CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 088 **May 10, 2018**

Deep Nets: What have they ever done for Vision?

Alan L. Yuille[1], **Chenxi Liu**[1]
1: Johns Hopkins University

Abstract

This is an opinion paper about the strengths and weaknesses of Deep Nets. They are at the center of recent progress on Artificial Intelligence and are of growing importance in Cognitive Science and Neuroscience since they enable the development of computational models that can deal with a large range of visually realistic stimuli and visual tasks. They have clear limitations but they also have enormous successes. There is also gradual, though incomplete, understanding of their inner workings. It seems unlikely that Deep Nets in their current form will be the best long-term solution either for building general purpose intelligent machines or for understanding the mind/brain, but it is likely that many aspects of them will remain. At present Deep Nets do very well on specific types of visual tasks and on specific benchmarked datasets. But Deep Nets are much less general purpose, flexible, and adaptive than the human visual system. Moreover, methods like Deep Nets may run into fundamental difficulties when faced with the enormous complexity of natural images. To illustrate our main points, while keeping the references small, this paper is slightly biased towards work from our group.

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

Deep Nets: What have they ever done for Vision?*

Alan L. Yuille[1,2] Chenxi Liu[2]
Department of Cognitive Science[1] & Computer Science[2]
Johns Hopkins University

May 9, 2018

This is an opinion paper about the strengths and weaknesses of Deep Nets. They are at the center of recent progress on Artificial Intelligence and are of growing importance in Cognitive Science and Neuroscience since they enable the development of computational models that can deal with a large range of visually realistic stimuli and visual tasks. They have clear limitations but they also have enormous successes. There is also gradual, though incomplete, understanding of their inner workings. It seems unlikely that Deep Nets in their current form will be the best long-term solution either for building general purpose intelligent machines or for understanding the mind/brain, but it is likely that many aspects of them will remain. At present Deep Nets do very well on specific types of visual tasks and on specific benchmarked datasets. But Deep Nets are much less general purpose, flexible, and adaptive than the human visual system. Moreover, methods like Deep Nets may run into fundamental difficulties when faced with the enormous complexity of natural images. To illustrate our main points, while keeping the references small, this paper is slightly biased towards work from our group.

Some History

We are in the third wave of neural network approaches. The first two waves – 1950s-1960s and 1980s-1990s – generated considerable excitement but slowly ran out of steam. Despite some notable exceptions, the overall performance of neural networks was disappointing both for machines (Artificial Intelligence/Machine Learning), brains (Neuroscience), and minds (Cognitive Science, Psychology). The third wave – 2000s-present – is distinguished because of the dramatic success of Deep Nets on many large benchmarked problems and their industrial application to real world tasks (e.g., face recognition on datasets containing tens of millions of different people). It should be emphasized that several of the most successful models were developed during the second wave, but their strengths were not appreciated until the third wave due to the availability of

*For those readers unfamiliar with Monty Python see: `https://youtu.be/Qc7HmhrgTuQ`

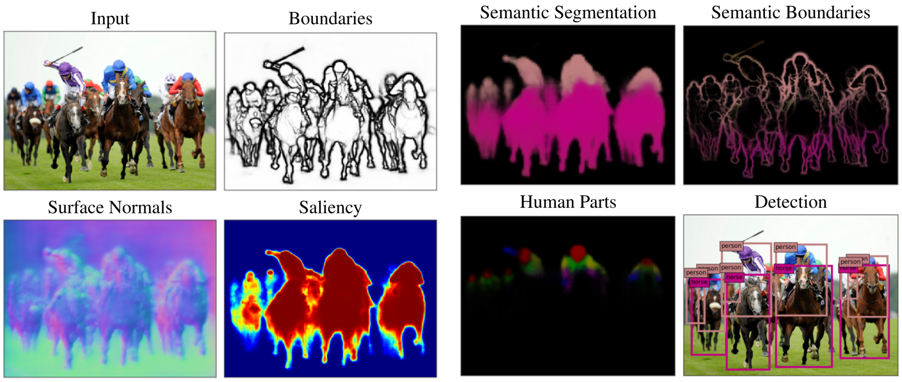
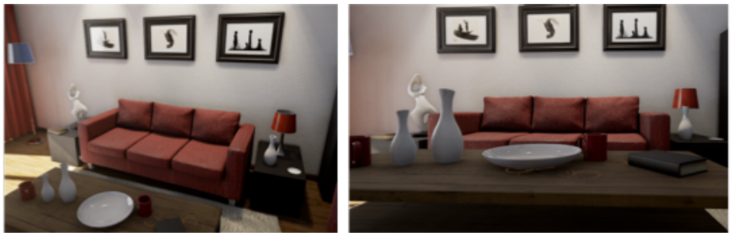


Figure 1: Figure taken from Kokkinos (2017). A wide variety of vision tasks can be performed by Deep Nets. These include: boundary detection, semantic segmentation, semantic boundaries, surface normals, saliency, human parts, and object detection.

big datasets and the ubiquity of powerful computers which can be used to train them (e.g., GPUs).

As a side comment, the history of neural networks is a good illustration of how science is often driven by intellectual fashions. Neural networks often served as a counter-point to the more logic and structured-representation based approaches to Artificial Intelligence and Cognitive Science which had dominated in the 1960s-1980s. When representations were popular then deep networks were not, and vice versa. Indeed it became hard to get neural network research published during the gap between the second and third wave, and credit is due to those neural network researchers who carried on despite discouragement. Conversely other researchers whose work did not fit with neural network, or connectionist ideas, had difficulty getting their work accepted during the second wave of neural networks. We would argue for a middle way which combines the strengths of both types of approaches.

The relationship between neural networks to real neurons in the brain is extremely interesting since understanding real neural networks is a holy grain of neuroscience. But this relationship should be treated with caution, unless expert neuroscientists are involved. Real neurons are much more varied and complex than artificial neurons (there are fifty different types of neurons in the retina alone and also a range of different morphological structures) and important properties such as the neural code are only partially understood. In the short term, it may be best to think of artificial neural networks as a way of doing statistics (as researchers at early neural network meetings started speculating as we shared ski lifts), for example by interpreting Deep Nets as a sophisticated form of probabilistic regression, instead of as models of real neurons. “Neuronal plausibility” is a desirable property of a cognitive system but it is not always easy to pin down.



Azimuth Elevation	90	135	180	225	270
0	-	0.713	0.769	0.930	0.319
30	0.900	1.000	0.588	1.000	0.710
60	0.255	0.100	0.148	0.296	0.649

Figure 2: Figure taken from Qiu and Yuille (2016). UnrealCV allows vision researchers to easily manipulate synthetic scenes, e.g. by changing the viewpoint of the sofa. We found that the Average Precision (AP) of Faster-RCNN (Ren et al., 2015) detection of the sofa varies from 0.1 to 1.0, showing extreme sensitivity to viewpoint. This is perhaps because the biases in the training cause Faster-RCNN to favor specific viewpoints.

1 The Successes: Specific vision tasks and on specific benchmarked datasets

From the vision perspective, the performance of Deep Nets for classifying objects (Krizhevsky et al., 2012) in ImageNet (Deng et al., 2009) was very dramatic and persuaded researchers that Deep Nets should be taken seriously (many computer vision researchers had previously been skeptical about neural nets). The object classification task assumes a foreground object which is surrounded by a limited background region and the goal is to classify the object. The input would be similar to one of the red boxes of the bottom right image in Figure 1. The performance of Deep Nets on ImageNet has kept increasing as researchers have explored variants of Deep Net architectures (Simonyan and Zisserman, 2015; He et al., 2016).

Deep Nets were soon adapted to other visual tasks such as object detection where the image contains one or more objects and the background is much larger, e.g, the PASCAL challenge (Everingham et al., 2010). For this task, Deep Nets were augmented by an initial stage which made proposals for possible positions and sizes of the objects and then applied Deep Nets to classify the proposals (current methods train the proposals and objects together in what is called “end-to-end”). These methods outperformed the previous best methods, the Deformable Part Models (Felzenszwalb et al., 2010), for the PASCAL object detection challenge (PASCAL was the main object detection and classification challenge before ImageNet). Other Deep Nets architectures also gave enormous performance jumps in other classic tasks like edge detection, semantic segmen-

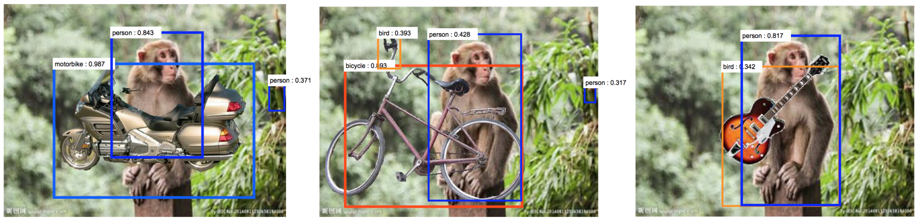


Figure 3: Figure taken from Wang et al. (2018). Adding occluders cause deep network to fail. Left Panel: The occluding motorbike turns a monkey into a human. Center Panel: The occluding bicycle turns a monkey into a human and the jungle turns the bicycle handle into a bird. Right Panel: The occluding guitar turns the monkey into a human and the jungle turns the guitar into a bird.

tation, occlusion detection (edge detection with border-ownership), symmetry axis detection. Major increases also happened for human joint detection, human segmentation, binocular stereo, 3D depth estimation from single images, and scene classification. Many of these tasks are illustrated in Figure 1.

Two main points should be made: (1) Deep Nets are designed for specific visual tasks (although "transfer learning" sometimes enables them to adapt to other tasks, see later section). There is no single Deep Net which does all of these tasks, but recent Deep Nets, like UberNet (Kokkinos, 2017), can do several tasks. A Deep Net designed for object classification on ImageNet cannot perform human parsing (i.e. joint detection) on the Leeds Sports Dataset (LSD). (2) Deep Net performance on benchmarked datasets, no matter how large, may fail to extend to good performance images outside the dataset. This is partly due to the enormous complexity of natural images and the difficulty of having datasets which are unbiased and which are representative of this complexity. This is a very important and fundamental issue, which we will return to in Section 6. As can be seen from Figure 2, a Deep Net trained to detect sofas on ImageNet may fail to detect them if shown from certain viewpoints which were underrepresented in the training dataset. In practice, when researchers want to apply Deep Nets for object classification to a new dataset like PASCAL then they will typically re-train the Deep Net (initializing its weights by the results from ImageNet). Deep Nets can also fail if the context is changed by, for example, giving a man-made object to a monkey as shown in Figure 3.

In short, when discussing Deep Net performance we need to take into account the dataset on which the Deep Net has been trained and tested as well as the visual task being performed. In general, the larger the training datasets the better the performance unless the dataset is badly biased. Hence high tech companies have big advantages over universities since they have access to enormous datasets and have the computer power to train increasingly complicated Deep Nets.

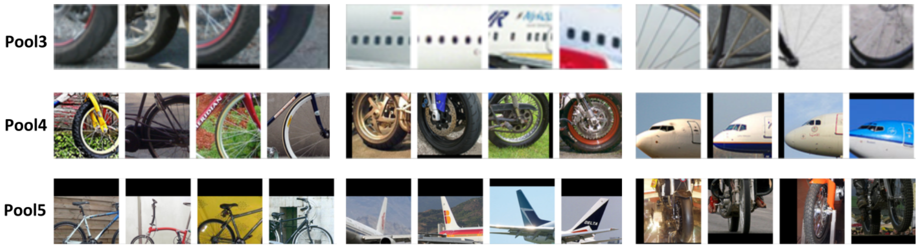


Figure 4: Figure taken from Wang et al. (2015). The visual concepts obtained by population encoding are visually tight and we can identify the parent object class pretty easily by just looking at the mid-level concepts.

2 Towards Understanding Deep Nets

Although Deep Nets are difficult to understand there has been some success at understanding the activities of the internal filters/features of the convolutional levels of Deep Nets (recall that most Deep Net architectures have a sequence of convolutional layers followed by a multi-layer perceptron). Early studies visualized the activities of Deep Net filters and showed that several appeared to be tuned to image properties (Zeiler and Fergus, 2014; Yosinski et al., 2015). In particular, if Deep Nets are trained for scene classification then some convolutional layer filters correspond to objects which appear frequently in the scene, while if the Deep Nets are trained for object detection, then some features correspond to parts of the objects (Zhou et al., 2015). In detailed studies of a restricted subset of objects (e.g., vehicles), researchers (Wang et al., 2015) discovered regular patterns of activity of the feature vectors, called visual concepts, which corresponded approximately to the semantic parts of objects (with sensitivity to viewpoint), see Figure 4.

This suggests the following rough conceptual picture of Deep Nets. The convolutional levels represent the manifold of intensity patterns at different levels of abstraction. The lowest levels represent local image patterns while the high levels represent larger patterns which are invariant to the details of the intensity patterns. From a related perspective, the feature vectors represent a dictionary of templates of image patterns. The multi-layer perceptron, at the top layers of the Deep Nets are harder to interpret, but it is plausible that they make decisions based on the templates represented by the convolutional layers. This "dictionary of templates" interpretation of Deep Nets suggests they are very efficient to learn and represent an enormous variety of image patterns, but cannot extrapolate much beyond the patterns they have seen in their training dataset. This also suggests that they are less useful when applied to model visual properties which are specified purely by geometry, where the input consists of sparse edges (or binary valued patterns). Although the number of possible sparse edge patterns are very large they are nowhere near as large as the possible number of image intensity patterns.

An alternative, and perhaps more standard, interpretation of Deep Nets is that they act as universal function approximators (Hornik et al., 1989). This interpretation that dates back to the second wave of neural networks is correct, provided there are enough hidden units, but such results offer little intuition and are of limited utility since for some functions the number of hidden units needed would be truly enormous.

3 Transfer Learning, Learning with fewer examples and less supervision

A disadvantage of Deep Nets is that they typically need a very large amount of annotated (i.e. fully supervised) training data, which restricts their use to situations where big data is available. But this is not always the case. In particular, "transfer learning" shows that the features of Deep Nets learned on annotated datasets for certain visual tasks can sometimes be transferred to novel datasets and related tasks. Thereby enabling learning with much less data and sometimes with less supervision.

For example, as mentioned earlier, Deep Nets were first successful for object classification on ImageNet but failed on object detection on the smaller PASCAL dataset. This was presumably because PASCAL was not big enough to train a Deep Net while ImageNet was (it was larger than PASCAL by almost two orders of magnitude). But researchers quickly realized that it was possible to train a Deep Net for object detection and semantic segmentation on PASCAL by initializing the weights of the Deep Net by the weights of a Deep Net trained on ImageNet (Girshick et al., 2014; Long et al., 2015; Chen et al., 2018). This also introduced a mechanism for generating proposals, see Figure 1 (bottom right).

More generally, researchers found that they could transfer features from Deep Nets trained on one task on one dataset to perform related tasks on a second dataset. In some cases, this consisted of simply using the first dataset to initialize the weights when training on the second (so that the final values of the weights, particularly for the higher levels, may have little to do with their initial values) while in other situations, the weights changed little and were similar for both tasks and/or datasets. For example, researchers showed that Deep Nets trained for face verification could be transferred to the related task of facial pain estimation (Wang et al., 2017). This is presumably because the two tasks required fairly similar image representations capturing the fine-scale appearance of facial features.

This ability to transfer Deep Net knowledge learned on another domain relates intuitively to how children learn. A child initially learns rather slowly compared to other young animals but at critical periods the child's learning accelerates very rapidly. From the "dictionary of templates" perspective, this could happen because after a child has learned to recognize enough objects he/she may have enough building blocks (i.e. deep network features/templates) to be able to represent new objects in terms of a dictionary of existing templates. If so, only a few examples of the new object may be needed in order to do few-shot learning.

Few-shot learning of novel object categories has been shown for Deep Nets provided they have first been trained on a large set of object categories (Mao et al., 2015; Vinyals et al., 2016; Qiao et al., 2018). Another strategy is to train a Deep Net to learn similarity (technically a *Siamese network*) on the set of object categories, hence obtaining a similarity measure for the new objects. For example, Lin et al. (2017) trained a Siamese network to learn similarity for objects in ShapeNet (Chang et al., 2015) and then this similarity measure was used to cluster objects in the Tufa dataset (Salakhutdinov et al., 2012). Other few-shot learning tasks can also be done by using features from Deep Nets trained for some other tasks as ways to model the visual patterns of objects.

An alternative way to make Deep Nets more generally applicable is to weaken the amount of supervision required. For example, to train object detection using images where only the names of the objects in the image are known but their locations and sizes are unknown. This is known as weakly supervised learning and it can be treated as missing/hidden data problem which can be addressed by methods such as Expectation-Maximization (EM) or Multiple Instance Learning (MIL). Performance of these types of methods is often improved by using a small amount of fully supervised training data which helps the EM or MIL algorithms converge to good solutions, e.g., see Papandreou et al. (2015).

4 Cognitive Science and Neuroscience

Deep Nets offer the possibility of developing computational theories for Cognitive Science and Neuroscience which can be tested on natural, or realistically synthetic, images. This enables researchers to take advantage of the advances in artificial intelligence and machine learning. It also offers the potential to make theories for complex phenomena, e.g., social interactions and for visual affordances, which can deal with the complexity of real world images.

In particular, Deep Nets have been used to predict brain activity, such as fMRI and other non-invasive measurements, for a range of visual tasks. A few examples are described in Cichy et al. (2016); Wen et al. (2017). Deep Nets have also been applied to predicting neural responses as measured by electro-physiology and, in particular, impressive results have been reported for predicting the response of neurons in the ventral stream (Yamins et al., 2014).

But despite these successes some caveats apply. The ventral stream of primates is complex and there is evidence that it estimates the three dimensional structure of objects and parts (Yamane et al., 2008), which relates to the classic theory of object recognition by component (Biederman, 1987). In general, primate visual systems must perform all the visual tasks listed in Section 1, namely edge detection, binocular stereo, semantic segmentation, object classification, scene classification, and 3D-depth estimation. The vision community has developed a range of different Deep Nets for these tasks so it is extremely unlikely, for example, for a Deep Net trained for object classification on ImageNet to be able to account for the richness of primate visual systems and, in particular, the ventral stream.

It should also be emphasized that while Deep Nets perform computations bottom-up in a feedforward manner there is considerable evidence of top-down processing in the brain (Lee and Mumford, 2003), particularly driven by top-down attention (Gregoriou et al., 2014). Researchers have also identified cortical circuits (McManus et al., 2011) which implement spatial interactions (though possibly in a bottom-up and top-down manner). These types of phenomena require other families of mathematical models, such as the compositional models described in Section 5.

It can also be questioned whether Deep Nets are really satisfactory as theories of Cognitive Science or Neuroscience. A common criticism is that they are merely black boxes and hence do not capture or explain internal representations or other cognitive processes. This echoes a similar criticism of their use for Artificial Intelligence applications. But, as with AI, this may only be a temporary limitation and that better understanding of Deep Nets, perhaps along the lines of Section 2, and the development of more interpretable, but equally effective, models will help alleviate it. Similar views were expressed in the second wave of neural networks (McCloskey, 1991). It remains possible that the best theory of the brain, or any really complex physical system, may only be a black box. But this seems too pessimistic and we would be very surprised if the long-term solution to AI or Neuroscience is an uninterpretable black box (partly due to scaling and diagnostic issues which will be discuss in Section 6).

One halfway measure, toward obtaining more interpretable theories, is to restrict Deep Nets to low-level visual tasks like the detection of object parts and use more explainable/understandable theories to model the relationships between parts. This is reminiscent of the early days of Artificial Intelligence when researchers like Minsky and Winston thought that high-level vision should be modeled symbolically but this required as a pre-requisite the ability to break the *signal-to-symbol barrier* to obtain elementary symbolic representations from realistic images (a very difficult tasks which Minsky severely underestimated and famously asked a summer student to solve). From this perspective, Deep Nets give a possible solution to the signal-to-symbol problem hence enabling researchers to build more interpretable symbolic models on top of them.

Perhaps most importantly, the use of Deep Nets offers the possibility to develop Cognitive Science and Neuroscience theories which can be tested on realistic images without having to restrict themselves to studying on simplified stimuli. Experimental and theoretical findings on simplified stimuli were historically very important, due to the need for experimental controls and the immense difficulty of working with natural stimuli. But they can also lead researchers to restrict themselves to working in *micro-worlds* and concentrate on issues that may not be most relevant when facing the core problem of vision: namely how does the visual system convert the incoming patterns of light rays into understanding the three-dimensional scene? Marr's criticism of early Artificial Intelligence – that it sometimes limited itself by a poor choice of micro-worlds

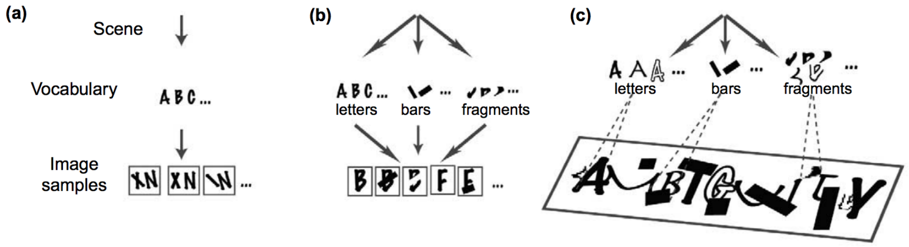


Figure 5: Figure taken from Yuille and Kersten (2006). From (a) to (b) to (c), an increasing level of variability and occlusion is used, yet humans can still do inference and correctly interpret the image.

– is relevant here. Despite their many limitations, the enormous progress in machine learning and Deep Nets in particular offer the possibility of developing visual theories that can address the core problem of vision directly. Moreover, it also offers tools to construct theories for complex visual phenomena, such as social interactions, which are otherwise seem very difficult to address.

5 Compositional Models: Grammars and Generative Models

There has been a long history of compositional and grammatical models of vision and the related idea of pattern theory (Zhu and Mumford, 2006; Mumford and Desolneux, 2010). These have many desirable theoretical properties, perhaps particularly from the Cognitive Science and Neuroscience perspective. But they have been less successful on vision benchmarked datasets and so have far less impact than Deep Nets.

Grenander (Grenander, 1996) was arguably the first to articulate modeling statistically the types of patterns that occur in images and to interpret an image by identifying the process that generated it, which he called analysis by synthesis. This idea can be illustrated by the simple microworlds from Yuille and Kersten (2006) shown in Figure 5. The three panels show microworlds of increasing complexity from left to right. For each microworld there is a grammar which specifies the possible images as constructed by compositions of the elementary components. In the left panel the elementary components are letters which do not overlap, and so interpreting the image is easy. The center and right panels are generated by more complicated grammars – letters of different fonts, bars, and fragments which can heavily occlude each other. Interpreting these images is much harder and seems to require the notion that letters are composed of elementary parts, that they can occur in a variety of fonts, and the notion of "explaining away" (to explain that parts of a letter are missing because they have been occluded by another letter).

Humans have little difficulty interpreting the image in Figure 5(c) even if the

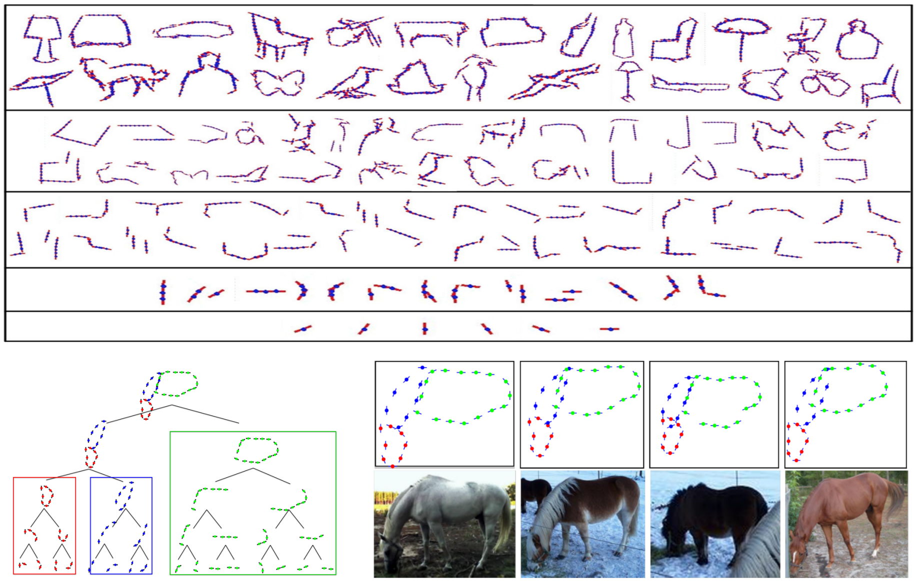


Figure 6: Top: Figure taken from Zhu et al. (2010). Mean shapes from Recursive Compositional Models at different levels. Bottom: Figure taken from Wang and Yuille (2015). One learned mixture with corresponding landmark localization of horse images.

letters are shown in fonts they have never seen before. Indeed images of these types are used on CAPTCHAs to distinguish between humans, who can interpret them, and robots who cannot. Recent work (George et al., 2017) describes a compositional model for reading CAPTCHAs which factorize geometry and appearances, enabling the geometry and appearance to be learned separately, hence saving on training data. Letters are represented explicitly in terms of parts so that they can still be detected even if some parts are occluded (this is harder for a "black box" like a Deep Net). In addition, the inference algorithm involves bottom-up and top-down processing which enables the algorithm to "explain away" missing parts of the letters and to impose "global consistency" of the interpretation to remove ambiguities. Intuitively, part detectors combine to make bottom-up proposals for letters which can be validated or rejected in the top-down stage. By contrast, the authors report that Deep Nets performed poorly on these tasks.

Compositional models have many desirable properties, such as being interpretable, and the ability to be *generative* so they can be samples from. This means that they know everything about the object (or whatever entity is being modeled) which makes them harder to fool than black box methods like Deep Nets. Learning compositional models is much harder than learning Deep Nets although researchers have been able to learn hierarchical dictionaries starting from basic elements (like edges) enable part-sharing and efficient learning, rep-

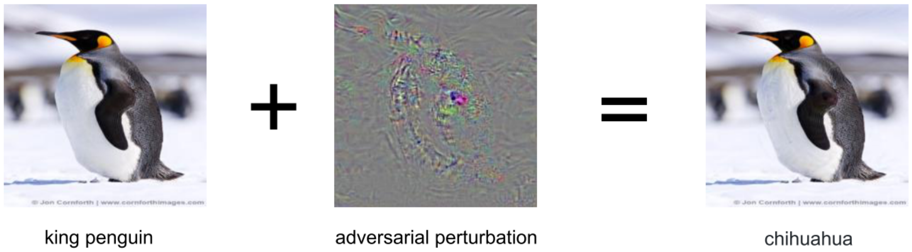


Figure 7: Figure taken from Xie et al. (2018). A deep network can correctly classify the left image as *king penguin*. The middle image is the adversarial noise magnified by 10 and shifted by 128, and on the right is the adversarial example misclassified as *chihuahua*.

resentation, and inference (Zhu et al., 2010) (see Figure 6). As we will discuss in Section 6, they offer a strategy to deal with the complexity problem of vision, i.e. the ability to perform an enormous number of visual tasks on a vast range of images. Theoretical studies (Yuille and Mottaghi, 2016) have analyzed the complexity of compositional models.

But many technical challenges remain. Addressing more challenging stimuli than letters, such as animals or vehicles, requires having richer appearance models which can factorize between shape and appearance, see Wang and Yuille (2015). But the increasing realism of virtual worlds constructed using computer graphics tools, e.g., see Figure 2, suggests we may soon have realistic generative models which make analysis by synthesis possible.

6 The Limits of Big Data? Adversaries and Exponential Explosions

As stated earlier, the evaluation of Deep Nets has been performed using Machine Learning evaluation protocols. These are based on the standard Statistics technique of checking performance of models by cross-validation (broadly speaking seeing if the results of the model trained on some of the data predict the results on the data on which the models have not been trained).

There are problems with these evaluation protocols which vision researchers have long been aware of. If the dataset is biased, i.e. unrepresentative of real world images, then algorithms which perform well on it will often fail to generalize to other datasets (unless they share the same biases). Several datasets have been discarded by the vision community after their biases became apparent (e.g., if object detection could be done by ignoring the objects and exploiting properties of the background) and when models which performed well on that dataset failed to generalize to other more challenging datasets (Torralba and Efros, 2011). A related problem is that datasets often represent typical events and do not adequately represent rare events. These rare events may be rela-

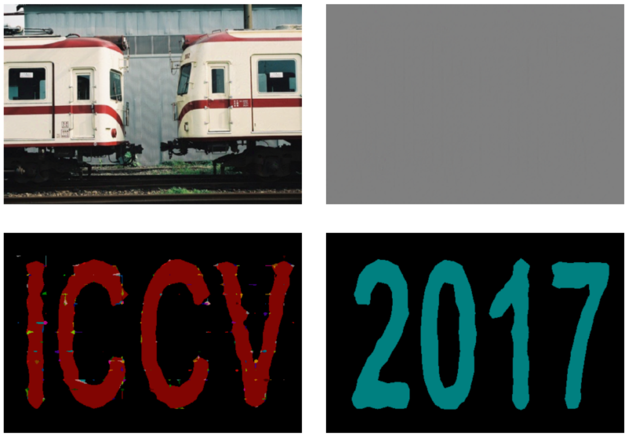


Figure 8: Figure taken from Xie et al. (2017). The top row is the input (adversarial perturbation already added) to the segmentation network, and the bottom row is the output. The red, blue and black regions are predicted as *airplane*, *bus* and *background*, respectively.

tively unimportant, like only having a few examples of a uncommon bird like a ptarmigan in a bird dataset. But they may also be highly important, like an infant running in front of a car in an autonomous driving dataset. Note that careful experimental design is difficult for vision datasets due to the complexity of visual stimuli and it is unclear how large datasets have to be to be representative of real world images. As shown earlier, even a very large dataset like ImageNet has biases and so “sofa-detectors” will fail if they are shown images from viewpoints under-represented in ImageNet, e.g., Qiu and Yuille (2016).

Another problem for evaluating vision algorithms are the recent studies showing that Deep Nets can be successfully attacked by tiny modifications of the images which nevertheless cause the Deep Nets to make major mistakes for object classification (Szegedy et al., 2014; Goodfellow et al., 2015), object detection, and semantic segmentation (Xie et al., 2017) (see Figure 7 and Figure 8). This problem partly arises because of the infinite space of possible images. These types of “attack images” almost never occur in real world datasets but these datasets, though enormous, are still only an infinitesimal subset of all images. To perform these attacks, the algorithm must exploit knowledge of the Deep Net and deliberatively search for images which cause these types of errors. (It is also possible that these attack images are extremely rare and only found because the researchers are looking for them). There are now strategies which defend against these attacks. One strategy is to treat these “attack images” as extra training data. A second recent alternative (Xie et al., 2018) is to introduce small random perturbations into the images, exploiting the assumption that the “attack images” are very unstable so small random perturbation will defend against them. It should be acknowledged that adversarial attacks can be mounted against any vision algorithm and that is a compliment to Deep Nets to study their ability to resist attacks (it would be much easier to successfully attack earlier vision algorithms).

But current adversary attacks may only be the tip of the iceberg and there may be even more serious problems in the evaluation of vision algorithms due to the enormous complexity of natural images and visual scenes. Recall that machine learning theory assumes that it is possible to have large, enormous benchmarked datasets for visual tasks to be representative of performing these tasks on the real world. But what happens if the nature of the tasks requires datasets which are exponentially large?

It is easy to see that a single object can be occluded in an exponential number of ways and real world scenes consist of large numbers of objects placed in arbitrary positions in a large set of possible three-dimensional scenes. Humans are very adaptive to changes in context and will have little difficulty detecting a monkey if it is given a guitar (as illustrated in Figure 3) or if it is put in a lecture room, or even hiding in a dinning room. But, by contrast, Deep Nets appear more sensitive to context. Realize that the context of any object can be changed in an infinite number of ways (although some contexts are certainly more common). Note that this enormous variability of context does not hold in certain applications, e.g., in medical images the different body organs have fairly standard context (e.g., the Pancreas is always very close to the Duodenum).

These complexity considerations mean that certain visual tasks require dealing with an exponential number of hypotheses. This is highly problematic from a machine learning perspective, because such algorithms may require, in principle, exponential amounts of data to train and test. In particular, standard evaluation methods like cross-validation will break down. From an intuitive perspective, there will be many rare events which will not be well represented in the evaluation datasets.

Recall this underlying theory of machine learning, informally known as Probably Approximately Correct (PAC) (Valiant, 1984; Vapnik, 1998; Poggio and Smale, 2003), gives theoretical bounds on the probability that a machine learning algorithm has learned the structure of the underlying data. A key insight is that the amount of training data must be much larger than the set of hypotheses that the learning algorithm is allowed to consider before seeing the data, otherwise one of the models might fit the data by chance (which is arguably how many conspiracy theories get started). This is one reason why experimental studies sometimes produce results which other researchers fail to replicate, particularly if researchers only decide on their hypotheses after first inspecting the data. Nevertheless it implies that if the set of hypotheses is exponentially large then the amount of data must also be exponential.

In short, the standard vision evaluation methods will start having problems as we develop increasingly complicated vision models. We may be faced with trying to learn models with exponential complexity requiring potentially exponential amounts of data. This is clearly impossible. It is time for a rethink to see how we can better learn and test complex vision algorithms. In particular, how we should evaluate them if the standard evaluation procedures become inappropriate.

There are two aspects to this problem. The first is how to learn models which are exponentially complex when there is only limited amounts of data available. The second is how to test these algorithms if we only have limited amounts of data to test (because it is impractical to test over the infinite set of all possible images).

A possible solution to the first problem is to use compositional models because of their ability to share parts and to factorize between geometry and appearance. They also seem to enable the types of "cognitive science" learning whereby humans can generalize without difficulty to novel environments which they have not been trained on. The work by George et al. (2017) gives an example of this. In addition, studies of human infants suggest that they learn causal models of the 3D world and exploit underlying knowledge of the physical properties of the real world. If these relationships can be captured then they enable true generalization to novel situations. Recall that the Ptolemaic model of the solar system gave very accurate predicts but required a large amount of data to determines its details (i.e. the epicycles). By contrast, a scientist knowing Newton's Laws could deduce that the orbits of the planets were roughly elliptical and could determine them from a much smaller number of observations. Moreover if the solar system was altered, e.g., due to a rogue planet entering it, Newton's Laws would enable us to predict what would happen.

A possible solution to the second problem, of how to test vision theories, is to magnify the importance of adversaries so that instead of corresponding merely to small perturbations of images they allow other more complex operations which cause reasonable changes to the image or scene, e.g., by occlusion, or changing the physical properties of the objects being viewed (Zeng et al., 2017), but without significantly impacting human perception. From another perspective "to let your worst enemy test your algorithm" instead of testing on a random set of images drawn from a dataset.

This, of course, is similar to how complex engineering (e.g., airplanes) or software structures are tested by systematically identifying their weak points. This is more reminiscent of Game Theory rather than decision theory (which focuses on the average loss and which underlies Machine Learning theory) because it suggests paying attention to the worst cases instead of the average cases. This makes sense if vision wants to develop algorithms for self-driving cars, or diagnosing cancer in medical images, where failures of the algorithms can have major consequences.

This also urges the development of vision theories which are understandable/interpretable because it will not only be easier to identify their failure modes but, due to their explicit structure, it will be easier to correct them. Scaling up attacks on vision algorithms is made easier by the growing availability of realistic synthetic visual stimuli where systematic changes of nuisance factors can stress test algorithms, as shown in Figure 2.

7 Conclusion

This opinion piece has been motivated by discussions about Deep Nets with several researchers in different disciplines. We have tried to strike a balance which acknowledges the immense success of Deep Nets but which does not get carried away by the popular excitement surrounding them. We have often used work from our own group to illustrate some of our main points and apologize to other authors whose work we would have cited in a more scholarly review of the field.

A few years ago Aude Oliva and the first author co-organized a NSF-sponsored workshop on the Frontiers of Computer Vision. The meeting was highly stimulating and there were some very frank exchanges about the future of vision and, in particular, there was considerable disagreement about the potential for Deep Nets. But a few years later, as Yann LeCun predicted, everybody is using Deep Nets to learn their features. The successes have been extraordinary and have helped vision become much more widely known, dramatically increased the interaction between academia and industry, lead to application of vision techniques to a large range of disciplines, and have many other important consequences. But despite their successes there remain enormous challenges to overcome in order to achieve the goal of general purpose Artificial Intelligence and to understand the Mind and the Brain. While Deep Nets, and other big data methods, will surely be part of the solution we believe that we will also need complimentary approaches which can build on their successes and insights.

Acknowledgment

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. We thank Kyle Rawlins and Tal Linzen for providing feedback.

References

Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40 (4):834–848, 2018.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 248–255. IEEE Computer Society, 2009.

Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, H. Wang, A. Lavin, and D. S. Phoenix. A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science*, 2017.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 580–587. IEEE Computer Society, 2014.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Georgia G Gregoriou, Andrew F Rossi, Leslie G Ungerleider, and Robert Desimone. Lesions of prefrontal cortex reduce attentional modulation of neuronal responses and synchrony in v4. *Nature neuroscience*, 17(7):1003–1011, 2014.

Ulf Grenander. *Elements of pattern theory*. JHU Press, 1996.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 770–778. IEEE Computer Society, 2016.

Kurt Hornik, Maxwell B. Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5454–5463. IEEE Computer Society, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Annual Conference on Neural Information Processing Systems*, pages 1106–1114, 2012.

Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

Xingyu Lin, Hao Wang, Zhihao Li, Yimeng Zhang, Alan L. Yuille, and Tai Sing Lee. Transfer of view-manifold learning to similarity perception of novel objects. In *International Conference on Learning Representations*, 2017.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3431–3440. IEEE Computer Society, 2015.

Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L. Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *IEEE International Conference on Computer Vision, ICCV*, pages 2533–2541. IEEE Computer Society, 2015.

Michael McCloskey. Networks and theories: The place of connectionism in cognitive science. *Psychological science*, 2(6):387–395, 1991.

Justin NJ McManus, Wu Li, and Charles D Gilbert. Adaptive shape processing in primary visual cortex. *Proceedings of the National Academy of Sciences*, 108(24): 9739–9746, 2011.

David Mumford and Agnès Desolneux. *Pattern theory: the stochastic analysis of real-world signals*. CRC Press, 2010.

George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *IEEE International Conference on Computer Vision, ICCV*, pages 1742–1750. IEEE Computer Society, 2015.

Tomaso Poggio and Steve Smale. The mathematics of learning: Dealing with data. *Notices of the AMS*, 50(5):537–544, 2003.

Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L. Yuille. Few-shot image recognition by predicting parameters from activations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, 2018.

Weichao Qiu and Alan L. Yuille. Unrealcv: Connecting computer vision to unreal engine. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, volume 9915 of *Lecture Notes in Computer Science*, pages 909–916, 2016.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Annual Conference on Neural Information Processing Systems*, pages 91–99, 2015.

Ruslan Salakhutdinov, Joshua B. Tenenbaum, and Antonio Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, volume 27 of *JMLR Proceedings*, pages 195–206. JMLR.org, 2012.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1521–1528. IEEE Computer Society, 2011.

Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

Vladimir Vapnik. *Statistical learning theory.* Wiley, 1998. ISBN 978-0-471-03003-4.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Annual Conference on Neural Information Processing Systems*, pages 3630–3638, 2016.

Feng Wang, Xiang Xiang, Chang Liu, Trac D. Tran, Austin Reiter, Gregory D. Hager, Harry Quon, Jian Cheng, and Alan L. Yuille. Regularizing face verification nets for pain intensity regression. In *IEEE International Conference on Image Processing, ICIP*, pages 1087–1091. IEEE, 2017.

Jianyu Wang and Alan L. Yuille. Semantic part segmentation using compositional model combining shape and appearance. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1788–1797. IEEE Computer Society, 2015.

Jianyu Wang, Zhishuai Zhang, Vittal Premachandran, and Alan L. Yuille. Discovering internal representations from object-cnns using population encoding. *CoRR*, abs/1511.06855, 2015.

Jianyu Wang, Zhishuai Zhang, Cihang Xie, Yuyin Zhou, Vittal Premachandran, Jun Zhu, Lingxi Xie, and Alan Yuille. Visual concepts and compositional voting. *Annals of Mathematical Sciences and Applications*, 2(3):4, 2018.

Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, pages 1–25, 2017.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. In *IEEE International Conference on Computer Vision, ICCV*, pages 1378–1387. IEEE, 2017.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

Yukako Yamane, Eric T Carlson, Katherine C Bowman, Zhihong Wang, and Charles E Connor. A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature neuroscience*, 11(11):1352–1360, 2008.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579, 2015.

Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006.

Alan L. Yuille and Roozbeh Mottaghi. Complexity of representation and inference in compositional models with part sharing. *Journal of Machine Learning Research*, 17: 11:1–11:28, 2016.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer, 2014.

Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L. Yuille. Adversarial attacks beyond the image space. *CoRR*, abs/1711.07183, 2017.

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *International Conference on Learning Representations*, 2015.

Long Zhu, Yuanhao Chen, Antonio Torralba, William T. Freeman, and Alan L. Yuille. Part and appearance sharing: Recursive compositional models for multi-view. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1919–1926. IEEE Computer Society, 2010.

Song-Chun Zhu and David Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2(4):259–362, 2006.