NSF Workshop on Integrating Approaches to Computational Cognition

Supplementary Material: Examples of Existing Work Bridging CS and ML

The report from this workshop[1] focuses on future gains that can be expected from cooperation between the machine learning (ML) and cognitive science (CS) research communities. There are numerous examples of recent research that point to this potential. Of many lines of research that could be used as illustrations, we give in this appendix a few examples selected by the workshop participants, noting that this is a small selection of both the research in the field and the research of the workshop members.

Griffiths

How should we represent the features of complex objects? When we look at the objects around us, the features that are relevant for deciding what an object is or how to use it seem obvious. But these features aren't as obvious to computers as they are to people, and the ease with which we identify the features of objects means that most psychological experiments assume the features of objects are simply available to observers. Griffiths and Ghahramani (2006) show how sophisticated statistical methods can be used to identify the features of objects from their appearance. This statistical approach solves two problems at once - identifying what the features are, and how many features need to be used. The model, named the Indian Buffet Process, was motivated by issues that come up in cognitive science, but has had a significant impact in machine learning. In subsequent work, we have examined how this idea can be used to explain how people identify the features of objects and how computers can do a better job of automatically forming representations of objects that are more similar to those formed by people.

[1] The main report can be obtained at http://matt.colorado.edu/compcogworkshop/report.pdf

Jones

A primary challenge in both human cognitive modeling and machine learning is acquisition or adaptation of representations that support efficient learning. The state of the two fields is complementary, in that psychology has insightful theories about representation learning and machine learning has sophisticated computational frameworks for expressing different representations and implementing them in complex tasks. Work in our lab has found that integrating these can produce models that both explain aspects of human learning in more complicated tasks than are normally modeled in psychology and produce performance better than standard machine learning approaches. Specifically, kernel methods from machine learning provide a formal framework for learning with complex representations that are high-dimensional nonlinear transformations of the input stimulus space. A major goal is automatic discovery of kernels (i.e., representations) that are adapted to individual tasks, but the space of possible kernels is so vast the search problem is underconstrained. Research on human perception of similarity (which can be identified with the kernel function) offers principles for how the search for effective representations can be more constrained and hence more efficient. We have applied this approach in two ways, one based on selective attention among stimulus dimensions (Canas & Jones, 2010; Jones & Canas, 2010), and the other based on analogical reasoning with structured relational representations (Foster, Canas, & Jones, 2012; Foster & Jones, 2013). In both cases, we have implemented these models of human representation learning as mechanisms for adaptive kernels in a reinforcement learning framework, yielding models that learn to perform complex dynamic tasks while adapting their representations to the structure of those tasks.

Kemp

Different kinds of representations are useful for capturing knowledge about different domains. For example, living kinds are usefully organized into a taxonomic tree, and politicians are usefully organized along a spectrum from liberal to conservative. Kemp and Tenenbaum (2008) developed a computational framework that discovers which kind of representation is best for a given domain. Their approach was inspired by previous psychological research on multidimensional scaling and structure learning, but also drew on recent developments in machine learning, including work on semi-supervised learning and Gaussian graphical models. Their work has subsequently been cited by machine learning researchers who develop methods for finding structure in large data sets, and by psychologists who study how knowledge structures emerge over the course of cognitive development.

LeCun

In the last several years, artificial vision systems for object recognition and scene understanding have become increasingly similar to computational models of biological vision. Standard models of the ventral pathway in the visual cortex and top-performing object recognition systems (such as convolutional networks) are "deep learning" architectures consisting of multiple stages. Each stage is composed of 4 layers: contrast normalization, a filter bank, non-linearity (similar to simple cells in the cortex), feature pooling and subsampling (similar to complex cells in the cortex). Unsupervised learning algorithms to train these layers can produce topographic maps of orientation-selective feature detectors similar to what is found in V1. Because of the simultaneous appearance of large datasets and fast GPUs, it is only in the last year that biological inspiration has lead to speech and image recognition systems that outperform previous approaches on standard benchmarks (Farabet, Couprie, Najman, & LeCun, in press; LeCun, 2012). The limitations of these models, such as the susceptibility to visual crowding, seem similar to that of the human visual system.

Love

1. Here is an example of a near miss between communities. I developed a model for why certain concepts are more central than others in a web of concepts (Love &

Sloman, 1995). The approach is formally identical to Google PageRank. It predates that work. To be fair, there are earlier examples in operations research.

2. Larkey and Love (2003) proposed a model of analogy and relational comparison that has been used in AI systems on story understanding.

3. Gureckis and Love (2009) is influenced by modern work in reinforcement learning.

4. Knox, Otto, Stone, and Love (2012) is a collaboration with machine learning researchers applying POMDP models to analyzing psychological data. We find that people show sophisticated patterns of exploration corresponding to belief updating. The work suggests some ways to make sophisticated exploration more tractable in machine systems.

5. Knox, Glass, Love, Maddox, and Stone (2012) is another collaboration with machine learning researchers exploring how humans can provide the reward signal (i.e., train) robots in reinforcement learning tasks.

6. Giguère and Love (2013) is heavily influenced by work in machine learning (e.g., SVMs). We work through the predictions of these models to develop predictions for human behavior. We identify a way in which humans differ from machine systems, advancing cognitive theory and suggesting ways to train people more effectively.

Lu

One of the hallmarks of human reasoning is the ability to form representations of relations between entities, and then to reason about the higher-order relations between these relations. By the time they reach school age, children have acquired the ability to accurately assess whether one object (e.g., bear) is “larger” or “smaller” than another (e.g., fox). Although other species have been shown to share similar mechanisms for comparative judgment with perceptual relations, human children go on to acquire a deeper understanding of comparative relations. Modeling work by Lu, Chen, and Holyoak (2012) indicates that what is special about human relational learning can be characterized as a capacity for relational reinterpretation: the ability to transform perceptually-grounded relations into explicit relational structures that distinguish the roles of relations from the objects that fill them. Children eventually understand that a pair of concepts like larger-smaller is related in basically the same way as the pair faster-slower, allowing them to see that such pairs of relations form analogies. Our results provide a proof-of-concept that structured analogies can be solved using representations induced from unstructured feature vectors by mechanisms that operate in a largely bottom-up fashion. Our findings also show how relation learning can move beyond perceptual inputs to create abstract higher-order relations.

Humans are capable of learning and reasoning based on relational roles, including higher-order relations such as cause and effect. By employing structured relations, it becomes possible to transfer knowledge across diverse situations by drawing analogies. The ability to make analogical inferences is the key to understanding and modeling the flexibility of human thinking. By contrast, current machine systems still lack the flexibility and adaptiveness of humans. They can perform extremely well, and even outperform human experts, in narrowly specified though highly complex domains (e.g., chess, Jeopardy). These successes are partly attributable to the fact that machine systems typically employ fairly simple representations. However, these simple forms of representations limit the generalizability of the system to other complex decision-making situations associated with high uncertainty. In addition, machine systems still have not mastered the task of learning relations from structured environments. Finally, the machine learning community needs to address how relational representations can be mapped to one another across superficially different situations.

McAllester

Felzenszwalb, Girshick, McAllester, and Ramanan (2010) introduced the object detection system known as the deformable part model (DPM) in computer vision. An object detection system finds instances of a certain kind of object in an image.

For example, digital cameras typically have face detectors (camera face detectors are typically Viola-Jones detectors which predate DPM detectors). A DPM detector can be trained for any class of object. For example, it can be trained to detect people, cars, or dogs. Machine learning, and latent support vector machines (LSVMs) in particular, are central to the DPM model. The DPM model has become the standard baseline system in the computer vision community for object detection. As of this writing the paper has 964 Google scholar citations. In spite of the success of this system in the computer vision community, its performance is nowhere near that of the human visual system. It seems that a deeper understanding of human vision could be leveraged to improve the performance of computer object detectors.

Koehn et al. (2007) describes the Moses system for machine translation between natural languages, for example the translation of Japanese into English. This system has become the standard baseline for machine translation. As of this writing the paper has 1772 Google scholar citations. The Moses system automatically synthesizes a mechanical translator from a corpus of translation pairs. Although it has become the standard baseline, its performance is far below that of human translators.

Mozer

Human memory is imperfect. Individuals of all ages and abilities gradually forget previously learned knowledge and skills. Robust, durable learning is achieved only through periodic review. Although academic curricula could benefit from incorporating review in a comprehensive, systematic manner, two challenges must be overcome. First, students at every educational level are faced with an ongoing imperative to master new material, which demands a time-efficient means of reviewing an ever-growing body of old material. Second, the effectiveness of review crucially depends on its timing, but efforts to predict the optimal timing have not adequately considered individual differences. To address these challenges, Lindsey, Shroyer, Pashler, and Mozer (2013) developed an adaptive method for personalizing study based on a Bayesian model of forgetting that leverages psychological theory and collaborative filtering. Here, collaborative filtering involves using data from a population of students studying a variety of material to infer the dynamic knowledge state of an individual student for specific material. The method was incorporated into a semester-long middle school foreign language course via retrieval-practice software. In a cumulative exam administered one month after the semester's end that compared time-matched study strategies, personalized review yielded a 16.5% boost in course retention over current educational practice (massed study) and a 10.0% improvement over a one-size-fits-all strategy for spaced study. Our results demonstrate that integrating adaptive, personalized software into the classroom is practical and yields appreciable improvements in long-term educational outcomes.

Salakhutdinov

The ability to learn abstract representations that support transfer to novel but related tasks, lies at the core of many problems in computer vision, natural language processing, cognitive science, and machine learning. In typical applications of machine classification algorithms today, learning a new concept requires tens, hundreds or thousands of training examples. For human learners, however, just one or a few examples are often sufficient to grasp a new category and make meaningful generalizations to novel instances. Clearly this requires very strong but also appropriately tuned inductive biases.

Salakhutdinov, Tenenbaum, and Torralba (2013) take a step towards this "one-shot learning" ability by learning several forms of abstract knowledge at different levels of abstraction, that support transfer of useful inductive biases from previously learned concepts to novel ones. In this work we propose compound HD (hierarchical-deep) architectures that integrate these deep models with structured hierarchical Bayesian models. In particular, we show how we can learn a hierarchical Dirichlet process (HDP) prior over the activities of the top-level features in a Deep Boltzmann Machine (DBM), coming to represent both a layered hierarchy of increasingly abstract features, and a tree-structured hierarchy of classes. Our model depends minimally on domain-specific representations and achieves state-of-the-art performance by unsupervised discovery of three components: (a) low-level features that abstract from the raw high-dimensional sensory input (e.g. pixels, or 3D joint angles) and provide a useful first representation for all concepts in a given domain; (b) high-level part-like features that express the distinctive perceptual structure of a specific class, in terms of class-specific correlations over low-level features; and (c) a hierarchy of super-classes or "superordinate" categories for sharing abstract knowledge among related classes via a prior on which higher-level features are likely to be distinctive for classes of a certain kind and are thus likely to support learning new concepts of that kind.

Schölkopf

Around 2005, we started working on understanding whether kernels as used in machine learning have any relevance for the issues of generalization and similarity in cognitive science. This was a process that took about three years and involved a psychologist, a cognitive scientist and a machine learning person. It led us to understand that most similarity measures considered by psychologists were examples of positive definite kernels, for which a rich body of mathematical theory exists. As a consequence, we were able to put forward kernel methods as a unifying theoretical tool showing how several competing and seemingly incommensurate theories in psychology (exemplar models vs. perceptron models) can be viewed as the same thing, linked by the so-called representer theorem of kernel methods. These connections were summarized in Jäkel, Schölkopf, and Wichmann (2009).

Shiffrin

This research and model was the first of several to follow that explain how the visual/cognitive system manages to infer, with high accuracy, the visual world after each eye movement, despite features that intrude from the previous eye fixation. Huber, Shiffrin, Lyle, and Ruys (2001) explained the accuracy of perception in terms of Bayesian adaptation and induction. Later research in this series explained the timing of these processes with use of a Hidden Markov Model (HMM). Another later article gave a neural network account based on synaptic depression. These models have proved predictive rather than descriptive: In case after case, the models have predicted correctly the outcomes of experiments yet to be done, even when the experimenters predicted alternative outcomes.

Thomas

One of the recurring themes at the workshop regarding the interface of CS and ML was the notion of optimality of a human "agent" in a broad sense. One version of this idea pondered the consequences of optimal processing in dynamic decision-making. Specifically, not only should the rewards and costs explicitly involved in choice be taken into account, but the "bounds" of human cognition and the costs associated with those bounds were deemed important areas that cognitive science can contribute to machine learning approaches to this problem. In addition, ML theorists indicated their strong interest in methodologies that could reveal these bounds or properties of the human mind deemed its architecture and capacity limitations (e.g., Sorg, Singh, & Lewis, 2010). Much work in cognitive modeling has been devoted to providing methodologies that jobbing experimental psychologists could use to reveal hidden architecture inside the black box of the human mind (e.g., Townsend & Nozawa, 1995). In that vein, one of my own publications, Thomas (2006) advanced our understanding of this experimental methodology to identify how systems of thought are organized in time and how these systems interact to produce observable behavior in simple and complex tasks and the role that optimality of the agent plays in determining the processing behavior of this architecture. One of the results of that work provided an explanation for a paradoxical finding that had been documented in the literature decades earlier but that could not be explained by models of the time (e.g., Miller & Pachella, 1973). Specifically, investigators had observed that the probability of an experience seemed to influence other aspects of cognition beyond weighting of decision criteria, something that most models had not deemed possible. In Thomas (2006), architectures were infused with explicit models of optimal decision processing and, as a result, naturally produced the pattern of data previously observed that had defied explanation. ML approaches are well suited to representing problems computationally that allow considerations of optimality with respect to the environment and task but these do not produce solutions that are flexible over the lifetime of an agent. Incorporating knowledge of how human cognition is organized in terms of architecture and capacity into these optimality analyses would allow programs to exhibit the kind of flexibility and robustness the human decision-maker has long enjoyed.

Yu

Through the interdisciplinary study of cognition in brains and machines, an interesting picture is beginning to emerge: The brain far surpasses state-of-the-art computer algorithms in certain tasks, such as planning and decision-making under complex conditions, and yet falls well short of a cell phone in other tasks, such as processing large quantities of data precisely. This contrast provides an opportunity for identifying the design principles of the brain, with many potential applications in science, medicine, and engineering. For example, we have used ML tools to understand how the healthy brain utilizes past experiences to anticipate situations that require the interruption of default processing. This feature turns out to be very useful for situations where past experiences are predictive of upcoming events, but amounts to nothing more than "superstition" in unpredictable situations. Interestingly, as we have shown, this property of the brain yields overall behavioral advantage. This modeling work has helped us to identify the neural processes underlying prediction and planning, and moreover discover significant neural alterations in occasional stimulant users, which accumulate with lifetime use of drugs (Ide, Shenoy, Yu, & Li, 2013). We believe that this is a promising line of research that will not only advance brain science, but also provide early biological and behavioral markers of clinical pathology associated with drug use, as well as aid the design of more robust and dynamic artificial agents that can use experiences to cope with noisy and constantly changing environments.

Zhang

Motivated by the need (from cognitive psychology) for removing the symmetry assumption in similarity judgments, Zhang, Xu, and Zhang (2009) extended the reproducing kernel methods in machine learning from the Hilbert-space setting to the more general Banach-space setting. It was proved (Zhang & Zhang, 2012) that the Representer Theorem, the key for linking exemplar-based with prototype-based categorization models, still holds in this situation. This opens the way for modeling similarity/generalization versus feature/attention in a unified computational framework.

Zhu

How can a computer classify a test item X if it does not look like any of the labeled training items at all? For example, X may be the side-view photo (taken at a 90-degree angle) of a suspect while all photos on file are frontal (taken at 0-degree). If the computer has photos of the suspect from many angles, it may reason that the photo taken at 10-degree looks very much like the known frontal photo and the two must be the same person. Similarly, the photo taken at 20-degree looks very similar to the 10-degree photo and should be the same person, and so on. In this way, the label (identity of the suspect) propagates through the series of photos taken from different angles until it reaches the side-view photo X. Of course, real data is more complex and items in general form a weighted similarity graph. Given that a few nodes in the graph are labeled, one wants to classify the remaining nodes using the same propagation idea. This is a challenging problem because on a graph different labels may propagate to the same node via competing paths. Zhu, Ghahramani, and Lafferty (2003) proposed an elegant mathematical solution to this problem. It significantly advanced ML research in the area of semi-supervised learning. Ten years later, the paper received a Classic Paper Prize from the same ML conference. In follow-up work, we showed that the same mathematical solution can also be applied to cognitive modeling to explain certain human categorization behaviors.

References

Cañas, F., & Jones, M. (2010). Attention and reinforcement learning: Constructing representations from indirect feedback. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (in press). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*, 1627-1645.

Foster, J. M., & Jones, M. (2013). Analogical reinforcement learning. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Foster, J. M., Cañas, F., & Jones, M. (2012). Learning conceptual hierarchies by iterated relational consolidation. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*.

Giguère, G. & Love, B.C. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America*, 110(19), 7613-7618.

Griffiths, T. L., & Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems, 18*.

Gureckis, T. M., & Love, B. C. (2009). Short term gains, long term pains: Reinforcement learning in dynamic environments. *Cognition, 113*, 293-313.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review, 108*(1), 149-182.

Ide, J. S., Shenoy, P., Yu, A. J., & Li, C-S. R. (2013). Bayesian prediction and evaluation in the anterior cingulate cortex. *Journal of Neuroscience, 33*, 2039-2047.

Jäkel, F., Schölkopf, B., Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences, 13*(9), 381-388.

Jones, M., & Cañas, F. (2010). Integrating reinforcement learning with models of representation learning. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

Kemp, C. & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*(31), 10687-10692.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology, 2*, 398.

Knox, W. B., Glass, B. D., Love, B. C., Maddox, W. T., & Stone, P. (2012). How humans teach agents. *International Journal of Social Robotics, 4*(4), 409-421.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. et al. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 177-180). Association for Computational Linguistics.

Larkey, L. B., & Love, B. C. (2003). CAB: Connectionist analogy builder. *Cognitive Science, 27*, 781-794.

LeCun, Y. (2012). Learning invariant feature hierarchies. In A. Fusiello, V. Murino, Vittorio, R. Cucchiara (Eds.), *European Conference on Computer Vision (ECCV 2012)*, *7583*, 496-505.

Lindsey, R. V., Shroyer, J. D., Pashler, H., and Mozer, M. C. (2013). Improving long-term knowledge retention through personalized review. *Manuscript submitted for publication*.

Love, B. C., & Sloman, S. A. (1995). Mutability and the determinants of conceptual transformability. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, 654-659.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review, 119*, 617-648.

Miller, J. O., & Pachella, R. G. (1973). Locus of the stimulus probability effect. *Journal of Experimental Psychology, 101*, 227–231.

Salakhutdinov, R., Tenenbaum, J. B., & Torralba, A. (2013). Learning with hierarchical-deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22*, 189-228.

Sorg, J., Singh, S., & Lewis, R. (2010). Internal rewards mitigate agent boundedness. *Proceedings of the 27th International Conference on Machine Learning* (Haifa, Israel).

Thomas, R. D. (2006). Processing time predictions of current models of perception in the classic additive factors paradigm. *Journal of Mathematical Psychology, 50*, 441-455.

Townsend, J. T., & Nozawa, G. (1995). Spatiotemporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology, 25*, 321-359.

Zhang, H., Xu, Y., & Zhang, J. (2009). Reproducing kernel Banach spaces for machine learning. *Journal of Machine Learning Research, 10*, 2741-2775.

Zhang, H., & Zhang, J. (2012). Regularized learning in Banach spaces as an optimization problem: representer theorems. *Journal of Global Optimization, 54*, 235-250.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. *Proceeding of the 20th International Conference on Machine Learning (ICML)*.