Are There Deep Reasons Underlying the Pathologies of Today's Deep Learning Algorithms?

Ben Goertzel(✉)

OpenCog Foundation, Tai Po, Hong Kong
ben@goertzel.org

Abstract. Some currently popular and successful deep learning architectures display certain pathological behaviors (e.g. confidently classifying random data as belonging to a familiar category of nonrandom images; and misclassifying miniscule perturbations of correctly classified images). It is hypothesized that these behaviors are tied with limitations in the internal representations learned by these architectures, and that these same limitations would inhibit integration of these architectures into heterogeneous multi-component AGI architectures. It is suggested that these issues can be worked around by developing deep learning architectures that internally form states homologous to image-grammar decompositions of observed entities and events.

1 Introduction

In recent years "deep learning" architectures – specifically, systems that roughly emulate the visual or auditory cortex, with a goal of carrying out image or video or sound processing tasks – have been getting a lot of attention both in the scientific community and the popular media. The attention this work has received has largely been justified, due to the dramatic practical successes of some of the research involved. In image classification, in particular (the problem of identifying what kind of object is shown in a picture, or which person's face is shown in a picture), deep learning methods have been very successful, coming reasonably close to human performance in various contexts. Current deep learning systems can be trained by either supervised or unsupervised methods, but it's the supervised-learning approaches that have been getting the great results and headlines. Two good summaries of the state of the art are Juergen Schmidhuber's recent review with 888 references [13], and the in-process textbook by Yoshua Bengio and his colleagues [1].

The precise definition of "deep learning" is not very clear, and the term seems to get wider and wider as it gets more popular. Broadly, I think it works to consider a deep learning system as a learning system consisting of adaptive units on multiple layers, where the higher level units recognize patterns in the outputs of the lower level units, and also exert some control over these lower-level units. A variety of deep learning architectures exist, including multiple sorts of neural

J. Bieger (Ed.): AGI 2015, LNAI 9205, pp. 70–79, 2015.
DOI: 10.1007/978-3-319-21365-1_8

nets (that try to emulate the brain at various levels of precision), probabilistic algorithms like Deep Boltzmann machines, and many others. This kind of work has been going on since the middle of the last century. But only recently, due to the presence of large amounts of relatively inexpensive computing power and large amounts of freely available data for training learning algorithms, have such algorithms really begun to bear amazing practical fruit.

A paper by Stanford and Google researchers [8], which reported work using a deep learning neural network to recognize patterns in YouTube videos, received remarkable press attention in 2012. One of the researchers was Andrew Ng, who in 2014 was hired by Baidu to lead up their deep learning team. This work yielded some fascinating examples most famously, it recognized a visual pattern that looked remarkably like a cat. This is striking because of the well-known prevalence of funny cat videos on Youtube. The software's overall accuracy at recognizing patterns in videos was not particularly high, but the preliminary results showed exciting potential.

Another dramatic success was when Facebook, in mid-2014, reported that they had used a deep learning system to identify faces in pictures with over 97% accuracy [15] – essentially as high as human beings can do. The core of their system was a Convolutional Neural Network (CNN), a pretty straightforward textbook algorithm that bears only very loose conceptual resemblance to anything "neural". Rather than making algorithmic innovations, the main step the Facebook engineers took was to implement their CNN on massive scale and with massive training data. A Chinese team has since achieved even higher accuracies than Facebook on standard face recognition benchmarks, though they also point out that their algorithm misses some cases that most humans would get correctly [16].

Deep learning approaches to audition have also been very successful recently. For a long time the most effective approach to speech-to-text was a relatively simple technique known as "Hidden Markov Models" or HMMs. HMMs appear to underlie the technology of Nuance, the 800-pound gorilla of speech-to-text companies. But in 2013 Microsoft Research published a paper indicating their deep learning speech-to-text system could outperform HMMs [2]. In December 2014 Andrew Ng's group at Baidu announced a breakthrough in speech processing – a system called Deep Speech, which reportedly gives drastically fewer errors than previous systems in use by Apple, Google and others [7].

With all these exciting results, it's understandable that many commentators and even some researchers have begun to think that current deep learning architectures may be the key to advanced and even human-level AGI. However, my main goal in this article is to argue, conceptually, why this probably isn't the case. I will raise two objections to the hypothesis:

1. Current deep learning architectures (even vaguely) mirror the structure and information-processing dynamics of – at best – only parts of the human brain, not the whole human brain
2. Some (and I conjecture nearly all) current deep learning architectures display certain pathological behaviors (e.g. confidently classifying random data as

belonging to a familiar category of nonrandom images; and misclassifying miniscule perturbations of correctly classified images), which seem to be traceable to the nature of their internal knowledge representation. In this sense they seem not to robustly mirror the information-processing dynamics of the parts of the brain they resemble most, the visual and auditory cortex

My core thesis here is that these two objections are interconnected. I hypothesize that the pathological behaviors are rooted in shortcomings in the internal (learned) representations of popular deep learning architectures, and these shortcomings also make it difficult to connect these architectures with other AI components to form integrated systems better resembling the architecturally heterogeneous, integrative nature of the human brain.

I will also give some suggestions as to possible remedies for these problems.

2 Broad and Narrow Interpretations of "Deep Learning"

In his book"Deep Learning" [12], cognitive scientist Stellan Ohlsson formulates the concept of deep learning as a general set of information-processing principles. He also makes clear that these principles could be implemented in many different kinds of systems, including neural networks but also including logic systems or production rule systems or many other possibilities:

- **Spontaneous activity:** The cognitive system is constantly doing things, always processing inputs if they are there, and always reprocessing various of its representation of its inputs
- **Structured, unbounded representations:** Representations are generally built out of other representations, giving a hierarchy of representations. The lowest level representations are not fixed but are ongoingly reshaped based on experience
- **Layered, feedforward processing:** Representations are created via layers of processing units, with information passing from lower layers up to higher layers
- **Selective, capacity-limited processing:** Processing units on each layer pass information upward selectively each one generally passes up less information than it takes in, and doesn't pass it everywhere that it could
- **Ubiquitous monotonic learning:** Some of the representations the system learns are stored in long term memory, others aren't
- **Local coherence and latent conflict:** The various representations learned by a system don't have to be consistent with each other overall. Consistency is worked toward locally when inconsistencies between elements are found; there's no requirement of global consistency.
- **Feedback and point changes:** Higher level processing units feed information down to lower level units, thus potentially affecting their dynamics
- **Amplified propagation of point changes:** A small change anywhere in the processing hierarchy might cause a large change elsewhere in the system – as typical of complex and "chaotic" dynamical systems

- **Interpretation and manifest conflict:** Conflict between representations may go unnoticed until a particular input comes in, which then reveals that two previously learned representations can be in conflict
- **Competitive evaluation and cognitive utility:** Conflict between representations are resolved broadly via "reinforcement learning", i.e. based on which representation proves most useful to the overall system in which context

In the context of my own AI work with the OpenCog AGI architecture [5] [6], I find it interesting to note that, of Ohlsson's principles of deep learning, only one ("Representations are created via layers of processing units") does not apply to OpenCog's AtomSpace knowledge store, a heterogeneously structured weighted, labeled hypergraph. So to turn OpenCog into a deep learning system in Ohlsson's sense, it would suffice to arrange some OpenCog Nodes into layers of processing units. Then the various OpenCog learning dynamics including, e.g. Probabilistic Logic Networks reasoning, which is very different in spirit from currently popular deep learning architectures would become "deep learning" dynamics.

Of course, restricting the network architecture to be a hierarchy doesn't actually make the learning or the network any more deep. A more freely structured hypergraph like the general OpenCog Atomspace is just as deep as a deep learning network, and has just as much (or more) complex dynamics. The point of hierarchical architectures for visual and auditory data processing is mainly that, in these particular sensory data processing domains, one is dealing with information that has a pretty strict hierarchical structure to it. It's very natural to decompose a picture into subregions, subsubregions and so forth; and to define an interval of time (in which e.g. sound or video occurs) into subintervals of times. As we are dealing with space and time which have natural geometric structures, we can make a fixed processing-unit hierarchy that matches the structure of space and time lower-down units in the hierarchy dealing with smaller spatiotemporal regions; parent units dealing with regions that include the regions dealt with by their children; etc. For this kind of spatiotemporal data processing, a fairly rigid hierarchical structure makes a lot of sense (and seems to be what the brain uses). For other kinds of data, like the semantics of natural language or abstract philosophical thinking or even thinking about emotions and social relationships, this kind of rigid hierarchical structure seems much less useful, and in my view a more freely-structured architecture may be more appropriate.

In the human brain, it seems the visual and auditory cortices have a very strong hierarchical pattern of connectivity and information flow, whereas the olfactory cortex has more of a wildly tangled-up, "combinatory" pattern. This combinatory pattern of neural connectivity helps the olfactory cortex to recognize smells using complex, chaotic dynamics, in which each smell represents an "attractor state" of the oflactory cortex's nonlinear dynamics (as neuroscientist Walter Freeman has argued in a body of work spanning decades [10]). The portions of the cortex dealing with abstract cognition have a mix of hierarchical and combinatory connectivity patterns, probably reflecting the fact that they do both hierarchy-focused pattern recognition as we see in vision and audition, and attractor-based pattern recognition as we see in olfaction. But this is largely speculation most likely, until we can make movies somehow of the neural dynamics corresponding to various kinds of cognition, we won't really know how these various structural and dynamical patterns come together to yield human thinking.

My own view is that for anything resembling a standard 2015-style deep learning system (say, a convolutional neural net, stacked autoencoder, etc.) to achieve anything like human-level intelligence, major additions would have to be made, involving various components that mix hierarchical and more heterogeneous network structures in various ways. For example: Take "episodic memory" (your life story, and the events in it), as opposed to less complex types of memory. The human brain is known to deal with the episodic memory quite differently from the memory of images, facts, or actions. Nothing, in currently popular architectures commonly labeled "deep learning", tells you anything about how episodic memory works. Some deep learning researchers (based on my personal experience in numerous conversations with them!) would argue that the ability to deal with episodic memories effectively will just emerge from their hierarchies, if their systems are given enough perceptual experience. It's hard to definitively prove this is wrong, because these models are all complex dynamical systems, which makes it difficult to precisely predict their behavior. Still, according to the best current neuroscience knowledge [3], the brain doesn't appear to work this way; episodic memory has its own architecture, different in specifics from the architectures of visual or auditory perception. I suspect that if one wanted to build a primarily brain-like AGI system, one would need to design (not necessarily strictly hierarchical) circuits for episodic memory, plus dozens to hundreds of other specialized subsystems.

3 Pathologies of Contemporary Deep Learning Architectures

Even if current deep learning architectures are limited in scope, they could still be ideal solutions for certain aspects of the AGI problem, e.g. visual and auditory data processing. In fact, though, they seem to be subject to certain pathologies – and these pathologies seem (though have not been demonstrated) to be related to properties that would make it difficult to integrate these architectures into multi-component AGI architectures.

In a paper titled "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images" [11], one group of researchers showed they could construct images that looked random to the human eye, but that were classified by a CNN deep learning vision network as representing particular kinds of objects, with high confidence. So, a picture that looks like random noise to any person, might look exactly like a frog or a cup to the CNN. We may call this the **random images pathology**.

Another group, in a paper titled "Intriguing properties of neural networks" [14], showed that by making a very small perturbation to a correctly classified

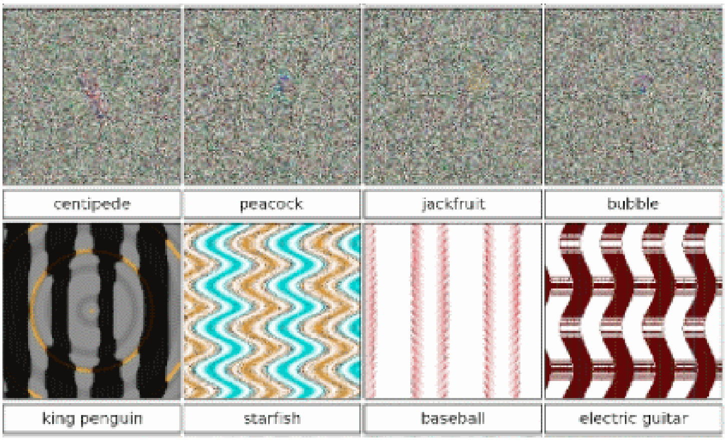

Fig. 1. From Examples of images that are unrecognizable to humans, but that state-of-the-art deep neural networks trained on the standard ImageNet image collection believe with $\geq 99.6\%$ certainty to be a familiar object. From [11].

image, they could cause the deep network to misclassify the image. The perturbations in question were so small that humans wouldn't even notice. We may call this the **brittleness pathology**.

Now, these two odd phenomena have no impact on practical performance of convolutional neural networks. So one could view them as just being mathematical pathologies found by computer science geeks with too much time on their hands. The first pathology is pragmatically irrelevant because a real-world vision system is very unlikely to ever be shown weird random pictures that just happen to trick it into thinking it's looking at some object (most weird random pictures won't look like anything to it). The second one is pragmatically irrelevant because the variations of correctly classified pictures that will be strangely misclassified, are very few in number. Most variations would be correctly classified. So these pathologies will not significantly affect classification accuracy statistics. Further, these pathologies have only been demonstrated for CNNs – I suspect they are not unique to CNNs and would also occur for other currently popular deep learning architectures like stacked autoencoders but this has not been demonstrated.

But I think these pathologies are telling us something. They are telling us that, fundamentally, these deep learning algorithms are not generalizing the way that people do. They are not classifying images based on the same kinds of patterns that people are. They are "overfitting" in a very subtle way not overfitting to the datasets on which they've been trained, but rather overfitting to the kind of problem they've been posed. In these examples, these deep networks have been asked to learn models with high classification accuracy on image databases and they have done so. They have not been asked to learn models

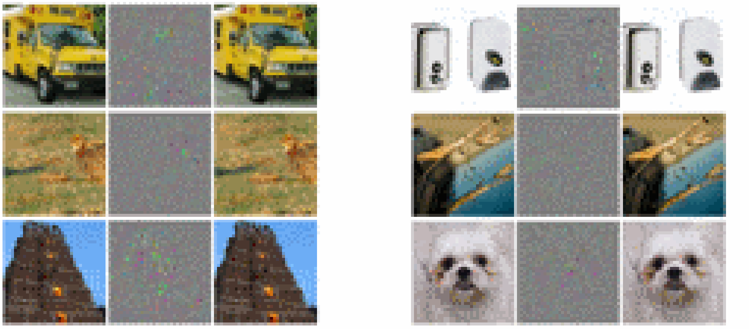

Fig. 2. All images in the right column are incorrectly classified as ostriches by the CNN in question. The images in the left column are correctly classified. The middle column shows the difference between the left and right column. From [14].

that capture patterns in images in a more generally useful way, that would be helpful beyond the image classification task and so they have not done that.

When a human recognizes an image as containing a dog, it recognizes the eyes, ears and nose and fur, for example. Because of this, if a human recognized the image on the bottom left of the right image array in Figure 188 as a dog, it would surely recognize the image on the bottom right of the right image array as a dog as well. But a CNN is recognizing the bottom left image differently than a human in a way that fundamentally generalizes differently, even if this difference is essentially irrelevant for image classification accuracy.

I strongly suspect there is a theorem lurking here, stating in some way that these kinds of conceptually pathological classification errors will occur if and only if the classification model learning algorithm fails to recognize the commonly humanly recognizable high level features of the image (e.g. eyes, ears, nose, fur in the dog example). Informally, what I suspect is: The reason these pathologies occur is that these deep networks are not recognizing the "intuitively right" patterns in the images. They are achieving accurate classification by finding clever combinations of visual features that let them distinguish one kind of picture from another but these clever combinations don't include a humanly meaningful decomposition of the image into component parts, which is the kind of "hierarchical deep pattern recognition" a human's brain does on looking at a picture.

There are other kinds of AI computer vision algorithms that do a better job of decomposing images into parts in an intuitive way. Stochastic image grammars [17] are one good example. However, these algorithms are more complicated and more difficult to implement scalably than CNNs and other currently popular deep learning algorithms, and so they have not yet yielded equally high quality image classification results. They are currently being developed only minimally, whereas CNNs and their ilk are being extremely heavily funded in the tech industry.

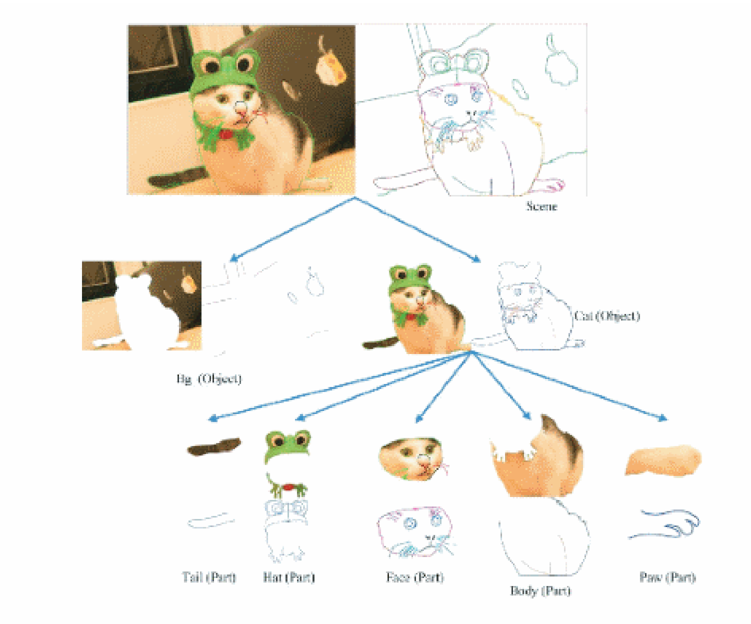

Fig. 3. Illustrative example of an image grammar for a simple object. Image grammar based methods have been used for object classification as well, though not yet with comparable accuracy to, say, CNNs or stacked autoencoders. From [11].

Connecting these different threads of research I suggest that the pathological results noted above would occur even on corpora generated by formal image grammars:

Proposition 1. *Suppose one generated a large corpus of images, falling into N commonsensical categories, based on a moderately complex, but formally defined image grammar. Then training current deep learning architectures on this corpus would yield the brittleness and random images pathologies.*

If true, this could be useful for studying the pathologies and how to eliminate them, especially in conjunction with the proposition suggested below.

4 A Possible Way Out

How then could these pathologies be avoided, staying within the general deep learning framework? And would avoiding these pathologies actually give any practical benefit?

I believe, but have not rigorously shown, that there is a sensible and viable way to bypass the random image and brittleness pathologies, not via any clever tricks but via modifying deep learning algorithms to make them create more sensible internal knowledge representations. Specifically, I suggest:

Proposition 2. *For a deep learning hierarchy to avoid the brittleness and random images pathologies (on a corpus generated from an image grammar, or on a corpus of natural images), there would need to be a reasonably straightforward mapping from recognizable activity patterns on the different layers, to elements of a reasonably simple image grammar, so that via looking at the activity patterns on each layer when the network was exposed to a certain image, one could read out the "image grammar decomposition" of the elements of the image. For instance, if one applied the deep learning network to a corpus images generated from a commonsensical image grammar, then the deep learning system would need to learn an internal state in reaction to an image, from which the image-grammar decomposition of the image was easily decipherable.*

As stated this is an intuitive rather than formal proposition. Approaches to formalization will be interesting to explore.

If this hypothesis is conceptually correct, then one interesting research direction might be to generate corpora using image grammars, and see what it would take to get a deep learning algorithm to learn the image grammar from the corpus, in the sense of emerging a structure in which the image grammar is observable. Once this worked, the same algorithm could be applied to natural-image corpora and the results analyzed.

My colleagues and I have pursued one approach to making a deep learning network capable of learning an internal image grammar. In this approach, reported in [4], the states of the DeSTIN deep learning algorithm are saved and frequent patterns in the state-set are mined. A DeSTIN network state may then be labeled with the frequent patterns from the learned pattern-library that are instantiated in that state. These labels, in simple cases, appear function like an image grammar. But it is not clear how general or robust this phenomenon is; this requires further study.

Another question is whether difference target propagation, as proposed in [9], might display the property suggested in Proposition 2. Difference target propagation seeks to minimize reconstruction error at each level in a deep hierarchy (as opposed to propagating error backwards from the top of a network as in standard gradient descent methods). Whether, and under what circumstances, this may cause formation of a meaningful image grammar inside a network's state, is a fascinating open question.

References

1. Bengio, Y., Goodfellow, I.J., Courville, A.: Deep learning (2015). http://www.iro.umontreal.ca/bengioy/dlbook, book in preparation for MIT Press
2. Deng, L., Li, J., Huang, J.T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y., Acero, A.: Recent advances in deep learning for speech research at microsoft. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2013)
3. Gazzaniga, M.S., Ivry, R.B., Mangun, G.R.: Cognitive Neuroscience: The Biology of the Mind. W W Norton (2009)

4. Goertzel, B.: Perception Processing for General Intelligence: Bridging the Symbolic/Subsymbolic Gap. In: Bach, J., Goertzel, B., Iklé, M. (eds.) AGI 2012. LNCS, vol. 7716, pp. 79–88. Springer, Heidelberg (2012)
5. Goertzel, B., Pennachin, C., Geisweiller, N.: Engineering General Intelligence, Part 1: A Path to Advanced AGI via Embodied Learning and Cognitive Synergy. Springer, Atlantis Thinking Machines (2013)
6. Goertzel, B., Pennachin, C., Geisweiller, N.: Engineering General Intelligence, Part 2: The CogPrime Architecture for Integrative, Embodied AGI. Springer, Atlantis Thinking Machines (2013)
7. Hannun, A.Y., Case, C., Casper, J., Catanzaro, B.C., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., Ng, A.Y.: Deep speech: Scaling up end-to-end speech recognition. CoRR abs/1412.5567 (2014). http://arxiv.org/abs/1412.5567
8. Le, Q.V., Ranzato, M., Monga, R., Matthieu Devin, K.C., Corrado, G.S., Dean, J., Ng., A.Y.: Building high-level features using large scale unsupervised learning. In: Proceedings of the Twenty-Ninth International Conference on Machine Learning (2012)
9. Lee, D., Zhang, S., Biard, A., Bengio, Y.: Target propagation. CoRR abs/1412.7525 (2014). http://arxiv.org/abs/1412.7525
10. Li, G., Lou, Z., Wang, L., Li, X., Freeman, W.J.: Application of chaotic neural model based on olfactory system on pattern recognition. ICNC **1**, 378–381 (2005)
11. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. CoRR abs/1412.1897 (2014). http://arxiv.org/abs/1412.1897
12. Ohlsson, S.: Deep Learning: How the Mind Overrides Experience. Cambridge University Press (2006)
13. Schmidhuber, J.: Deep learning in neural networks: An overview. CoRR abs/1404.7828 (2014). http://arxiv.org/abs/1404.7828
14. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. CoRR abs/1312.6199 (2013). http://arxiv.org/abs/1312.6199
15. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
16. Zhou, E., Cao, Z., Yin, Q.: Naive-deep face recognition: Touching the limit of lfw benchmark or not? (2014). http://arxiv.org/abs/1501.04690
17. Zhu, S.C., Mumford, D.: A stochastic grammar of images. Found. Trends. Comput. Graph. Vis. **2**(4), 259–362 (2006). http://dx.doi.org/10.1561/0600000018