International Journal of Early Childhood Special Education (INT-JECSE)
DOI:10.9756/INTJECSE/V14I5.73 ISSN: 1308-5581 Vol 14, Issue 05 2022

RECENT TRENDS IN NEURAL NETWORK-BASED APPROACHES TO VISUAL PROCESSING VIS-A-VIS THEIR BIOLOGICAL CORRELATES IN THE BRAIN

Sujatha Ramesh,
Post Graduate Student, Lovely Professional University (LPU), Phagwara, Punjab
Sanjana Iyer
Hiranandani Foundation School, Powai, Mumbai, Maharashtra
K. Natarajan
ISRO (Retd.), Tiruvananthapuram, Kerala

Abstract

The rapid pace of recent advances in the realm of Artificial Neural Networks (ANN) and the simultaneous developments in the space of brain function study, especially, during early childhood, have led to a dual-development of neurosciences and neural networks in tandem. Several instances exist where the two realms have taken inspiration from each other to make large leaps of progress. In this paper, we have analyzed the recent convolutional neural network (CNN) algorithms that provide insights into the workings of the biological brain in visual perception, starting from the early childhood through different stages. We have also identified the challenges and possibilities of interdisciplinary work through symbiotic collaboration between these disparate fields of research.

Introduction

The challenge with modeling the brain's functions is that we only are beginning to know *what* it has within it. We don't really know most of *how* it works. While the biochemistry and electrophysiology insights emerging now help us create structural models at a smaller scale, and we can recreate even some very local biochemical processes such as neuron-to-neuron communication, the dynamic functioning of network-level effects largely confounds our understanding. The basic functions of gross brain networks that we can *describe* well - such as memory, learning, visual perception, and auditory perception - are still hard to *explain*. Some help in modeling brain networks has come in from the world of computer science and mathematics, where artificial intelligence and neural network models have impressively outdone even human capacity in realms like playing Chess or Go. The hope that these binary-world digital models will soon provide insights into the Hebbian biological model of the brain has only been slightly rewarding in the last two decades. Independently, on the one hand, the structural understanding of the brain is improving through core neuroscience and neurobiological research, on the other hand, the artificial intelligence in the computing world is improving in leaps and bounds. Yet, the twain fail to be in a hurry to meet. Nevertheless, there is some hope from very recent developments to find common ground, and the highest success in aligning in-silico and in-vivo results have come in the realm of image processing.

The human brain embodies several functional processes, of which sensory perception, memory encoding, memory recall, concept formation, understanding, creative insight and consciousness are broadly well-understood in common parlance. Yet, defining them well,

describing their working, and subsequently modeling them or recreating them in the lab have been fiendishly out of reach. Several strides have been made in understanding the modalities of sensory perception – especially visual perception, auditory perception – and the other cognitive processes of memory, attention, and learning. Today, we understand much better than ever before the brain areas involved in these processes and have some confidence in the step-by-step process that the brain follows in effecting these functions. The progress comes from our continually improving understanding of neuroanatomy, functional parcellation and connectome definition of the brain, neurochemistry, electrochemistry of the brain, neuroplasticity and neural oscillations. Insight and consciousness still entirely evade us even in their definition, so we have some ways to go there.

Separately in the realm of computer science, increasingly sophisticated artificial neural networks (ANNs) are driving the growth of greater and more powerful artificial intelligence (AI). From simple architectures comprising a few layers of neurons that were trained by supervision to classify images in the 1990s, we have come a long way now to having systems that can identify hitherto unknown images of a learned class, generate new images, recognize audio and comprehend and generate language. Machines learn certain classes of problems better and faster than humans do. Essentially, they find patterns and deviations from pattern in large data sets much faster than we do – and this is what we mean by 'learning' in the world of neural networks. This is especially evident in the realm of visual processing. Therefore, visual perception modelling has been highly convergent in the biological and computational fields.

Meanwhile, the success stories of the AI community have been resonating when game after strategic game such as Chess and Go were spectacularly won by computers against the most trained grandmasters. Indeed, the exponential growth of ANNs and AI has generated fears about "machine overtaking man" or even enslaving humans! These may be sensational or even ominous at first sight, but when you dig deeper, it is evident that machine intelligence has a long way to catch up with not just man, but even basic organisms with a brain. This is because none of the neural networks have yet to master the most basic ability of creating concepts, storing memories in the long run, and manipulating them in the context of sensory and emotional inputs.

Reproducing architectures mimicking human memory systems has had less pronounced successes. After training a neural net (NN) on a task, if a new task is to be learned by the same network, it needs to necessarily forget everything it has painstakingly learnt so far and start from scratch for the new problem space. This significant lacuna of neural networks is so frustrating to its developers that it is nicknamed "catastrophic forgetting". This is something humans do not have to worry about. A mother may teach a toddler about fruits and then turn around and teach numbers too without worrying about the latter wiping out the former. This is the challenge in creating computational systems that mimic human memory.

The comparative landscape of biological networks and artificial networks can be best understood by comparing the latest models in either space across different brain functions, viz. visual perception, auditory perception, memory and learning. A more comprehensive review would include a comparison of algorithm across all these functional spaces, but in the interest of focus, we limit our review here to only the visual perception area of study. In this

paper, we compare developments in the field of neural networks solving object recognition problems with their biological correlates in primate visual perception.

In this connection, the impact of neural networks vis-à-vis early childhood requires specific mention. It is necessary to understand the brain power during childhood phase: particularly, early childhood; because, between the ages of 2 and 3 the child's brain power enhances rapidly. Apparently, the improvements in respect of thinking, memory, skills and learning give the child new ways to move, play, and express its feelings, in addition to demands for affection and comforts. It is now known that, by the age of 3, an infant has trillions of brain connections or synapses: the most they will ever have in their entire life. From the age of 5, a child's brain starts developing more than what happens at any other time in life. Clearly, early brain development has considerable impact on a child's capability to learn and progress not only in school, but also in life. Neurons have the capability to grow longer dendrites as well as axons: these allow them to make a number of connections (synapses) with other cells. It may be mentioned that the number as well as the density of synapses increase very fast during the first few years of life. A 2-year-old's brain is perhaps 20% smaller than that of an adult brain; but yet, has a large number of synapses. The first stage: 13-15 years of age: increases the size and function of the brain; particularly, of motor areas as well as spatial perception-activity. The second stage: 17 years onwards: increases frontal lobe size in addition to its connections with all parts of the brain. The development of the brain starting from early childhood depends on:

- Adequate nutrition during pregnancy.
- Exposure to bacteria, infections and toxins.
- The infant's environment and experiences with others

Methodology

This investigation is based on the published data relating to neuroscience, AI and NN. A total of 37 papers were studied in this endeavor, mostly from the period of 2019 and thereafter, to account for the most recent advances in either realm of knowledge. The advantages and disadvantages of the biological and digital systems have been collated.

Analysis

We analyze here the latest developments in visual process under purview and the closest ANN that matches that function. Visual Perception has been the favorite playground where neurobiologists and computer scientists like to meet. Indeed most of the advances in ANNs have been in the realm of image processing, classification and reconstruction. Likewise, in the biological realm, one of the best understood sensory systems is the visual system, especially during the early childhood. It has the most likeness to the physical world outside – be it in the eye's analogous behavior to a DSLR camera, or in the striking similarities in the image processing algorithms of computers and living creatures.

This bonhomie goes back in history, as Convolutional Neural Networks (CNNs) originated from being inspired by biological vision. Biological vision comprises two pathways – the 'what' pathway and the 'how' pathway. [1] The 'what' pathway, more formally known as the Ventral Pathway captures and recognizes the form of the image and 'where' it is spatially

located. It generates internal representation of visual inputs to be stored in long-term memory. The 'how' pathway relates to motion and is more formally known as the Dorsal Pathway, which we will not discuss in this paper. The Ventral Pathway consists of areas in the brain labelled V1, V2, V3, V4, which start next to the primary visual cortex (V1) at the back of the head and progressively largely flow through to the more anterior and ventral portions of the temporal cortex. This may be referred to as the Inferior Temporal Cortical zone. As information flows from one region to the next on the Ventral Pathway, the representation of the visual input that was originally received increasingly becomes more abstract, as the brain extracts invariant features at each step to form generalized representations of the input. These feedforward projections are also backed up by feedback projections, the nature and purpose of which are not well known. In addition, it is important to note that the encoding of visual information in memory is thought to be modulated by subcortical neurochemical projections which inform about the level of arousal related to the image and influence its long-term storage priority. Significantly, from the inferior temporal cortex, there are projections to frontal cortical areas (working memory to store the image briefly even after it is gone), hippocampus (to encode into long-term memory) and the amygdala (where the emotional valence of the stimulus is attached).[2a][2b]

The information processing complexity is captured in the HMAX model [3] . Fig. 1. Illustrates the Visual Pathway. The primary visual cortex (V1) captures low-level visual features such as edges and contrasts. At this stage, object categorizations are linearly separable.[6] Subsequent stages contain increasingly larger receptive fields, as they capture more abstract and complex representations. Kravitz et. al [4] have summarize that as per the current model, RF size, the complexity in the representation, invariance to matters like visual transformations, etc. escalate from the early to late units by means of iterative sum plus max operations as applied by each and every input to their units.

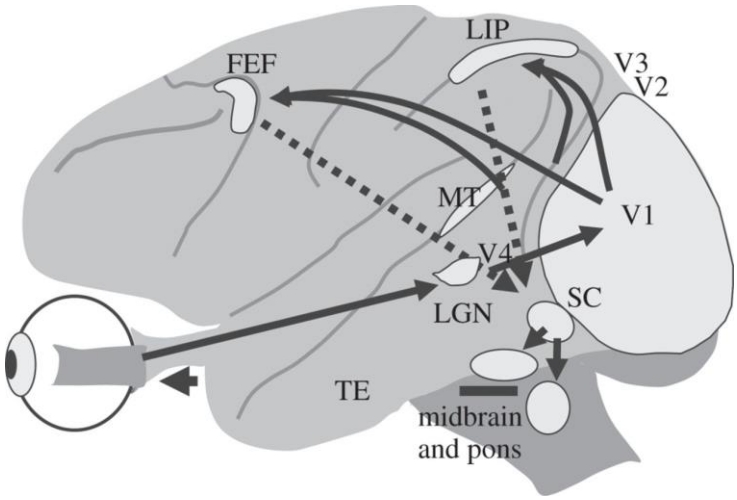

Fig. 1. Visual Patthway

Image Source: Reproduced from [16]

Object recognition in the human brain, therefore, is thought to be processed as a 2D input being transformed into an invariant 3D space, where the target space has a 2D map of edges and regions, with textures added into a 2.5D map, and finally with prior information (memory) added onto a 3D space.[10] CNNs, or more recently, Deep CNNs, have successfully modeled the hierarchical, feedforward structure of the visual ventral pathway. There is a strong correlation between the model's performance in categorization and the response expected at each neural layer. At the point of V4, from which emerge the projections to the various memory and emotional valence systems, the output of the neural network was highly predictive of the V4 neural response.

Although CNNs have been remarkably successful in automatic feature detection, they have nevertheless suffered from the criticism that they do not model the biological brain architecture and are too computationally intensive to be viable as a true brain model. More critically, CNNs are unable to deal with objects that are presented from a different view-point in the image. This is different from recognizing objects that have been translated to a different *position*, which CNNs can handle. When a picture of a dog is taken from front, versus if it is taken from a top-view, the perspective change causes the dog to not look like a dog at all. Yet, the human brain can readily infer that it is a dog, seen from above. CNNs can identify a dog in any part of a room, but not one that has been photographed from above.

Recently, a specific class of unsupervised CNNs have emerged to solve this problem, by using contrastive embedding objectives. The key idea of contrastive embedding is creating augmented versions of the image and then keeping augmentations of similar objects together. "Given a list of input samples $\{x_i\}$, each has corresponding label $y_i \in \{1,\ldots,L\}$ among L classes. We would like to learn a function $f_\theta(.):X \rightarrow R^d$ that encodes x_i into an embedding vector such that examples from the same class have similar embeddings and samples from different classes have very different ones. Thus, contrastive loss takes a pair of inputs (x_i,x_j) and minimizes the embedding distance when they are from the same class but maximizes the distance otherwise." [13][14] The core idea here is that various views of the image (v(x)) may be generated, which are augmented versions of the sample.

The goal is to identify an embedding function f(v(x)) which places these embeddings as close as possible to their original view (v(x)), but far from other samples. This means that every embedding will maintain a distance from unrelated images, but remain close to related views. In contrastive objective, the loss function teaches the neural network to place images with the same labels clustered, while dissimilar images are kept far apart. In this, the mutual information is maximized and the higher layers support any natural statistic that reliably distinguishes between two sets of inputs. [5]

The power of CNNs is that the output from each convolutional layer can be max-pooled to compare with neuronal activity in primate brains that represent the equivalent visual processing region. [7][8] The contrastive embedded deep CNNs have proved to display significantly high predictive capacity of the activity in the corresponding visual layer of the primate brain, for the same image.

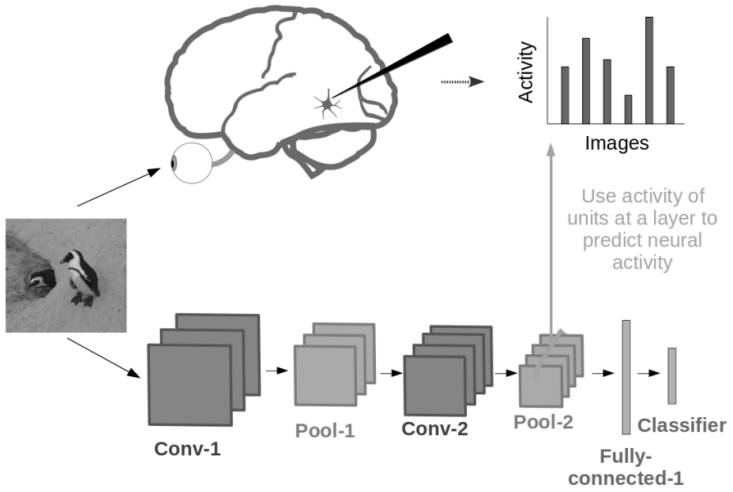

Fig. 2. Comparison Between in-silico and in-vivo Responses

[Reproduced From: Grace Lindsay[15]]

A further problem with CNNs has been that while they could recognize objects they were trained for, they could not recognize faces. In the human visual system, the Fusiform Face Area (FFA) – an area near inferior temporal cortex – is specialized just for face recognition. Lesions in this area affect specifically a person's inability to recognize faces, even if they can recognize other things like a flower vase or a traffic light. Research by Nancy Kanwisher's team at MIT [18] identified that dually trained CNNs optimized for both tasks (objects and faces) spontaneously segregate themselves into two different specialized classification networks that differentially categorize objects and faces. This gives a powerful fillip to the idea that the same must have happened in the evolution of the human cortex, with areas specialized for face recognition spontaneously branching off at a downstream stage of visual processing, forming distinct physical structural regions.

A further concern about the artificiality of deep CNNs is that it does not capture the selective attention focus that our eyes afford in visual processing. When we focus our eye on a particular part of the scene and this drives perception differently from if we look at the scene in its entirety in a neutral manner. Selective attention networks are being currently incorporated into deep CNNs, dubbed Visual Attention Networks (VAN) [19]. In this, discriminative features are selected and noisy responses are ignored to form an attention map that captures the relative importance of different parts of an image.

It may be mentioned that developing infants during the first few months of life can considerably shift fixation from a somewhat central target to a different target that may be appearing in the periphery, provided that these two targets are not visible together; and that there are no other visual or different type of 'distractors' in the rest of the visual field. The capability to assess children's attention development at an early age will be of most value if it can be used to target effective interventions. There is much interest in procedures for training attention and executive function. Some studies have evaluated an educational program called 'Tools for the Mind'. This aims to improve self-regulation in Kindergarten-level children,

and has found that it improves 5-year-olds' basic performance relating to attentional and inhibitory assignments without much training on these. Further, recently, eleven-month-old infants' experiences with certain types of displays have encouraged them to steadfastly sustain fixation after ignoring distractions. Thus, some fixation control has been achieved in children. Atkinson and Braddick[20] have given an account of visual attention and child neurology. It is suggested that attention is influenced by the following considerations:

- Novel methods make it possible to facilitate attention in infants and pre-school children.
- Distinct attention profiles can be associated with varied developmental disorders.
- Some developmental disorders relating to attention are connected with the dorsal-cortical stream.

Some of the relevant observations are as follows:

- Biological viability and Computational complexity management:
- While CNNs mimic the behavior of the visual pathway, the artificial net often needs a larger receptor field size than is biologically viable. Nevertheless, vNet architectures with plausible receptor field sizes adequately approximate the visual activations at each neuronal layer.
- A point of difference between neural networks in general and brain architectures is that the objective function is optimized using a mechanism known as backpropagation. There is not much evidence for backprop in the human brain, or at least, not in the form that traditional ANNs did it. However, it has been found that there are indeed some potential biological correlates to backpropogation. Pyramidal neurons are known to have two types of dendrites – apical dendrites and basal dendrites – permitting them to propagate inferences forward and errors backwards, as do deep neural networks.[17]
- Adding attentional networks increases computational complexity. VANs, while closer to the model of the human visual model, are still only in their infancy, needing to be optimized for a viable computational complexity.
- Attention modulation: It was found in a study that vNet architectures[11], which are inherently closer to the human visual field, utilizes spatial priorities (constructing salience maps) similar to those of human observers during feed-forward object recognition. [9] Spatial priority of information processing in creatures is done by eye scanning – we focus on some areas of an image more than others, giving it more attentional resources and priority. In neural network it is achieved by constructing salience maps.
- Attentional visual-processing neural networks are still in their early days. Attention in the human eye field is adaptive and changes over time. Further, the reason for attentional direction towards a particular part of space is not necessarily just novelty or texture-driven, but might be from emotional or fear stimuli arising bottom-up from ventral systems of the brain or for other reasons that evolve spontaneously at that moment. These are being studied as part of video processing research, but are not mature.
- Emotional modulation: The major difference between the artificial and biological visual network is the modulating effect of valence appraisal in creature brains.

Evolution has tuned us to avoid or approach certain types of stimuli based on prior experience, and to prioritize some information over others dynamically. However, CNNs do not have this quality, and have been found to have a consistent strong bias towards texture (based on ImageNet trained networks). [9]

- A baby's vision needs to be investigated too as part of this study to understand attention during early childhood. A baby's eyesight matures over many months, and he or she is able to see quite well near and far and even focus on quickly moving objects. For older preschoolers (4–5 years old), certain skills will develop: Older preschoolers would start counting numbers and can easily answer: how many: whenever they are shown a group of objects. However, older pre-schooling children will be able to group few objects such as blocks, cups and plates.

Conclusion

This paper has investigated the recent research in Convolutional and Deep Neural Networks (CNN, DNN) towards imitating visual image processing in the brain. The early childhood phase has considerable influence on the child's learning and memory skills. Advances in Convolutional Neural Networks (CNN) research have successfully created highly comparable versions of neural network architectures in-silico to what is observable in primates in-vivo. With agreement being found on receptor field size, spatial information prioritization, intermediate activations of responses and eventually information classification itself, within reasonable time-periods, the artificial neural networks may serve to better understand how our brains work where we cannot realistically probe for information. On the other hand, biological phenomena such as the modulation of affective input, recurrent processing and the use of pre-processed information to inform current learning can guide the development of artificial networks. Such symbiotic collaborations across the disciplines of mathematical sciences, computational research and neuroscience can bring forth a speed of advancement in the understanding of human cognitive processes: starting from early childhood: that has never been possible before. It can in turn also accelerate the growth of artificial intelligence systems that are powered by such algorithms that bring the best of man and machine together.

References:

[1] Goodale MA, Milner, AD. Separate visual pathways for perception and action. Trends *Neurosci*. 1992. 15 (1): 20-5. Doi:1016/0166-2236(92)90344-8.PMD. 1374953.S2CID793980.

[2a] http://www.scholarpedia.org/article/What_and_where_pathways

[2b] Ungerleider LG, Mishkin M. Two Cortical Visual Systems. In: Ingle DJ, et al., editors. Analysis of Visual Behavior. The MIT Press; 1982. pp. 549–586.

[3] *Serre T, Oliva A, Poggio T.* A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci U S A. 2007 Apr 10; 104(15):6424-9.*

[4] Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in cognitive sciences*, *17*(1), 26–49. https://doi.org/10.1016/j.tics.2012.10.011

[5] Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(3), e2014196118. https://doi.org/10.1073/pnas.2014196118

[6] Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science (New York, N.Y.)*, *310*(5749), 863–866. https://doi.org/10.1126/science.1117593

[7] Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

[8] Lindsay G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of cognitive neuroscience*, *33*(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

[9] Van Dyck Leonard Elia, Kwitt Roland, Denzler Sebastian Jochen, Gruber Walter Roland; Comparing Object Recognition in Humans and Deep Convolutional Neural Networks—An Eye Tracking Study; *Frontiers in Neuroscience;* 2021; Vol. 15, https://www.frontiersin.org/article/10.3389/fnins.2021.750639 DOI=10.3389/fnins.2021.750639; ISSN=1662-453X

[10] Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information.* San Francisco, CA: W. H. Freeman.

[11] Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2011417118. doi: 10.1073/pnas.2011417118

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738, 2020.

[13] https://lilianweng.github.io/posts/2021-05-31-contrastive/

[14] S. Chopra, R. Hadsell and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," ***2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)***, 2005, pp. 539-546 vol. 1, doi: 10.1109/CVPR.2005.202.

[15] https://arxiv.org/ftp/arxiv/papers/2001/2001.07092.pdf

[16] Robert H. Wurtz Using perturbations to identify the brain circuits underlying active vision Phil. Trans. R. Soc. B 2015 370 20140205; DOI: 10.1098/rstb.2014.0205. Published 3 August 2015 http://rstb.royalsocietypublishing.org/content/370/1677/20140205

[17] J Guerguiev, TP Lillicrap, BA Richards. Towards deep learning with segregated dendrites. eLife 2017;6:e22901 DOI: 10.7554/eLife.22901

[18] Dobs, Katharina & Martinez, Julio & Kell, Alexander & Kanwisher, Nancy. (2021). Brain-Like Functional Specialization Emerges Spontaneously In Deep Neural Networks. 10.1101/2021.07.05.451192.

[19] Guo, M. H., Lu, C. Z., Liu, Z. N., Cheng, M. M., & Hu, S. M. (2022). Visual attention network. *arXiv preprint arXiv:2202.09741*.

[20] Atkinson, J., and Braddick, O.; Visual attention in the first years: typical development and developmental disorders; *Developmental Medicine and Child Neurology*; 08 May 2012; https://doi.org/10.1111/j.1469-8749.2012.04294.x