How infant-inspired mechanisms could improve deep learning as a model for human vision

Word count: 5049

Much of what we know is learned from what we see. Our early experience is awash with sensory input; before we can even speak we are observing and learning from the world around us. Focusing on the emergence of semantic knowledge from visual experience, I will review where the fields of cognitive neuroscience, psychology and computer vision stand in terms of accounting for our robust visual understanding. Deep learning networks continue to excel as models for human vision, but are they truly accounting for the wealth of semantic information present in our neural representations? The human vision literature is expansive, but are we missing key insights from naturalistic, contextualised experiments that will delineate how to build robust semantic knowledge? Finally, I argue that progress can be made in both computer and biological vision research by looking for inspiration in the most efficient learning systems that we know of: the human infant. Through an exploration of the downfalls of current vision research and computational modelling I discuss how infant-inspired machine learning just might improve our current best models of the brain and, in turn, enlighten our knowledge of how abstract semantic knowledge is built from the physical world.

How does a system learn to know what it sees?

The question of semantic knowledge could be approached from many angles, but often is formalised in psychological research as categorisation behaviour or concept formation. This goes hand-in-hand with knowledge representation in computer vision, where (sometimes to a fault) successful classification performance is taken as confirmation that a system has successfully learned what things are. This can be further distilled down to the behaviour of object recognition: the ability to attach a label to a visual percept. The site of object recognition in human visual cortex is on the ventral occipitotemporal pathway known as the ventral stream. This 'what' pathway of vision is tasked with recognising objects in space and is separate to the 'where' dorsal stream which instead focuses on tasks related to motion or vision for action (Goodale & Milner, 1992). Structured along a posterior to anterior gradient, earlier regions in this hierarchical ventral pathway such as V1 and V2 respond to simpler

critical features than more anterior regions (V4, IT) (Kobatake & Tanaka, 1994), and invariance to transformations in the input image increases along the pathway to enable robust recognition performance (Rust & DiCarlo, 2010). The complexity of the information also increases from posterior to anterior. Early visual cortex (EVC) processes features like luminance or orientation and more abstract features are processed later in the pathway with anterior regions such as inferotemporal cortex (IT) in macaques and lateral or ventral occipitotemporal cortex (LOT/VOT) in humans being the site of more high-level, semantic representations for objects (Orban et al., 2004; Tanaka, 1996). Taken together, this hierarchical pathway is confirmed to be the cortical site of extracting meaningful representations from an object's visual input.

Object recognition is an equally relevant task in computer vision systems. Inspired by the ventral stream's hierarchical architecture as well as the pooling from complex to simple receptive field sizes identified by seminal visual cortex experiments (Hubel & Wiesel, 1962), the early Neocognitron model (Fukushima & Miyake, 1982) first introduced the idea of brain-inspired computer vision. However, it wasn't until 2012 when the field experienced a revolution with the introduction of AlexNet (Krizhevsky et al., 2012), a deep convolutional neural network (CNN) that was capable of state-of-the-art performance on a common computer vision benchmark: the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015). Such networks have even gone on to outperform humans at object recognition and labelling tasks (He et al., 2015). Deep learning was revolutionary for the field, and the CNNs that facilitated innovative leaps within the computer vision community emerged to be more than simply inspired by the brain; they are in fact predictive of the very neural structures from which they were inspired.

Many influential studies have now established firm links between human neuroimaging data and CNNs. Within the past decade, new methods such as representational similarity analysis (RSA) (Kriegeskorte et al., 2008) have enabled valid comparisons across systems as disparate as the brain and computers. With the advent of CNNs, computational models for neural data were now being optimised on the *task* of object recognition rather than being directly constrained by physiological recordings. This shift towards a focus on modelling goal-driven behaviour lead Yamins and colleagues to describe the first quantitatively accurate image

computable model of spiking responses in IT cortex (Yamins et al., 2014), following up with a study showing that the then emerging deep neural networks rivalled the representational performance of IT in a visual recognition task (Cadieu et al., 2014). Further elegant studies went on to affirm the power of these CNNs in modelling neural responses (Cichy et al., 2016; Eickenberg et al., 2017; Khaligh-Razavi & Kriegeskorte, 2014) and deep learning is now well-established as a leading method for constructing models within cognitive computational neuroscience (Storrs & Kriegeskorte, 2019).

A recent paper replicated the findings of previous studies, confirming the links between the brain and CNNs, but the authors went further in testing not only *if* the CNNs were predictive of neural responses but *how well* they were predicting the brain (Xu & Vaziri-Pashkam, 2021). New fMRI data was collected with improved signal to noise, and 14 popular CNNs were tested which had been pretrained on the standard ImageNet dataset (Deng et al., 2009). Once again, RSA was employed to show the impressive mapping of lower neural network layers to earlier regions in the ventral stream as was shown by Güçlü and van Gerven (2015), illustrating a definite correspondence between the brain and the models. However, by calculating the noise ceiling on the fMRI data Xu and Vaziri-Pashkam found that while higher layers of the networks were predictive of brain responses in regions such as LOT and VOT, the quality of the prediction was poor with the highest amount of explainable variance being 60%. In contrast, many of the CNNs tested could fully capture the RDM variance of lower visual areas (e.g. V1-V4). This finding raises an interesting limitation of the deep learning models. It appears that their learned representations are missing something in the high-level information that would be present in semantic, anterior visual cortex. Of course there are numerous differences between the CNN models and the brain that could explain the dissimilarities, but perhaps this particular shortcoming is a surmountable issue that lies in the way the models learn. While anterior visual regions do receive top-down influence from frontal cortex (Bar et al., 2006), the input is largely fed forward from earlier visual regions. Perhaps there is sufficient structure in naturalistic perceptual input that enables learning of more abstract representations, and there is something more fundamental missing from the images input to CNNs that precludes learning of the more high-level information that is widespread in naturalistic experience.

Downfalls of CNNs as models of anterior visual cortex

When we consider the manner in which these neural network models are trained, it's not surprising that they fail to fully capture representations in the brain's semantic loci. All of the networks found to be predictive of neural data in the above studies were trained with a supervised machine learning approach using millions of images and their associated labels. This does not align with how humans learn to recognise. We spend our early days observing and interacting with the world around us, with very little in the way of supervision except for when our caregivers explicitly name things for us, which (in comparison to experience as a whole) is not very often. Human learning is more akin to unsupervised or semi-supervised machine learning methods (Zaadnoordijk et al., 2020), and efforts to follow an infant-inspired self-supervised learning curriculum have been shown to improve predictions of macaque neural data (Zhuang et al., 2019). There are arguments from deep learning researchers that taking a more biologically-plausible approach will reap benefits for their engineering goals (Sinz et al., 2019). The real-life training dataset of our visual environment houses much richer information that provide cues to guide learning of semantic knowledge.

A current limitation of CNNs that prevents them from capturing this high-level information is a bias towards local features and an inability to capture global relational structure. In a study that constructed a CNN using a method that ignores spatial relations between the parts of images by considering an image a simple "bag of features", it was shown that CNNs are robust to large transformations of images that would otherwise greatly impair human recognition (Brendel & Bethge, 2019). This more traditional approach to learning did not preclude the network from reaching high ImageNet classification performance, operating at an accuracy comparable to state-of-the-art models at the time. The result illustrates that CNNs do not use large-scale spatial regularities or global shape integration of the input images to form representations, but instead focus on regularities at the scale of individual features. While Brendel and Bethge's network was explicitly designed to be a feature-focused model, its still-strong performance highlights that CNNs with high classification accuracy don't necessarily rely on visual cues at the whole image level. In contrast, human scene recognition can proceed rapidly by focusing on the global image features that provide a summary of the spatial layout of the visual input in a low dimensional code, allowing a rapid "gist" of the scene to be

understood and thereby constrain the subsequent local feature analysis for more specific object recognition (Oliva & Torralba, 2006). In this way, the scene as a whole is first perceived as a single entity and then the local details are processed at a finer scale. This disparity in the visual processing between computational and neural systems is further exemplified by the sensitivity of deep neural networks to adversarial attack. Even a single pixel can fool the network into misclassifying, when such a difference is virtually undetectable by the human eye (Yuan et al., 2019). Clearly, the over-reliance of these models on local information is detrimental under certain conditions, limiting their capacity to grasp an overall understanding of the patterns and regularities that make an image what it is.

Given that this difference in processing focus seems to be so pronounced, it is curious that the CNN models are still so predictive of brain responses. Recall that the models are significantly predicting fMRI data from visual regions across the posterior to anterior hierarchy, but in semantic regions the prediction is simply incomplete (Xu & Vaziri-Pashkam, 2021). What is it about the lack of global information in an image that limits the models' ability to fully account for the concepts in high-level visual regions, but does not significantly impair their ability to recognise objects? A huge difference between the two systems is the nature of their image inputs. Deep learning networks are trained on highly-curated still images whereas visual experience is defined by more slowly-evolving inputs with consistent spatiotemporal regularities. Perhaps the patterns and distributional properties of naturalistic experience is informative for object semantics, and by incorporating this into CNN learning we may build more holistic models of human visual regions.

Meaningful relational structure emerges from naturalistic co-occurrence statistics

Elements of visual input hold informative statistical regularities that guide representation learning. In a recent review, Hafri and Firestone (2021) inadvertently highlighted the exact local-feature bias pitfall of CNN models stating *"the world is more than a bag of objects: it contains not only isolated entities and features (red apples, glass bowls) but also relations between them (red apples in glass bowls). These relations are rich, abstract, categorical and structured"* (Hafri & Firestone, 2021). Indeed, they highlight in this review how the informative relations between objects in space and across wider visual input are actually

perceived quite automatically by the visual system to result in representations that contain abstract relational information. While typical computer vision CNNs are guilty of the local bias, the idea of extracting meaning from surrounding context does exist in the machine learning literature. Distributional semantic models from natural language processing (NLP) learn a word's embedding based on its surrounding words in a large text corpus, and can even perform interesting analogy tasks (Mikolov et al., 2013) with transformer networks such as BERT capable of changing an embedding for a word given its context (Devlin et al., 2019). The latest-and-greatest in artificial intelligence is GPT-3, generating significant buzz for its impressive ability to perform seemingly complex writing tasks (Floridi & Chiriatti, 2020). Each of these models has a unique technical profile but they all have one thing in common in that they attempt to learn what a word means by learning how it relates to other things in context. Perhaps the same principle can be applied to visual models to improve semantic understanding.

Although these NLP models don't have quite the same reach into cognitive science as CNNs do for vision, there are papers that connect the two in an attempt to generate more semantically meaningful machine learning models. It has been shown that incorporating typical distributional semantic models with pixel-based CNNs can provide better prediction of neural responses, with later semantic layers being more correlated to anterior neural responses (Devereux et al., 2018). Many studies investigate multimodal computational models and reveal significant correlations to the brain (Anderson et al., 2013; Bruni et al., 2014; Derby et al., 2018; Rotaru & Vigliocco, 2020). These are all exciting efforts and there is a theoretical case for the different models – pixel-based and text-based – being analogous to multimodal processing within the human brain. However, each model referenced above requires explicit coding of the semantic element, with none being able to extract meaning from a purely pixel-based input. As reviewed by Hafri and Firestone (2021), there is mounting evidence for relations being rapidly processed by the perceptual system and not only something that requires our careful and slow reasoning. Much like how a word appears in a sentence, they suggest that objects appear in their contexts with useful statistical information.

Taking the spatial configuration of objects as an example, it has been shown through continuous flash suppression that objects in configurations that would typically occur in the world access awareness more quickly than when the same objects are shown in unexpected spatial layouts (Stein et al., 2015). Furthermore, recent EEG analyses reveal that expected spatial configurations of objects facilitate the extraction of contextual associations between objects that tend to co-occur, with a larger signal arising in occipitotemporal cortex when associated objects are typically positioned (Quek & Peelen, 2020). This importance on general spatial configuration ties in with the aforementioned gist perception, put forward by Oliva & Torralba (2006). The global statistics of a scene are crucial for human understanding of the input's general meaning, as expected spatiotemporal configurations appear to confer an ability to rapidly understand what is being perceived. Through sensitivity to the statistical co-occurrence structure beyond just that at the feature level, a more holistic and semantic understanding of the scene can be learned. This extends to the relationship between objects in a scene and the context itself, with Palmer showing in early psychological experiments that the successful recognition of an object is facilitated having first seen its congruent context (Palmer, 1975) and Biederman and colleagues showing through behavioural experiments that semantic relations between entities in a scene are rapidly accessed to facilitate identification (Biederman et al., 1982). Clearly, there are additional benefits for recognition once things are placed in their typical settings. This idea is reviewed well by Willems & V. Peelen (2021) who also point out that perceptual research in a contextualised setting is scant, and something to be developed in coming years.

A few neuroimaging experiments follow this line of thinking and investigate the extent to which constituent objects of a scene are encoded in neural representations. Using an encoding model based on an NLP method called Latent Dirichlet Allocation it was shown that by applying this algorithm to the frequency counts of object labels in a large visual dataset, the learned co-occurrence statistics of the scenes were predictive of typical anterior functional regions of interest including retrosplenial cortex (RSC), parahippocampal place area (PPA), lateral occipital (LO) and others (Stansbury et al., 2013). The authors claim this as quantitative evidence for the findings of Palmer and Biederman and it is interesting that these are the exact areas that were not fully explained by the CNNs tested in Xu & Vaziri-Pashkam (2021). Furthermore, the encoding model put forward by Stansbury et al. was not significantly

predictive of earlier visual regions V1 to V4, reiterating the importance of co-occurrence statistics for anterior visual responses. Further results show how particular anterior regions in the lateral occipital complex use the combination of object representations to form scene understanding in an experiment that classifies fMRI responses to scene images from a linear combination of object-based predictors (MacEvoy & Epstein, 2011), providing converging evidence for the idea that there is something unique to anterior regions in their ability to extract statistics that are reflective of the environment. This appears to form the basis of how meaningful structure manifests in neural responses. Perhaps the use of more naturalistic datasets with typical statistical regularities in CNN training would facilitate learning of relational structure and improve their potential as models for the ventral pathway.

Psychological accounts of concepts being structured by co-occurrence statistics

The idea that concepts are emergent from relational or statistical structure is prevalent in psychological literature. Eleanor Rosch's prototype theory takes a probabilistic rather than the original definitional approach to defining a concept, such that category membership – a common operationalisation of concept formation – is decided by 'family resemblance' to a typical member of the category (Rosch & Mervis, 1975). According to this theory, probabilistic structure at the feature level of category members define membership which, despite falling short in explaining the spread of highly variable concept examples (Storms et al., 2000) presents an appealing intuition that ties into the above arguments. This is further formalised in the Conceptual Structure Account where the internal structure of a concept is defined by relations between features and their degree of correlation (Taylor et al., 2007), highlighting once again the expansive explanatory power of co-occurrences that exist in naturalistic statistics.

Note that traditional views of knowledge representation in the psychological domain consider the representation to be an amodal, internal symbol that lies beyond sensory modalities in the brain; for example, that the anterior temporal lobe (ATL) is the sole site of semantic convergence. However, theories in grounded or embodied cognition state the opposite and claim that all representations must involve a sensory component that is re-activated in the brain once the knowledge is needed (see Barsalou (2008) for review of this theory). While I

would argue that an entirely amodal system is unlikely given the first cortical port-of-call for any information is sensory areas, at some level a representation must become abstracted to something that one would define as a symbol. However, the symbol itself may be reflective of the statistics and activation patterns that were present during experience, similar to the ideas proposed by Barsalou but not as prescriptive in the need for replay. Lambon-Ralph and colleagues present the idea that modal-specific inputs are integrated in their hub-and-spoke model of semantic cognition. According to this view, the ATL is an integrator of modality specific information which may still house semantic content – to form robust, generalisable concepts (Lambon Ralph et al., 2017). While I would argue against the idea of a single site of semantic knowledge within the temporal lobes, there is a definite role for ATL in semantic cognition as evidenced by the condition of semantic dementia which is marked by highly-specific atrophy of this region alongside cognitive deficits in recall of semantic knowledge (Chen et al., 2020; Czarnecki et al., 2008).

Despite ATL playing a role in semantic convergence and cognition, there are numerous statistical regularities in visual input that can guide formation of semantically meaningful representations in a purely perceptual manner, often discussed in terms of statistical learning (Turk-Browne, 2012). While signatures of statistical learning are present in numerous brain areas including medial temporal lobe and frontal cortex, there is evidence for learning from concurrent sequences in as early as V1 (Rosenthal et al., 2018) and signals of suppressed activation to predictable object pairs are present throughout the ventral stream (Richter et al., 2018). Recent work in macaques used a behavioural training paradigm followed by fMRI recordings to show that more regularly recurring pairs of objects resulted in different processing patterns to random pairs of objects in occipitotemporal cortex and EVC (Vergnieux & Vogels, 2020). Indeed, the relevance of statistical patterns in the environment for learning concepts is well accounted for, and the idea does find its way into CNN models even if it is at too small a scale to learn big-picture relational structure. Turning to the infant literature, we can explore how this statistical information is used early in life to facilitate conceptual learning, and how this may improve computational models of object recognition.

Inspiration for better semantic learning from developmental science

Studies of looking time habituation have firmly established that infants are sensitive to the statistical regularities of the environment. This has been shown to be a domain general learning mechanism that spans modalities (Fiser & Aslin, 2002; Kirkham et al., 2002; Saffran et al., 1996). Evidence even exists for statistical learning in as young as new born infants, albeit highly constrained by limited cognitive resources (Bulf et al., 2011). While there are meaningful spatial regularities within typical CNN inputs, say single ImageNet exemplars, the wider context and scene where one may find an object is certainly lacking from traditional computer vision datasets. Even in scene-focused MSCOCO (Lin et al., 2014) there is no consideration of temporal consistencies, an important contextual cue that has been shown to affect recognition in parahippocampal cortex causing two images to be represented as more similar when preceded by similar versus different stimuli (Turk-Browne et al., 2012). Undoubtedly, the impoverished and highly-curated nature of such datasets (which are undeniably beneficial for their designed engineering purposes) can only prevent these models from learning more robust, generalisable conceptual structure. While the general goal of unsupervised methods is comparable to infant statistical learning on the surface – both find patterns in the data that are useful for specific tasks - the two fields remain quite separate. This is despite early enthusiasm from influential computational neuroscientist Horace Barlow in relating unsupervised visual recognition to connectionist models (Barlow, 1989), and subsequent explorations of his theory in behavioural statistical learning paradigms (Fiser & Aslin, 2001). We know that infants are sensitive to these patterns, but to expose machine learning models to entirely naturalistic experience is unfeasible. However, there is compelling evidence that elements of infants' visual input is tailored in specific ways towards facilitating better object representation learning and these may be worthy lines of inquiry to follow in computer vision.

Using headcams attached to infants in a naturalistic play setting, Linda Smith and colleagues have revealed the interesting ways that young children create their own "training data" in a manner that best facilitates learning (Pereira et al., 2014). The authors show that when playing with toys and having parents assign labels or names to the objects, those words that were successfully learned post-play had a distinct visual signature. Infants manipulated the

object so that it was centred, taking up a large portion of the visual field and sustained in view before and after word naming. These types of manipulations are increasingly common with development, and prevalent in results from head cam studies in older toddlers (Yu & Smith, 2012). This raises an interesting suggestion that infants are creating the types of visual input that best enables their learning for assigning labels to objects. This input is starkly different to that given to machine learning models. When examining head-cam data from meal times, it was shown that despite a typically cluttered visual environment a very small set of objects are present much more often; few things in the environment are very common and many things are rare with the most frequently encountered objects going on to become earlier learned words (Clerkin et al., 2017). In contrast, machine learning models receive huge numbers of image examples and learn numerous different types of things from the get-go. As argued in Smith & Slone (2017), perhaps building these developmental considerations into CNNs may lead to benefits in machine learning. Infants appear to learn completely about a few things first which then enables later rapid generalisation to the wider space of knowledge from very limited experience; maybe by taking the opposite approach in computer vision models we are limiting their potential to form robust, generalisable representations that can fully account for anterior visual cortex.

Bambach *et al.* (2018) develop this fascinating connection between infant-derived views of the world and CNN training. Once again using headcam data, including eye tracking and a model of foveated vision, the authors show that the views from infant data lead to more robust learning and generalisation in the CNN than adult-derived images (Bambach et al., 2018). This suggests that the subset of data selected by the infants through their self-generated views of the world through play actually provide more suitable training inputs for learning about objects in general. This once again highlights how the data we are currently using to train CNNs is impoverished; not only is it highly unnaturalistic in its content and distribution of categories, it has also been subject to adult judgements of what is a "good image" when really those seen by infants are inherently more instructive to visual learning. The clear discrepancies between the types of training data from which humans and CNNs learn make obvious why these models are failing to fully capture the explainable variance in the semantic visual regions LOT and VOT (Xu & Vaziri-Pashkam, 2021). The highly-curated

machine learning datasets simply don't possess the wealth of statistical information to which we have access in a naturalistic setting early in our lives.

Bringing better semantic understanding into CNN models

There are some studies developing the idea that visual co-occurrence statistics lead to improved semantics in object representations, often by taking the distributional hypothesis of NLP algorithms as a starting point. Using latent similarity analysis it was shown that non-verbal co-occurrences of objects in a visual dataset allow for concepts to be meaningfully categorised, a proof of principle that visual perceptual patterns hold merit for constructing conceptual representations (Sadeghi et al., 2015). In work that linked this idea to neural data, Bonner and Epstein (2020) constructed a word2vec-inspired model called object2vec and used this deep learning network to perform voxel-wise encoding as was introduced in previous papers (Bonner & Epstein, 2020). They show that the distributional statistics of the objects in scenes that are captured by object2vec embeddings are predictive of anterior PPA, and that those derived from the text-based word2vec were also predictive of these semantic regions. In an unexpected but satisfying result, the authors show that those regions best predicted by the object-based model tended to be scene-selective whereas those predicted by the language models were object-selective. This indicates that there may be some added weight on the distributional statistics between objects in a scene for understanding the overall image, as was put forward by previous research (Biederman et al., 1982; Oliva & Torralba, 2006; Palmer, 1975). Note that the object2vec model proposed here and the latent similarity analysis performed by Sadeghi et al. still use text-based object labels instead of a purely pixel-based input. A true account of relations emergent from visual input demands efforts to be made from extracting this information from purely visual input, and not being reliant on manual annotations.

The relevance of capturing meaningful relations within a perceptual system for forming concepts is brought together in elegant modelling work that aligns unsupervised embeddings (Roads & Love, 2020). The authors show that the distinct signature of a concept in one system (say a visual CNN) is recapitulated in another (a text-based model), such that the idiosyncratic signals between the concepts within each can be used to align the two representations,

forming a more holistic and meaningful representation of knowledge. What's really key here is that the relational structure guides understanding; if everything was perceived as equidistant then alignment across systems would not be possible. This makes clear the importance of fully capturing the global structure that supervised CNNs are biased against. Further interesting results from Roads and Love (2020) show how a system comprised of few concepts can be better aligned across modalities by restricting efforts to those words learned earlier in life. This is a fascinating finding when taken into consideration with the above work by Smith and colleagues that suggests infants and toddlers are creating for themselves better training data from which to learn. Are the optimal representations learned by infants, enabled by their innate tendency to create the most beneficial inputs, those that best facilitate alignment across unimodal systems like in Roads and Love's modelling work? Moreover, is this computational account of semantic learning akin to the hub and spoke theory of semantic cognition (Lambon Ralph et al., 2017) whereby important structure is learned in sensory cortices and alignment then occurs in ATL? Note that this does not remove the onus on perceptual regions to encode meaningful semantics. In fact, it emphasises a need to extract meaningful relations from the perceptual input so that robust alignment of concepts can occur, as was discussed by Hafri & Firestone (2021). This is only something that can be achieved by a system that has been given meaningful inputs, and is sensitive to the co-occurrence patterns within the data.

Conclusion

Deep learning, and specifically CNNs, hold their weight as models for human vision. However, their unnaturalistic learning methods inhibits learning of robust, generalisable and semantically meaningful embeddings that fully capture anterior visual cortex responses. To overcome this problem, inspiration can be found in infant development in the form of unsupervised learning mechanisms, inputs that are tailored for object representation learning and naturalistic datasets that preserve important statistical regularities in spatiotemporal co-occurrences. By implementing and exploring these methods in machine learning, innovative progress can be made towards improved computer vision that better models neural processes.

Bibliography

Anderson, A. J., Bruni, E., Bordignon, U., Poesio, M., & Baroni, M. (2013). Of Words, Eyes and Brains: Correlating Image-Based Distributional Semantic Models with Neural Representations of Concepts. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1960–1970. https://www.aclweb.org/anthology/D13-1202

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2018). *Toddler-Inspired Visual Object Learning*. https://www.researchgate.net/publication/328789075_Toddler-Inspired_Visual_Object_Learning

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449–454. https://doi.org/10.1073/pnas.0507062103

Barlow, H. B. (1989). Unsupervised Learning. *Neural Computation*, *1*(3), 295–311. https://doi.org/10.1162/neco.1989.1.3.295

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-X

Bonner, M. F., & Epstein, R. A. (2020). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *BioRxiv*, 2020.03.09.984625. https://doi.org/10.1101/2020.03.09.984625

Brendel, W., & Bethge, M. (2019). Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. *ArXiv:1904.00760 [Cs, Stat]*. http://arxiv.org/abs/1904.00760

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47. https://doi.org/10.1613/jair.4135

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127–132. https://doi.org/10.1016/j.cognition.2011.06.010

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, *10*(12). https://doi.org/10.1371/journal.pcbi.1003963

Chen, Y., Huang, L., Chen, K., Ding, J., Zhang, Y., Yang, Q., Lv, Y., Han, Z., & Guo, Q. (2020). White matter basis for the hub-and-spoke semantic representation: Evidence from semantic dementia. *Brain*, *143*(4), 1206–1219. https://doi.org/10.1093/brain/awaa057

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. https://doi.org/10.1038/srep27755

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160055. https://doi.org/10.1098/rstb.2016.0055

Czarnecki, K., Duffy, J. R., Nehl, C. R., Cross, S. A., Molano, J. R., Jack, C. R., Jr, Shiung, M. M., Josephs, K. A., & Boeve, B. F. (2008). Very Early Semantic Dementia With Progressive Temporal Lobe Atrophy: An 8-Year Longitudinal Study. *Archives of Neurology*, *65*(12), 1659–1663. https://doi.org/10.1001/archneurol.2008.507

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Derby, S., Miller, P., Murphy, B., & Devereux, B. (2018). Using Sparse Semantic Embeddings Learned from Multimodal Text and Image Data to Model Human Conceptual Knowledge. *ArXiv:1809.02534 [Cs]*. http://arxiv.org/abs/1809.02534

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports*, *8*(1), 10636. https://doi.org/10.1038/s41598-018-28865-1

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194. https://doi.org/10.1016/j.neuroimage.2016.10.001

Fiser, J., & Aslin, R. N. (2001). Unsupervised Statistical Learning of Higher-Order Spatial Structures from Visual Scenes. *Psychological Science*, *12*(6), 499–504. https://doi.org/10.1111/1467-9280.00392

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822–15826. https://doi.org/10.1073/pnas.232472899

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Fukushima, K., & Miyake, S. (1982). Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. In S. Amari & M. A. Arbib (Eds.), *Competition and Cooperation in Neural Nets* (pp. 267–285). Springer. https://doi.org/10.1007/978-3-642-46466-9_18

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Guclu, U., & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Hafri, A., & Firestone, C. (2021). The Perception of Relations. *Trends in Cognitive Sciences*, *25*(6), 475–492. https://doi.org/10.1016/j.tics.2021.01.006

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 1026–1034. https://openaccess.thecvf.com/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. https://doi.org/10.1016/S0010-0277(02)00004-5

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, *71*(3), 856–867. https://doi.org/10.1152/jn.1994.71.3.856

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran Associates, Inc. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 740–755). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48

MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience*, *14*(10), 1323–1329. https://doi.org/10.1038/nn.2903

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *ArXiv:1310.4546 [Cs, Stat]*. http://arxiv.org/abs/1310.4546

Oliva, A., & Torralba, A. (2006). Chapter 2 Building the gist of a scene: The role of global image features in recognition. In S. Martinez-Conde, S. L. Macknik, L. M. Martinez, J.-M. Alonso, & P. U. Tse (Eds.), *Progress in Brain Research* (Vol. 155, pp. 23–36). Elsevier. https://doi.org/10.1016/S0079-6123(06)55002-2

Orban, G. A., Van Essen, D., & Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends in Cognitive Sciences*, *8*(7), 315–324. https://doi.org/10.1016/j.tics.2004.05.009

Palmer, tephen E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*(5), 519–526. https://doi.org/10.3758/BF03197524

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A Bottom-up View of Toddler Word Learning. *Psychonomic Bulletin & Review*, *21*(1), 178–185. https://doi.org/10.3758/s13423-013-0466-4

Quek, G. L., & Peelen, M. V. (2020). Contextual and Spatial Associations Between Objects Interactively Modulate Visual Processing. *Cerebral Cortex*, *30*(12), 6391–6404. https://doi.org/10.1093/cercor/bhaa197

Richter, D., Ekman, M., & Lange, F. P. de. (2018). Suppressed Sensory Response to Predictable Object Stimuli throughout the Ventral Visual Stream. *Journal of Neuroscience*, *38*(34), 7452–7461. https://doi.org/10.1523/JNEUROSCI.3421-17.2018

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, *2*(1), 76–82. https://doi.org/10.1038/s42256-019-0132-2

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Rosenthal, C. R., Mallik, I., Caballero-Gaudes, C., Sereno, M. I., & Soto, D. (2018). Learning of goal-relevant and -irrelevant complex visual sequences in human V1. *NeuroImage*, *179*, 215–224. https://doi.org/10.1016/j.neuroimage.2018.06.023

Rotaru, A. S., & Vigliocco, G. (2020). Constructing Semantic Models From Words, Images, and Emojis. *Cognitive Science*, *44*(4), e12830. https://doi.org/10.1111/cogs.12830

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *Journal of Neuroscience*, *30*(39), 12978–12995. https://doi.org/10.1523/JNEUROSCI.0179-10.2010

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, *76*, 52–61. https://doi.org/10.1016/j.neuropsychologia.2014.08.031

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., & Tolias, A. S. (2019). Engineering a Less Artificial Intelligence. *Neuron*, *103*(6), 967–979. https://doi.org/10.1016/j.neuron.2019.08.034

Smith, L. B., & Slone, L. K. (2017). A Developmental Approach to Machine Learning? *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02124

Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. *Neuron*, *79*(5), 1025–1034. https://doi.org/10.1016/j.neuron.2013.06.034

Stein, T., Kaiser, D., & Peelen, M. V. (2015). Interobject grouping facilitates visual awareness. *Journal of Vision*, *15*(8), 10–10. https://doi.org/10.1167/15.8.10

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and Exemplar-Based Information in Natural Language Categories. *Journal of Memory and Language*, *42*(1), 51–73. https://doi.org/10.1006/jmla.1999.2669

Storrs, K. R., & Kriegeskorte, N. (2019). Deep Learning for Cognitive Neuroscience. *ArXiv:1903.01458 [Cs, q-Bio]*. http://arxiv.org/abs/1903.01458

Tanaka, K. (1996). *Inferotemporal Cortex and Object Vision*. 31.

Taylor, K. I., Moss, H. E., & Tyler, L. K. (2007). The conceptual structure account: A cognitive model of semantic memory and its neural instantiation. In *Neural basis of semantic memory* (pp. 265–301). Cambridge University Press. https://doi.org/10.1017/CBO9780511544965.012

Turk-Browne, N. B. (2012). Statistical Learning and Its Consequences. In M. D. Dodd & J. H. Flowers (Eds.), *The Influence of Attention, Learning, and Motivation on Visual Search* (pp. 117–146). Springer. https://doi.org/10.1007/978-1-4614-4794-8_6

Turk-Browne, N. B., Simon, M. G., & Sederberg, P. B. (2012). Scene Representations in Parahippocampal Cortex Depend on Temporal Context. *Journal of Neuroscience*, *32*(21), 7202–7207. https://doi.org/10.1523/JNEUROSCI.0942-12.2012

Vergnieux, V., & Vogels, R. (2020). Statistical Learning Signals for Complex Visual Images in Macaque Early Visual Cortex. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00789

Willems, R. M., & V. Peelen, M. (2021). How context changes the neural basis of perception and language. *IScience*, *24*(5), 102392. https://doi.org/10.1016/j.isci.2021.102392

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, *12*(1), 2065. https://doi.org/10.1038/s41467-021-22244-7

Yamins, D. L. K., 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*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2). https://doi.org/10.1016/j.cognition.2012.06.016

Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(9), 2805–2824. https://doi.org/10.1109/TNNLS.2018.2886017

Zaadnoordijk, L., Besold, T. R., & Cusack, R. (2020). The Next Big Thing (s) in Unsupervised Machine Learning: Five Lessons from Infant Learning. *ArXiv Preprint ArXiv:2009.08497*.

Zhuang, C., Yan, S., Nayebi, A., & Yamins, D. (2019). Self-supervised Neural Network Models of Higher Visual Cortex Development. *2019 Conference on Cognitive Computational Neuroscience*. 2019 Conference on Cognitive Computational Neuroscience, Berlin, Germany. https://doi.org/10.32470/CCN.2019.1393-0