Neurdiness
Thinking about brains

May 17, 2018 / neurograce

Deep Convolutional Neural Networks as Models of the Visual System: Q&A

As with most of my recent blogging, I was moved to write this post due to a recent twitter discussion (https://twitter.com/dlevenstein/status/994716148578037760), specifically about how to relate components of a deep convolutional neural network (CNN) to the brain. However, most of the ideas here are things I've thought and talked about quite a bit. As someone who uses CNNs as a model of the visual system, I frequently (in research talks and other conversations) have to lay out the motivation and support for this choice. This is partly because they are (in some ways) a fairly new thing in neuroscience, but also because people are suspicious of them. Computational models generally can catch slack in neuroscience, largely (but not exclusively) from people who don't use or build them; they're frequently painted as too unrealistic or not useful. Throw into that atmosphere a general antipathy towards techbros and the over-hyping of deep learning/AI (and how much $$ it's getting) and you get a model that some people just love to hate.

So what I'm trying to do here is use a simple (yet long…) Q&A format to paint a fairly reasonable and accurate picture of the use of CNNs for modeling biological vision. This sub-field is still very much in development so there aren't a great many hard facts, but I cite things as I can. Furthermore, these are obviously *my* answers to these questions (and my questions for that matter), so take that for what it's worth.

I've chosen to focus on CNNs as model of the visual system—rather than the larger question of "Will deep learning help us understand the brain?"—because I believe this is the area where the comparison is most reasonable, developed, and fruitful (and the area I work on). But there is no reason why this general procedure (specifying an architecture informed by biology and training on relevant data) can't also be used to help understand and replicate other brain areas and functions. And of course it has been (https://neuroecology.wordpress.com/2018/05/12/what-hasnt-deep-learning-replicated-from-the-brain/). A focus on this larger issue can be found here (https://www.frontiersin.org/articles/10.3389/fncom.2016.00094/full).

(I'm hoping this is readable for people coming either from machine learning or neuroscience, but I do throw around more neuroscience terms without definitions.)

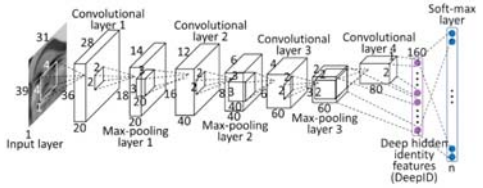
1. What are CNNs?

Convolutional neural networks are a class of artificial neural networks. As such, they are comprised of units called neurons, which take in a weighted sum of inputs and output an activity level. The activity level is always a nonlinear function of the input, frequently just a rectified linear unit ("ReLu") where the activity is equal to the input for all positive input and 0 for all non-positive input.

What's special about CNNs is the way the connections between the neurons are structured. In a feedforward neural network, units are organized into layers and the units at a given layer only get input from units in the layer below (i.e. no inputs from other units at the same layer, later layers, or—in most cases—layers more than one before the current layer). CNNs are feedforward networks. However unlike standard vanilla feedforward networks, units in a CNN have a spatial arrangement. At each layer, units are organized into 2-D grids called feature maps. Each of these feature maps is the result of a *convolution* (hence the name) performed on the layer below. This means that the same convolutional filter (set of weights) is applied at each location in the layer below. Therefore a unit at a particular location on the 2-D grid can only receive input from units at a similar location at the layer below. Furthermore, the weights attached to the inputs are the same for each unit in a feature map (and different across feature maps).

After the convolution (and nonlinearity), a few other computations are usually done. One possible computation (though no longer popular in modern high-performing CNNs) is cross-feature normalization. Here the activity of a unit at a particular spatial location in a feature map is divided by the activity of units at the

same location in other feature maps. A more common operation is pooling. Here, the maximum activity in a small spatial area of each each 2-D feature map grid is used to represent that area. This shrinks the size of the feature maps. This set of operations (convolution+nonlin[—->normalization]—> pooling) is collectively referred to as a "layer." The architecture of a network is defined by the number of layers and choices about various parameters associated with them (e.g. the size of the convolutional filters, etc).



(https://neurdiness.files.wordpress.com/2018/05/12.png)

Most modern CNNs have several (at least 5) of these layers, the final of which feeds into a fully-connected layer. Fully-connected layers are like standard feedforward networks in that they do not have a spatial layout or restricted connectivity. Frequently 2-3 fully connected layers are used in a row and the final layer of the network performs a classification. If the network is performing a 10-way object classification, for example, the final layer will have 10 units and a softmax operation will be applied to their activity levels to produce a probability associated with each category.

These networks are largely trained with supervised learning and backpropagation. Here, pairs of images and their associated category label are given to the network. Image pixel values feed into the first layer of the network and the final layer of the network produces a predicted category. If this predicted label doesn't match the provided one, gradients are calculated that determine how the weights (i.e. the values in the convolutional filters) should change in order to make the classification correct. Doing this many, many times (most of these networks are trained on the ImageNet database which contains over 1 million images from 1000 object categories) produces models that can have very high levels of accuracy on held-out test images. Variants of CNNs now reach 4.94% error rates (https://arxiv.org/abs/1502.01852) (or lower), better than human-level performance. Many training "tricks" are usually needed to get this to work well such as smart learning rate choice and weight regularization (mostly via dropout, where a random half of the weights are turned off at each training stage).

Historically, unsupervised pre-training was used to initialize the weights, which were then refined with supervised learning. However, this no longer appears necessary for good performance.

For an in-depth neuroscientist-friendly introduction to CNNs, check out: Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing (2015) (https://www.annualreviews.org/doi/10.1146/annurev-vision-082114-035447)

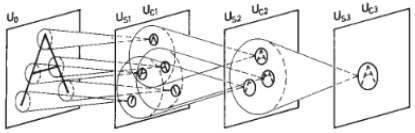
2. Were CNNs inspired by the visual system?

Yes. First, artificial neural networks as whole were inspired—as their name suggests—by the emerging biology of neurons being developed in the mid-20th century. Artificial neurons were designed (http://www.cse.chalmers.se/~coquand/AUTOMATA/mcp.pdf) to mimic the basic characteristics of how neurons take in and transform information.

Second, the main features and computations done by convolutional networks were directly inspired by some of the early findings about the visual system. In 1962 Hubel and Wiesel discovered that neurons in primary visual cortex respond to specific, simple features in the visual environment (particularly, oriented edges). Furthermore, they noticed two different kinds of cells (http://fourier.eng.hmc.edu/e180/lectures/v1/node7.html): simple cells—which responded most strongly to their preferred orientation only at a very particular spatial location—and complex cells—which had more spatial invariance in their response. They concluded that complex cells achieved this invariance by pooling over inputs from multiple simple cells, each with a different preferred location. These two features (selectivity to particular features and increasing spatial invariance through feedforward connections) formed the basis for artificial visual systems like CNNs.

Neocognitron

This discovery can be directly traced to the development of CNNs through a model known as the Neocognitron



(https://www.rctn.org/bruno/public/papers/Fukushima1980.pdf). This model, developed in 1980 by Kunihiko Fukushima, synthesized the current knowledge about biological visual in an attempt to build a functioning artificial visual system. The neocognitron is comprised of "S-cells" and "C-cells" and learns to recognize simple images via unsupervised learning. Yann LeCun, the AI researcher who initially developed CNNs, explicitly states (https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf) that they have their roots in the neocognitron.

3. When did they become popular?

Throughout the history of computer vision, much work focused on hand-designing the features that were to be detected in an image, based on beliefs about what would be most informative. After filtering based on these handcrafted features, learning would only be done at the final stage, to map the features to the object class. CNNs trained end-to-end via supervised learning thus offered a way to auto-generate the features, in a way that was most suitable for the task.

The first major example of this came in 1989. When LeCun et al. (https://ieeexplore.ieee.org/document/6795724/) trained a small CNN to do hand-written digit recognition using backprop. Further advances and proof of CNN abilities came in 1999 with the introduction of the MNIST dataset (http://Gradient-based learning applied to document recognition). Despite this success, these methods faded from the research community as the training was considered difficult and non-neural network approaches (such as support vector machines) became the rage.

The next major event came in 2012, when a deep CNN trained fully via supervised methods won the annual ImageNet competition. At this time a good error rate for 1000-way object classification was ~25%, but AlexNet (https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf)achieved 16% error, a huge improvement. Previous winners of this challenge relied on older techniques such as shallow networks and SVMs. The advance with CNNs was aided by the use of some novel techniques such as the use of the ReLu (instead of sigmoid or hyperbolic tangent nonlinearities), splitting the network over 2 GPUs, and dropout regularization, This did not emerge out of nothing however, as a resurgence in neural networks can be seen as early as 2006. Most of these networks, however, used unsupervised pre-training. This 2012 advance was definitely a huge moment for the modern deep learning explosion.

Resources: Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review (2017) (https://www.mitpressjournals.org/doi/abs/10.1162/neco_a_00990)

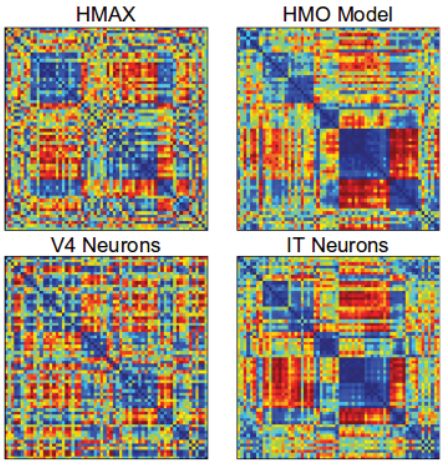
4. When was the current connection between CNNs and the visual system made?

Much of the hullabaloo about CNNs in neuroscience today stems from a few studies published in ~2014. These studies explicitly compared neural activity recorded from humans and macaques to artificial activity in CNNs when the different systems were shown the same images.

The first is Yamins et al. (2014) (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4060707/). This study explored many different CNN architectures to determine what leads to a good ability to predict responses of monkey IT cells. For a given network, a subset of the data was used to train linear regression models that mapped activity in the artificial network to individual IT cell activity. The predictive power on held-out data was used to assess the models. A second method, representational similarity analysis

(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605405/), was also used. This method does not involve direct prediction of neural activity, but rather asks if two systems are representing information the same way. This is done by building a matrix for each system, wherein the values represent how similar the response is for two different inputs. If these matrices look the same for different systems, then they are representing information similarly.

By both measures, CNNs optimized for object recognition outperformed other models. Furthermore, the 3rd layer of the network better predicted V4 cell activity while the 4th (and final) layer better predicted IT. Indicating a correspondence between model layers and brain areas.

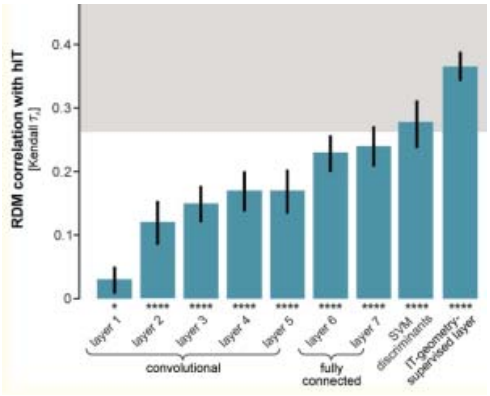


Representational Dissimilarity Matrices for different systems

Another finding was that networks that performed better on object recognition also performed better on capturing IT activity, without a need to be directly optimized on IT data. This trend has largely held (https://arxiv.org/pdf/1609.03529.pdf) true for larger and better networks, up to some limits (see Q11).

Later layers of the CNN have a more similar representation to human IT

Another paper, Khaligh-Razavi and Kriegeskorte (2014)



(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4222664/), also uses representational similarity analysis to compare 37 different models to human and monkey IT. They too found that models better at object recognition better matched IT representations. Furthermore, the deep CNN trained via supervised learning ("AlexNet") was the best performing and the best match, with later layers in the network performing better than earlier ones.

5. Did neuroscientists use anything like CNNs before?

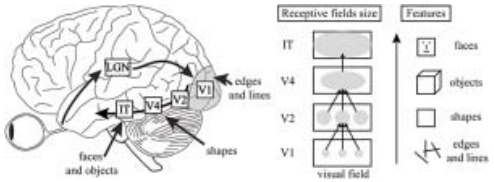
Yes! The neocognitron model mentioned in Q2 was inspired by the findings of Hubel and Wiesel and went on to inspire modern CNNs, but it also spawned a branch of research in visual neuroscience recognized perhaps most visibly in the labs of Tomaso Poggio (http://cbcl.mit.edu/people/poggio/poggio-new.htm), Thomas Serre (http://serre-lab.clps.brown.edu/), Maximilian Riesenhuber (http://maxlab.neuro.georgetown.edu/index.html), and Jim DiCarlo (http://dicarlolab.mit.edu/), among others. Models based on stacks of convolutions and max-pooling (http://www.pnas.org/content/104/15/6424.long) were used to explain various properties of the visual system. These models tended to use

different nonlinearities than current CNNs and unsupervised training of features (as was popular in machine learning at the time as well), and they didn't reach the scale of modern CNNs.

The path taken by visual neuroscientists and computer vision researchers has variously merged and diverged, as they pursued separate but related goals. But in total, CNNs can readily be viewed as a continuation of the modeling trajectory set upon by visual neuroscientists. The contributions from the field of deep learning relate to the computational power and training methods (and data) that allowed these models to finally become functional.

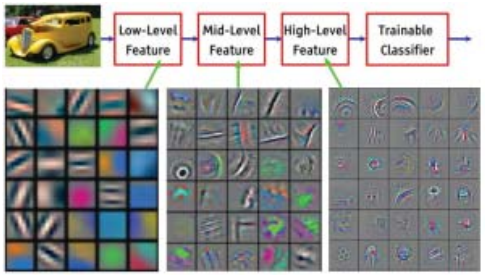
6. What evidence do we have that they "work like the brain"?

Convolutional neural networks have three main traits that support their use as models of biological vision: (1) they can perform visual tasks at near-human levels, (2) they do this with an architecture that replicates basic features known about the visual system, and (3) they produce activity that is directly relatable to the activity of different areas in the visual system.



(https://neurdiness.files.wordpress.com/2018/05/fncom-08-00135-g001.jpg)
Features of the visual hierarchy (https://neurdiness.files.wordpress.com/2018/05/fncom-08-00135-g001.jpg)

To start, by their very nature and architecture, they have two important components of the visual hierarchy. First, receptive field sizes of individual units grow as we progress through the layers of the network just as they do as we progress from V1 to IT. Second, neurons respond to increasingly complex image features as we progress through the layers just as tuning goes from simple lines in V1 to object parts in IT. This increase in feature complexity can be seen directly through visualization techniques (https://distill.pub/2017/feature-visualization/) available in CNNs.

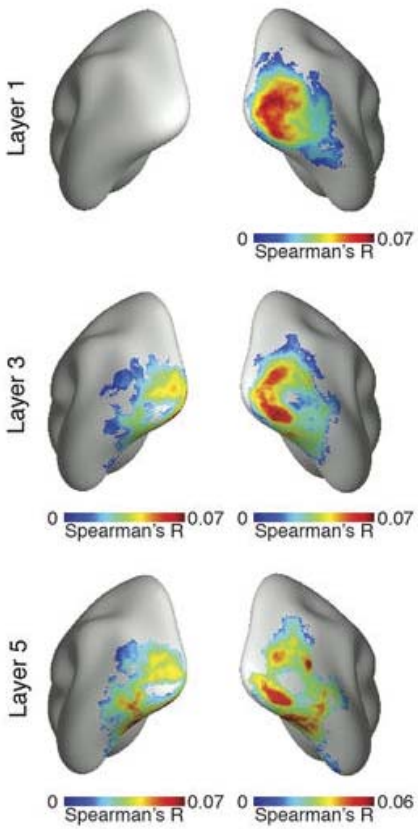


Visualizations of what features the network learns at different layers

Looking more deeply into (3), many studies subsequent to the original 2014 work (Q4) have further established the relationship between activity in CNNs and the visual systems. These all show the same general finding: the activity of artificial networks can be related to the activity of the visual system when both are shown the same images. Furthermore, later layers in the network correspond to later areas in the ventral visual stream (or later time points in the response when using methods such as MEG).

Many different methods and datasets have been used to make these points, as can be seen in the following studies (amongst others): Seibert et al. (2016) (https://www.biorxiv.org/content/early/2016/01/12/036475.full.pdf+html) Cadena et al. (2017) (https://www.biorxiv.org/content/early/2017/10/11/201764.full.pdf+html), Cichy et al. (2016) (https://www.nature.com/articles/srep27755), Wen et al. (2018) (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5830584/), Eickenberg et al. (2017) (https://www.sciencedirect.com/science/article/pii/S1053811916305481), Güçlü and van Gerven (2015) (http://www.jneurosci.org/content/35/27/10005), and Seeliger et al. (2017) (https://www.ncbi.nlm.nih.gov/pubmed/28723578).

Correlation between the representations at different CNN layers and brain areas (from Cichy et al.)



The focus of these studies is generally on the initial neural response to briefly-presented natural images of various object categories. Thus, these CNNs are capturing what's been referred to as "core object recognition (https://www.cell.com/neuron/fulltext/S0896-6273(12)00092-X)," or "the ability to rapidly discriminate a given visual object from all other objects even in the face of identity-preserving transformations (position, size, viewpoint, and visual context)." In general, standard feedforward CNNs best capture the early components of the visual response, suggesting they are replicating the initial feedforward sweep of information from retina to higher cortical areas.

The fact that the succession of neural representations created by the visual system can be replicated by CNNs suggests that they are doing the same "untangling" process (http://dicarlolab.mit.edu/sites/dicarlolab.mit.edu/files/pubs/dicarlo%20and%20cox That is, both systems take representations of different object categories that are inseparable at the image/retinal level and create representations that allow for linear separability.

In addition to comparing activities, we can also delve deeper into (1), i.e., the performance of the network. Detailed comparisons of the behavior of these networks to humans and animals can further serve to verify their use as a model and identify areas where progress is still needed. The findings from this kind of work have shown that these networks can capture patterns of human classification behavior better than previous models in multiple domains (and even predict/manipulate it), but fall short in certain specifics such as how performance falls off with noise, or when variations in images are small.

Such behavioral effects have been studied in: Rajalingham et al. (2018) (https://www.biorxiv.org/content/early/2018/01/01/240614), Kheradpishesh et al. (2015) (https://www.nature.com/articles/srep32672), Elsayed et al. (2018) (https://arxiv.org/pdf/1802.08195.pdf), Jozwik et al. (2017) (https://www.ncbi.nlm.nih.gov/pubmed/29062291), Kubilius et al. (2016) (https://www.ncbi.nlm.nih.gov/pubmed/27124699), Dodge and Karam (2017) (https://arxiv.org/pdf/1705.02498.pdf), Berardino et al. (2017) (https://nips.cc/Conferences/2017/Schedule?showEvent=10028), and Geirhos et al. (2017). (https://arxiv.org/pdf/1706.06969.pdf)

Whether all this meets the specification of a good model of the brain is probably best addressed by looking at what people in vision have said they wanted out of a model of the visual system:

"Progress in understanding the brain's solution to object recognition requires the construction of artificial recognition systems that ultimately aim to emulate our own visual abilities, often with biological inspiration (e.g., [2–6]). Such computational approaches are critically important because they can provide experimentally testable hypotheses, and because instantiation of a working recognition system represents a particularly effective measure of success in understanding object recognition." – Pinto et al., 2007 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2211529/)

From this perspective it's clear that CNNs do not represent a moving of the target in vision science, but more a reaching of it.

7. Can any other models better predict the activity of visual areas?

Generally, no. Several studies have directly compared the ability of CNNs and previous models of the visual system (such as HMAX (http://maxlab.neuro.georgetown.edu/hmax.html)) to capture neural activity. CNNs come out on top. Such studies include: Yamins et al. (2014) (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4060707/), Cichy et al. (2017) (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5542416/), and Cadieu et al. (2014). (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4270441/)

8. Are CNNs *mechanistic* or *descriptive* models of the visual system?

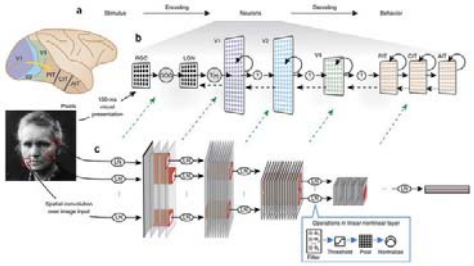
A reasonable definition of a mechanistic model is one in which internal parts of the model can be mapped to internal parts of the system of interest. Descriptive models, on the other hand, are only matched in their overall input-output relationship. So a descriptive model of the visual system may be one that takes in an image and outputs an object label that aligns with human labels, but does so in a way that has no obvious relation to the brain. As described above, however, layers of a CNN can be mapped to areas of the brain. Therefore, CNNs are mechanistic models of the representational transformation carried out by the ventral system as it does object recognition.

For a CNN to, as a whole, be a mechanistic model does not require that we accept that all sub-components are mechanistic. Take as an analogy, the use of rate-based neurons in traditional circuit models of the brain. Rate-based neural models are simply a function that maps input strength to output firing rate. As such, these are descriptive models of neurons: there are no internal components of the model that relate to the neural processes that lead to firing rate (detailed bio-physical models such as Hodgkin-Huxley neurons would be mechanistic). Yet we can still use rate-based neurons to build mechanistic models of circuits (an example I'm fond of (https://www.ncbi.nlm.nih.gov/pubmed/25611511)). All mechanistic models rely on descriptive models as their base units (otherwise we'd all need to get down to quantum mechanics to build a model).

So are the components of a CNN (i.e. the layers—comprised of convolutions, nonlinearities, possibly normalization, and pooling) mechanistic or descriptive models of brain areas? This question is harder to answer. While these layers are comprised of artificial neurons which could plausibly be mapped to (groups of) real neurons, the implementations of many of the computations are not biological. For example, normalization (in the networks that use it) is implemented with a highly-parameterized divisive equation. We believe that these computations can be implemented with realistic neural mechanisms (see the above-cited example network), but those are not what are at present used in these models (though I, and others, are working on it… see Q12).

9. How should we interpret the different parts of a CNN in relationship to the brain?

For neuroscientists used to dealing with things on the cellular level, models like CNNs may feel abstracted beyond the point of usefulness (cognitive scientists who have worked with abstract multi-area modeling for some time though may find them more familiar).



(https://neurdiness.files.wordpress.com/2018/05/screenshot-from-2018-05-17-20-24-45.png)
Relating CNNs to brain areas and processing (https://neurdiness.files.wordpress.com/2018/05/screenshot-from-2018-05-17-20-24-45.png)

But even without exact biological details, we can still map components of the CNN to components of the visual system. First, inputs to a CNN are usually 3-D (RGB) pixel values that have been normalized or whitened in some way, roughly corresponding to computations performed by the retina and lateral geniculate nucleus. The convolutions create feature maps that have a spatial layout, like the retinotopy found in visual areas, which means that each artificial neuron has a spatially-restricted receptive field. The convolutional filter associated with each feature map determines the feature tuning of the neurons in that feature map. Individual artificial neurons are not meant to be mapped directly to individual real neurons; it may be more reasonable to think of individual units as cortical columns.

Which layers of the CNN correspond to which brain areas? The early work using models that only contained a small number of layers provided support for a one layer to one brain area mapping. For example, in Yamins et al. (2014), the final convolutional layer best predicts IT activity and the second to last best predicts V4. The exact relationship, however, will depend on the model used (with deeper models allowing for more layers per brain area).

The fully connected layers at the end of a convolutional network have a more complicated interpretation. Their close relationship to the final decision made by the classifier and the fact that they no longer have a retinotopy makes them prefrontal cortex-like. But they also may perform well when predicting IT activity.

10. What does the visual system have that CNNs don't have?

Lots of things. Spikes, saccades, separate excitatory and inhibitory cells, dynamics, feedback connections, feedforward connections that skip layers, oscillations, dendrites, [***inhale****] cortical layers, neuromodulators, fovea, lateral connections, different cell types, binocularity, adaptation, noise, and probably whatever your favorite detail of the brain is.

Of course these are features that the most standard CNNs used as models today don't have by default. But many of them have already been worked into newer CNN models, such as: skip connections (https://arxiv.org/pdf/1608.06993.pdf), feedback connections (http://papers.nips.cc/paper/5276-deep-networks-with-internal-selective-attention-through-feedback-connections.pdf), saccades (http://proceedings.mlr.press/v37/xuc15.pdf), spikes (https://arxiv.org/pdf/1804.08150.pdf), lateral connections (https://www.cv-foundation.org/openaccess/content_cvpr_2015/app/2B_004.pdf), and a fovea (http://vislab.isr.ist.utl.pt/wp-content/uploads/2017/11/aalmeida-robot2017.pdf).

So clearly CNNs are not direct replicas of primate vision. It should also be clear that this fact is not disqualifying. No model will be (or should be) a complete replica of the system of interest. The goal is to capture the features necessary to explain what we want to know about vision. Different researchers will want to know different things about the visual system, and so the absence of a particular feature will matter more or less to one person than to someone else. Which features are required in order to, say, predict the response of IT neurons averaged over the first ~100ms of image presentation? This is an empirical question. We cannot say a priori that any biological feature is necessary or that the model is a bad one for not having it.

We can say that a model without details such as spiking, E-I types, and other implementation specifics is more *abstract* than one that has those. But there's nothing wrong with abstraction. It just means we're willing to separate problems out into a hierarchy and work on them independently. One day we should be able

to piece together the different levels of explanation and have a model that replicates the brain on the large and fine scale. But we must remember not to make the perfect the enemy of the good on that quest.

11. What do CNNs do that the visual system doesn't do?

This, to me, is the more relevant question. Models that use some kind of non-biological magic to get around hard problems are more problematic than those that lack certain biological features.

First issue: convolutional weights are positive and negative. This means that feedforward connections are excitatory and inhibitory (whereas in the brain connections between brain areas are largely excitatory) and that individual artificial neurons can have excitatory and inhibitory influences. This is not terribly problematic if we simply consider that the weights indicate net effects, which may in reality be executed via feedforward excitatory connections to inhibitory cells.

Next: weights are shared. This means that a neuron at one location in a feature map uses the exact same weights on its inputs as a different neuron in the same feature map. While it is the case that something like orientation tuning is represented across the retinotopic map in V1, we don't believe that a neuron that prefers vertical lines in the one area of visual space has the *exact same* input weights as a vertical-preferring neuron at another location. There is no "spooky action at a distance" that ensures all weights are coordinated and shared. Thus, the current use of weight sharing to help train these networks should be able to be replaced by a more biologically plausible way of creating spatially-invariant tuning.

Third: what's up with max-pooling? The max-pooling operation is, in neuroscience terms, akin to a neuron's firing rate being equal to the firing rate of its highest firing input. Because neurons pool from many neurons, it's hard to devise a neuron that could straightforwardly do this. But the pooling operation was inspired by the discovery of complex cells and originally started as an averaging operation, something trivially achievable by neurons. Max-pooling, however, has been found (http://yann.lecun.com/exdb/publis/pdf/boureau-icml-10.pdf) to be more successful in terms of object recognition performance and fitting biological data (https://www.cell.com/neuron/abstract/S0896-6273%2811%2900876-2) and is now widely used.

The further development of CNNs by machine learning researchers have taken them even farther away from the visual system (as the goal for ML people is performance alone). Some of the best performing CNNs now have many features that would seem strange from a biological perspective. Furthermore, the extreme depths of these newer models (~50 layers) has made their activity less relatable (https://www2.securecms.com/CCNeuro/docs-0/5928796768ed3f664d8a2560.pdf) to the visual system.

There is also of course the issue of how these networks are trained (via backpropagation). That will be addressed in Q13.

12. Can they be made to be more like the brain?

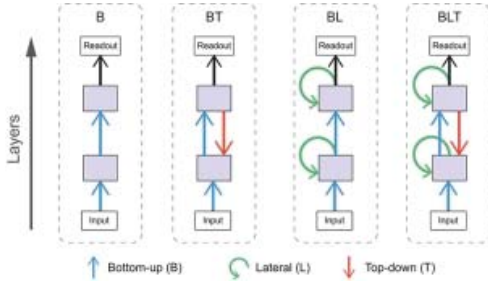
One of the main reasons I'm a computational neuroscientist is because (without the constraints of experimental setups) we can do whatever we want. So, yes! We can make standard CNNs have more biology-inspired features. Let's see what's been done so far:

As mentioned above in Q10, many architectural elements have been added to different variants of CNNs, which make them more similar to the ventral stream. Furthermore, work has been done to increase the plausibility of the learning procedure (see Q13).

In addition to these efforts, some more specific work to replicate biological details includes:

Spoerer et al. (2017) (https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01551/full) which, inspired by biology, shows how adding lateral and feedback connections can make models better at recognizing occluded and noisy objects.

Some of my own work (presented at Cosyne 2017 and in preparation for journal submission) involves putting the stabilized supralinear network



Adding biologically-inspired connections, in Spoerer et al.

(https://www.ncbi.nlm.nih.gov/pubmed/25611511) (a biologically-realistic circuit model that implements normalization) into a CNN architecture. This introduces E and I cell types, dynamics, and recurrence to CNNs.

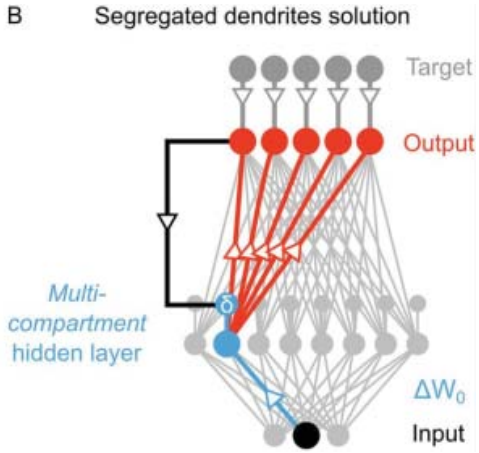
Costa et al. (2017) (http://papers.nips.cc/paper/6631-cortical-microcircuits-as-gated-recurrent-neural-networks) implemented long-short-term-memory networks using biologically-inspired components. LSTMs are frequently used when adding recurrence to artificial neural networks, so determining how their functions could be implemented biologically is very useful.

13. Does it matter that CNNs use backpropagation to learn their weights?

Backpropagation involves calculating how a weight anywhere in the network should change in order to decrease the error that the classifier made. It means that a synapse at layer one would have some information about what went wrong all the way at the top layer. Real neurons, however, tend to rely on local learning rules (such as Hebbian plasticity), where the change in weight is determined mainly by the activity of the pre- and post-synaptic neuron, not any far away influences. Therefore, backprop does not seem biologically realistic.

This doesn't need to impact our interpretation of the fully-trained CNN as a model of the visual system. Parameters in computational models are frequently fit using techniques that aren't intended to bear any resemblance to how the brain learns (for example Bayesian inference to get functional connectivity (https://ieeexplore.ieee.org/abstract/document/4703283/)). Yet that doesn't render the resulting circuit model uninterpretable. In an extreme view, then, we can consider backpropagation as merely a parameter-fitting tool like any other. And indeed, Yamins et al. (2014) did use a different parameter fitting technique (not backprop).

However taking this view does mean that certain aspects of the model are not up for interpretation. For example, we wouldn't expect the learning curve (that is, how the error changes as the model learns) to relate to the errors that humans or animals make when learning.



Local error calculations with segregated dendrite in Guerguiev et al (https://elifesciences.org/articles/22901).

While the current implementation of backprop is not biologically-plausible, it could be interpreted as an abstract version of something the brain is actually doing. Various efforts to make backprop biologically plausible by implementing it with local computations and realistic cells types are underway (for example, this

(https://elifesciences.org/articles/22901) and this (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5467749/)). This would open up the learning process to better biological interpretation. Whether the use of more biologically plausible learning procedures leads to neural activity that better matches the data is an as-yet-unanswered empirical question.

On the other hand, unsupervised learning seems a likely mechanism for the brain as it doesn't require explicit feedback about labels, but rather uses the natural statistics of the environment to develop representations. Thus far, unsupervised learning has not been able to achieve the high object categorization performance reached by supervised learning. But advances in unsupervised learning and methods to make it biologically plausible may ultimately lead to better models of the visual system.

14. How can we learn about the visual system using CNNs?

Nothing can be learned from CNNs in isolation. All insights and developments will need to be verified and furthered through an interaction with experimental data. That said, there are three ways in which CNNs can contribute to our understanding of the visual system.

The first is to verify our intuitions. To paraphrase Feynman "we don't understand what we can't build." For all the data collected and theories developed about the visual system, why couldn't neuroscientists make a functioning visual system? That should be alarming in that it signifies we were missing something crucial. Now we can say our intuitions about the visual system were largely right, we were just missing the computing power and training data.

The second is to allow for an idealized experimental testing ground. This is a common use of mechanistic models in science. We use existing data to establish a model as a reasonable facsimile of what we're interested in. Then we poke, prod, lesion, and lob off parts of it to see what really matters to the functioning. This serves as hypothesis generation for future experiments and/or a way to explain previous data that wasn't used to build the model.

The third way is through mathematical analysis. As is always the case with computational modeling, putting our beliefs about how the visual system works into concrete mathematical terms opens them up to new types of investigation. While doing analysis on a model usually requires simplifying it even further, it can still offer helpful insights about the general trends and limits of a model's behavior. In this particular case, there is extra fire power here because some ML researchers are also interested in mathematically dissecting these models. So their insights can become ours in the right circumstance (for example (http://www.cs.toronto.edu/~wenjie/papers/nips16/top.pdf)).

15. What have we learned from using CNNs as a model of the visual system?

First, we verified our intuitions by showing they can actually build a functioning visual system. Furthermore, this approach has helped us to define the (in Marr's terms) computational and algorithmic levels of the visual system. The ability to capture so much neural and behavioral data by training on object recognition suggests that is a core computational role of the ventral stream. And a series of convolutions and pooling is part of the algorithm needed to do it.

The success of these networks has also, I believe, helped allow for a shift in what we consider the units of study in visual neuroscience. Much of visual neuroscience (and all neuroscience…) has been dominated by an approach that centers individual cells and their tuning preferences. Abstract models that capture neural data without a strict one neuron-to-one neuron correspondence put the focus on population coding. It's possible that trying to make sense of individual tuning functions would someday yield the same results, but the population-level approach seems more efficient.

Furthermore, viewing the visual system as just that—an entire system—rather than isolated areas, reframes how we should expect to understand those areas. Much work has gone into studying V4, e.g., by trying to describe in words or simple math what causes cells in that area to respond. When V4 is viewed as a middle ground on a path to object recognition, it seems less likely that it should be neatly describable on its own. From this review (https://pdfs.semanticscholar.org/ec26/83205fa146b8873d3611650b51a73f67f533.pdf "A verbal functional interpretation of a unit, e.g., as an eye or a face detector, may help our intuitive understanding and capture something important. However, such verbal interpretations may overstate the degree of categoricality and localization, and understate the statistical and distributed nature of these representations." Indeed, analysis of trained networks (https://openreview.net/pdf?id=r1iuQjxCZ) has indicated that strong, interpretable tuning of individual units is not correlated with good performance, suggesting the historic focus on that has been misguided.

Some more concrete progress is coming from exploring different architectures. By seeing which details are required for capturing which elements of the neural and behavioral response, we can make a direct connection between structure and function. In this study (https://www.biorxiv.org/content/biorxiv/early/2017/08/17/177196.full.pdf), lateral connections added to the network did more to help explain the time course of the dorsal stream's response than the ventral stream's. Other (https://www2.securecms.com/CCNeuro/docs-0/5928ba3f68ed3f0d4a8a2589.pdf) studies (https://scholar.google.com/citations?hl=en&user=BMterywAAAAJ&view_op=list_works&sortby=pubdate#d=gs_md_cid&p=&u=%2Fcitations%3Fview_op%3Dview_citation%26hl%3Den%26user%3DBM60) have suggested that feedback connections will be important for capturing ventral stream dynamics. It's also been shown (https://www.nature.com/articles/srep27755) that certain components of the neural response can be captured in a model with random weights, suggesting the hierarchical architecture alone can explain them. While other components requiring training on natural and valid image categories.

Furthermore, observing that certain well-performing CNNs (see Q11) are not capable of accurately predicting neural activity is important because it indicates that not all models that do vision will be good models of the brain. This lends credence to the idea that the architectures that we have seen predict neural activity well (with a correspondence between brain areas and layers) do so because they are indeed capturing something about the transformations the brain does.

Because CNNs offer an "image computable" way of generating realistic neural responses, they are also useful for relating lesser understood signals to visual processing, as has been done here (https://www2.securecms.com/CCNeuro/docs-0/5915c06468ed3f5f1fec4f1a.pdf) and here (https://www.biorxiv.org/content/early/2018/02/09/133694) for contextualizing oscillations.

Using CNNs as a model of the visual system, my own work (https://www.biorxiv.org/content/biorxiv/early/2017/12/20/233338.full.pdf) has focused on demonstrating that the feature similarity gain model (which describes the neural impacts of attention) can explain attention's beneficial performance effects.

Finally, some studies have documented neural (https://www.biorxiv.org/content/biorxiv/early/2017/11/20/221630.full.pdf) or behavioral elements (see Q6) not captured by CNNs. These help identify areas that need further experimental and computational exploration.

And there are more examples (for a complete collection of papers that compare CNNs to the brain see this list (https://docs.google.com/document/d/1qil2ylAnw6XrHPymYjKKYNDJn2qZQYA_MaQ/edit) from Martin Hebart). All in all I would say not a bad amount, given that much of this has really only been going on in earnest since around 2014.

Filed under Uncategorized

7 Comments **leave a comment**

1. **drbabinski** / May 18 2018 5:07 pm
   This is simply amazing! Thank you for taking the time to do this, and putting all the references together in one place.

   reply
2. **Alex Hernández-García** / May 25 2018 2:07 pm
   Great review! I like the question & answer format and I also found some very interesting references I was not aware of ☺

reply

3. **Lula Memirira** / Jun 5 2018 2:50 am

   A great deal of skepticism can probably be boiled down to the belief that "backprop is not biologically-plausible", and therefore that artificial neural networks aren't worthy of the time of the skeptical neuroscientist. This belief has an impeccable pedigree, as it was argued by Francis Crick in a 1989 article in Nature to dismiss "the recent excitement about neural networks".

   Since then, there has been considerable progress in our understanding of neuroscience, and it's now rather apparent that the belief would eventually be shown to be mistaken. Unfortunately, it's also been almost 3 decades since this belief was advanced, and it'd probably take another 3 decades before people are completely disabused of the idea.

   reply

4. **ckolluru** / Jun 20 2018 8:56 pm

   Great review. I'd love to hear your thoughts on capsule networks. Do you think that they are another step towards reaching target of visual recognition or are they aiming at a completely different target?

   reply

   - **neurograce** / Jun 25 2018 9:05 am

     Yea that's an interesting question. From what I understand they do well on image recognition and with fewer parameters, but their relationship to biology is less straightforward to me (there is a paper that compares them to fMRI data but I don't care for the methodology: https://arxiv.org/pdf/1801.00602.pdf).

     Of course computer vision needn't follow biology; if they work, they work and thats fine. I don't see them being the new thing in modeling of the visual system though.

     reply

5. **Franklin** / Mar 6 2019 7:59 pm

   Very good work.

   However, I think, what you mean by descriptive model here is what is called predictive model. And what you call mechanistic model is what is called explanatory model. So the adjectives descriptive and mechanistic are not suitable.

   reply

Trackbacks

1. Human-Like Machine Hearing With AI (1/3) – Towards Data Science – RE-WORK