a model for the classification of new studies as either “fracture” or “no fracture”. The model was trained on a total of 11,112 images, after an eightfold data augmentation technique, from an initial set of 1,389 radiographs (695 “fracture” and 694 “no fracture”). The training data set was split 80:10:10 into training, validation, and test groups, respectively. An additional 100 wrist radiographs, comprising 50 “fracture” and 50 “no fracture” images, were used for final testing and statistical analysis.

RESULTS: The area under the receiver operator characteristic curve (AUC) for this test was 0.954. Setting the diagnostic cut-off at a threshold designed to maximise both sensitivity and specificity resulted in values of 0.9 and 0.88, respectively.

CONCLUSION: The AUC scores for this test were comparable to state-of-the-art providing proof of concept for transfer learning from CNNs in fracture detection on plain radiographs. This was achieved using only a moderate sample size. This technique is largely transferable, and therefore, has many potential applications in medical imaging, which may lead to significant improvements in workflow productivity and in clinical risk reduction.

Introduction

Demands on radiology services have increased dramatically in recent years causing a considerable strain on the workforce. In the UK, the number of computed tomography (CT) examinations increased by 29% between 2012 and 2015, whilst recruitment lagged behind, leaving 9% of consultant radiology posts vacant in 2015. Furthermore, in the same year nearly all radiology departments declared a failure to meet their reporting requirements.[1] In February 2016, there was an estimated backlog of 200,000 plain radiographs and 12,000 cross-sectional studies.[2] It is clear from these figures that improvements in reporting efficiency and workflow management are desperately needed if we are to avoid patient harm from delayed or missed diagnosis.

Artificial intelligence (AI) has the potential to address these issues. The wide adoption of electronic picture

* Guarantor and correspondent: D. H. Kim, Royal Devon and Exeter Hospital, Barrack Road, Exeter EX2 5DW, UK. Tel.: 01392 411611; Mob.: 07799113665.

E-mail addresses: Daniel.kim@nhs.net, dan_kim92@hotmail.com (D.H. Kim).

https://doi.org/10.1016/j.crad.2017.11.015
0009-9260/

archiving and communications systems (PACS) has resulted in the development of one of the largest image data sets in existence. In the UK, there were 41 million imaging studies performed in 2016 alone.[3] These data lend themselves perfectly to machine learning.

Machine learning is a form of AI, which uses algorithms that iteratively improve, or learn, in response to training data in order to make autonomous predictions.[4] Supervised machine learning is one subtype, which relies on the provision of pre-labelled training data. In the field of general imaging and computer vision, deep learning is the leading machine learning tool.[5] Deep learning refers to techniques that build on developments in artificial neural networks in which multiple network layers are added to increase the levels of abstraction and performance.[6]

A simplified overview of the mechanics behind deep convolutional neural networks (CNNs) is provided in Fig 1. Neural networks aim to mimic the structure of the human brain. They use a series of interconnected "neurons", which collect, summarise, and/or transform data before passing the values to the next neuron in the sequence. This process culminates in an output layer, which can be used to formulate a prediction.

Constructing and training an effective neural network from scratch requires huge amounts of data. State-of–the-art image classification networks are frequently trained on data sets containing millions of images, facilitated by multiple computer servers, running continuously for several weeks.[9] This is not feasible for the majority of medical researchers. One method for overcoming this is to use a process called transfer learning. This is the process of adopting powerful and highly refined features from large existing pre-trained CNNs and using these as a starting point in training a new model for a different task. These low-level features, such as those illustrated in Fig 2, can be thought of as the basic building blocks of images such as lines and curves, and have been shown to be applicable to many different image-recognition tasks.[10] This technique can vastly reduce the computational requirements needed for network training and can deliver substantial performance benefits compared with training a CNN from scratch.[11]

Transfer learning has been relatively underutilised in the clinical setting despite the availability of vast amounts of image data. Two recent publications have demonstrated that transfer learning from pre-trained CNNs can produce human-level diagnostic results in the categorisation of skin lesions and in defining disease on digital retinal images.[12,13] Similar levels of diagnostic accuracy have yet to be reported in the analysis of plain radiographs; however, a few promising studies have recently been published. For example, one study retrained the GoogLeNet CNN for the detection of pathology in plain frontal chest radiographs, resulting in an area under the curve of between 0.861 and 0.964 for different chest radiograph features.[14] A different study used transfer learning from a pre-trained ImageNet CNN for the automated categorisation of osteoarthritis in knee radiographs.[15]

Plain radiographs are the most common radiological test with over 22 million studies performed in the UK in 2016.[3] A high proportion of these are plain extremity radiographs in the context of trauma. There is a strong case for developing automated strategies to improve efficiency and workflow management in this area considering the backlog in unreported studies and the fact that over £88 million was spent on outsourcing radiology reports in 2014–2015.[1] It is surprising therefore that, to the author's knowledge, there are currently no studies in the literature that have successfully applied transfer learning from pre-trained CNNs to the problem of fracture detection on plain radiographs. This proof of concept study aims to establish to what extent this is possible.

Materials and methods

This National Health Service (NHS)-based study was granted approval by the Health Research Authority in England. This study was retrospective and used only anonymised data, and therefore, ethics approval was not required.

Imaging study selection

Anonymised lateral wrist radiographs were obtained from the Royal Devon & Exeter Hospital for studies performed between January 2015 and January 2016. Radiographs were excluded if there was a plaster cast in place, if the growth plates of the wrist had not yet fused, or if the study demonstrated any fracture other than a fracture of the distal radius or ulna. The images were classified as either "fracture" or "no fracture" on the basis of the radiological report. This classification was checked and verified by a radiology registrar competent in the reporting of plain radiographs and with 3 years radiology experience. Images were also excluded if the single lateral projection was inconclusive for the presence or absence of fracture.

Image pre-processing

This resulted in a preliminary data set of 695 wrist radiographs demonstrating a fracture and 694 wrist radiographs demonstrating no fracture. The images were converted to JPEG by a trained radiologist ensuring the most appropriate windowing was selected. Images were then pre-processed by removing additional annotations such as the commonly used "Red Spot" annotation applied by radiographers at the time of image acquisition in order to reduce the possibility of over-fitting. Over-fitting describes the process by which the learning algorithm learns features that are not truly representative of the image category in an attempt to reduce the error in the learning process.

A data augmentation technique was used to amplify the data. This involved making a number of non-exact copies, or transformations, of each image. This served to provide the CNN with more training examples by incorporating the salient features in multiple orientations. The aim was to better reflect the real-world population of wrist radiographs,

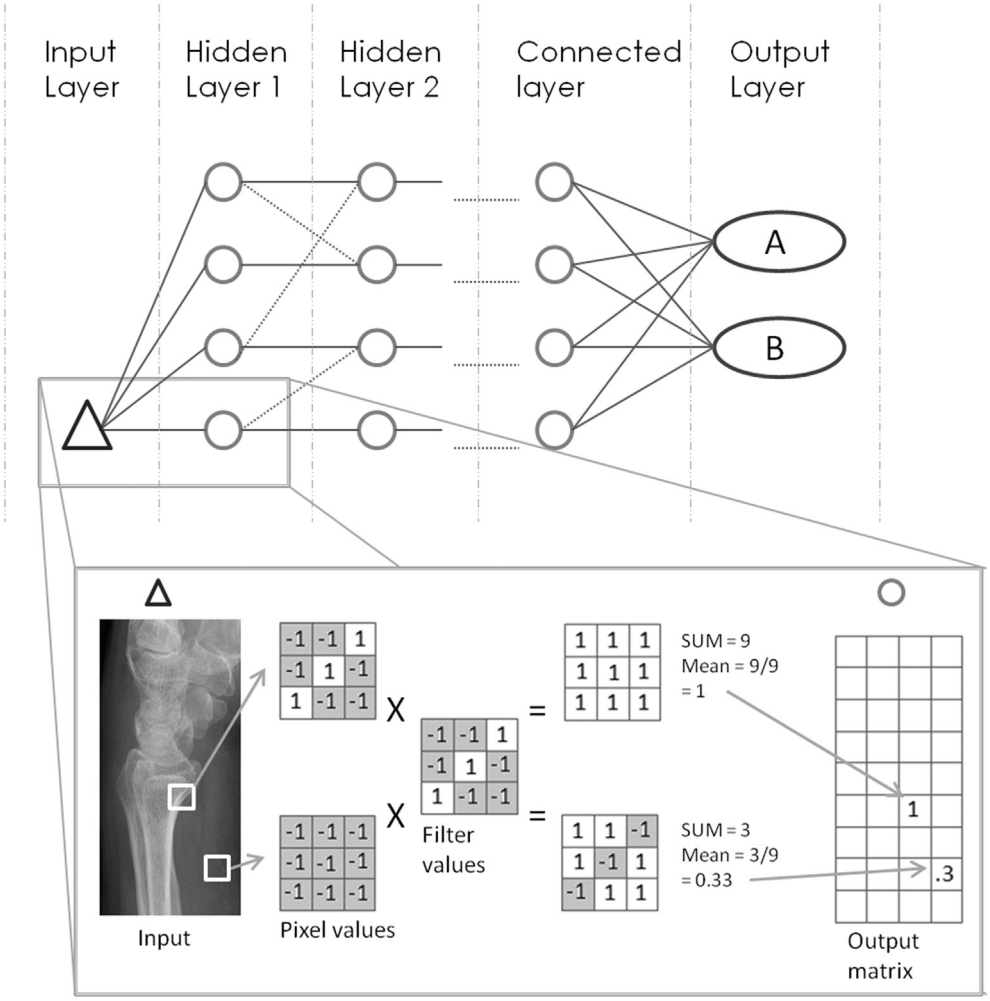

Figure 1 Basic mechanics underlying CNNs. An abstract CNN architecture is illustrated (top). A simplified example is shown (expanded box) demonstrating how a filter is used in the identification of an image feature, in this case a specifically orientated diagonal line. The input image is first converted to a matrix of pixel values which can then be sampled using multiple filters. These filters are developed by the training process using a technique called back propagation, which uses gradient descent algorithms to minimise error in the network.[7] The diagonal line filter is an example of a typical first layer filter derived from training a basic CNN. These low-level filters frequently represent basic image features such as lines and textures. Higher-level filters represent more complex features, such as complex shapes or anatomical structures. The white boxes overlying the "Input" image represent a snapshot of the convolution process in which the filters roll, or convolve, across the input matrix, sampling every possible position in turn. Expanded from the white boxes are representations of the actual matrix values being sampled by the filter window, or receptive field, of the filter. The example filter is shown to the right of this, which happens to be a detector for diagonal lines orientated from bottom left to top right. Multiplying the filter with the pixel values and then taking the average produces a score which forms a single value in the output matrix (beneath the circle in the figure). The higher the value the more likely there was to be a similarly orientated diagonal line at that position. This calculation must be performed for every possible receptive field in the input image. The process is repeated in subsequent layers with different filters convolving across the output matrix from the previous layer. The result of this is the identification of features of increasing complexity as a result of combining low level features.[8] Additional intervening steps are incorporated for simplifying the distribution of values in order to improve computational efficiency and interpretability, often referred to as pooling layers.[7]

so that factors, such as handedness, limb size, and slight variations in wrist positioning, would be better accounted for in the model. All images were first horizontally flipped through 180° to double the sample size. A random transformation was then applied consisting of rotation of between 0 and 25°, width and height shift by a factor of 0–15%, shearing of between 0 and 10% and zoom of between 0 and 15%. This resulted in an overall amplification by a factor of 8. Transformations were computationally applied using the Keras library for Python (version 3.5). This resulted in a final sample size of 5,560 in the fracture group and 5,552 in the no fracture group.

CNN training

Models were developed using Tensorflow 1.0, an open-source software library for machine intelligence. The inception v3 network,[16] trained for the IMageNet Large Visual Recognition Challenge,[17] and based on analysis of non-radiological images, was computationally adapted to

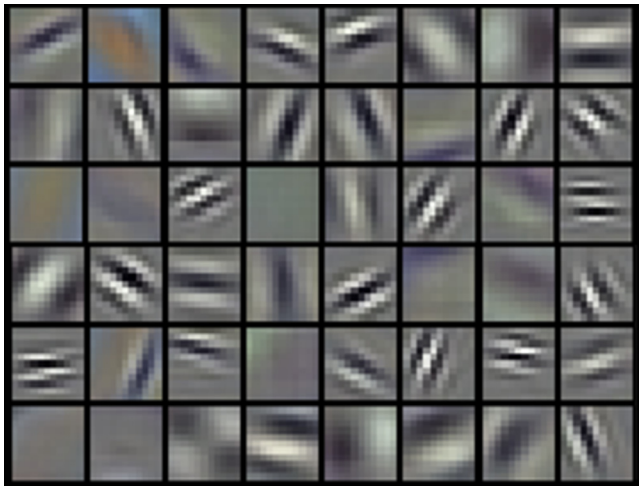

Figure 2 Visualisation of first layer filters from a trained version of the "AlexNet" CNN.[24] Permission for free use of this image has been granted by the authors and image files are freely available and labelled for re-use at: https://www.cs.toronto.edu/~kriz/.

produce models for fracture recognition. This was achieved by using the "fracture" and "no fracture" training data sets to re-train the top layer of the inception v3 network.

Hyper-parameters, such as learning rate and cycle number, were optimised iteratively. Learning rate refers to the magnitude of changes made to the model parameters after each cycle, or iteration. Learning rate decay was also implemented, which reduces the magnitude of parameter changes as the training progresses. This decreases training time early in the process whilst maintaining accuracy gains from smaller refinements later on. The final model was trained using an initial learning rate of 0.02 and learning rate decay by factor 0.67 every 1,800 iterations. Code was written and adapted in the Python (version 3.5) programming language.

The training data were split into training, validation, and final testing sets with an 80:10:10 split ratio, respectively. This ratio offered a balance between maximising the amount of training data and minimising the variance in performance testing. This is a similar split ratio to that used elsewhere.[9] An additional 100 images were kept completely separate for final testing and statistical analysis. The network was trained using 20,000 iterations. The final test accuracy, representing how well the network performed on 10% of the withheld training data, was used as a guide for iterative hyper-parameter optimisation.

Testing the model and statistical methods

Final analysis of the fracture detection model was performed using a new, previously unused data set of 100 wrist radiographs, with a 1:1 split ratio between "fracture" and "no fracture" classes. These were obtained by selecting studies meeting the inclusion criteria from a consecutive set of plain radiograph studies until the quota of 50 fracture and 50 non-fracture radiographs was reached. These testing images had not been used in the machine learning process and were used only once. The exclusion criteria were the same as those applied to the training data. Each study was reviewed by a radiology registrar competent in the assessment of plain radiographs.

Each image was analysed using the final fracture detection model resulting in a score representing the likelihood that the image should be classified as "fracture" or "no fracture". This score was a continuous value of between 0 and 1. The area under the receiver operator characteristic (ROC) curve (AUC) was calculated using the ROC Analysis tool, a web-based Calculator for ROC curves.[18] Sensitivity and specificities were calculated using the optimum cut-off threshold as derived from the AUC analysis.

Results

Training progress against iteration number is illustrated in Fig 3. The ROC curve for the test output is depicted in Fig 4. The AUC was 0.954. Setting the diagnostic cut-off at a threshold designed to maximise sensitivity and specificity resulted in values of 0.9 and 0.88, respectively. Fig. 5 illustrates the distribution of test scores with respect to the true classification.

Discussion

An AUC of 0.954 for this model demonstrates that transfer learning from deep CNN, pre-trained on non-medical images, can be successfully applied to the problem of fracture detection on plain radiographs. This level of accuracy surpasses previous computational methods for automated fracture analysis based on methods, such as segmentation, edge detection, and feature extraction. Such studies reported sensitivities and specificities in the range of 80–85%.[19,20] Similarly, a 2016 study using shape and texture features (not based on transfer learning), reported an AUC accuracy of 0.886 for the classification of fractures on anterior–posterior (AP) wrist radiographs.[21] Studies combining these different techniques and studies analysing both AP and lateral projections would likely yield increased accuracy and may prove to be a fertile area of future study.

This study demonstrates that transfer learning from deep CNN pre-trained on non-medical images can readily be applied to the analysis of plain radiographs. Furthermore this can be achieved even with modest sample sizes as demonstrated here. Whilst a computer-aided diagnostic test with AUC accuracy of 0.954 may not currently boast the accuracy needed to replace human interpretation, this level of accuracy could be very useful in applications, such as workflow prioritisation and minimisation of error. For example, a similar system could automatically assess the numerous unreported studies languishing in radiology silos for the presence of pathology. If positive, these studies could then be flagged up to the radiology team, thus expediting the identification of potentially serious diagnoses and improving the productivity of the radiologist. The small fraction of incorrectly flagged studies would be a minor issue as all studies would continue to be assessed by a

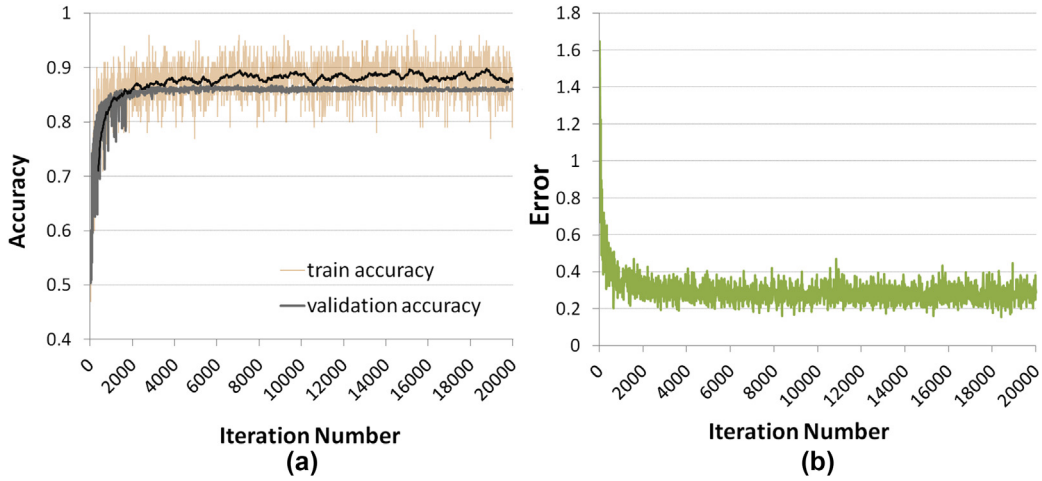

Figure 3 (a) Training accuracy (light orange erratic line) and validation accuracy (smooth grey line) with respect to iteration number. A 40-point average trend line is included for training accuracy values (black). (b) Network error with respect to iteration number.

trained radiologist. In a second example, pathologies such as scaphoid fractures can have severe long-term consequences if not identified. A system that identifies scaphoid fractures automatically with a sensitivity and specificity of around 90% would flag up the majority of these cases for second review. This would have the potential to vastly reduce the chance of patient harm due to reporting error. This is especially true for visualisation errors where the feature is simply missed rather than interpreted incorrectly.

Limitations and future directions

Although this study provides proof of concept, a number of limitations remain. The reference standard, or ground proof, for the training and testing images was based on the assessment of a human radiologist. This meant that, in this example, the model could never outperform the human. It could be predicted that future studies using extremely large data sets pooled from multiple centres could result in vast training sets too large for assimilation by a single human assessor. This is less relevant for the fairly basic example of wrist fracture detection, but could prove invaluable in the assessment of conditions with more complex imaging appearances. Another way in which the performance of automated systems could rival that of humans would be to use higher-level reference standards, such as biopsy results, or in the fracture detection example, cross-sectional data. In this way, leveraged by a superior knowledge of the ground truth, machine learning algorithms have the potential to identify additional abstract features that may not have been apparent to the human assessor.

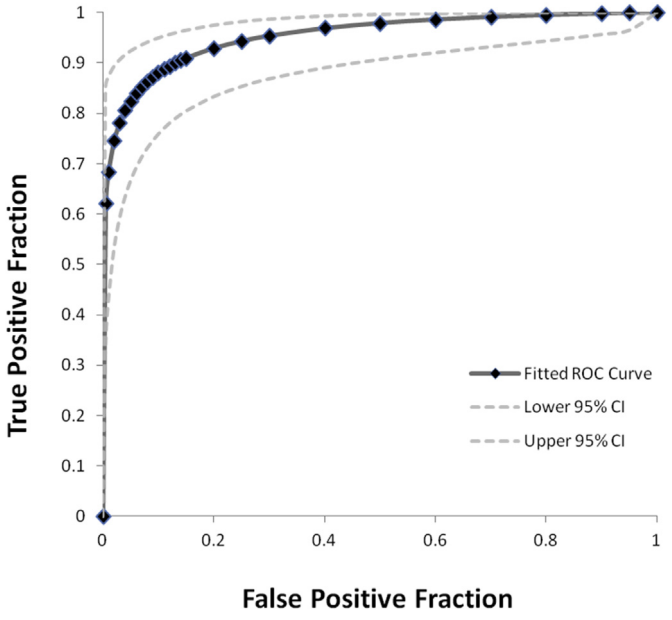

Figure 4 ROC curve. Fitted values are displayed (dark grey with diamond data points) with 95% confidence intervals (dotted light grey).

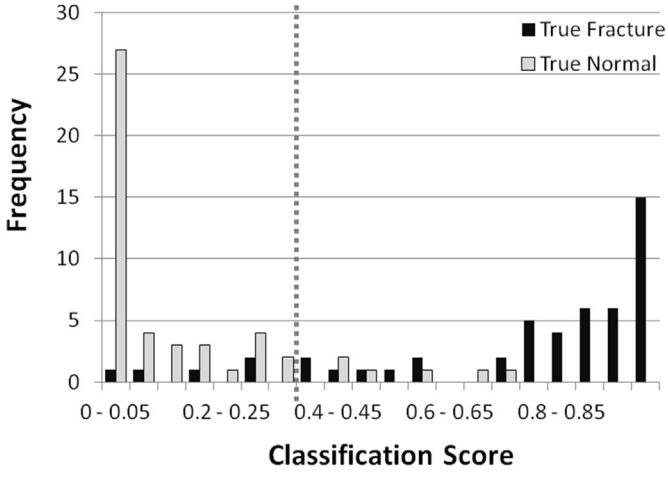

Figure 5 Histogram illustrating the distribution of classification scores separately for the true fracture images (black) and the true normal images (grey). A higher score indicates a higher suggestion of a fracture. The dotted grey line represents the optimal cut-off for maximisation of both specificity and sensitivity.

The need to remove studies from the training data that were inconclusive to the radiologist highlights some of the issues automated systems have with regards to dealing with uncertainty. Clearly incorporating an image with an inaccurate label would not be conducive to the training of an automated system. One method that could be explored in future studies would be to include a third diagnostic category of "inconclusive". Having been given multiple examples of inconclusive studies the machine learning process could extract features that are overrepresented in this subgroup, and therefore, identify studies that are unlikely to be classified accurately. This could then prompt second review by a human assessor or perhaps suggest further investigation, but this area requires more research. Incorporating Bayesian inferences to produce more meaningful probability predictions is another area of potential that needs further study. In this strategy, the baseline known incidence of fracture occurrence in the population could be combined with the test accuracy metrics of the model to produce an overall probability score more accurately reflecting the confidence of the output prediction.

As illustrated in Fig 3, a small discrepancy was seen between the training accuracy and the validation accuracy at the end of the training process. This was likely to reflect over-fitting. There are several strategies that could be used to minimise over-fitting in future studies. One strategy would be to use automated segmentation of the most appropriate region of interest, which in this study would be the distal radius and ulna. The pixels outside of the region of interest would be cropped from the image so that irrelevant features could not influence the training process. Another option would be to explore alternative network architectures, for example optimising drop-out of random neurons, a strategy that has been shown to be effective in reducing over-fitting.[22]

Sample size is often the limiting factor in machine learning studies. A larger sample corresponds to a more accurate reflection of the true population. Wrist fractures often demonstrate fairly similar appearances, such as those described eponymously as Colles and Smith fractures (Fig 6). It is yet to be seen whether this technique produces similar levels of accuracy in the analysis of more varied pathology, which may require a correspondingly larger sample size. Amplifying small data sets with data augmentation is one method of addressing sample size issues and has been shown to reduce over-fitting and improve performance.[23]

An often quoted limitation in the use of CNNs is the inability to confidently dissect the mechanisms through which the conclusions are drawn. As a result, such techniques are occasionally referred to as "black-box" methods. As data passes through layers of the network the levels of abstraction increase and these become increasingly difficult to comprehend. On a superficial level this may be unsettling; however, it is argued here that as long as robust testing methods are applied that prove that acceptable levels of safety and efficacy are met, there is no reason to reject so-called black-box methods. Although every aspect of the inner workings of a network may be illusive, there are various techniques that do shed some light; for example, as illustrated in Fig 2, visualisation of CNN filters is being increasingly used. Visualisation allows researchers to better understand and improve models, but also provides a fascinating insight into the way the model "sees" the image.[24–26] Filter visualisation may be a valuable area of further research for the current study. Nevertheless, as computing power and algorithm sophistication continues to improve, these techniques will have the capacity to process information to degrees of complexity that far exceed the comprehension of the human mind. It is within this realm that the major technological advances will be seen.

Figure 6 Lateral wrist radiographs demonstrating typical examples of the eponymously named Colles (a) and Smith (b) fractures. Both are transverse fractures of the distal radial metaphyses but with dorsal angulation in the Colles type and palmer angulation in the Smith type.

In conclusion, this study provides proof of concept for the use of transfer learning from deep CNNs pre-trained on real world images for the analysis of plain medical radiographs. The results were comparable to state-of-the-art for automated fracture detection having trained the model using only a modest sample size. The methodology used here is extremely transferable. The possible applications for this kind of technique are many and varied given the enormity of the medical image data set. If properly utilised this has the potential to significantly improve workflow productivity, minimise the risk of error, and prevent patient harm by reducing diagnostic delays.

Acknowledgements

The authors would like to thank Dr Mark Thurston and Mr David Batey for programming and technical advice.

This research was sponsored by The Royal Devon and Exeter NHS Foundation Trust (RD&E) and received no external funding.

References

1. Faculty of Clinical Radiology. *Clinical radiology UK workforce census 2015 report*. 2016. Available at: www.rcr.ac.uk. [Accessed 25 May 2017].
2. Cliffe H, Liu D, Wykes V, *et al*. *Summary of the Royal College of Radiologists reporting backlog survey and assessment of potential causes and solutions*. 2016. Available at: https://www.rcr.ac.uk/sites/default/files/reporting_backlog_surveys_potential_causes_solutions.pdf. [Accessed 25 May 2017].
3. NHS England. *Diagnostic imaging dataset statistical release v1.0*. May 2017. Available at: https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/. [Accessed 25 May 2017].
4. Kohli M, Prevedello LM, Filice RW, *et al*. Implementing machine learning in radiology practice and research. *AJR Am J Roentgenol* 2017;**208**:754–60.
5. Greenspan H, Ginneken B, Summers RM. Deep learning in medical imaging: overview and future promise of an exciting new technique. *IEEE Trans Med Imaging* 2016;**35**:1153–9.
6. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–44.
7. Stanford University. CS231n: Convolutional neural networks for visual recognition. Available at: http://cs231n.stanford.edu/. Accessed 25/5.17.
8. Nielsen MA. *Neural networks and deep learning*. Determination Press; 2015. Available at: http://neuralnetworksanddeeplearning.com/chap6.html. [Accessed 25 May 2017].
9. Szegedy C, Liu W, Jia Y, *et al*. *Going deeper with convolutions*. arXiv; 2015. arXiv:1409.4842.
10. Yosinski J, Clune J, Bengio Y, *et al*. How transferable are features in deep neural networks? *Adv Neural Inf Process Syst* 2014;**27**:3320–8.
11. Razavian AS, Azizpour H, Sullivan J, *et al*. CNN Features off-the-shelf: an astounding baseline for recognition. In: *CVPRW'14. Proceedings of the 2014 IEEE conference on computer vision and pattern recognition workshops*; 2014. p. 512–9. arXiv;arXiv:1403.6382.
12. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;**542**:115–8.
13. Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;**316**:2402–10.
14. Cicero M, Bilbily A, Colak E, *et al*. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Investig Radiol* 2017;**52**:281–7.
15. Antony J, McGuinness K, Connor NEO, *et al*. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In: *ICPR 2016 proceedings. arXiv*; 2016. arXiv:1609.02469.
16. Szegedy C, Vanhoucke V, Loffe S, *et al*. *Rethinking the inception architecture for computer vision*. arXiv; 2015. arXiv:1512.00567.
17. Russakovsky O, Deng J, Su H, *et al*. Imagenet large scale visual recognition challenge. *Int J Comput Vis* 2015;**115**:211–52.
18. Eng J. ROC analysis: web-based calculator for ROC curves. Baltimore: Johns Hopkins University. Updated 2014 March 19; Accessed 25/05/17. Available from: http://www.jrocfit.org.
19. Anu TC, Mallikarjunaswamy MS, Raman R. Detection of bone fracture using image processing methods. In: *IJCA proceedings on national conference on power systems and industrial automation NCPSIA*, vol. 3; 2015. p. 6–9.
20. Cao Y, Wang H, Moradi M, *et al*. Fracture detection in X-ray images through stacked random forests feature fusion. In: *12th international symposium on biomedical imaging (ISBI)*, vol. 84. Piscataway, NJ: IEEE; 2015. p. 801–5.
21. Ebsim R, Naqvi J, Cootes T. Detection of wrist fractures in X-ray images. In: Shekhar R, Wesarg S, Ballester M, *et al*., editors. *Clinical image-based procedures. Translational research in medical imaging. CLIP*, vol. 9958. Cham: Springer; 2016. p. 1–8. Lecture Notes in Computer Science.
22. Hinton G, Krizhevsky A, Sutskever I, *et al*. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929–58.
23. Wong SC, Gatt A, Stamatescu V, *et al*. Understanding data augmentation for classification: when to warp? In: *International conference on digital image computing: techniques and applications. arXiv*; 2016. arXiv:1609.08764.
24. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *NIPS'12 proceedings of the 25th international conference on neural information processing systems, Lake Tahoe, Nevada, Dec 03–06*; 2012. p. 1097–105.
25. Zeiler MD, Fergus R. *Visualizing and understanding convolutional networks*. arXiv. 2013. arXiv:1311.2901.
26. Simonyan K, Vedaldi A, Zisserman A. *Deep inside convolutional networks: visualising image classification models and saliency maps*. arXiv. 2013. arXiv:1312.6034.