Multi-modal Sense-Making

Alessandro OLTRAMARI[a,1]
[a] *Bosch Research and Technology Center, Pittsburgh, PA, USA*

By learning how to make sense of the environment we live in, as humans we survived in the wilderness, escaping predators, enduring natural catastrophes, epidemics, and overcoming the intrinsic limitations of our own species. But what is this "sense-making" capability, anyway? Although at first sight the notion may sound naive, it can be traced back to Newell and Simon's theory of cognition [1, 2]: through sensory stimuli, we cumulate experiences, generalize and reason over them, "storing" the resulting knowledge in long-term memory; the dynamic combination of live experiences and knowledge during task execution, enables us to make time-sensitive *rational* decisions, evaluating how good (or bad) a decision was by factoring in the feedback from the environment.

Do you see any artificial being around, exhibiting these properties? Of course you don't: despite the progress that robotics has made over the last decade (e.g., some of the most dexterous results being BigDog [3] and Baxter [4]) embodied AI is still in its infancy. Let me manage the expectations then, and reframe the question without the *burden of embodiment*: are you aware of any AI system capable of processing multi-modal sensor data [6] in real time, and identify with high degree of accuracy the events represented in the data (e.g., gunshot, people taking cover), and their context of reference (armed robbery in a bank)? Well, if you attended TriCoLoRe 2018, you may have heard my own answer to that question: we are getting there, but there are still significant gaps that need to be filled.

First of all, the explosion of deep neural networks, namely networks with a high number of intermediate layers, has augmented the breadth and depth of machine learning, to the point that these algorithms, running on powerful GPU clusters, play a major role in the "artificial brains" of self-driving cars [7]. But, regardless of hype [8] and dramatic setbacks [9], these systems are *de facto* not reliable: weather conditions, anomalous behavior of vehicles and pedestrians, street lightning and all sort of adversarial situations that the environment can naturally present, have experimentally demonstrated how error-prone deep learning solutions still are. These limits, though, shouldn't really surprise if we recognize that autonomous vehicles lack of "sense-making" capabilities. No human driver exclusively relies on her senses behind the wheel! The decisions we make are the result of a continuous evaluation of the context, where perceptual cues are constantly (and seemingly unconsciously) combined with background knowledge of the surroundings, and common sense: for instance, driving in an area where college students go clubbing on a Friday night requires extra attention to erratic behavior of possibly intoxicated jaywalkers. But if you don't brave the weekend nightlife, the following examples may sound more familiar: residents generally know which area of the neighborhood might have icy road conditions in a frigid winter day,

[1] Alessandro Oltramari, Bosch Research and Technology Center, Pittsburgh, USA; E-mail: Alessandro.Oltramari.ext@us.bosch.com.

or where in the city flooding is more frequent after a powerful storm, which streets are more likely to have kids playing around after school, and which intersections tend to have poor lighting. Currently, this type of common knowledge is not being used to assist self-driving cars.

Like humans, machines can only make sense of the intricacies of physical and social reality by combining perception and knowledge. This is not just a theoretical tenet: from an empirical standpoint, the performance of purely data-driven AI is close to reach a plateau. Knowledge is required not only for complex tasks like autonomous driving [10], or natural language understanding [11], but also for relatively simpler applications. For instance, the company Vicarious[2] showed that a system trained on only 1406 images, but endowed with spatial knowledge, can break captchas with significantly higher accuracy than state of the art deep neural nets trained with ~8 million images.

Far from being a new idea, but definitely boosted by the technological breakthroughs of our era, the integration between symbolic knowledge and sub-symbolic learning is poised to become important in AI again. I firmly believe that multi-modal sense making will serve as testbed for such integration.

References

[1] Newell, A. and Simon, H.A., 1972. Human problem solving (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-Hall.

[2] Newell, A., 1994. Unified theories of cognition. Harvard University Press.

[3] https://www.bostondynamics.com/bigdog

[4] https://robots.ieee.org/robots/baxter/

[5] https://www.heykuri.com/

[6] Baltrušaitis, T., Ahuja, C. and Morency, L.P., 2019. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), pp.423-443.

[7] Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J. and Zhang, X., 2016. End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316.

[8] https://www.nytimes.com/2018/11/27/business/self-driving-cars-autonomous-vehicles.html

[9] https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html

[10] Zhao L, Ichise R, Mita S, Sasaki Y. Core Ontologies for Safe Autonomous Driving. In International Semantic Web Conference (Posters & Demos) 2015.

[11] McShane M. Natural language understanding (NLU, not NLP) in cognitive systems. AI Magazine. 2017 Dec 1; 38(4):43-56.

[2] https://www.vicarious.com/