Radiology

REVIEWS AND COMMENTARY • EDITORIAL

Deep Learning for Triage of Chest Radiographs: Should Every Institution Train Its Own System?

Bram van Ginneken, PhD

From the Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen, the Netherlands. Received October 5, 2018; revision requested October 10; revision received October 10; accepted October 11. **Address correspondence to** the author (e-mail: *bram.vanginneken@radboudumc.nl*).

Conflicts of interest are listed at the end of this article.

See also the article by Dunnmon et al in this issue.

Radiology 2019; 290:545–546 • https://doi.org/10.1148/radiol.2018182318 • Content code: CH •

"This paper will describe a concept of converting the visual images on roentgenograms into numerical sequences that can be manipulated and evaluated by the digital computer."

The above statement by Lodwick et al (1) was the opening sentence of a visionary article in the August 1963 issue of *Radiology*. The authors analyzed images from 514 chest radiography examinations in patients with lung cancer. They developed what they called a coding system, which was a set of features specifically designed to analyze lung cancer on radiographs and that were visually scored for each examination by a radiologist. They showed that by using these numbers as input, a computer system could predict 1-year survival. The authors realized that ultimately, the computer could compute these coding features itself, "through direct optical scanning of roentgenograms," something that was not yet possible in 1963. The main task ahead was to determine which features to extract from the images, as they wrote: "It would be unrealistic to assume that the development of a coding system, as described here, is a simple task. Yet, if the computer is to be developed to its full potentiality, the establishment of such coding systems for most diagnostic examinations may be required."

Fast-forward to 2018, and chest radiography is still the most commonly performed radiologic examination. Software products exist to remove rib shadows and indicate possible locations of pulmonary nodules, but chest radiographs are still read exclusively by radiologists, almost always without support from computers. This could change, however, because promising results in automatic interpretation of medical images have been achieved in recent years by using deep learning, or, more precisely, convolutional neural networks with many layers. The number of publications on this topic is rapidly increasing (2). New conferences and journals are being founded, including *Radiology: Artificial Intelligence*, a subspecialty journal that will be published by the Radiological Society of North America. There is immense interest from both start-up and established companies to bring artificial intelligence to radiology. We now, finally, may be close to achieving what Lodwick et al envisioned.

In one respect, deep learning is fundamentally different from the more traditional and established approaches of rule-based image processing, machine learning, radiomics, and computer-aided detection and diagnosis (3). These deep convolutional networks operate directly on the image input data. They do not rely on a handcrafted set of features, the so-called coding system that was described by Lodwick et al. In its training process, the network continuously adjusts the weights in all layers, making sure that the training images—the input—are mapped to the correct output. At the end of this long computational process, the coding system has been established, inculcated in the network weights. The hard task identified by Lodwick et al is now easy; it is a byproduct of the training process. All that is needed to solve a certain task is a large corpus of images with the corresponding output, the choice for a network architecture and its hyperparameters, and a computer with a graphics processing unit, a powerful chip with many processors that run together the endless forward and backward passes that batches of images make through the network during training.

In this issue of *Radiology*, Dunnmon and colleagues (4) addressed automatic interpretation of chest radiographs as normal or abnormal. The approach taken to develop the system was deliberately straightforward. The authors collected a set of 200 000 frontal radiographs obtained at their institution over a 15-year period in patients who had not undergone prior imaging. They used the summary labels from the original report to determine if images were normal or abnormal. They applied three well-known deep networks: AlexNet, ResNet-18, and DenseNet-121. To show how widely used these three standard architectures are, the conference papers from 2012, 2016, and 2017 that describe them have already garnished more than 43 000 citations. Code for these off-the-shelf networks can simply be downloaded from open-source repositories—if desired, already "pretrained" on the ImageNet data set, millions of photographs downloaded from the internet and labeled by human readers. These networks require low-resolution (224 × 224 pixels) images to be fed into them. Thus, Dunnmon et al subsampled the radiographs substantially; nevertheless, they reported an impressive classification performance, with an area under the receiver operating characteristic curve (AUC) of 0.96. DenseNet-121 performed the best. AlexNet, the oldest architecture, performed the worst, but the difference between network performance was very small. Interestingly, the study also reported on the results of a classic machine-learning approach, bag-of-words with a support vector machine. This technique performed

reasonably well (AUC = 0.93), but its performance was substantially inferior to that of the convolutional networks.

How could one use such a system? The authors suggest that it could be used for triage in areas without access to trained radiologists and for workflow prioritization in clinics with staff shortages. A recent study (5) presented a similar analysis system that was trained on Chest X-ray 14, a publicly available data set of more than 112 000 chest radiographs (6). It was suggested that this system could take over the role of radiographers in "red dotting." The "red dot" system means that radiographers communicate the presence of potential abnormalities on a film hard copy by placing a round red sticker on the abnormal image. Some modern picture archiving and communications systems use digital superimposition of the words *red dot* on such images.

Alternatively, the output of the network can be averaged with a rating provided by a human reader. Dunnmon et al showed that such a combined human and artificial intelligence system achieves an AUC of 0.98, which is significantly better than the computer system alone, and achieves a higher accuracy than human reading alone (the best network alone still has slightly lower accuracy than human reading).

Deep-learning networks are thought to be very data hungry. An excellent aspect of this study is that the effect of the size of the training data is analyzed in detail. Experimental results with training sets of 2000, 20 000, and 200 000 images are compared. Results when using only 2000 images are substantially worse, but the difference between 20 000 and 200 000 training images is insignificant, as measured in a hold-out test set of 1000 images that were carefully reannotated by expert readers. Dunnmon et al (4) conclude that "while carefully adjudicated image labels are necessary for evaluation purposes, prospectively labeled single-annotator data sets of a scale modest enough (approximately 20 000 samples) to be available to many institutions are sufficient to train high-performance classifiers for this task."

This is a thought-provoking statement. It suggests that to deal with the large variety of imaging protocols, equipment, patient populations, and maybe even reporting guidelines between institutions, every institution could curate its own training data sets and train its own deep-learning systems for a wide variety of tasks. This runs counter to the notion generally accepted both by this journal (7) and by the regulatory authorities that software used to analyze medical images should be validated in large multicenter data sets. It is my expectation that a frontal chest radiograph triaging system that works well on data from many institutions and is thus generally applicable would benefit from training on a multicenter data set much larger than 20 000 or even 200 000 examinations. This larger size is needed to capture the diversity of data from different centers and to ensure that there are enough examples of relatively rare abnormal findings so that the network learns not to miss them. Such a large-scale system should be based on newly designed network architectures that take the full-resolution images as input. It would be advisable to train systems not only to provide a binary output label but also to detect specific regions in the images with specific abnormalities. This would require annotation of such regions on many training images. This would be a step toward a deep-learning network that explains to the user why it arrived at the overall conclusion that examination results might be abnormal.

It will be interesting to see if there is a role in radiologic clinical practice for simpler "homegrown" networks. The study by Dunnmon et al shows that off-the-shelf networks have great promise in automated chest radiograph triage. When trained with modestly sized data sets from one hospital using the noisy labels derived from radiology reports, they have already achieved compelling results.

Disclosures of Conflicts of Interest: B.v.G. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution receives funding from Thirona, Delft Imaging Systems, Canon Medical, Philips Healthcare, and Siemens Healthineers; is a cofounder of and shareholder in Thirona; receives royalties from Thirona, Delft Imaging Systems, and MeVis Medical Solutions. Other relationships: disclosed no relevant relationships.

References

1. Lodwick GS, Keats TE, Dorst JP. The coding of roentgen images for computer analysis as applied to lung cancer. Radiology 1963;81(2):185–200.
2. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
3. van Ginneken B. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. Radiol Phys Technol 2017;10(1):23–32.
4. Dunnmon JA, Yi D, Langlotz CP, et al. Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology 2019;290:537–544.
5. Yates EJ, Yates LC, Harvey H. Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. Clin Radiol 2018;73(9):827–831.
6. Wang X, Peng Y, Lu L, et al. Chest X-ray 8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE CVPR, 2017; 3462–3471.
7. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018;286(3):800–809.