Pediatric Radiology (2019) 49:1066–1070
https://doi.org/10.1007/s00247-019-04408-2

ORIGINAL ARTICLE

Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning

Paul H. Yi[1,2] · Tae Kyung Kim[1,2] · Jinchi Wei[2] · Jiwon Shin[2] · Ferdinand K. Hui[1,2] · Haris I. Sair[1,2] · Gregory D. Hager[2] · Jan Fritz[1,2]

Received: 10 January 2019 / Revised: 24 March 2019 / Accepted: 11 April 2019 / Published online: 30 April 2019

Abstract

Background An automated method for identifying the anatomical region of an image independent of metadata labels could improve radiologist workflow (e.g., automated hanging protocols) and help facilitate the automated curation of large medical imaging data sets for machine learning purposes. Deep learning is a potential tool for this purpose.

Objective To develop and test the performance of deep convolutional neural networks (DCNN) for the automated classification of pediatric musculoskeletal radiographs by anatomical area.

Materials and methods We utilized a database of 250 pediatric bone radiographs (50 each of the shoulder, elbow, hand, pelvis and knee) to train 5 DCNNs, one to detect each anatomical region amongst the others, based on ResNet-18 pretrained on ImageNet (transfer learning). For each DCNN, the radiographs were randomly split into training (64%), validation (12%) and test (24%) data sets. The training and validation data sets were augmented 30 times using standard preprocessing methods. We also tested our DCNNs on a separate test set of 100 radiographs from a single institution. Receiver operating characteristics (ROC) with area under the curve (AUC) were used to evaluate DCNN performances.

Results All five DCNN trained for classification of the radiographs into anatomical region achieved ROC AUC of 1, respectively, for both test sets. Classification of the test radiographs occurred at a rate of 33 radiographs per s.

Conclusion DCNNs trained on a small set of images with 30 times augmentation through standard processing techniques are able to automatically classify pediatric musculoskeletal radiographs into anatomical region with near-perfect to perfect accuracy at superhuman speeds. This concept may apply to other body parts and radiographic views with the potential to create an all-encompassing semantic-labeling DCNN.

Keywords Artificial intelligence · Children · Deep learning · Machine learning · Musculoskeletal · Radiography · Semantic labeling

Introduction

Radiologist workflow in the picture archiving and communication system (PACS) depends heavily on accurate identification of image modalities and anatomical areas for various tasks, including hanging protocols and identifying relevant comparison exams. Although the Digital Imaging and Communications in Medicine (DICOM) format stores metadata, including image modality and anatomical area, the inclusion of this metadata is inconsistent, can vary between equipment manufacturers and can be inaccurate [1]. Inaccurate or variable metadata, such as from studies from an outside facility, could result in omission of relevant comparison studies from being automatically recognized, with downstream technical issues; for example, if an outside facility hand radiograph was not recognized as such due to inaccurate metadata, the interpreting radiologist's hanging protocol may erroneously omit this radiograph as a pertinent prior examination, and impair the radiologist's ability to appropriately compare studies. An automated method for

✉ Jan Fritz
jfritz9@jhmi.edu

[1] The Russell H. Morgan Department of Radiology
and Radiological Science,
Johns Hopkins University School of Medicine,
601 N. Caroline St., Room 4223, Baltimore, MD 21287, USA

[2] Radiology Artificial Intelligence Lab (RAIL),
Malone Center for Engineering in Healthcare,
Johns Hopkins University Whiting School of Engineering,
Baltimore, MD, USA

identifying the anatomical area of an image independent of metadata labels could thus improve radiologist workflow, as well as help facilitate the automated curation of large medical imaging data sets for machine learning purposes [1].

Deep learning is a machine learning technique that utilizes a deep convolutional neural network (DCNN) to recognize image features, and has emerged as a promising method for automated medical image classification [2, 3]. By loosely modeling the structure of the brain, DCNNs can effectively teach themselves the features needed to classify images, given an appropriately large data set with accurate labels [2–4]. One particular use of interest for deep learning is for the automated semantic labeling of images by modality view [1], and anatomical area [2, 5], thereby obviating the need for often-unreliable metadata labels. Prior work has demonstrated the ability of DCNNs to automatically distinguish between chest and abdominal radiographs [2], as well as chest radiograph views (frontal vs. lateral) [1] with 100% accuracy. However, the ability of DCNNs to automatically label pediatric musculoskeletal radiographs by anatomical region has not been evaluated.

The purposes of our study were to develop and define the diagnostic performance of DCNNs for the automated classification of pediatric musculoskeletal radiographs by anatomical area. We hypothesized that different DCNNs would have high diagnostic accuracy to distinguish between five different anatomical regions.

Materials and methods

All images used to develop our DCNNs were part of the public domain and obtained through internet search engines, including Google (http://www.google.com) and Bing (http://www.bing.com). We utilized a second data set to test generalizability of our DCNNs comprised of previously de-identified radiographs obtained at our institution. All images were de-identified and compliant with the Health Insurance Portability and Accountability Act (HIPAA). In accordance with 45 CFR 46.102(f), our institutional review board approved this study and did not require informed consent Our study was compliant with HIPAA. All images were de-identified.

Data sets

To develop our DCNNs, we used 250 radiographs, which were separated into 5 data sets, each consisting of 50 radiographs of the anteroposterior (AP) shoulder, lateral elbow, posteroanterior (PA) hand, AP pelvis and AP knee, respectively, performed in pediatric patients. These radiographs were of patients of varying ages, ranging from newborn to 17 years old. The image quality was considered diagnostic or near-diagnostic, although there were some projectional differences; two representative elbow radiographs are presented in Fig. 1, where the first radiograph shows a well-positioned lateral elbow radiograph and the second shows one with a large amount of internal rotation.

We also curated a second data set comprised of 20 radiographs each of the above 5 anatomical areas obtained from our institution as a clinical test set for our DCNNs spanning ages newborn to 17 years old (100 total test images).

Ground truth was established by two observers, including a board-certified musculoskeletal radiologist (J.F., with 6 years of experience) and a second-year radiology resident with 2 years of orthopedic surgery residency training (P.H.Y.).

Computer hardware and software specifications

All images were saved in the Portable Network Graphics (PNG) format, resized to a 256×256 matrix, and loaded onto a personal computer with a Linux operating system. The computer was equipped with a Core i5 central processing unit

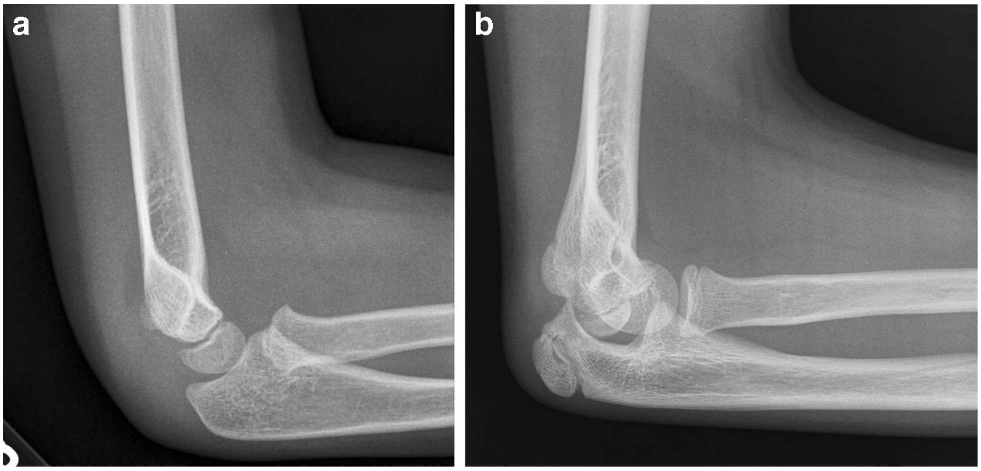

Fig. 1 Representative lateral elbow radiographs demonstrate the elbow well-positioned (**a**) and with a large degree of internal rotation (**b**)

(CPU) (Intel Corporation, Santa Clara, CA), 8 GB RAM and a GeForce GTX 1050 graphics processing unit (GPU) (Nvidia Corporation, Santa Clara, CA). This computer was connected remotely to a computing facility with CPU and GPU nodes utilizing a dual-socket 14-core 2.6 GHz CPU (Intel), 128 GB RAM, and 2-T K80 GPUs (Nvidia Corporation), respectively. All computing work was performed using 6 CPU nodes and 1 GPU node. All computer programming activity was performed using the PyTorch deep learning framework (Version 0.3.1, https://pytorch.org).

Deep learning system development

A DCNN is a complex computational model that uses multiple algorithm layers to create high-level interpretations of data (e.g., classifying images), as opposed to performing single, specific tasks (e.g., detecting a line or an edge on an image) [1, 2]. In developing our deep learning system, we utilized the ResNet-18 DCNN pretrained on 1.2 million color images of everyday objects (1,000 categories) from ImageNet (http://www.image-net.org/) before training on the images. The last linear layer of the pretrained ResNet-18 DCNN was redefined to have 2 outputs instead of the default 1,000. During the training and validation parts of our study, all model parameters were fine-tuned using our data set. This "transfer learning" technique [1, 2] allowed for modification of established ("pretrained") neural network architectures to be optimized for classification of novel data sets not used in training of the original network, which can result in superior performance in medical image classification compared to those without pretraining [2]. The solver parameters used for our DCNN training were 49 epochs, stochastic gradient descent with a learning rate of 0.001, momentum of 0.9 and weight decay of 1×10^5.

Five separate DCNNs were developed, one to detect each anatomical region. For each DCNN, the 250-image data set was divided into training (64% of total data set; 160 images [20 for region of interest, 140 others]), validation (12% of total data set; 30 images [10 for region of interest, 20 others]), and testing (24% of total data set; 60 images [20 for region of interest, 40 others]) sets. The training and validation data sets were augmented by standard techniques, including random cropping, randomly flipping and random rotations (between 30 and 330 degrees) and affine transformations [2], ultimately resulting in a 30 times image augmentation, which has been shown to improve DCNN performance, especially when using small data sets [2]. No augmentation was performed for the testing data sets. We subsequently tested each DCNN on a clinical test set of radiographs obtained at our institution as part of clinical practice to evaluate external generalizability.

To identify the distinguishing features of each radiograph that the DCNN used to classify each image, we created heat maps through *class activation mapping* [6], a technique that visually highlights the importance of various parts of an image in the classification decision through different colors. Red signifies the increasing importance of an image feature in the decision rendered by a DCNN.

Statistical analysis

Statistical analyses were performed using VassarStats (http://vassarstats.net). For each DCNN testing data set, receiver operating characteristics (ROC) with area under the curve (AUC) were generated to define test accuracy (0.9–1=excellent, 0.8–0.9=good, 0.7–0.8=fair, 0.6–0.7=poor, 0.5–0.6=fail). Optimal diagnostic thresholds determined with the aid of Youden J-statistics were used to calculate test sensitivity and specificity. The DeLong non-parametric method was used to statistically compare performance parameters of the DCNN. *P*-values of 0.05 and less were considered statistically significant.

Results

All 5 DCNNs trained to classify the radiographs into anatomical region achieved AUC ROCs of 1, with sensitivity of 100% and specificity of 100% for all. There was no significant difference in AUC ROC between any of the DCNNs (*P*=1 for all pair-wise comparisons of different DCNNS). At optimal diagnostic thresholds, sensitivity and specificity were 100% for all DCNNs. Classification of the test radiographs occurred at a rate of 33 radiographs per s.

On the clinical test set of 100 radiographs obtained at our institution, all 5 DCNNs once again achieved AUC ROCs of 1, with sensitivity and specificity of 100% for all, and unchanged radiograph classification rate of 33 images per s.

Heat maps of the radiographs of both test sets demonstrated that the DCNNs emphasized unique anatomical features of each osseous anatomical region. For example, the flaring of the distal femur metadiaphysis was emphasized to identify the knee, the sacrum was emphasized to identify the pelvis and the antecubital fossa was emphasized to identify the elbow (Fig. 2).

Discussion

We demonstrate that DCNN trained on a small set of images with 30 times augmentation through standard computer processing techniques are able to automatically classify pediatric musculoskeletal radiographs into the anatomical region with perfect accuracy at superhuman speeds. As current methods for semantic labeling through metadata are often unreliable, this technique could contribute to reliable automatic semantic labeling of medical images, radiologist workflow optimization in the PACS, and curating large data sets for machine learning purposes.

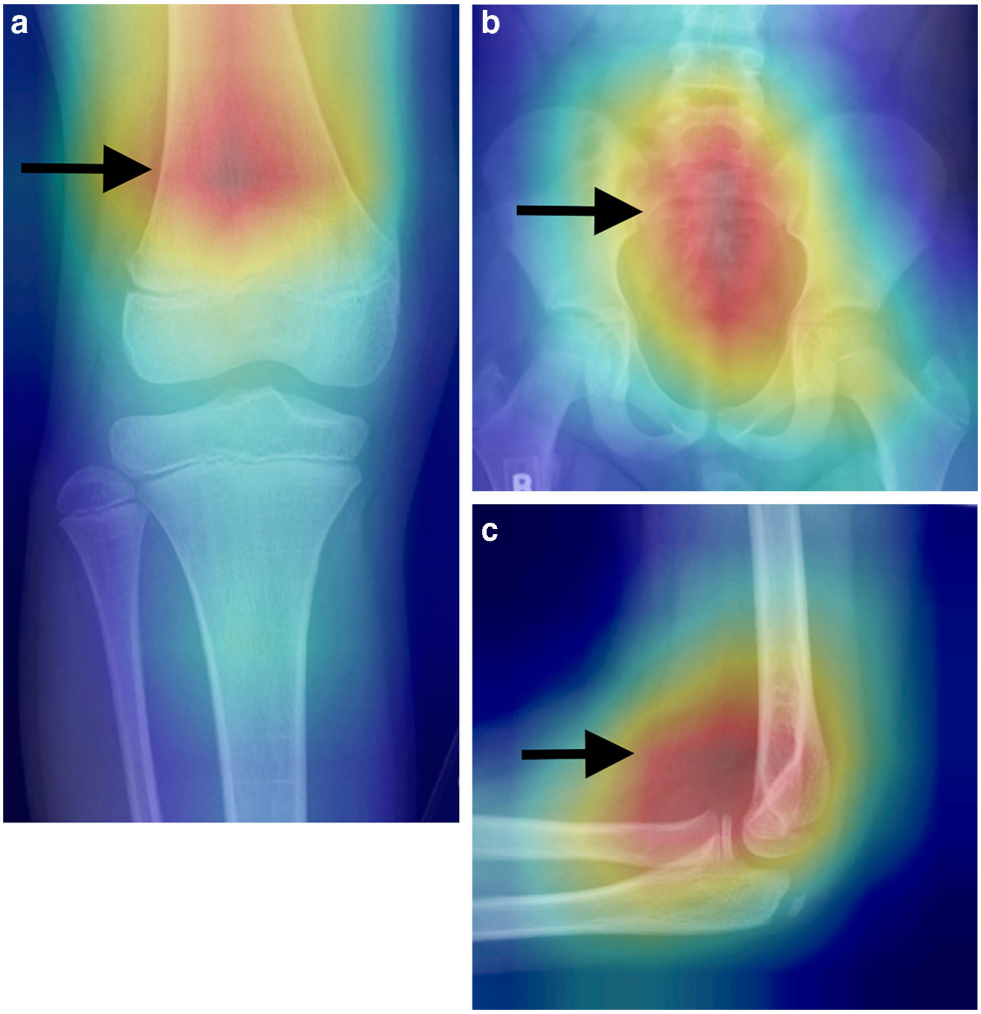

Fig. 2 Deep convolutional neural networks (DCNNs) correctly identify an anteroposterior knee (**a**), AP pelvis (**b**) and lateral elbow (**c**) in radiographs with heat maps demonstrating the algorithm that utilizes the imaging features (*arrows*) of the distal femoral metadiaphyseal flaring, sacrum, and antecubital fossa, respectively, to make the correct classification. Red signifies the increasing importance of an image feature in the decision rendered by a DCNN

Our DCNNs demonstrated AUCs of 1 for all five musculoskeletal anatomical regions, despite using small data sets, and were externally valid on data from our institution. By applying standard image augmentation techniques and increasing our testing and validation data set size by 30-fold, as well as by utilizing transfer learning, we have confirmed the efficacy of this approach to develop high-performing DCNN for semantic labeling, previously used to develop DCNN with 100% accuracy for labeling radiographs of the chest and abdomen [2], as well as frontal and lateral chest radiographs [1]. Interestingly, Lakhani et al. [2] previously demonstrated that 45 chest and 45 abdominal radiographs with similar augmentation technique were sufficient to train a DCNN to distinguish the two anatomical regions. Our results are in accordance with those prior results in that 50 images per anatomical region with augmentation were sufficient for training.

In addition to excellent-to-perfect accuracy of our DCNN for semantic labeling of pediatric musculoskeletal radiographs by anatomical region, the DCNN demonstrated image classification at superhuman speeds with 33 radiographs per s, which is similar to 38 radiographs per s for chest radiograph view classification [1]. Although classification speed would likely vary based on particular computer hardware specifications, our computer hardware specifications are similar to those utilized in prior studies [2, 5]. We note that comparison of our results with Rajkomar et al. [1] is limited due to the lack of reporting of their computer hardware specifications. Nevertheless, the rapid rate of image classification demonstrated in our study and others suggest that DCNN not only can provide accurate semantic labeling of medical imaging, but can do so at rates that exceed human capabilities, and which may facilitate rapid curation of large medical image databases for machine learning purposes, as well as for PACS workflow optimization.

Interestingly, the heat maps in our study reliably demonstrated appropriate identification of unique anatomical features of each bony region. For example, flaring of the distal femur metadiaphysis was consistently focused on to identify the knee, which is consistent with intuition and common approaches utilized by human radiologists. It is also interesting to note that the heat maps focused on the central portions of the images, e.g., the sacrum in the pelvis, which may suggest that larger, more central features of a specific anatomical region are easier for a DCNN to detect than smaller, more peripheral ones.

Our study has limitations. First, our sample sizes were small, theoretically limiting DCNN performance and introducing the possibility of overfitting. However, because the images in our database came from a heterogeneous group of sources, this increased the diversity of images available for DCNN training. Our DCNNs demonstrated no loss in performance when tested on data from our own institution, and our heat map analysis demonstrated appropriate and consistent focus on unique anatomical features for a given joint. Additionally, we attempted to account for these small sample sizes through standard data augmentation methods to increase the total number of images and diversity of imaging presentations. Furthermore, as these images were obtained from the internet, which is in accordance with the technique that the ImageNet database used to pretrain DCNN [2, 3, 5, 7, 8], they were of lower resolution than would be expected for images obtained from an institutional PACS in DICOM format, and differences in technique and modality utilized to acquire images is unknown, which could potentially have an impact on model performance. Accordingly, training a DCNN to classify images using these lower-resolution images would be expected to be a more difficult task than using higher-resolution images, which suggests that better results could be obtained using PACS-derived DICOM images, perhaps with even fewer images than we used. Nevertheless, the approach of using internet-derived images is, in fact, the same approach used to create the ImageNet database, which is the current standard of reference for image classification DCNNs and the dominant database currently used for DCNN pretraining. Second, we only included five osseous anatomical regions with five radiographic views because we sought to develop a proof-of-concept of deep learning networks toward musculoskeletal anatomical classification spanning the upper and lower extremities. Subsequent studies will be required to test the ability of the DCNN to identify other bony regions and radiographic views, especially anatomical regions that are different, but similar, such as the femur and tibia. Third, we utilized a single DCNN only, as opposed to multiple DCNNs used in prior studies [2]. Other DCNN architectures may have higher performance, which is a topic for future study. Fourth, we are unable to determine precisely how a deep learning system makes a diagnosis, raising concerns that a deep learning system may function as a “black box.” Nevertheless, visualization techniques such as class activation mapping [6], which creates visual heat maps for distinguishing features of images used by DCNN for classification decisions, may facilitate further understanding of the mathematical modeling of classification performed by DCNNs. Finally, despite the high performance of our DCNNs, our study represents a mere proof-of-concept, which requires further fine-tuning and prospective clinical validation, as well as expansion to other anatomical regions and radiographic views.

Conclusion

In conclusion, DCNN have good-to-perfect accuracy for automatically classifying pediatric musculoskeletal radiographs into the anatomical region at superhuman speeds, which may enhance radiologist workflow, as well as facilitate rapid curation of large medical imaging databases for machine learning purposes. The proof-of-concept from our work may apply to other body parts and radiographic views to create an all-encompassing semantic labeling DCNN, although further fine-tuning and clinical validation are required.

Acknowledgments The authors acknowledge Tae Soo Kim, MSE, for technical advising.

Compliance with ethical standards

Conflicts of interest None

References

1. Rajkomar A, Lingam S, Taylor AG et al (2017) High-throughput classification of radiographs using deep convolutional neural networks. J Digit Imaging 30:95–101
2. Lakhani P (2017) Deep convolutional neural networks for endotracheal tube position and X-ray image classification: challenges and opportunities. J Digit Imaging 30:460–468
3. Lakhani P, Sundaram B (2017) Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 284:574–582
4. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521: 436–444
5. Cheng PM, Malhi HS (2017) Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. J Digit Imaging 30:234–243
6. Zhou B, Khosla A, Lapedriza A et al (2016) Learning deep features for discriminative localization. In: 2016 IEEE Conf Comput Vis Pattern Recognit, pp 2921–2929
7. Ting DSW, Cheung CY, Lim G et al (2017) Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 318:2211–2223
8. Kim DH, MacKinnon T (2018) Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 73:439–445