CTN Seminar: Yann LeCun

Report: 1 of 10 Submitted: Jan. 30, 2009 Author: Marc Hurwitz

Learning Hierarchies of Invariant Visual Features

Speaker: Yann Lecun *Jan. 27, 2009*

Summary

The talk was about how people learn invariant representations. For example, the number ‘8’ can be written as *8* or 8 or **8** or 8 or 8. How, then, can we recognize each of these as the same number? How might we learn that these are all the same or, if we haven’t seen a version of 8 before, recognize it as an 8? Finally, how can we separate this object from others that might appear simultaneously in a noisy environment?

Dr. Lecun noted that most models of how this works are composed of two parts. The first tends to be fairly specific to the object(s) being learned, while the second is more generic. In general, models to date have mostly used supervised learning even though only a small percentage of visual learning is likely to be supervised (reading is an example where learning the alphabet is definitely supervised and, even after considerable time, some people still have trouble distinguishing p from q, for example). As well, the algorithms tend to be fairly slow and do not yet have the accuracy of a normal human.

He introduced the notion of a convolutional net:

- Much faster and more accurate than other methods
- Still requires more nodes than would be expected form the brain
- Still uses supervised learning.

Next, Dr. Lecun discussed a number of other topics including minimizing the ‘energy function’ (because it is the same formulation as the Gibbs free energy from statistical mechanics) of discrepancies between the input and the current representation. Further, he explained that harsh non-linearities such as the abs function were needed to dramatically improve the accuracy of convolutional nets (i.e., some form of rectification).

Comments, Observations, Questions

I had a question about his energy minimization technique. Basically, the approach he takes in some sense minimizes the L2-normed distance on a pixel-by-pixel basis between the input image and the current representation. However, does the brain compare the veridical input to an existing representation? I believe that it is one possibility, though not the only one. For example, error monitoring (which is presumably what is meant by the energy function) could as easily compare the contents of working memory with that of long term memory. In other words, the Lecun idea is that the brain is computing Min $\{E(Y,Z_i)\}$ where Y is the (veridical) input and Z_i the current representation. The hope, of course, is the $Z_i \rightarrow Y$ as $i \rightarrow \mathbb{N}$. If Y is input veridically, however, why does the brain need to store a representation? The answer is that really we are interested in a sequence such as $Z_i \rightarrow Y^*$ where Y^* is an object and not the input data. Specifically, Y is a collection of input waveforms impinging on the retina, but Y^* is the object called ‘8’. This requires a different formulation of his algorithm. One suggestion is that we replace Y with Z_{i-1} (or, possible, by some function of previous estimates of Z). If this were done, then the criterion of energy minimization would, in fact, refer to perceptual stability. Would an algorithm based on perceptual stability learn objects? Would it learn efficiently? Would it work effectively in a multi-layered network (which is pretty clearly the way the brain processes vision)? I don’t know, but it feels like a more realistic model.