EDITORIAL

Translating Artificial Intelligence Into Clinical Care

Andrew L. Beam, PhD; Isaac S. Kohane, MD, PhD

Artificial intelligence has become a frequent topic in the news cycle, with reports of breakthroughs in speech recognition, computer vision, and textual understanding that have made their way into a bevy of products and services that are used every day. In contrast, clinical care has yet to reach the much lower bar of automating health care information transactions in the form of electronic health records. Medical leaders in the 1960s and 1970s were already speculating about the opportunities to bring automated inference methods to patient care,[1] but the methods and data had not yet reached the critical mass needed to achieve those goals.

Editorial and Viewpoint

Related article

The intellectual roots of "deep learning," which power the commodity and consumer implementations of present-day artificial intelligence, were planted even earlier in the 1940s and 1950s with the development of "artificial neural network" algorithms.[2,3] These algorithms, as their name suggests, are very loosely based on the way in which the brain's web of neurons adaptively becomes rewired in response to external stimuli to perform learning and pattern recognition. Even though these methods have had many success stories over the past 70 years, their performance and adoption in medicine in the past 5 years has seen a quantum leap. The catalyzing event occurred in 2012 when a team of researchers from the University of Toronto reduced the error rate in half on a well-known computer vision challenge using a deep learning algorithm.[4] This work rapidly accelerated research and development in deep learning and propelled the field forward at a staggering pace. With the increased availability of digital clinical data, it remains to be seen how these deep learning models might be applied to the medical domain.

In this issue of *JAMA*, Gulshan and colleagues[5] present findings from a study evaluating the use of deep learning for detection of diabetic retinopathy and macular edema. To build their model, the authors collected 128 175 annotated images from the EyePACs database. Each image was rated by 3 to 7 clinicians for referable diabetic retinopathy, diabetic macular edema, and overall image quality. Each rater was selected from a panel of 54 board-certified ophthalmologists and senior ophthalmology residents. Using this data set, the algorithm learned to predict the consensus grade of the raters along each clinical attribute: referable diabetic retinopathy, diabetic macular edema, and image quality. To validate their algorithm, the authors assessed its performance on 2 separate and nonoverlapping data sets consisting of 9963 and 1748 images. On the validation data, the algorithm had high sensitivity and specificity. Only one of these values (sensitivity on the second validation data set) failed to be superior at a statistically significant level. The other performance metrics (eg, area under the receiver operating characteristic curve, negative predictive value, positive predictive value) were likewise impressive, giving the authors confidence that this algorithm could be of clinical utility.

This work closely mirrors a recent "Kaggle" contest in which 661 teams competed to build an algorithm to predict the grade of diabetic retinopathy, albeit on a smaller data set with fewer grades per image. Kaggle is a website that hosts machine learning and data science contests. Companies and researchers can post their data to Kaggle and have contestants from around the world build predictive models. In the diabetic retinopathy contest, nearly all of the top teams used some form of deep learning and had little to no knowledge of the eye or ophthalmology. The first-place team[6] and second-place team[7] both used standard deep learning models and were data science practitioners, not medical professionals. Gulshan et al correctly pointed out that a prerequisite for a successful deep learning model is access to a large database of images with high-quality annotations. Accordingly, the investigators increased both the number of images available and the number of ratings per image, which allowed them to improve on the existing state of the art with respect to both Kaggle and the existing scientific literature.

To build their algorithm, Gulshan et al leveraged a workhorse model in deep learning known as a convolutional neural network that has been critically important to recent advances in automatic image recognition. The convolutional neural network model used by the authors is known as the Inception-V3 network,[8] which was developed by Google for entry in the Large Scale Visual Recognition Challenge, which it won in 2014. In this contest, known as ImageNet,[9] researchers were given 1.2 million images that involve 1000 different categories that cover a wide variety of everyday objects, such as cats, dogs, automobiles, and different kinds of food. The goal of the contest was to build a classifier that could automatically recognize which object was present in an image and to identify which region of the image contained the object. This challenge was broad so that it covered many types of objects that a computer vision system could encounter in the real world.

As a result of this contest, several techniques[10-12] have been pioneered that improved the accuracy of these models immensely. As with the study by Gulshan et al, these improvements are beginning to trickle into other areas of computer vision, including medical image processing. For example, Gulshan et al not only used the same network that was originally built for ImageNet, they also used that network

configuration to initialize their model for this study. This is often known as "transfer learning" and occurs when a network trained for one task (eg, ImageNet recognition) is used to bootstrap a network to be used for a different task (eg, detection of diabetic retinopathy). Gulshan et al observed a boost in performance when they used the parameters learned on ImageNet to initialize their model, thus demonstrating how progress in one domain can be used to accelerate progress in another.

Stepping back, one can consider how these results might affect medicine and, in particular, areas of medicine that involve the analysis of images such as pathology, radiology, and dermatology.[13] It seems likely that these algorithms will reshape specific aspects of these specialties as more algorithms are developed to address a wider range of medical imaging tasks. Because these algorithms are by their nature standardized, repeatable, and scalable, they can be deployed to analyze a large number of images in hospitals around the world once an algorithm has been developed and validated, enabling clinicians to focus on other aspects of their practice.

A simple cost-benefit analysis reveals some interesting implications. Once a model has been "trained," it can be deployed on a relatively modest budget. Deep learning uses a specialized type of computer chip known as a graphics processing unit to process data at high speeds. A modern graphics processing unit costing approximately $1000 can be added to most existing computer systems with little difficulty and can process about 3000 images per second[14] depending on the complexity of the underlying deep learning model. This translates to an image processing capacity of almost 260 million images per day (because these devices can work around the clock), all for the cost of approximately $1000. How will practice and clinical training adapt to refocus if initial screening of images is delegated to a machine with a learning algorithm? How will these capabilities mesh with current regulatory and reimbursement policies, or will these have to be modified?

Finally, the commercial efforts to push this technology into clinical care are becoming apparent, as several companies have begun to translate these research advancements to commercial applications. For example, one company is using deep learning models to improve cancer detection,[15] while another company uses deep learning to read radiology images.[16] Outside of imaging, other companies using artificial intelligence have started to help manage care, predict patient outcomes, or monitor patients through wearable devices, all in an attempt improve health care delivery. Given that artificial intelligence has a 50-year history of promising to revolutionize medicine and failing to do so, it is important to avoid overinterpreting these new results. However, given the rapid and impressive progress in other areas of artificial intelligence, along with results such as those presented by Gulshan et al, there are valid reasons to remain cautiously optimistic that the time could now be right for artificial intelligence to transform the clinic into a much higher-capacity and lower-cost information processing care service.

ARTICLE INFORMATION

Author Affiliations: Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts.

Corresponding Author: Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck St, Boston, MA 02115 (isaac_kohane@harvard.edu).

Published Online: November 29, 2016. doi:10.1001/jama.2016.17217

Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Medicine and the computer: the promise and problems of change. In: Schwartz W. *Use and Impact of Computers in Clinical Medicine*. New York, NY: Springer-Verlag; 1970:321-335. http://link.springer.com/chapter/10.1007/978-1-4613-8674-2_20. Accessed November 10, 2016.

2. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*. 1943;5(4):115-133.

3. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*. 1958;65(6):386-408.

4. Krizhevsky A, Sutskever I, Hinton GE. *ImageNet Classification With Deep Convolutional Neural Networks*. Vol 1. La Jolla, CA: Neural Information Processing Systems Foundation Inc; 2012:4.

5. 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*. doi:10.1001/jama.2016.17216

6. Kaggle. Diabetic retinopathy winner's interview: 1st place, Ben Graham. http://blog.kaggle.com/2015/09/09/diabetic-retinopathy-winners-interview-1st-place-ben-graham/. Accessed November 9, 2016.

7. Kaggle. Team o_O solution summary. https://www.kaggle.com/c/diabetic-retinopathy-detection/forums/t/15617/team-o-o-solution-summary. Accessed November 9, 2016.

8. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. *Rethinking the Inception Architecture for Computer Vision*. December 2015. http://arxiv.org/pdf/1512.00567v3.pdf. Accessed November 7, 2016.

9. Deng J, Dong W, Socher R, Li L, Li K. ImageNet: a large-scale hierarchical image database. 2009. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5206848. Accessed October 19, 2016.

10. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. *Improving Neural Networks by Preventing Co-adaptation of Feature Detectors*. July 3, 2012. https://arxiv.org/abs/1207.0580. Accessed November 14, 2016.

11. Simonyan K, Zisserman A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. September 4, 2014. https://arxiv.org/abs/1409.1556. Accessed November 7, 2016.

12. He K, Zhang X, Ren S, Sun J. *Deep Residual Learning for Image Recognition*. December 10, 2015. https://arxiv.org/abs/1512.03385. Accessed November 7, 2016.

13. Jha S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *JAMA*. doi:10.1001/jama.2016.17438

14. Andersch M. Inference: the next step in GPU-accelerated deep learning. November 11, 2015. https://devblogs.nvidia.com/parallelforall/inference-next-step-gpu-accelerated-deep-learning/. Accessed November 9, 2016.

15. PathAI. https://www.pathai.com. Accessed November 9, 2016.

16. Enlitic. http://www.enlitic.com. Accessed November 9, 2016.