Toward General Intelligence: The Role of Inhibition and Iteration

Abel Torres Montoya, abel@dataveras.com, DataVeras, Brussels, Belgium

1. The Challenge of Generalization

Despite significant achievements and unprecedented interest in ML and AI, there is also a growing debate about the limitations of the existing solutions for building artificial general intelligence (AGI) (Jordan, 2018; Marcus, 2018). Key missing functionalities include: common sense reasoning, unsupervised learning and zero-shot learning among others.

Several research groups are addressing this challenge working on interpretability (Zhang & Zhu, 2018), visual reasoning (Perez et al., 2017), zero-shot learning (Santoro et al., 2016), intuitive physics (Chang et al., 2016) and other topics. Still, generalization properties of deep learning are poorly understood (Zhang et al., 2016), brittle (Azulay & Weisss, 2018; Rosenfeld et al., 2018; Alcorn et al., 2018) and vulnerable to attacks (Brown et al., 2017; Li et al., 2018). This leads to questioning whether the DL paradigm is sufficient for implementing reasoning and general problem solving.

In alignment with Lake et. al. (2016) and Gershman et. al. (2015) we believe that the research on core mechanisms and theory of intelligence will be crucial for making progress towards generalization. In this work, we take a fresh look at the challenge and find that two commonly ignored information processing mechanisms in the brain, inhibition and iteration, are critical for building good representations and performing basic logical operations between elements of the network. By scaling those mechanisms in time and over the network, we can achieve advanced reasoning capabilities.

2. A Different Paradigm

External inputs are stored system's internal representations. Those inputs are our only factual evidence and we want to keep the representations and further processing tightly linked to them.

Principle I. Mirrored Representations: *The internal representation of one input is a hierarchical compositional structure containing all extracted properties. The topmost element matches the input in a generative way, i.e. by activating it top down we can reconstruct the input.*

We refer to this representation as *grounded* for it correlation with the input. To build such representation we will start by applying a Hebbian association rule: neighbors with similar activation patterns are associated under a parent node. This rule creates abstractions that capture the spatiotemporal consistency of the input. Every time a parent node is created it propagates a cancellation signal downstream that removes its child nodes from the pattern seeking candidates. *Synchronized loops of bottom-up activation followed by top-down cancellation create a compositional representation that is generative and that recreates the input top-down*. This mechanism is unsupervised, learns from a single input and allows continuous learning, reusing without distortion previous knowledge.

2.1. Integrating Learning and Problem Solving

Usually, learning and problem solving mechanisms are separated but, in practice, we have to do problem solving during learning (e.g. a known object is presented rotated) or learn during problem solving (e.g. to abstract a sequence of repeated actions). While in learning we had to create an internal representation matching the input, for problem solving we will have to create an external input matching the internal representation of the goal (Figure 1).

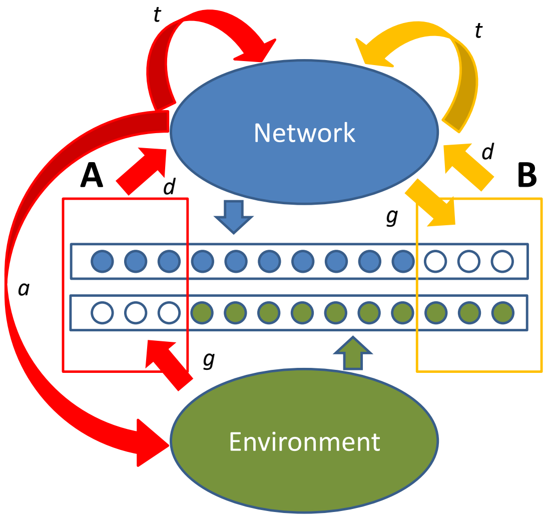

Figure 1. Grounded Intrinsic Motivation: Input (green) versus internal representation (blue) as the mechanism of information processing. **A**. *Problem solving* (red) takes place when the internal representation is not matched by the input. **B**. *Learning* (yellow) happens when the input has no matching representation. The discrepancy $\boldsymbol{d}$ between the input and the representation is propagated into the system, triggering transformations $\boldsymbol{t}$ and actions $\boldsymbol{a}$ to achieve the goal $\boldsymbol{g}$ of filling the blanks.

Principle II. Grounded Intrinsic Motivation: *Discrepancies between the bottom level activations from the external input and active internal representations drive the information processing in the system. Restoring the equilibrium at the bottom layer leads to learning, problem solving and imagination.*

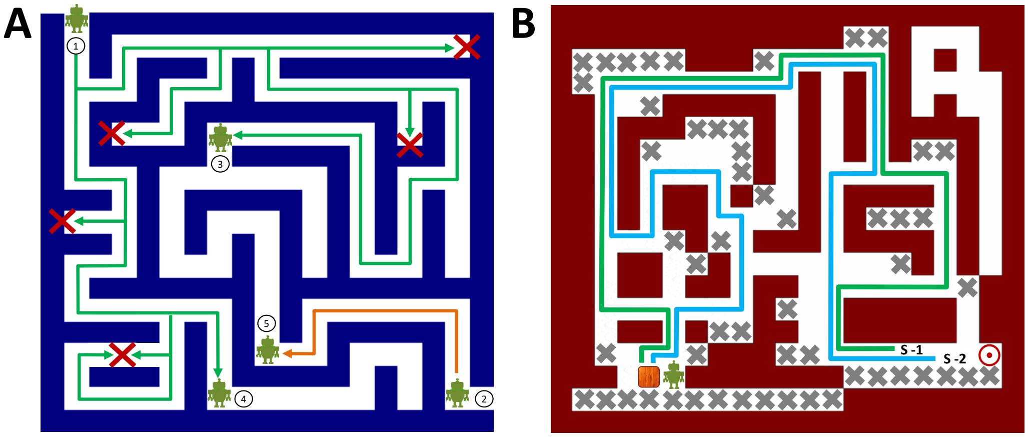

Figure 2. **A** The agent starts from location **1** and has to reach exit **2**. Directions reaching dead ends or already transited paths are inhibited (red). Currently active paths are **3**, **4** and current target is **5** by abductive reasoning. **B** Common sense reasoning removes deadlock states making then inaccessible to the planning (gray crosses). Even in scenarios considered difficult by RL implementations the resulting space of actions can be covered with few logical paths. To find an alternative solution **S-2** to the existing **S-1** we restart the process after inhibiting **S-1** and, consequently, all paths attached only to it.

2.2. Reasoning: Common Sense and Counterfactuals

Reasoning is a cornerstone of intelligence and it requires performing logical operations over network elements. In the logical domain, an element is present or 'true' when it triggers; equally, the inhibition of an element indicates its negation, i.e. the element is 'false' if prevented from triggering by inhibition. The following extension of the Hebbian rule will be used to include the interaction between contradictory statements: *if the occurrence of one element takes place at the same time as the removal of another, then an inhibitory link is established between them.*

Principle III. Adaptive Logic: *The computational logic of the system is constantly changing during information processing. These transient changes cummulate and decay in time. A particular set of* ***Inhibition rules*** *describe its impact on the network:* ***a)*** *Parent nodes of an inhibited node are also inhibited.* ***b)*** *If all higher level abstractions containing a given element are inhibited then the element is also inhibited.* ***c)*** *In the state/action space, if a given non-target state can only do a transition to inhibited states, then such state is also inhibited (i.e. an action that has no impact on the system is inhibited).*

All necessary logical operations for reasoning can be obtained by combining activation and inhibition, adding the benefits of classical logic based processing to the neural networks context. The changes in the network information processing capabilities resulting from spreading the logical impact of each active element are the basis of common sense (e.g. two mutually exclusive concepts cannot be active simultaneously).

Due to the lack of common sense RL requires a large amount of iterations to achieve only partial solution of exploratory challenges. This proposal solves the exploratory challenge in unsupervised manner using few data. It learns past trajectories and removes the explored paths by inhibition, bifurcating in crossroads into parallel solution candidates (Figure 3). The model not only finds the solution to the general case of maze/Sokoban challenges: it can detect the cases where there is no solution possible or can compute all existing solutions. The first case corresponds to the situation when all paths are inhibited without reaching the goal. The second case results from counterfactual reasoning: by inhibiting a found solution at top level we can exclude it from reach in the system and find alternatives by simply restarting the process.

One important differences between this paradigm and current ML solutions is in the goal formulation: traditionally it is set as maximizing performance on a given task, while in our proposal is to create a model of the environment. We believe that adaptive and iterative reasoning is required for extracting the maximum value from small data and dealing with unexpected deviations. We see in this model a good candidate for autonomous robot exploration, protein discovery and other applications.

References

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. *arXiv preprint*, 2018. arXiv:1811.11553.

Azulay, A. and Weisss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint*, 2018. arXiv:1805.12177.

Brown, T. B., Mané, D., Roy, A., Abadi, M., and Gilmer, J. Adversarial patch. *arXiv preprint*, 2017. arXiv:1712.09665.

Chang, M. B., Ullman, T., Torralba, A., and Tenenbaum, J. B. A compositional object-based approach to learning physical dynamics. *arXiv preprint*, 2016. arXiv:1612.00341.

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278, 2015.

Jordan, M. Artificial intelligence – The Revolution Hasn't Happened Yet. *https://medium.com/p/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7*, 2018.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *arXiv preprint*, 2016. arXiv:1604.00289.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Y., Bian, X., and Lyu, S. Attacking object detectors via imperceptible patches on background. *arXiv preprint*, 2018. arXiv:1809.05966.

Marcus, G. Deep learning: A critical appraisal. *arXiv preprint*, 2018. arXiv:1801.00631.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. Film: Visual reasoning with a general conditioning layer. *arXiv preprint*, 2017. arXiv:1709.07871.

Rosenfeld, A., Zemel, R., and Tsotsos, J. K. The Elephant in the Room. *arXiv preprint*, 2018. arXiv:1808.03305.

Santoro, A., Bartunov, S., Botvinick, M., and D. Wierstra, T. L. One-shot learning with memory-augmented neural networks. *arXiv preprint*, 2016. arXiv:1605.06065.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint*, 2016. arXiv:1611.03530.

Zhang, Q. and Zhu, S. Visual interpretability for deep learning: a survey. *arXiv preprint*, 2018. arXiv:1802.00614.