Deep neural network models of visual cognition

Sörensen, L.K.A.

Publication date
2023

Chapter 7

Discussion

In what follows, I will first revisit the motivation for using DCNNs to study visual cognition. Subsequently, I will discuss the challenges and opportunities connected to this approach, using the chapters of this thesis as examples throughout.

7.1 Motivation

Why should we bother with complicated models such as DCNNs when we want to understand something about visual cognition? After all, theories of visual cognition are known for their elegant abstractions and computational subdivisions (e.g., Rensink, 2000; Wolfe, 2021). Traditional models of visual cognition are straightforward to reason about and therefore potentially easier to understand. Based on the research in this thesis, I argue that the main reason to embrace DCNNs for studying visual cognition is that they provide a good proxy of visual representations. This makes it possible to not only model visual cognition in the abstract, but to directly link it to the substrate of visual cognition in the brain — that is, visual representations.

In addition to this, DCNNs also have superior predictive ability compared to abstract models of visual cognition. That is, since DCNNs are effectively a weighted mapping between an image and an output label, a model's representations and outputs can be directly accessed and used to predict neural data and behavioural responses to the same stimuli. This

contrasts with abstract models, which usually cannot process images and therefore lack image-specific predictions.

This is not to say that more abstract models are not helpful — they are very much necessary for theorizing and integrating across empirical studies. Yet, the motivation for the work in this thesis is to implement mechanisms often developed in abstract models, and to evaluate them when faced with rich sensory representations and challenging tasks. Establishing a link to task performance and visual representation is critical to ensure that our modelling efforts lead us to recover the behavioural hallmarks of visual cognition: an effective integration of visual processing with our goals, internal state and knowledge of the world.

7.2 Challenges

While my outlook for using DCNNs as a framework for studying visual cognition is optimistic, there are some specific challenges associated with this approach: (1) the difficulty of building task-performing models of biological processes, (2) the challenge of understanding the driving factors behind the representational alignment of DCNNs and humans during object recognition and (3) the challenge to establish a comprehensive mapping between model and human behaviour.

The first challenge is at the heart of every modelling effort, namely: what is a good neural mechanistic model for a given behavioural phenomenon? Consider selective attention, often described as a spotlight or selective filter in its effect on object recognition. How should we set up a model to reproduce such selective processing in its performance? Asking this question naturally leads us to see the many choices and trade-offs researchers make when modelling neural processing, sometimes favouring interpretability over complexity or abstraction over realism. For instance, one could argue that a good model of the visual ventral stream should share as many aspects as possible with biological brains. Yet, this goal quickly conflicts with other modelling goals such as a model's interpretability and abstraction. In practice, model realism also oftentimes is at odds with a model's performance on a challenging task such as object recognition. That is, simulating every (known) aspect of neural processing is computationally intensive, and for some biological details (e.g., oscillations) we do not even know yet whether they are functionally relevant,

or rather reflect some other property of the system. In contrast, the simplified neural units of DCNNs work so well partly because they can be scaled to hundreds of thousands of neurons and billions of connections. Beyond interpretability and abstraction, there thus exists another trade-off between task performance and biological realism. Throughout this thesis, I chose for the simplest model possible that allowed me to still observe task performance on images as well as to implement the mechanism of interest into the model. As a result, I chose a different model family in almost every chapter. Most of them were feedforward convolutional neural networks (e.g., Chapter 3) enabling task performance, and some had more biologically inspired components, such as spiking neurons (Chapter 2), or recurrent connectivity (Chapter 4 and 5).

Computational modelling of neural processing has the difficulty that it abstracts away and is disconnected from the biological organism, in contrast to neural or behavioural data. Thereby, our modelling choices determine how valid our conclusions will be for the modelled system and its phenomena. Still, researchers can make a virtue of necessity by being transparent about their modelling choices and promoting the reproducibility of their results. For one, by being explicit about a model's components and the conditions during which certain phenomena arise, we can contextualize our findings in light of the modelling choices made. Furthermore, by sharing code, parameter choices and stimuli to reproduce the gained insights, we can enable other researchers to explore the same theoretical space or make predictions for new situations and stimuli. Together, this may remedy the shortcomings of individual studies and help to identify productive trade-offs between realism, task performance, and abstraction for modelling of visual cognition.

The research in this thesis builds on the representational parallels observed between the visual ventral stream and DCNNs optimized for object recognition. Understanding the nature and origins of these parallels is the second challenge that will inform and shape the use of DCNNs for studying visual cognition. Currently, it is not clear what the key factors for an alignment with neural representations and behaviour are. While initially the architecture, learning function and training objective were deemed the most important factors for the alignment with neural processing (Kriegeskorte, 2015b; Marblestone et al., 2016; Richards et al., 2019), more recent empirical studies observe that a model's ability to predict fMRI activity is only partly explained by these factors (Conwell et

al., 2022; Storrs et al., 2021), emphasizing the importance of the mapping strategy between model and neural activations, as well as of the training data and objective instead (Cadena et al., 2022; Dwivedi et al., 2021; Lindsay et al., 2021; Nayebi et al., 2021; Zhuang et al., 2021). Still others argue that the higher latent dimensionality that some DCNNs exhibit is the decisive factor for yielding an alignment with neural representations (Elmoznino & Bonner, 2022). Like neural representations (e.g., Xu & Vaziri-Pashkam, 2021), the behaviour of human observers can also differ qualitatively from those of DCNNs (Baker et al., 2018; Geirhos et al., 2018; Jacob et al., 2021) and here as well, the best predictor for a fit to human data along a series of image transformations was the size of the dataset that a model was trained on rather than other architectural constraints or the learning rule (Geirhos et al., 2021). Together, these studies make it clear that understanding the driving factors behind the representational parallels in brains and models will continue to be an active area of research in the future and will present a challenge for using DCNNs as mechanistic models of the visual ventral stream (see also Cao & Yamins, 2021a, 2021b for an extensive discussion on this).

As research continues and gains new insight into the alignment between DCNNs and the visual ventral stream, it will be critical to also update models of visual cognition to incorporate DCNNs with an improved fit to visual processing and human behaviour. Morre recent studies already emphasize the importance of rich statistical features derived from large, diverse training datasets, combined with highly parametrized models and semantic objectives as the core driver for an effective mapping between models, brains and behaviour. For instance, the currently most predictive model for human behaviour (*CLIP*, Radford et al., 2021) was jointly optimized on images and text descriptions, presumably teaching the model more about the visual world than can be seen in an image. For studying visual cognition, models like this will be essential to further define the boundaries and computational task of visual cognition. For instance, how much of the adaptive, goal-driven visual recognition associated with visual cognition in humans can be achieved by learning from large-scale multi-modal training data with these models? Are there task objectives that will lead a model to develop selective attention, working memory or arousal state as a by-product of solving another task, given the right architecture and training data? One example that such a strategy can be effective is the optimization for next-word prediction in language models (e.g., Brown et al., 2020; Vaswani et al., 2017), which was

not only shown to subsume various other computational problems (e.g., compositionality), but also to result in model features highly predictive of language processing in the brain (Caucheteux & King, 2022; Schrimpf et al., 2021). For studying human visual cognition, aiming to model complex visual tasks using DCNNs is therefore a natural steppingstone.

A third challenge for modelling visual cognition using DCNNs is that current DCNN models are optimized for accurate performance, whereas human behaviour is much more flexible and varies along many output dimensions depending on the context. For instance, consider how humans flexibly adjust their reaction times and accuracy in response to an instruction or some other contextual factor. This flexibility makes that simply comparing between DCNN and participant responses with regard to their accuracy is not sufficient, since it might not generalize to another behavioural experiment with different instructions (for instance, emphasizing accuracy over speed). Instead, a more principled way to translate model outputs to human performance measures would be desirable. This problem is also illustrated by the results in Chapter 4 and 5 where I linked the same model to multiple output measures. I observed that by using multiple mapping methods I can come to conflicting conclusions across performance measures. That is, for example in Chapter 5, I found that while some models were a good fit with human reaction times, they did not match the accuracy levels of human participants. This suggests that some of our strategies for translating model outputs to human-like performance measures may be overfitting on the chosen behavioural measure or task.

For future efforts, it will be critical to develop a more comprehensive mapping approach between human and model performance that takes various behavioural measures into account. Inspiration for this could come from sequential sampling models, a well-established cognitive modelling framework that accommodates both accuracy as well as reaction time distributions (Forstmann et al., 2016; Ratcliff & McKoon, 2008). Augmenting DCNNs with a principled way to generate various behavioural responses may be an attractive avenue to marry these different modelling traditions, as well as to have a more stringent evaluation of the match between model and human behaviour along a multitude of behavioural measures.

7.3 Opportunities

While using DCNNs as a framework for studying visual cognition does not come without its challenges (as described above), this approach also brings about clear opportunities: (1) DCNNs enable us to build and test more complex sensory simulations of cognitive modulations; (2) DCNNs can serve as a testbed to provide a proof-of-principle as well as compare different mechanisms for their effects on performance; (3) DCNNs afford the modelling of inter-stimulus interactions, such as similarity, a core dimension in different frameworks of visual cognition.

First, sensory-enriched models provide the opportunity to make more far-reaching predictions about complex interactions between mechanisms and general principles in visual processing. Mechanisms proposed for cognitive modulation are often derived from observations in neural data. But these observations cannot cover all circumstances and areas in which such a modulation can occur. This also limits our theorizing about such a modulation to a set of neurons, or a cortical area. Yet, cortical processing is more complex, spanning millions of neurons, organized along various functional principles. For instance, neuroscientists have identified the functional hierarchy in visual processing as one such principle (Felleman & Van Essen, 1991). Other examples of such general principles are neural tuning (Hubel & Wiesel, 1959) and divisive normalization (Carandini & Heeger, 2011). DCNNs were designed to share some of these principles with the visual ventral stream. In Chapter 2, I took advantage of this and systematically explored how a global gain modulation — previously mainly observed in early visual cortices — may interact with a hierarchy of sensory representations. This was motivated by the idea that global gain changes linked to changes in arousal state occur brain-wide and are thus not exclusive to early visual processing. This approach enabled us to test the functional implications of this proposed interaction for object recognition. It also allowed us to understand that a global gain mechanism acting on such a sensory hierarchy readily gives rise to a commonly observed interaction between arousal state and task difficulty, known as the Yerkes-Dodson effect (Yerkes & Dodson, 1908). This interaction in performance was not previously linked to an effect of arousal state on the sensory hierarchy. While it was certainly not a new idea to model global gain in a neural network (e.g., Servan-Schreiber et al., 1990), in this case it was critical to simulate these effects in a complex and task-performing model, such as a DCNN, in order to clarify the

link between a mechanism, a model's functional hierarchy, and its object recognition performance. This example thus illustrates how DCNNs can be used to simulate more general interactions between a mechanism and sensory processing.

Second, DCNNs are a testbed for comparing different mechanisms with regard to their effects on object recognition performance. That is, using DCNNs as a base model and augmenting the same model with different mechanisms, we can compare different implementations of a mechanism or different mechanisms altogether, while tracking changes in task performance. Critically, DCNNs offer researchers an unprecedented degree of control: of a model's visual diet during training, its objective, its access to certain representations when performing classification, as well as its degree of plasticity during training. This has the marked advantage that these factors can be excluded as an explanation when observing differences between mechanisms during experiments. In this thesis, Chapter 3, 4 and 5 contain examples of the effectiveness of this approach. In Chapter 3, I used a spiking DCNN as a base model and compared three different mechanisms implementing selective spatial attention within the same model. When the model received a valid or invalid cue, performance changed for all three mechanisms, but not to a similar extent — some mechanisms were more effective at biasing object recognition towards the attended location. This adds a nuance to the evaluation of a mechanism: not simply assessing *whether* a mechanism can work, but rather *how well* it works. In a similar vein, in Chapter 4, I also pursued this approach by successively adding different mechanisms (i.e., lateral recurrence and adaptation) to the same model while tracking its ability to perform object detection on RSVP sequences. This way, synergies between mechanisms became apparent, such as reported for sensory adaptation and lateral recurrence during dynamic object recognition. Taken together, adopting DCNNs as a testbed for mechanisms and their implementations can thereby enrich our understanding of the functional contributions of a mechanism to object perception and visual cognition.

Third, DCNNs enable us to assess interactions between sensory stimuli. This aspect is of particular relevance for visual cognition, where visual relationships such as similarity have long been known (Koffka, 1935; Wertheimer, 1923) to affect visual search and perceptual grouping (e.g., Duncan & Humphreys, 1989). However, estimates of similarity are often not derived from image-computable models, but rather based on dimensions considered relevant for human cognition a priori, such as letters,

shapes or colour categories (but see Hebart et al., 2020 for an alternative approach). DCNNs provide an opportunity to assess these similarity relationships in an image-computable fashion, based on features dedicated to object recognition (but see Jozwik et al., 2017, 2022). As an example, consider how in Chapter 4 I linked models of sequential stimulus processing to a participant's report rate for a given trial. This pointed to a systematic link between the visual representations evoked by a sequence of images, the timing of the sequence, and the ability to report on the presence of a target category.

Whereas this example describes how stimulus similarity modulates object recognition over time, it is easily conceivable that such interactions based on visual and semantic similarity may extend to domains such as working memory, selective attention, search templates and categorical learning. For instance, a previous study has shown that implementing attentional selection into a DCNN according to the feature similarity gain principle observed in neural data is an effective way to mimic the magnitude of behavioural attentional modulation (Lindsay & Miller, 2018). Moreover, a number of studies showcase that there is a shared notion of visual and semantic similarity between perceptual judgements and processing in higher visual cortices (Jozwik et al., 2016; Mur et al., 2013; Op de Beeck et al., 2008; Wardle et al., 2016). In line with that, the dissimilarity of neural activations derived from higher visual cortices has also been shown to predict reaction times on a categorical visual search task (Cohen et al., 2017). Together, these examples show how notions of neural similarity directly affect visual cognition and highlight the appeal of capturing such similarity using DCNNs when studying visual cognition.

7.4 Concluding remarks

Understanding the interaction between sensation and cognition is challenging, partly because it is hard to know where sensation ends, and cognition begins. In recent years, DCNNs and their rich statistical sensory features have unequivocally demonstrated that learning from sensory inputs can be a powerful technique for building models that are able to solve challenging problems, such as object recognition. In this thesis, I have explored how cognitive modulation, a hallmark of human behaviour, may add to such a capable sensory system to support a wider and more flexible set of behaviours.