What Possible Use Could Consciousness Be?

Joshua Bensemann JOSH.BENSEMANN@AUCKLAND.AC.NZ
Michael Witbrock M.WITBROCK@AUCKLAND.AC.NZ
The Broad AI Lab, School of Computer Science, The University of Auckland, New Zealand

Abstract

Consciousness is a widely studied phenomenon, but is still poorly understood. In this very preliminary work, we consider properties of consciousness in the context of systems made up of largely uniform computational elements, such as neurons or Deep Learning units. We argue that the phenomenology of consciousness may be, in part, a consequence of this uniformity and the brain's need to compensate for it in identifying sources of information. We performed some initial experiments to show that the resulting representations can be useful in improving downstream tasks. This utility may explain in part why consciousness has been conserved in humans.

1. Introduction

There has been a great deal of philosophical, psychological and, to a lessor extend, artificial intelligence (AI) research into the phenomenon of consciousness (see Gamez, 2008; Reggia, 2013). In this working paper, we discuss an effort to describe some properties of consciousness as a consequence of the physical indistinguishability of external from internal signals within the brain, the utility of learning to make those distinctions, and the consequential utility of having done so. While the first two points are speculative, we develop a simple experimental setup that shows that developing a perception-like representation[1] of the task being performed by a Deep Neural Network (DNN) significantly improves performance in multi-modal perception.

1.1 Signal Indistinguishability

To a very great extent, neural signals incident on a neuron are indistinguishable as to origin; despite this limitation, multiple types of information can be contained within aggregations of these signals. These signals are then somehow decoded to serve as representations, some of which act like symbols, for example, sound, vision, thoughts. Additionally, signals can be representations of other representations, (i.e., higher-order thought or HOT; see Lau & Rosenthal, 2011). While it is apparent that HOT allows for more complex representations of information, what is less understood is the benefit of having multi-modal conscious experiences of these representations (i.e., phenomenal consciousness or P-Consciousness; Block, 1995). The key difference between HOT and P-Consciousness is that the former is the manipulation of information without any awareness of the information, whereas the latter is the experience of that information as such. In other words,

1. A representation that, in an act of hubris, we might liken to a quale.

humans and DNNs can use images to make decisions but only the humans visually experience (, or *see*,) the images.

1.2 Purpose of Phenomenal Consciousness

The purpose of P-Consciousness has long been the subject of debate. Some speculated that it creates stability within cognitive processes (Ramachandran & Hirstein, 1997). For example, having a common representation of colour allows that same representation to be used in future. Others suggest that P-Consciousness can be used to determine what is occurring in real-time, such as images captured by the eyes, compared to mental images of prior events (Gregory, 1998). Others suggest that P-Consciousness could have some predictive function; the experience of being in control your body appears to be a contrast between what you intend to do and the feedback you receive after attempting it (Hohwy & Frith, 2004). Or perhaps P-Consciousness has no purpose (see Dennett, 2016).

All speculation about phenomenal consciousness mentioned thus-far applies to how it benefits humans. However, by adding components of P-Consciousness to existing AI systems, it is possible to observe whether the systems performance is enhanced. Research of this type is known as synthetic phenomenology (see Chrisley, 2009) and provides a stepping stone on the path to identifying the utility of such a function in humans. Here, we have used DNNs, whereas related research has used other AI systems (Arrabales et al., 2011; Zaadnoordijk & Besold, 2019).

1.3 The Experiment

Signals in DNNs are the values being propagated throughout the network and are also indistinguishable from one another. An individual node does not contain enough information to identify the input received, as different inputs may lead to similar values being propagated through that node. Although the values arriving at the output layers often provide enough of a signal to complete classification tasks, we were interested in determining whether knowing what kind of processing was being done in earlier layers of the network could be used to improve overall performance.

First, we trained a DNNs to classify both sound and image inputs. Without explicitly training the network to differentiate between the two file types we were able to recover whether the input was originally a sound or an image using the information present in the network's hidden layer, approximating a conscious symbol for the type of processing the network was performing. Second, we trained DNNs using both the original inputs and this recovered "experiential" information. By doing this, we were able to improve performance of the networks. In terms of P-Consciousness, this experiment was to see whether being able to distinguish between (,or *"experience"*) the processing of two types of information could make an impact on cognitive performance.

2. Method

2.1 Data

The dataset consisted of 77 classes; 47 of image data (balanced EMNIST dataset; Cohen et al., 2017) and 30 of sound data (Speech commands dataset; Warden, 2017). Several steps were required to

convert the sounds into a format compatible with the images. First, each sound was converted into a Mel spectrogram (sampling rate 22050 per second) producing 128 x 44 representations. To match the 28 x 28 size of images, the sounds were zero-padded to a size of 140 x 56, and max pooling (pool size 5 x 2) was then applied. 20% of the sound data were then reserved for validation. The images and sounds were normalised to values between 0 and 1 and combined to make a 77 class dataset.

2.2 Models

The base model consisted of 3 convolution layers followed by two dense layers and an output layer. Hyperband optimisation (Li et al., 2018) was used to determine the best parameters for each layer.

To determine whether the models were developing "concepts" for input type, a Hebbian learner (Hebb, 1949) was attached to each of the base model's dense layers. These learners received the linear outputs from their layer. The outputs were divided by their absolute value, converting each into 1 or -1. The Hebbian learners were then trained using the output data as well as a 1 or -1 label representing whether the layers output was generated via a sound or image file, respectively.

The aided model was almost identical the base model. The difference being the output from a Hebbian learner of a pre-trained base model was concatenated on to the data entering the first dense layer of the aided model. This provided the aided models with the equivalent of one bit of additional information, from the internally developed 'quale' representation of input type, during training.

2.3 Training

Ten base models were trained using the Adam optimizer (Kingma & Ba, 2015). Each training run lasted for 500 epochs of 1000 steps. A batch of 128 training examples was created each step by sampling uniformly between the images and sounds. Each sound and image sample appeared approximately 1.2 times and 0.5 times per epoch, respectively. This resulted in a balanced sound: image ratio instead of an even distribution of all 77 classes.

Ten aided models were also trained. Training began by pre-training a base model until any of the attached Hebbian learners achieved greater than 99% accuracy on the validation data. Once achieved, the base model's weights were frozen along with the more accurate Hebbian learner. After pre-training, training of the aided model proceeding like base training. The exception being that the training batch was first passed through the pre-trained base model and Hebbian learner. The output of the Hebbian learner was then input to the aided model along with the original batch and concatenated onto flattened outputs of the convolutional layers.

3. Results

3.1 Hebbian Learners

The Hebbian learners were able to accurately classify whether a base model's input was an image or sound. After 1 epoch, the accuracy of the learner attached to the 1st layer always exceeded the 99% validated accuracy criterion (mean 99.7%, SD .2%). The 2nd layer's learner was less accurate at this point. (mean 96.7%, SD 1.6%).

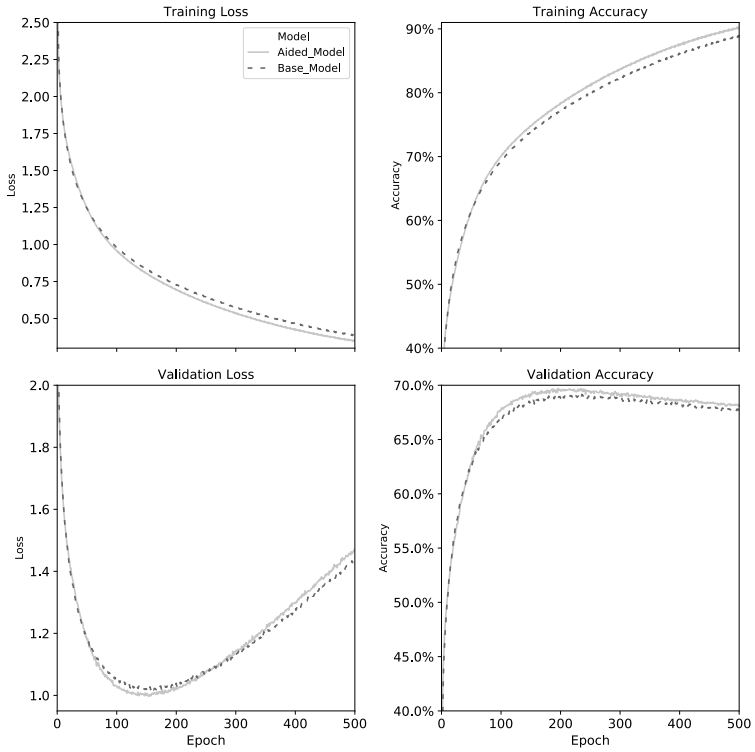

Figure 1. The averages of training loss and accuracy, as well as validation loss and accuracy during training of the base and aided models. The average aided model outperformed the average base model in all cases.

3.2 Training Performance

Visual inspection of Figure 1 shows that average training loss was lower for the aided model resulting in higher average training accuracy. The same was observed in the validation data, until training exceeded approximately 150 epochs and the models began to over-fit. The maximum validation accuracy obtained from each training run were compared using the Mann-Whitney U test. This confirmed that there was a significant difference between the validation accuracy produced by both models, providing evidence that the inclusion of additional information extracted from a pre-trained model lead to an improvement in performance ($P = .011$). Tests performed on the number of epochs required to reach best performance were non-significant, suggesting that the same number of training epochs were required for both models to obtain the best results on the validation dataset.

4. Discussion

4.1 Results so far

The research presented here are the preliminary findings from our attempts to 1) describe phenomenal consciousness as a consequence of the physical indistinguishability of neural signals, 2)

describe the utility of learning to make those distinctions, and 3) show the consequential utility of having done so. So far we have shown that it is possible to use Hebbian learners (Hebb, 1949) to distinguish between signals and that this distinction lead to significant improvements in a DNNs performance when learning a multi-modal classification task.

4.2 Consciousness and Artificial Intelligence

Although our ultimate goal is to determine the utility of P-Consciousness, it would be premature to make conclusions based on our preliminary data. However, for the sake of scientific discourse, we shall make limited comparisons between our results and selected literature.

In our experiment we emulated one property of P-Consciousness by providing DNNs with a representation of whether the current training sample was a sound or image. While we are not claiming to have produced P-Conscious machines, this did gave the model the ability to separate the modalities while learning a multi-modal classification task. This is analogise to humans learning to identify new sounds and objects; initially they will not be able to distinguish between each sound and each object, but they can use the differences in visual and auditory phenomenal experiences to separate sounds from objects.

Why the ability to separate modalities lead to improved performance is the important question. One possibility is that phenomenal experiences act as shortcut to important information. Such an idea was raised by Zaadnoordijk and Besold (2019) when discussing phenomenal experience in AI. In their view, phenomenal experiences are a direct mapping from sensory inputs to mental representations which bypass many high-level functions to provide information more efficiently. This is effectively what is happening in our experiment as the Hebbian learners mapped the original 28 x 28 inputs on to 1 bit of information, to create the most efficient representation of a binary label. With all that said, knowing that information acts as a shortcut does not specify how that shortcut improves performance.

Mapping high dimensional data on to lower dimensional representations has also been employed in other computational frameworks of consciousness. Although it does not contain a phenomenal component, Bengio's (2017) *The Consciousness Prior* conceptualises conscious states as lower-dimensional representations of selected unconscious states. While Bengio's proposed model differs greatly from our own, it should be noted that experiments inspired by *The Consciousness Prior* have achieved state-of-the-art results (Xu et al., 2019). This means that mapping information onto lower dimensions appears to cause improved performance in more than just our preliminary results.

4.3 Limitations

We have shown that the aided model out performed the base model. However, we have only tested one model architecture and this does not yet prove the existence of a general pattern. We are currently conducting future studies with wider ranges of architectures to test the robustness of the effect described here.

References

Arrabales, R., Ledezma, A., & Sanchis, A. (2011). Simulating visual qualia in the cera-cranium cognitive architecture. In *From brains to systems*, 223–238. Springer.

Bengio, Y. (2017). The consciousness prior. *arXiv preprint arXiv:1709.08568*.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and brain sciences*, *18*, 227–247.

Chrisley, R. (2009). Synthetic phenomenology. *International Journal of Machine Consciousness*, *1*, 53–70.

Cohen, G., Afshar, S., Tapson, J., & Van Schaik, A. (2017). EMNIST: Extending MNIST to handwritten letters. *Proceedings of the International Joint Conference on Neural Networks*, *2017-May*, 2921–2926.

Dennett, D. C. (2016). Illusionism as the obvious default theory of consciousness. *Journal of Consciousness Studies*, *23*, 65–72.

Gamez, D. (2008). Progress in machine consciousness. *Consciousness and cognition*, *17*, 887–910.

Gregory, R. (1998). *Flagging the present with qualia*, (pp. 200 – 209). Allen Lane/Penguin.

Hebb, D. (1949). *The organization of behavior*. New York: Wiley & Sons.

Hohwy, J., & Frith, C. (2004). Can neuroscience explain consciousness? *Journal of consciousness studies*, *11*, 180–198.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, (pp. 1–15).

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in cognitive sciences*, *15*, 365–373.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, *18*, 1–52. From `http://jmlr.org/papers/v18/16-558.html`.

Ramachandran, V. S., & Hirstein, W. (1997). Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of Consciousness Studies*, *4*, 429–457.

Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, *44*, 112–131.

Warden, P. (2017). Speech commands: A public dataset for single-word speech recognition. *Dataset available from http://download.tensorflow.org/data/speech*$_c ommands_v 0.01.tar.gz$.

Xu, X., Feng, W., Sun, Z., & Deng, Z.-H. (2019). Neural consciousness flow. *arXiv preprint arXiv:1905.13049*.

Zaadnoordijk, L., & Besold, T. R. (2019). Artificial phenomenology for human-level artificial intelligence. *AAAI Spring Symposium: Towards Conscious AI Systems*.