The Consciousness Prior

Yoshua Bengio
Université de Montréal, MILA

September 26, 2017

Abstract

A new prior is proposed for representation learning, which can be combined with other priors in order to help disentangling abstract factors from each other. It is inspired by the phenomenon of consciousness seen as the formation of a low-dimensional combination of a few concepts constituting a conscious thought, i.e., consciousness as awareness at a particular time instant. This provides a powerful constraint on the representation in that such low-dimensional thought vectors can correspond to statements about reality which are either true, highly probable, or very useful for taking decisions. The fact that a few elements of the current state can be combined into such a predictive or useful statement is a strong constraint and deviates considerably from the maximum likelihood approaches to modeling data and how states unfold in the future based on an agent's actions. Instead of making predictions in the sensory (e.g. pixel) space, the consciousness prior allow the agent to make predictions in the abstract space, with only a few dimensions of that space being involved in each of these predictions. The consciousness prior also makes it natural to map conscious states to natural language utterances or to express classical AI knowledge in the form of facts and rules, although the conscious states may be richer than what can be expressed easily in the form of a sentence, a fact or a rule.

1 Introduction

We propose here a new kind of prior for top-level abstract representations, inspired by our understanding of consciousness as a form of awareness (van Gulick, 2004), i.e., as defined by Locke, consciousness is "the perception of what passes in a man's own mind", or awareness of an external object or something within oneself (Wikipedia definition). This proposal is based on a regularization term which encourages the top-level representation (meant to be at the most abstract level) to be such that when a sparse attention mechanism focuses on a few elements of the state representation (factors, variables or concepts, i.e. a few axes or dimensions in the representation space), that small set of variables of which the agent is aware at a given moment can be combined to make a useful statement about reality or usefully condition an action or policy. We do not refer here to more elusive meanings that have been attributed to the word "consciousness" (like qualia (Kriegel, 2014)), sticking instead to the notion of attentive awareness in the moment, our ability to focus on information in our minds which is accessible for verbal report, reasoning, and the control of behaviour.

2 Consciousness Prior Theory

The following points can be derived from the basic idea introduced above, in the context of a learning agent, where we refer the reader to standard notions (Sutton and Barto, 1998) of reinforcement learning (RL).

2.1 Subnetworks

Let s_t be the **observed state** at time t and let h_t be the high-level representation derived from s_t (and from past observed values s_{t-k} in the partially observable case). For example, h_t could be the output of some kind of RNN (with whatever architecture is appropriate) that reads the sequence of s_t as input and produces an output h_t at each time step:

$$h_t = F(s_t, h_{t-1}) \tag{1}$$

where we call F the **representation RNN** or encoder and h_t the **representation state**. A core objective is to learn good representations in h_t, which disentangles abstract explanatory factors, in the sense that there exist a simple transformation of h_t which can select the information about a single factor (its value or uncertainty about it).

We can think of the representation RNN as the content of almost the whole brain at time t, i.e., the representation state h_t is a very high-dimensional vector (and probably sparse if we want to imitate biology), which is an abstract representation of the full current information available to the agent (beyond what is stored in the weights), thus summarizing the current and recent past observations.

In contrast, we will define the **conscious state** c_t as a very low-dimensional vector which is derived from h_t by a form of attention mechanism applied on h_t, taking into account the previous conscious state as context.

$$c_t = C(h_t, c_{t-1}, z_t) \tag{2}$$

where z_t is a random noise source. The cognitive interpretation is that the value of c_t corresponds to the content of a thought, a very small subset of all the information available to us unconsciously, but which has been brought to our awareness by a particular form of attention which picks several elements or projections from h_t. The function C is the **consciousness RNN** and because of its random noise inputs, produces a random choice of the elements on which the attention gets focused. This is useful if we think of the consciousness RNN as a tool for exploring interpretations or plans or to sample predictions about the future. We can also think of the consciousness RNN as the tool to isolate a particular high-level abstraction and extract the information about it (its value, uncertainty about it or even the fact that it is unobserved). This would happen if we think about a single factor, but in general C will aggregate a few (e.g. a handful) of such factors into a more complex and composed thought.

2.2 Training Objectives

To capture the assumption that a conscious thought can encapsulate a statement about the future, we introduce a **verifier network** which can match a current representation state h_t with a past conscious state c_{t-k}:

$$V(h_t, c_{t-k}) \in \mathbb{R} \tag{3}$$

which should be structured so that $V(h_t, c_{t-k})$ indicates the consistency of c_{t-k} with h_t, e.g., estimating the probability of the corresponding statement being true, given h_t.

More generally, we would like to define an objective (or reward) function which embodies the idea that the attended (conscious) elements are used in some way whose value can be quantified and optimized, i.e., that the representation RNN is trained to optimize this objective function, as well as possibly other objectives such as being able to reconstruct the raw input or any other supervised, RL, or unsupervised objective which we want to throw in, such as the independently controllable factors prior (Bengio *et al.*, 2017).

There are two distinct mechanisms at play which contribute to map the high-level state representation to the objective function: (1) the attention mechanism (e.g. the consciousness RNN) which selects and combines a few elements from the high-level state representation into a low-dimensional "conscious sub-state" object (the current content of our consciousness), and (2) the predictions or actions which are derived from the sequence of these conscious sub-states. The second mechanism is easy to grasp and frame in standard ML practice, either in deep learning or RL, e.g. for supervised or unsupervised or RL tasks. For example, the attention mechanism could select elements B from the current representation state and choose to make a prediction about future elements A. Then to improve the quality of the prediction mechanism we may just want to maximize $logP(A|B)$ or some proxy for it, e.g., using a variational auto-encoder (Kingma and Welling, 2014) objective or a a conditional GAN (Mirza and Osindero, 2014) if one wants to sample accurately an A from B. Note again that such an objective function is not just used to learn the mapping from B to A (or to probabilities over the space of A values), but also drives the learning of the representation function itself, i.e., is back-propagated into the representation RNN). However, this part of the objective function (e.g. predictive value, computed by V above) is not sufficient and in fact is not appropriate to train the attention mechanism itself (which variables A and B should be selected?). Indeed, if that was the driving objective for attention, the learner would always pick a B which is trivially predictable (and there are such aspects of reality which are trivially predictable yet do not help us to further understand the world and make sense of it or achieve our goals). It remains an open question what other objectives would be appropriate for learning how to attend to the most useful elements, but ultimately we should be able to use the actual RL reward of the learning agent for that purpose (though some shorter-term proxy might be welcome). Some form of entropy or diversity may be needed so that the attention mechanism is stochastic and can choose a very diverse set of possible attended elements, so as to cover widely the possible variables A on which a prediction is made.

2.3 Naming Variables and Indirection

It would be very convenient for the consciousness attention mechanism and for the verifier network to be able to refer to the "names" of variables on which a prediction is made. In some models, we already distinguish keys and values in variations of memory augmented neural networks (Weston *et al.*, 2014; Graves *et al.*, 2014). The conscious state must indirectly *refer* to some of the aspects or dimensions computed in the representation h. Wether this should be done explicitly or implicitly remains to be determined. A key-value mechanism also makes it easier for the verifier network to do its job because it must match just the *key* of the predicted variable with its instances in a future representation state (with that variable becomes observed). If the key and value are mixed up and the predicted value differs substantially from the observed value, a simple associative process might miss the opportunity to match these and thus provide a strong training signal (to correct the predictor).

2.4 Connection to Language and Symbolic Knowledge Representation

Linked to this is the interesting property of the conscious state that there is a fairly simple transformation of it into a natural language sentence (possibly taking into account previously uttered sentences). An externally provided sentence could also elicit an associated conscious state, although we postulate that the conscious state is generally a richer object than the uttered sentence, i.e., mapping from conscious states to sentences loses information (think about visual imagery, or artistic expression, which are difficult to put in words), and the same sentence could thus be interpreted differently depending on context and the particulars of the agent who reads that sentence. Formally, we could use another RNN to map a conscious state to an utterance u_t:

$$u_t = U(c_t, u_{t-1}). \tag{4}$$

A learning agent which uses language could thus benefit from an additional regularization term: the set of currently consciously attended elements can often be mapped to something like a sentence in natural language which may be uttered by another agent, such as a human teacher. A sentence focuses on just a handful of elements and concepts, unlike our full internal state. This imposes further constraints on the representation function in that its individual elements or dimensions are more likely to correspond to concepts which can typically be expressed by a single word or phrase. Based on these arguments, it is reasonable to hypothesize that language may actually help humans build sharper internal representations (which are better disentangled) as well as facilitate learning – see the arguments around curriculum learning (Bengio *et al.*, 2009) and cultural learning (Bengio, 2014) – and enable collaborative task-solving.

Along the same line, this research opens the door to the possibility of better connecting deep learning with classical symbolic AI and cognitive science, and move deep learning from perception (where it shines) to higher-level cognition and knowledge representation (where many questions remain open). For example, knowledge is classically represented by facts and rules: each of them is a very sharp statement about reality involving just a few concepts. Such a nugget of information or knowledge seems to fit well as a conscious state. Combining such conscious states sequentially in order to make more complex predictions and inferences or actions is basically what reasoning is about. However, I am not implying that we should return to the symbolic forms of knowledge representation, which have their well-known limitations. Instead, we may consider a form of regularization on the representations captured by deep learning agents, which could have many of attributes of classical AI facts and rules, while keeping a richer representation needed for inference and planning in the presence of uncertainty and non-discrete aspects of the world. Progress in this direction would also address the often expressed concern about obtaining explanations from deep nets, since the approach proposed here would make it easier for a trained agent to communicate verbally its high-level state.

3 Considerations for Experimenting with the Consciousness Prior

Because this is a novel theory which may be developped in many different ways, it is important to start with simple toy experiments allowing one to test and evaluate qualitatively different approaches, such that the turnaround time for each experiment is very short and the analysis of the representations learned very easy (because we already have a preconceived idea of what concepts would be the most appropriate to disentangle).

Although working with natural language input would be likely to help the agent learn better and more abstract representations, it would be better to start with experiments with no linguistic input, to make sure that it is the training objective and the training framework alone which are leading to the discovery of the appropriate high-level features. For example, learning some form of intuitive physics is done by babies without the need for linguistic guidance. Similarly, although the consciousness prior could be used in supervised learning or task-oriented RL, testing its ability alone to discover high-level abstractions would be best done in the context of unsupervised RL, e.g., using an intrinsic reward which favours the discovery of how the environment works.

It would be more interesting for the learning task to involve meaningful abstractions which have a high predictive power. For example, consider predicting whether a pile of blocks will fall on or off a table. It involves a high-level discrete outcome which can be predicted easily, even if the details of where the blocks will fall is very difficult even for humans to predict. In that case, predicting the future at the pixel level would be extremely difficult because future states have high entropy, with a highly multi-modal distribution. However, some aspects of the future may have low entropy. If in addition, these aspects have a big impact on predicting what will come next (or on taking the right decisions now), then the consciousness prior should be very useful.

In terms of experimental comparisons, it would be good to compare systems based on the consciousness prior with systems based on more common RL approaches such as policy gradient deep RL on one hand, or model-based RL on the other hand (still with neural nets to learn the transition operator in sensory space, as well as the reward function). Even better, in toy problems we can compute the oracle solution, so we can get an upper bound on the best achievable performance.

Acknowledgements

The author wants to thank Philippe Beaudoin, Gerry (Tong) Che, William Fedus, Devon Hjelm and Anirudh Goyal for preliminary discussions about the consciousness prior, as well as funding from NSERC, CIFAR, the Canada Research Chairs, and the Open Philanthropy Project.

References

Bengio, E., Thomas, V., Pineau, J., Precup, D., and Bengio, Y. (2017). Independently controllable features. *arXiv e-prints*, **1703.07718**.

Bengio, Y. (2014). Deep learning and cultural evolution. In *Proceedings of the Companion Publication of the 2014 Annual Conference on Genetic and Evolutionary Computation*, pages 1–2. ACM.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *ICML'09*.

Graves, A., Wayne, G., and Danihelka, I. (2014). Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kriegel, U. (2014). *Current Controversies In Philosophy of Mind*.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

van Gulick, R. (2004). Consciousness. In *Stanford Encyclopedia of Philosophy*.

Weston, J., Chopra, S., and Bordes, A. (2014). Memory networks. *arXiv preprint arXiv:1410.3916*.