Learning Beyond Limited Labels

By Discovering Disentangled and Compact Representations of Abstract Concepts

As most machine learning practitioners know, the primary challenge for machine learning as a field continues to be ensuring that trained learning systems are capable of generalizing beyond limited training datasets - and humans are much better at this than current state-of-the-art learning systems. As a member of the applied machine learning research group at Intuit, I have been toying with the idea of combining domain knowledge and human interactions with data-driven approaches to teach learning systems to learn more effectively.

At ICML (the International Conference on Machine Learning) this year, I came across an inspiring talk that proposed innovative ideas to tackle this problem. It was an invited lecture in the Saturday Workshop on "Toward Learning with Limited Labels" entitled "Towards disentangling underlying explanatory factors" by Dr. Yoshua Bengio. Given the speaker's reputation, it's not surprising that this talk attracted a lot of interest. What did surprise me was that the overflow was so significant that halfway through the talk, much of the audience had to be escorted out in order to comply with the facility's safety regulations. Here's what those of us lucky enough to stay for the whole talk got to hear.

What's Missing with Deep Learning: Deep Understanding

In the first part of the talk, Dr. Bengio provided concrete evidence that current state-of-the-art deep learning-based models tend to capture "statistical regularities" in the training data sets, rather than the underlying causal mechanisms. In standard machine learning practice, training, validation, and testing data sets are sampled from the same pool (i.e., the same statistical distributions). In many real-world applications such as autonomous driving, the statistical distributions of actual data inputs can deviate vastly from the training data sets. This discrepancy explains why models trained using static data sets appear to generalize so well at test time, but remain vulnerable to adversarial attacks when deployed in the real world. While this observation would not surprise any seasoned machine learning/deep learning practitioners, the fact that Dr. Bengio's group set out to measure the extent to which this happens is impressive[1]. (Their approach is summarized in the appendix.)

Toward A Machine That Plans

Dr. Bengio spoke next about the tried and true strategies that encourage learning systems to "learn how the world ticks," including equivariance/invariance, symmetries, and leveraging spatial and temporal scales. Inspired by the neuroscientific approach to understanding how human brains work, these strategies attempt to:

- Capture "good" internal, high-level, low-dimensional representations

- Maximize independence among controllable external factors, which leads to the concept of agents and actions

Why does capturing meaningful low-dimensional representation allow for generalization? From a neuroscientific perspective, there is evidence that human thought is inherently low-dimensional, hierarchical, and attentional. The human brain can store about seven items in working memory, which is small compared to the hundreds of dimensions of internal representations used in state-of-the-art machine learning models.

While limiting, this low-dimensional content-based attention allows humans to focus on a few elements out of a large set. In recent findings, neuroscientists have shown that the human brain integrates innumerable cognitive tasks into structured low-dimensional hierarchies. This allows a human to shift fluidly within changing but related tasks[10]. The low-dimensional core dynamic of the human brain is analogous to a meta-learner of related tasks. Such meta-learning allows learning systems to generalize and adapt to new tasks in a data-efficient manner (i.e., few shots)[11]. More importantly, the seemingly vast diversity of natural-world scenarios actually arises from a small set of coherent rules (in physics and chemistry for example) that can be explained through abstract concepts. By adding a low-dimensionality constraint into the model architecture, learning systems can be trained to learn such meaningful hierarchical lower-dimensional representations[3].

Why might maximizing independence among external factors improve generalization? It could be argued that the human brain can come up with control policies that influence independent and distinct aspects of the world. For example, consider an object that we can manipulate/control - let's say a coffee mug on a desk. One can choose to perform two independent actions on the mug, for instance: 1. turning the mug at its base by 90 degrees, and 2. filling up the mug with coffee to its rim. The outcome of this first action could be called "observing how the coffee mug looks from a 90-degree clockwise viewpoint," while the outcome of the second is "observing the volume capacity of the mug." These independent action-outcome pairs are maximally independent factors that explain how our perception of the coffee mug transitions from the original state (of sitting on a desk) to the two final states (turned by 90 degrees, and then filled with coffee). By mentally abstracting these action-outcome pairs, a human can generalize and predict the final states of the mug after these independent actions are taken in any sequence or combination. More importantly, humans can extend such generalization beyond coffee mugs to any object of similar physical properties.

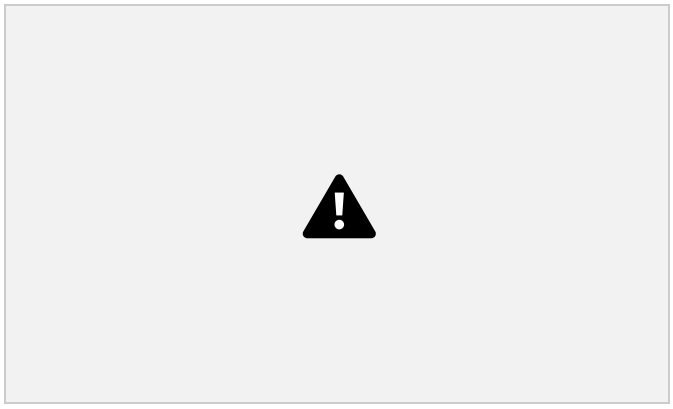
The internal representation of action-outcome pairs is what we call the "mental model" of the world. We use this mental model to plan, act, and predict the outcomes of our actions. In the realm of machine learning, the closest analogs to this idea are the advances in reinforcement learning that show how acting in the real world can help guide representation learning and disentangling[4, 5]. (Disentangling being the separation of independent causes.) Having the "right" representation of the environment and the policy to control/affect aspects of it *"are two*

important ingredients that occur in parallel that allow human minds to discover what's controllable in the world." And, as Dr. Bengio pointed out, *"This naturally leads to the emergence of the notion of objects and agents."*

The Two Priors

Next came the key ideas of the talk: the two types of priors that encourage learning systems to learn deeper "concepts"--the *Conscious Prior*[6] and the *Independent Controllable Explanatory Factors*[7].

The Conscious Prior



In brief, a conscious representation can be viewed as a low-dimensional combination of a few (concrete or abstract) concepts constituting a "thought." State-of-the-art learning systems are often trained to capture a high-dimensional representation of the input space. With the Conscious Prior, there are two levels of representation: a high-dimensional "unconscious (hidden)" representation (*h*) and a low-dimensional "conscious" representation (*c*). The low dimensionality of the conscious representation serves as a strong regularization for the learning problem. Attention can be applied over the conscious and unconscious state in such a way that the learning system output/prediction synthesized from the unconscious representation maps to a simple combination in the conscious representation. We can think of attention as a model-training mechanism that encourages a learning system to discover a conscious representation in which learned concepts can be "mentally" manipulated and/or referred to compactly. It should be noted that merely encouraging lower-dimensional representation does not guarantee that the discovered/learned representation is "conscious" in the neuroscience sense.

The Independent Controllable Explanatory Factors

The idea behind Independent Controllable Explanatory Factors is that by acting in the real world, a learning system (or agent) can learn "disentangled representations." Because an agent can act on a particular aspect of the environment while leaving others unaffected, the learned internal representation must be able to distinguish perturbations coming from independent factors. As in the example of manipulating the coffee mug, an agent learns to associate its

actions in the environment with internal representations that are unique to the action and the aspect of the environment being acted upon (i.e., Independent Controllable Explanatory Factors).

Final Thoughts

Among all the ICML talks I attended this year, this one has been the most thought-provoking. Shortly after wrapping up for the day, I hurried back to my hotel room, pulled up the papers[1, 6, 7, 8] referenced in the talks, and started immersing myself in them. While I'm still trying to figure out how to apply these concepts to my current work, I'm keeping them in my machine learning toolbox, hoping they'll become useful one day.

In closing, here are my thoughts on the two priors. Reflecting on my understanding of the Conscious Prior, the idea of using a lower-dimensional conscious space to represent a "concept" seems to correspond well to the current understanding of how thoughts occur in human brains. This type of machine training should lead to models that are not only more easily trainable and generalizable, but also more explainable (e.g., model decisions explained by activation of a few meaningful concepts vs. activation of a large sparse array of neural weights). Fewer and more coherent explanatory factors, plus making these factors maximally independent from one another, makes the factors' effects easier to tease apart. I see a parallel between Independent Controllable Explanatory Factors and the idea of learning with equivariance, which Dr. Bengio pointed out has led to discoveries of good representations in many visual learning tasks. Notably, not all features of an aspect of environments - such as the aspect of color - are information that an agent can act upon (i.e., controllable). So there must be other strategies the human brain uses to learn those concepts in generalizable ways.

Applying the Conscious Prior and Independent Controllable Factors even to toy problems (such as navigating in a simple grid world [7, 9]) provides a good example of how machine learning as a field can benefit from ideas borrowed from neuroscience. I'm hopeful that there are many more insights we can draw from our growing understanding of the human brain to help build more robust machine learning systems.

Appendix

How did Bengio's group measure the tendency of complex models to latch on to surface statistical regularities rather than generalizable features?

Jason Jo and Yoshua Bengio[1] implemented perturbation maps (Random and Radial Fourier filtering schemes) that qualitatively changed statistical regularities while preserving object recognizability, and applied them to well-known image data sets (SVHN and CIFAR-10). A state-of-the-art Convolutional Neural Network (Preact ResNet with Bottleneck architecture 200[2]) was trained on one data set, and then tested across differently-perturbed sets. With this

setup, the models exhibited up to a 28 percent generalization gap across the various test sets[1].

References:

[1] Jo J. and Bengio Y. (2017). "Measuring the tendency of CNNs to Learn Surface Statistical Regularities" https://arxiv.org/abs/1711.11561
[2] K. He, X. Zhang, S. Ren, and J. Sun. "Identity Mappings in Deep Residual Networks" https://arxiv.org/abs/1603.05027
[3] Higgins I. et al (2017). "SCAN: Learning Hierarchical Compositional Visual Concepts" https://arxiv.org/abs/1707.03389
[4] Lesort T. et al (2018). "State Representation Learning for Control: An Overview" https://arxiv.org/abs/1802.04181
[5] Raposo D. et al (2017). "Discovering objects and their relations from entangled scene representations" https://arxiv.org/abs/1702.05068
[6] Bengio Y. (2017). "The Consciousness Prior" https://arxiv.org/abs/1709.08568
[7] Thomas V. et al (2017). "Independently Controllable Factors" https://arxiv.org/abs/1708.01289
[8] Bengio Y. (2018) "Towards disentangling underlying explanatory factors" http://www.iro.umontreal.ca/~bengioy/talks/ICMLW-limitedlabels-13july2018.pptx.pdf
[9] Sukhbaatar S. et al (2015). "MazeBase: A sandbox for learning from games" https://arxiv.org/abs/1511.07401
[10] Shine et al (2018) "The low dimensional dynamic and integrative core of cognition in the human brain" https://doi.org/10.1101/266635
[11] Finn C., Abbeel P. and Levine S. (2017) "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks" https://arxiv.org/abs/1703.03400

Bio:

Joy Rimchala is a Data Scientist in Applied Machine Learning Research in Intuit's Technology Futures Group. She has implemented a synthetic data approach to overcome the lack of limited label data in computer vision and natural language settings, and has built model pipelines with TensorFlow and AWS SageMaker. Joy is currently leading the initiative on information extraction from images of structured documents using ideas from computer vision, natural language models, and representation learning. Joy holds a PhD from MIT, where she spent five years doing biological object tracking experiments, and modeling them using Markov Decision Processes.

Acknowledgment:

First and foremost, I would like to thank Heather White for her coaching and for constructive feedbacks. Heather's comments and suggestions brought much clarity and readability to the article. I also would like to thank Alex Gude, Andrew Mattarella-Micke, Conrad DePeutor, Riley Edmunds, Sricharan Kumar, Sumayah Rahman and Yang Li and for advice on the technical aspects of the article.