MEDICAL IMAGING—ORIGINAL ARTICLE

Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures

Matthew Adams,[1]* Weijia Chen,[2] David Holcdorf,[1] Mark W McCusker,[1,3] Piers DL Howe[2] and Frank Gaillard[1,3]

[1] Radiology Department, Royal Melbourne Hospital, Melbourne, Victoria, Australia
[2] School of Psychological Sciences, University of Melbourne, Melbourne, Victoria, Australia
[3] Radiology Department, University of Melbourne, Melbourne, Victoria, Australia

M Adams MBBS (Hons), BE (Hons); **W Chen** BSc (Hon), LLB, BCom; **D Holcdorf** BVSc (Hons), MBBS; **MW McCusker** MB, BCh, MRCPI, FFR(RCSI), FRANZCR; **PDL Howe** PhD, MPhys; **F Gaillard** MBBS (Hons), M.Med, FRANZCR.

Correspondence

Dr Matthew Adams, Department of Radiology, Waikato Hospital, Pembroke Street, Hamilton West, Hamilton, New Zealand.
Email: matthew.adams@waikatodhb.health.nz

Conflict of interest: Associate Professor Frank Gaillard is the founder & CEO of Radiopaedia.org. None of the other authors have any relevant relationships to disclose.

*Since undertaking this research at The Royal Melbourne Hospital, Dr. Matthew Adams has moved to New Zealand as a radiology registrar at Waikato Hospital.

Submitted 19 July 2018; accepted 7 October 2018.

doi:10.1111/1754-9485.12828

Abstract

Introduction: To evaluate the accuracy of deep convolutional neural networks (DCNNs) for detecting neck of femur (NoF) fractures on radiographs, in comparison with perceptual training in medically-naïve individuals.
Methods: This study extends a previous study that conducted perceptual training in medically-naïve individuals for the detection of NoF fractures on a variety of dataset sizes. The same anteroposterior hip radiograph dataset was used to train two DCNNs (AlexNet and GoogLeNet) to detect NoF fractures. For direct comparison with perceptual training results, deep learning was completed across a variety of dataset sizes (200, 320 and 640 images) with images split into training (80%) and validation (20%). An additional 160 images were used as the final test set. Multiple pre-processing and augmentation techniques were utilised.
Results: AlexNet and GoogLeNet DCNNs NoF fracture detection accuracy increased with larger training dataset sizes and mildly with augmentation. Accuracy increased from 81.9% and 88.1% to 89.4% and 94.4% for AlexNet and GoogLeNet respectively. Similarly, the test accuracy for the perceptual training in top-performing medically-naïve individuals increased from 87.6% to 90.5% when trained on 640 images compared with 200 images.
Conclusions: Single detection tasks in radiology are commonly used in DCNN research with their results often used to make broader claims about machine learning being able to perform as well as subspecialty radiologists. This study suggests that as impressive as recognising fractures is for a DCNN, similar learning can be achieved by top-performing medically-naïve humans with less than 1 hour of perceptual training.

Key words: femoral neck fractures; learning; radiology; supervised machine learning; X-rays.

Introduction

Artificial intelligence has gained great momentum in recent years with studies showing its ability to perform complex interpretation at the level of healthcare specialists.[1–6] These studies have fuelled hope and concern that AI systems will replace radiologists in the near future. Little study, however, has focused on what degree of human training this actually represents, with the assumption being that only a highly trained radiologist with years of experience can perform at these levels. We sought to explore how machine learning compares to the training of novice individuals on the same number of images with perceptual training. A novel, and perhaps counter-current, application of this would be enabling the medical sector to take some of the resources used for machine learning to improve the quality of doctors.

Much of AI success in radiology has been through machine learning and the branch of Deep Convolutional Neural Networks (DCNNs) for interpreting images. In computer science, DCNNs have become state of the art for interpreting images and are the model of choice for the annual ImageNet Large Scale Visual Recognition Competition.[7] AI has already shown potential across healthcare[1] with examples including dermatology for skin lesion identification,[2] ophthalmology for the detection of diabetic retinopathy,[3] orthopaedics for hand and ankle fractures[4] and in radiology for interpreting chest X-rays for tuberculosis[5] and interpreting CT and MRI for detecting strokes[8] to name a few.

Our study was a continuation of one in which we performed perceptual training on medically-naïve undergraduate students to detect neck of femur (NoF) fractures.[9] Perceptual training is the method of improving one's perceptual skills to allow for an enhanced capacity to identify or categorise images without being provided explicit rules.[10,11] An example of a perceptual skill is gender identification based on facial appearance.[12,13] Although ordinarily easy to perform, it can be difficult to verbalise and explicitly teach. The prior study found that top-performing students, with less than 1 hour of training, had comparable accuracy to board-certified radiologists in the detection of NoF fractures.

From a learning perspective, perceptual training and supervised machine learning operate in a very similar manner.[14] The subject, human or DCNN, is presented with a large number of training images, one-by-one, attempts to classify each image according to a pre-set criterion (e.g. whether it contains a NoF fracture) and is informed whether they were correct. The subject then learns, based on this feedback. Learning can be increased by increasing the number of training images.[14]

The purpose of the study is to evaluate the accuracy of DCNNs for detecting NoF fractures on radiographs, in comparison with perceptual training in medically-naïve individuals when trained on the same set of radiographic images.

Methods

Prior perceptual training study

The study followed on from one by Chen *et al.*[9] that conducted perceptual training in 142 medically-naïve undergraduate students for the detection of neck of femur fractures on a variety of dataset sizes. The students had no prior knowledge of X-ray interpretation. The study found that top-performing medically-naïve students could detect NoF fractures as accurately as radiology residents and board-certified radiology consultants with 48 minutes and 52 minutes of training respectively.

Dataset

The same dataset that was originally produced for the perceptual training study was used for this study. The images were from The Royal Melbourne Hospital radiographic archive for emergency presentations where patients had surgically confirmed NoF fractures. All presentation AP pelvic X-rays were de-identified and digitised using FujiFilm Synapse PACS v4.5 software. As seen in Figure 1, two non-overlapping hip region images (for left and right) were cropped from the pelvic radiograph allowing one image to have a fracture and the other without. Radiographs were excluded if there was an underlying pathology (excluding osteoporosis or osteoarthritis) or metalware in either the fractured or unfractured hip. In total there were 805 images in the dataset, 403 with a fracture and 402 without.

Deep learning methods

Processing environment

The image dataset was handled in 8-bit Portable Network Graphics format. The processing of the images was performed on a Windows 10 (Microsoft Corportation, Redmont, WA, USA) operating system using MATLAB R2018a (The MathWorks Inc, Natick, MA, USA) with its Neural Network Toolbox and Parallel Processing Toolbox. To adequately handle processing of the DCNN, the computer had a GeForce GTX 1070 graphics processor (Nvidia Corporation, Santa Clara, CA, USA).

Dataset sizes

To allow direct comparison with the perceptual training study by Chen *et al.*,[9] deep learning sample sizes were 200, 320 and 640 images, with images split into training (80%) and validation (20%). A separate group of 160 images were used as the final test set. Training, validation and test sets were randomised using inbuilt randomisation functions in MATLAB with equal proportions of fractured and non-fractured images in each.

Pre-processing

All dataset images underwent pre-processing to standardise them to be of equivalent size and proportions. All images of left-sided hips were mirrored to be right-sided. The images were also manually cropped to roughly standardise the size and location of the femoral head (see Fig. 1).

Image augmentation

To reduce the potential of over fitting of the trained network to the training data, MATLAB's inbuilt image database augmentation algorithms were used. Training images were randomly mirrored (horizontally and/or vertically) and randomly rotated up to 15 degrees bidirectionally.

DCNN architecture

Two pre-trained DCNN architectures, with prior training using the ImageNet database,[7] were used for the study.

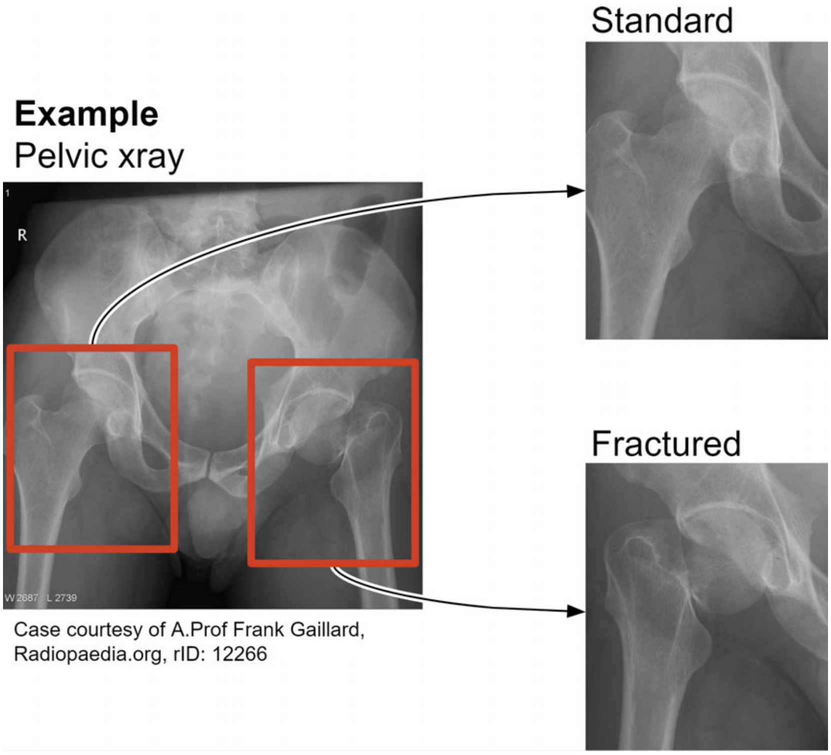
Example
Pelvic xray
R
W 2687 : L 2739
Case courtesy of A.Prof Frank Gaillard,
Radiopaedia.org, rID: 12266
Standard
Fractured

Fig. 1. Original AP pelvic radiographs were selected from the Royal Melbourne Hospital PACS archive for isolated neck of femur fractures. As shown in the figure, both hip regions were manually cropped from the pelvic radiograph to create two separate hip radiographs. Hips were labeled as either fractured or standard (non-fractured) and flipped as required to mimic right-sided orientation.

AlexNet[15] and GoogLeNet[16] are two open-source frameworks that have been developed and used in a variety of research scenarios, including the detection of tuberculosis changes in chest X-rays.[3] The final three layers were replaced to allow the pre-trained networks to be applied to detect NoF fractures.

Solver parameters

Hyperparameters were optimised using a grid search approach. The following solver parameters were selected: 50 Epochs; stochastic gradient descent; initial learning rate of 0.001 with 20% drop every 10 epochs. Each pre-trained DCNN was trained 5 times for each training size with the optimal model chosen based on highest validation accuracy obtained.

Statistical methods

All statistical analysis was performed using MedCalc (MedCalc v. 18.5, MedCalc Software, Ostend, Belgium). Comparison of accuracy based on sample size, augmentation or DCNN used 'N-1' Chi-squared test.[17] The receiver operating characteristic curves and AUC were determined for the test dataset for each DCNN.[18] Ninety five percent confidence intervals and comparisons of AUCs for receiver operating characteristic curves, were calculated using a nonparametric approach.[19]

Results

Deep learning results

Across all training dataset sizes, the GoogLeNet DCNN outperformed the AlexNet DCNN (overall accuracy 90.6% vs 85.3%, respectively, $P < 0.01$). Larger training sample sizes improved accuracy with 640 images in the training set outperforming 200 images (90.9% vs 85.5%, respectively, $P < 0.01$). However, the image augmentation that was implemented did not provide any significant difference (accuracy with augmentation vs accuracy without augmentation, 89.1% vs 86.9%, respectively, $P = 0.14$).

A summary of the Area Under Curve (AUC) results for the two DCNNs is shown in Table 1. Figure 2 shows a comparison of Receiver Operating Characteristic (ROC) curves for AlexNet and GoogLeNet DCNNs.

Deep learning benchmarking

Figure 3 compares AlexNet and GoogLeNet results with those previously obtained from perceptual training and

Table 1. Area Under Curve test dataset

	Without augmentation			With augmentation†		
	Dataset size‡			Dataset size‡		
	200	320	640	200	320	640
AlexNet	0.91 (0.86, 0.95)	0.92 (0.87, 0.96)	0.95 (0.91, 0.98)	0.89 (0.84, 0.94)	0.91 (0.86, 0.95)	0.94 (0.89, 0.97)
GoogLeNet	0.93 (0.88, 0.96)	0.96 (0.92, 0.99)	0.98 (0.94, 1.00)	0.94 (0.90, 0.97)	0.96 (0.92, 0.99)	0.98 (0.94, 0.99)

Data in parentheses are the 95% confidence interval.
†Additional augmentation of random mirroring of images (horizontally and/or vertically) and random rotation up to 15 degrees bidirectionally.
‡Includes the combined training and validation dataset size with a ratio of 80:20, respectively.

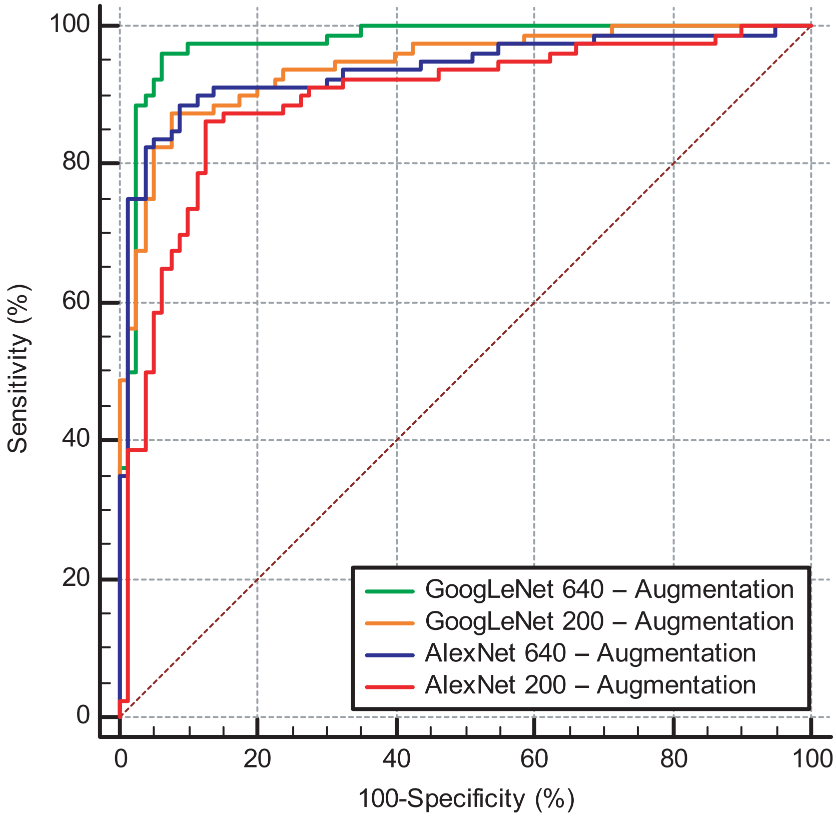

Fig. 2. Comparison of Receiver Operating Characteristic (ROC) Curves for AlexNet and GoogLeNet models with a combined training and validation dataset of 200 and 640 images with augmentation. There was no statistically significant difference between AlexNet with 640 images and GoogLeNet with 200 images ($P = 0.66$). Paired comparisons between all the other ROC curves in this graph were statistically significant ($P < 0.01$).

radiology specialists and trainees. The test accuracy for the perceptual training in top-performing medically-naïve individuals increased from 87.6% to 90.5% when trained on 640 images compared with 200 images. When these top-performing individuals were allowed to train on the 640-image dataset twice (1280 images in total), they achieved an average accuracy of 94.5% and exceeded the average accuracy achieved by board-certified radiologists (93.5%) and radiology residents (92.9%).

Discussion

The results obtained from the study in regard to deep learning when comparing the pre-trained AlexNet and GoogLeNet models were, by and large, as predicted. Accuracy improved through increasing the sample size, augmentation played a minor role in improving the models and GoogLeNet outperformed AlexNet. Impressively, the pre-trained DCNNs were able to detect neck of femur fractures with similar levels as radiology residents and

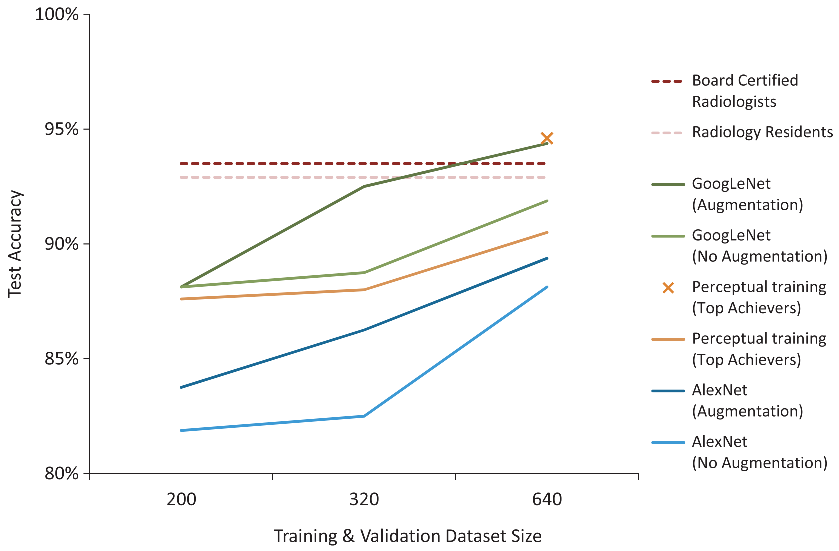

Fig. 3. Comparison of NoF fracture detection accuracy between deep learning and perceptual training when trained on the same dataset sizes. The perceptual training results were based on top achievers with no prior knowledge of radiography. Board-Certified Radiologists and Radiology Residents were used as the performance benchmark as they are the current gold standard for assessing hip radiographs (they did not undergo any perceptual training prior to assessment). Perceptual training results increased from an average of 90.5% to 94.6% when individuals were allowed to train on 640 images twice.

board-certified radiologists, with the top-performing GoogLeNet model outperforming them both.

Although this is certainly an impressive result, especially considering that this was achieved with a relatively small sample size and using generic network architectures, it is important to frame the achievement in terms of how much human training this represents. In this case, a more accurate description is that these DCNNs performed as well as some medically-naïve humans trained for less than an hour.

We believe that it is important to reframe the success of AI achievement in this way, not to devalue the impressive strides forward that AI has accomplished of late, but to put it in the context of normal human learning.

Deep learning's full capability at image detection and classification are almost certainly underrepresented in this study. Firstly, the image dataset is magnitudes smaller than what would be usual when training deep learning models. Ideally testing and validation datasets should be at least in the thousands, if not tens of thousands of images. This would have been unrealistic for our study. To allow for direct comparison to perceptual learning, we would have needed to train humans on the same number of images. To put this in perspective, if each image took 2.77 seconds (the average time per image taken during perceptual training in the current study), it would take 7 hours and 42 minutes without breaks to train each individual with 10,000 images. The increased image dataset would also rely on there being sufficient radiographs in our hospital to create the training set. The DCNN models chosen in the study, were selected as they are relatively easily implemented, have been well researched and have been applied in medical imaging studies prior.

This study highlights the importance of pathology detection in radiology, but that it is only one, albeit important, aspect to the interpretation of medical imaging. The ability of medically-naïve individuals, through perceptual training, and computers, through deep learning, to detect fractures at an accuracy matching or exceeding that of radiology residents and board-certified radiologists only highlights that more can be done to improve specialists' ability to detect pathology. Radiology needs to mitigate the well documented high levels of systemic radiological error (between 2–20%)[20] and embrace utilising tools like machine learning and concentrated perceptual training sessions to improve radiologists' diagnostic accuracy.

In conclusion, single detection tasks in radiology are commonly used in DCNN research with their results often used to make broader claims about machine learning being able to perform, as well as, subspecialty radiologists. This study suggests that as impressive as recognising fractures is for a DCNN, similar learning can be achieved by top-performing medically-naïve humans with less than 1 hour of perceptual training.

Acknowledgements

This study was approved by The University of Melbourne Human Research Ethics Committee (Ethics ID 1545113) and all experiments were performed in accordance with the ethical standards laid down in the amended Declaration of Helsinki.[21]

The funding for the initial study[9] was from a Royal Australian and New Zealand College of Radiology (RANZCR) research grant in 2015 to Frank Gaillard and Piers D. L. Howe (no grant number available, URL https://www.ranzcr.com/college/awards-and-prizes/research-awards-and-grants), and the University of Melbourne Engagement Grants Scheme 2015 to Frank Gaillard and Piers D. L. Howe (no available grant number or URL). There was no involvement by the funders as to the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Lee J-G, Jun S, Cho Y-W *et al.* Deep learning in medical imaging: general overview. *Korean J Radiol* 2017; **18**: 570–84.
2. Esteva A, Kuprel B, Novoa RA *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542**: 115–8.
3. 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.
4. Olczak J, Fahlberg N, Maki A *et al.* Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop* 2017; **88**: 581–6.
5. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017; **284**: 574–82.
6. Tang A, Tam R, Cadrin-Chenevert A *et al.* Canadian association of radiologists white paper on artificial intelligence in radiology. *Can Assoc Radiol J* 2018; **69**: 120–35.
7. Russakovsky O, Deng J, Su H *et al.* ImageNet large scale visual recognition challenge. *Int J Comput Vision* 2015; **115**: 211–52.
8. Lee EJ, Kim YH, Kim N, Kang DW. Deep into the brain: artificial intelligence in stroke imaging. *J Stroke* 2017; **19**: 277–85.
9. Chen W, HolcDorf D, McCusker MW, Gaillard F, Howe PDL. Perceptual training to improve hip fracture identification in conventional radiographs. *PLoS ONE* 2017; **12**: e0189192.
10. Sowden PT, Davies IR, Roling P. Perceptual learning of the detection of features in X-ray images: a functional role for improvements in adults' visual sensitivity? *J Exp Psychol Hum Percept Perform* 2000; **26**: 379–90.
11. Kellman PJ, Garrigan P. Perceptual learning and human expertise. *Phys Life Rev* 2009; **6**: 53–84.
12. Burton AM, Bruce V, Dench N. What's the difference between men and women? Evidence from facial measurement *Perception* 1993; **22**: 153–76.
13. O'Toole AJ, Deffenbacher KA, Valentin D, McKee K, Huff D, Abdi H. The perception of face gender: the role of stimulus structure in recognition and classification. *Mem Cognit* 1998; **26**: 146–60.
14. Haykin S. Neural Networks: A Comprehensive Foundation: Prentice Hall PTR; 1998. 842 p.
15. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems* – Volume 1; Lake Tahoe, Nevada. 2999257: Curran Associates Inc.; 2012. p. 1097–105.
16. Szegedy C, Wei L, Yangqing J *et al.*, editors. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2015 7-12 June 2015.
17. Richardson JT. The analysis of 2 × 2 contingency tables – yet again. *Stat Med* 2011; **30**: 890; author reply 1–2.
18. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology* 2003; **229**: 3–8.
19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**: 837–45.
20. Fitzgerald R. Error in radiology. *Clin Radiol* 2001; **56**: 938–46.
21. World Medical A. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 2013; **310**: 2191–4.