SPOTLIGHT ON TRANSACTIONS

Representation Learning

David Forsyth, University of Illinois at Urbana–Champaign

This installment of Computer's series highlighting the work published in IEEE Computer Society journals comes from IEEE Transactions on Pattern Analysis and Machine Intelligence.

How should a computer program represent important effects in pictures without getting distracted by details? Humans are good at focusing on what is important and ignoring what isn't. Human viewers aren't confused by small changes in light, texture, color, or configuration—but we can't say the same for computers.

There are good tricks for building image representations that aren't always confused by these effects, but object detectors seem to have strong performance limits. Systems using deep networks have recently smashed these limits, likely because their image representations are learned. In "Representation Learning: A Review and New Perspectives" (*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, 2013, pp. 1798–1828), Yoshua Bengio and his colleagues review recent work in the area of unsupervised feature learning and deep learning.

Deep representations have already had a major impact on practice in areas such as speech understanding.

Practice shows that to learn image representation for object detection, we should focus on representing any image, rather than on representing the particular objects of interest. First, modern methods generally do better with more data; there are always relatively few pictures of the objects of interest because labeling images is hard, and there is an immense number of unlabeled images. Second, we might change our mind about which objects are of interest. If our representation were specialized to one particular set of objects, we'd have a problem.

A representation is generally considered good if we can roughly reconstruct the image from the representation. We achieve this by training two functions at the same time: one constructs the representation from the image, and the other constructs the image from the representation. The representation must be small and the reconstruction must be "fairly accurate."

Good representations tend to summarize an image at longer and longer spatial scales. Imagine we have learned a representation that describes local patches of an image, called a *layer*. The layer will produce a set of numbers that we can organize like an image. We apply our method for learning a representation to this object, and the next layer will summarize larger patches. We do this again and again, stacking layers upon layers. The result is a deep representation.

I have described this method using images as an example for concreteness, but deep representations have already had a major impact on practice in areas such as speech understanding. There are major ongoing efforts to apply deep representations to a wide variety of practical problems. This paper covers big representational questions, rather than focusing on a particular problem; there's a good chance it'll be useful in your field. As a bonus, it appears in a special issue on learning deep architectures—I recommend reading the whole special issue, as well as the many papers in *TPAMI*'s pipeline describing recent successes of the deep learning approach.

DAVID FORSYTH is a professor of computer science at the University of Illinois at Urbana–Champaign. Contact him at daf@illinois.edu.

0018-9162/15/$31.00 © 2015 IEEE