ORIGINAL ARTICLE

Deep Learning in Radiology: Does One Size Fit All?

Bradley J. Erickson, MD, PhD, Panagiotis Korfiatis, PhD, Timothy L. Kline, PhD, Zeynettin Akkus, PhD, Kenneth Philbrick, PhD, Alexander D. Weston, BS

Abstract

Deep learning (DL) is a popular method that is used to perform many important tasks in radiology and medical imaging. Some forms of DL are able to accurately segment organs (essentially, trace the boundaries, enabling volume measurements or calculation of other properties). Other DL networks are able to predict important properties from regions of an image—for instance, whether something is malignant, molecular markers for tissue in a region, even prognostic markers. DL is easier to train than traditional machine learning methods, but requires more data and much more care in analyzing results. It will automatically find the features of importance, but understanding what those features are can be a challenge. This article describes the basic concepts of DL systems and some of the traps that exist in building DL systems and how to identify those traps.

Key Words: Deep learning, machine learning, computer-aided diagnosis

J Am Coll Radiol 2017;■:■-■.

Traditional machine learning identifies patterns that are present in training sets. In those traditional approaches, it is necessary to compute "features" that are thought to be important factors, which are then used as input to train the system how to classify images as positive or negative. The meaning of positive and negative, as well as the features selected, depends on the type of image (eg, CT, MR, PET) as well as the task (eg, determine if cancer is present, or if the lesion is benign versus malignant).

An early form of machine learning was the artificial neural network (ANN) that was based on the human brain. It had inputs (each input was called a node) with multiple connections from one layer of nodes to the next, just as neurons have dendrites that have multiple inputs and pass their signal on to the next neuron, finally outputting to a muscle. However, ANNs did not work very well and fell into disuse.

Radiology Informatics Laboratory, Department of Radiology, Mayo Clinic, Rochester, Minnesota.

Corresponding author and reprints: Bradley J. Erickson, Radiology Informatics Laboratory, Department of Radiology, Mayo Clinic, Mayo Building East 2, 200 First St SW, Rochester, MN 55905; e-mail: bje@mayo.edu.

Supported by the National Cancer Institute (NCI), Grant No. CA160045. The authors have no conflicts of interest related to the material discussed in this article.

Deep learning (DL) is a new form of machine learning that has dramatically improved performance of machine learning tasks. DL is interesting not only because its level of performance is greater, but also because it does not require a human to identify and compute the critical features. Instead, during training, DL algorithms "learn" discriminatory features that best predict the outcomes. This means that the amount of human effort required to train DL systems is less (because no feature engineering or computation is required) and may also lead to the discovery of important new features that were not anticipated.

IMPORTANT DRIVERS OF DL

Moore's law states that the performance of computing technology doubles about every 18 months. The world of DL has seen a much faster advance in computing speed. The main reason is that graphics processing units (GPUs) used for displaying graphics on computer screens can also be used to perform DL calculations. GPUs typically have hundreds or thousands of processing units that perform multiple calculations simultaneously. It is therefore typical to see DL algorithms run 10 to 100 times faster when GPUs are used compared with traditional CPUs. Another important factor is that the memory on the GPU

1546-1440/17 ■ https://doi.org/10.1016/j.jacr.2017.12.027

cards is increasing rapidly, which is critical for DL applications.

Billions of dollars have been invested in DL technologies across their scope of application, which includes Internet search, social media, and self-driving vehicles. Recent versions of GPUs now have special adaptations to further improve DL performance, and special computers with multiple GPUs designed specifically for DL have been built and sold. Companies are now building special purpose processors that can perform the DL computations even more efficiently than GPUs. These devices are focused both on being efficient in the training of the models as well as inference (making the prediction after being trained).

Although the rapid advance of technology is often cited as *the* reason for DL success, advances in algorithms are probably more responsible for the recent success. These advances addressed the problems seen in ANNs. DL networks get their name because they have many layers; most systems now have 30 to 150 layers, compared with traditional ANNs that would fail if they had more than about 3 layers.

One important advance that enabled DL is improved activation functions. An activation function converts the weighted sum of inputs into an output value that is then conveyed to nodes in the next layer. Activation functions add nonlinearity, which is a critical element of learning. Early neural networks used sigmoidal activation functions because neurons had sigmoidal activation functions, but they are susceptible to phenomena that can halt network learning or lead to network instabilities. When there are millions of nodes that can be altered, it is critical to identify those nodes that can produce the greatest improvement. In some cases, it can be hard to identify which ones to alter, a situation known as the vanishing gradient problem.

For DL systems, other simpler activation functions, such as rectified linear unit (ReLU), have been proven to work better than sigmoidal functions. The ReLU takes the weighted sum of inputs, and, if that is less than zero, it outputs a zero; otherwise, it outputs the input [1-3]. More recent variants (eg, leaky ReLU [4]) produce small negative values for negative inputs but give the same output as input when positive. Figure 1 shows sigmoidal, ReLU, and leaky ReLU activation functions. These activation functions are actually much simpler to compute than sigmoidal functions—evidence that computing power is not the only factor in DL success.

Overfitting poses another challenge to training deep neural networks. During training, networks can learn the specific examples in the training set that are not broadly predictive features for the problem of interest if they have enough parameters compared with the number of

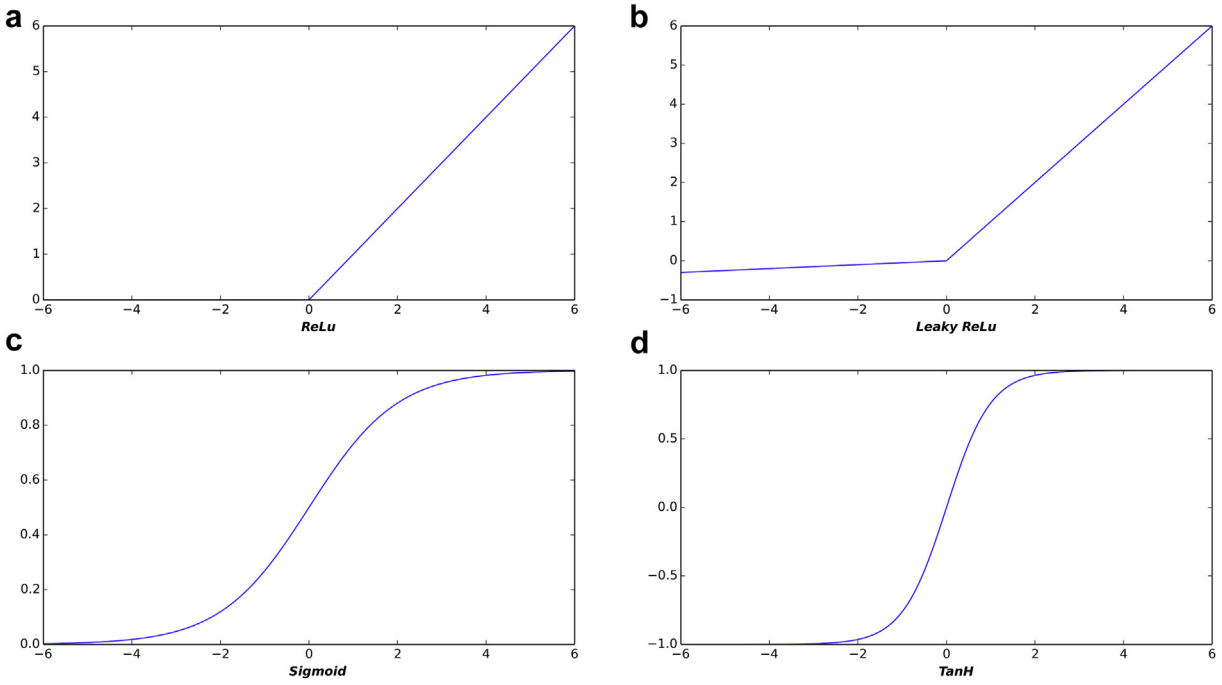

Fig 1. Example of three activation functions used in neural networks: (a) rectified linear unit (ReLU), (b) leaky ReLU, (c) sigmoid, and (d) Tanh. Traditional neural networks used sigmoidal functions that simulated actual neurons, but are less effective in current networks, likely because they do not adequately reward very strong activations.