DOI:10.1145/3065384

To view the accompanying paper, visit doi.acm.org/10.1145/3065386 rh

Technical Perspective
What Led Computer Vision to Deep Learning?

By Jitendra Malik

WE ARE IN the middle of the third wave of interest in artificial neural networks as the leading paradigm for machine learning. The first wave dates back to the 1950s, the second to the 1980s, and the third to the 2010s. The following paper by Krizhevksy, Sutskever and Hinton (henceforth KSH) is the paper most responsible for this third wave. Here, I sketch the intellectual history surrounding this work.

The current wave has been called "deep learning" because of the emphasis on having multiple layers of neurons between the input and the output of the neural network; the main architectural design features, however, remain the same as in the second wave, the 1980s. Central to that era was the publication of the back-propagation algorithm for training multilayer perceptrons by Rumelhart, Hinton and Williams.[7] This algorithm, a consequence of the chain rule of calculus, had been noted before, for example, by Werbos.[8] However, the Rumelhart et. al. version was significantly more impactful as it was accompanied by interest in distributed representations of knowledge in cognitive science and artificial intelligence, contrasted with the symbolic representations favored by the mainstream researchers.

The second intellectual strand comes from neuroscience, most specifically from Hubel and Wiesel's studies of cat and monkey visual cortex.[4,5] They developed a hierarchical model of the visual pathway with neurons in lower areas such as V1 responding to features such as oriented edges and bars, and in higher areas to more specific stimuli ("grandmother cells" in the cartoon version). Fukushima[2] proposed a neural network architecture for pattern recognition explicitly motivated by Hubel and Wiesel's hierarchy. His model had alternating layers of simple cells and complex cells, thus incorporating down sampling, and shift invariance, thus incorporating convolutional structure. LeCun et al.[6] took the additional step of using backpropagation to train the weights of this network, and what we today call convolutional neural networks were born.

The 1990s and 2000s saw diminished interest in neural networks. Indeed, one of the inside jokes was that having the phrase "neural networks" in the title of a paper was a negative predictor of its chance of getting accepted at the NIPS conference!

A few true believers such as Yoshua Bengio, Geoffrey Hinton, Yann LeCun, and Juergen Schmidhuber persisted, with a lot of effort directed towards developing unsupervised techniques. These did not lead to much success on the benchmark problems that the field cared about, so they remained a minority interest. There were a few technical innovations—max-pooling, dropout, and the use of half-wave rectification (a.k.a ReLU) as the activation function nonlinearity—but before the publication of the KSH paper in 2012, the mainstream computer vision community did not think that neural network based techniques could produce results competitive with our hand designed features and architectures. I was one of those skeptics, and I recall telling Geoff Hinton that convincing the computer vision community would require results on the real-world datasets that we used. Geoff did take this advice to heart and I like to think that conversation was one of the inspirations behind KSH.

It is my opinion the following paper is the most impactful paper in machine learning and computer vision in the last five years.

What was the secret sauce behind KSH's success? Besides the technical innovations (such as the use of ReLUs), we must give a lot of credit to "big data" and "big computation." By big data here I mean the availability of large datasets with category labels, such as ImageNet from Fei-Fei Li's group, which provided the training data for these large, deep networks with millions of parameters. Previous datasets like Caltech-101 or PASCAL VOC did not have enough training data, and MNIST and CIFAR were regarded as "toy datasets" by the computer vision community. This strand of labeling datasets for benchmarking and for extracting image statistics itself was enabled by the desire of people to upload their photo collections to the Internet on sites such as Flickr. The way big computation proved most helpful was through GPUs, a hardware development initially driven by the needs of the video game industry.

Let me turn now to the impact of the KSH paper. As of this writing, it has 10,245 citations on Google Scholar, remarkable for a paper not yet five years old. I was present at the ECCV ImageNet workshop where the KSH results were presented. Everyone was impressed by the results, but there was debate about their generality. Would the success on whole image classification problems extend to more tasks such as object detection? Was the finding a very fragile one, or was it a robust one that other groups would be able to replicate? Stochastic gradient descent (SGD) can only find local minima, so what is the guarantee the minima we find will be good?

In the true spirit of science, many

of us, skeptics and believers, went back to our laboratories to explore these questions. Within a year or two, the evidence was quite clear. For example, the R-CNN work of Girshick et al.[3] showed the KSH architecture could be modified, by making use of computer vision ideas such as region proposals, to make possible state of the art object detection on PASCAL VOC. Getting SGD to work well is an art, but it could be mastered by students and researchers and corporate employees and yield reproducible results in many different settings. We do not yet have convincing theoretical proof of the robustness of SGD but the empirical evidence is quite compelling, so we leave it to the theoreticians to find an explanation while experimentalists forge ahead. We have realized that generally deeper networks work better, and that overfitting fears are overblown. We have new techniques such as "batch normalization" to deal with regularization, and dropout is not so crucial any more. Practical applications abound.

It is my opinion the following paper is the most impactful paper in machine learning and computer vision in the last five years. It is the paper that led the field of computer vision to embrace deep learning. C

References

1. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Li, F.-F. ImageNet: A Large- scale hierarchical image database. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, (June 20–25, 2009).
2. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern 34*, 4 (1980), 193–202.
3. Girshick, R., Donahue, J., Darrell, T. and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, (2014).
4. Hubel, D.H. and Wiesel, T.N. Receptive fields, binocular interactions and functional architecture in the cat's visual cortex. *J. Physiology 160*, 1 (Jan. 1962), 106–154.
5. Hubel, D.H. and Wiesel, T.N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiology 195*, 1 (Mar. 1968), 215–243.
6. LeCun, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation 1* (1989), 541–551.
7. Rumelhart, D.E., Hinton G.E, and Williams R.J. Learning representations by back-propagating errors. *Nature 323* (Oct. 9, 1986), 533–536.
8. Werbos P. Beyond regression: New tools for prediction and analysis in the behavioral sciences. Ph.D. thesis, Harvard University, 1974.

Jitendra Malik is the Arthur J. Chick Professor of EECS at the University of California at Berkeley.