Project Ideas for 2015 Summer Course

Over the three weeks of the Brains, Minds, and Machines summer course, students engage in open-ended projects that provide an opportunity to explore course topics more deeply and apply new computational or empirical methods learned in the tutorials. To facilitate the design of individual projects, faculty instructors and teaching assistants provide initial project ideas in five broad research areas:

(1) development of intelligence
(2) neural circuits for intelligence
(3) visual intelligence
(4) social intelligence, and
(5) theories for intelligence.

This page describes the general project ideas provided for the 2015 summer course, including supplementary references and pointers to relevant code and data. These descriptions are not intended as detailed project specifications; rather, they are aimed at highlighting interesting problem areas that may trigger ideas for specific projects. Ongoing research projects in each of the above areas are described on the research pages of the Center for Brains, Minds, and Machines website.

For additional ideas, see the abstracts of projects (PDF) completed by students who attended the 2014 Brains, Minds, and Machines summer course.

1. Development of Intelligence

Instructors: Josh Tenenbaum, Tomer Ullman

Project 1.1 – Learning physical properties from video using top-down (physics engine) and bottom-up (neural network) analysis

Infants learn a great deal about the physical world over the first few months of life. In particular, they seem to learn about the physical properties of objects, such as their mass, friction, and elasticity, and relate this to the size, texture, and material of the object. Computational modeling can help to better understand how this learning happens and what representations are present earliest in human development. In this project, you will use convolutional neural networks combined with a generative physics engine to infer hidden physical properties from perceptual data (e.g. objects of different perceptual properties colliding and interacting).

References

Battaglia, P. W., J. B. Hamrick, and J. B. Tenenbaum. "Simulation as an engine of physical scene understanding (PDF)." *Proc. National Academy of Science* 110, no. 45 (2013): 18327-18332 *(article includes supporting information with details of experiments and model simulations)*

Teglas, E., E. Vul, V. Girotto, M. Gonzalez, J. B. Tenenbaum, and L. L. Bonatti. "Pure reasoning in 12-month-old infants as probabilistic reasoning." *Science* 332 (2011): 1054-1059.

For a more computationally oriented project, see the following papers and example from the probmods.org electronic text:

- Wu, J., I. Yildirim, J. J. Lim, W. T. Freeman, and J. B. Tenenbaum. "Galileo: Perceiving physical object properties by integrating a physics engine with deep learning (PDF)." *Proc. Neural Information Processing Systems conference* (2015).

- Zhang, R., J. Wu, C. Zhang, W. T. Freeman, and J. B. Tenenbaum." A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding (PDF)." *Proc. Cognitive Sciences Society conference* (2016).

- https://probmods.org/chapters/02-generative-models.html#example-intuitive-physics

Project 1.2 - Combining intuitive physics and psychology to explain agent intelligence

Early in development children expect other people to have goals and to act efficiently to achieve those goals. For example, young children expect animate beings to take a direct route to their target and hypothesize a barrier or obstruction if someone takes the long way around. Infants also understand more relational/social goals like chasing and fleeing, and helping and hindering. In order to understand all of this, infants need to represent people as rational agents, capable of planning, reasoning and understanding the consequences of their actions. Still, other agents can be seen as smart or dumb to the degree that they correctly understand the world or seem to learn from their failures. See, for example, the 'chasing' video, where children and adults see this as chasing only when physical obstacles are present, otherwise this is really inefficient chasing and fleeing. Also, the larger agent is seen as 'dumber' by adults. In this project you will build a model of a reasoning agent in a social-physical setting, define levels of its intelligence and competence, and relate those to potential infant preference studies.

References

Baker, C. L., J. B. Tenenbaum, and R. R. Saxe. "Action understanding as inverse planning (PDF)." *Cognition* 113.3 (2009): 329-349.

Gergely, G. and G. Csibra. "Teleological reasoning in infancy: The naïve theory of rational action." *Trends in Cognitive Sciences* 7, no 7 (2003): 287-292.

Southgate, V. and G. Csibra. "Inferring the outcome of an ongoing novel action in 13 months." *Developmental Psychology* 45, no. 6 (2009): 1794-1798.

A variation on this project was begun during the 2015 Brains, Minds, and Machines summer course and later developed into a conference paper:

- Kryven, M., T. Ullman, W. Cowan, and J. B. Tenenbaum. "Outcome or strategy? A Bayesian model of intelligence attribution (PDF)." *Proc. Cognitive Science conference* (2016).

Project 1.3 - Updating belief, perceptual access, and eye-tracking

One current model for understanding the beliefs, desires and actions of others suggests that we conceive of other people as rational planning agents (action understanding as inverse planning). This model includes a notion of rational updating and perceptual access; that is, we take into account what other people see and we expect them to take what they see into account. Infants also seem sensitive to perceptual access, although the degree to which they can entertain false belief (thinking other people think something that is not true) is a hotly debated topic in development.

Infants' understanding in this domain can potentially be explored with eye tracking. Recent research into adult counter-factual reasoning has used eye-tracking methods to show that adults 'simulate' things that did not happen in order to make sense of things that did. For example, when asked if Ball A caused Ball B to go into a goal, adult eye movements suggest a simulation of Ball's B trajectory, had Ball A not existed. In this project you will propose extensions to eye-tracking methods into the social-psychological domain, to see whether infants' eye movements track the line-of-sight of others, and how that affects the infants' expectations.

For reference, see Baker, Tenenbaum, and Saxe (2009) and Gergeley and Csibra (2009) above.

Project 1.4 - Force violations as a basis for physical expectation violation

Infants know a lot about physics. At one year old, they know that unstable things fall down, stable things stay stable, and heavy-looking things are heavy. They know a bit about weight, solidity, gravity, fluids, friction, and soft-bodies. Most experiments showing that infants 'know' these things work by setting up an expectation (e.g. a hand lets go of an object), showing a violation (e.g. the object floats in mid-air) and checking for surprise via looking-time measures. In this project, you will use 3D graphics (or 2D graphics) to create novel "physically surprising" stimuli and use the amount of force necessary to create that surprise to predict the degree of infant surprise.

For reference, see Battaglia, Hamrick, and Tenenbaum (2013) and Teglas et al. (2011) above.

2. Circuits for Intelligence

Project 2.1 - What is there? Representations of visual information in neuronal responses and computer vision models

Instructors: Gabriel Kreiman, Leyla Isik, Bill Lotter

Being able to understand the world we see requires constructing visual representations where behaviorally important variables can be "read-out." The brain implements such representations along the ventral visual stream where visual input is transformed into patterns of neural responses that contain stimulus specific information. There is much interest in developing computer algorithms that can similarly understand the visual world, and state-of-the-art computer vision models are largely based on how we believe the brain works. In this project, you will see how information can be decoded from neural responses, as well as computer vision models (HMAX and/or deep convolutional neural networks), and compare the two. Specifically,

- **Subproject 2.1.1:** Consider a set of single-object images $\{p_1,\ldots,p_N\}$ for which we have behavioral data, neurophysiological data (in monkeys and humans), and computational model responses. Use machine learning classifiers to evaluate how well we can discriminate among them in single trials based on physiological data, or based on computational models

- **Subproject 2.1.2:** Examine the tolerance to image transformations (scale, position, viewpoint)

- **Subproject 2.1.3:** Behavioral experiments, physiology and computational models to understand the mechanisms underlying pattern completion

References

Serre, T. M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. "A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex." MIT CSAIL Memo 2005-036, CBCL Memo 259 (2005).

Liu, H., Y. Agam, J. R. Medsen, and G. Kreiman, "Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex." *Neuron* 62, no. 2 (2014): 281-290.

Isik, L., E. M. Meyers, J. Z. Leibo, and T. Poggio. "The dynamics of invariant object recognition in the human visual system." *J Neurophsiology* 11 (2014): 91-102.

Tang, H., C. Buia, R. Madhavan, J. Madsen, W. Anderson, N. Crone, and G. Kreiman. Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron* 83 (2014): 736-748.

Resources

See Kreiman Lab Code/Data/Databases website for additional resources.

Project 2.2 - The role of STDP in neural networks

Instructors: Gabriel Kreiman, Joseph Olson

Spike-timing dependent plasticity (STDP) is a fundamental ingredient for learning processes and for neural circuit development. STDP is a refinement of the Hebbian learning rule, “neurons that fire together wire together.” The classical STDP model is that synapse strength increases (called long-term potentiation or LTP) when a pre-synaptic neuron fires right before a post-synaptic neuron; meanwhile, synapse strength weakens (long-term depression or LTD) when the reverse firing order occurs (Markram et al. 1997). However, more exotic STDP rules have been biologically observed in both different brain regions (see Abbott and Nelson (2000) for a good review) and even along the same dendrite (see Caporale and Dan (2008) for a review). We are currently researching how different STDP rules may interact together to develop local circuits such as those reported in Douglas and Martin (2004). These local circuits are likely the building blocks for cognitive functions. This project will further investigate, via computer simulations, how the various types of STDP protocols induce different network structures and circuitry.

References

Abbott, L. F. and S. B. Nelson. Synaptic plasticity: Taming the beast. *Nature Neuroscience Supplement* 3 (2000): 1178-1183.

Caporale, N. and Y. Dan. Spike timing-dependent plasticity: A Hebbian learning rule. *Annual Reviews of Neuroscience* 31 (2008): 25-46.

Douglas, R. J. and K. A. C. Martin. Neuronal circuits of the neocortex. *Annual Reviews of Neuroscience* 27 (2004): 419-451.

Burbank, K. and G. Kreiman. "Temporally reversed STDP is required for learning stable, diverse, weak feedback connections." Paper presented at the 2011 COSYNE conference.

3. Visual Intelligence

Instructors: Boris Katz, Andrei Barbu

Project 3.1 - Language and vision, and perhaps robotics

Recently there has been great interest in combining vision and language into unified systems. There are many potential approaches and the space of designs is not well explored. In this project you will implement one or more such models and attempt to perform some new tasks with them. For example, you could train language-vision models and use them to translate between languages, or use them to understand and generate plans. Alternatively, you could build a model that combines language, vision, and robotics; one can imagine that machines that can interact with objects would be able to learn about the world more quickly than passive observers.

- Implement a language-vision model
- Use it to recognize and describe images or videos
- Attempt a new task such as: translation, planning, integration with robotics, kinematics

References

Siddharth, N., A. Barbu, and J. M. Siskind. "Seeing What You're Told: Sentence-Guided Activity Recognition in Video (PDF)." *IEEE Conference on Computer Vision and Pattern Recognition,* (2014).

Yu, H., N. Siddharth, A. Barbu, and J. M. Siskind. A compositional framework for grounded language inference, generation, and acquisition in video (PDF). J Artificial Intelligence Research 52 (2015): 601-713.

Resources

See Andrei Barbu's Visual Language and Video Events research pages for additional resources, including source code available on github.

Project 3.2 - Integrated vision

Humans do not just perform individual computer vision tasks like object detection or determining the color of an object. Our vision is in some sense integrated and attempts to come to a somewhat more global understanding of a scene that combines information about color, shading, shadows, illumination, depth, object identity, matching and tracking objects over time, etc. Approaches to perform many of these tasks individually exist and it would be interesting to see if combining them together produces better performance. For example, we would expect a system that jointly segments objects and identifies them to perform better than a system that does these two steps separately.

- Pick two computer vision tasks, perhaps out of the list above
- Implement both in the same framework, like deep learning, graphical models, probabilistic programming, etc.
- Combine and train the two models jointly
- Examine whether performance improves and what the joint model does better than the individual models

4. Social Intelligence

Instructors: Nancy Kanwisher, Alex Kell, Leyla Isik

Project 4.1 - Data driven fMRI analysis of social video stimuli

fMRI research over the last 15 years has successfully identified several dozen robust functional divisions of the brain through the use of traditional hypothesis driven methods, which test specific mental functions hypothesized in advance. However, there is no guarantee that all of the important ways the brain carves up the problem of cognition correspond to subdivisions scientists will think to test. To address this problem, a newer set of data-driven fMRI analysis methods have been devised. These analysis methods entail the collection of fMRI responses to large, relatively unconstrained stimulus sets designed to broadly sample the space of stimuli/mental processes, rather than to test any particular hypothesis. Methods to discover structure in the resulting neural activations include clustering, in which sets of voxels are identified that have similar profiles of response across stimuli (or vice versa) and a variety of linear analysis methods including PCA and ICA, which respectively identify primary dimensions of variation and the most statistically independent dimensions in the fMRI response across stimuli.

In this project you will analyze one of two datasets. The first contains two subjects' patterns of response across voxels to a two-hour commercial movie stimulus. The movies have labels containing which character was present and what action they were performing in each frame. The second dataset contains three subjects' fMRI responses to 300 clips of movies (including 282 clips of various kinds of social interactions, and 18 non-human control stimuli), as well as a handful of useful localizers. The stimuli in this second dataset have extensive ratings on a variety of social dimensions, acquired with Mechanical Turk. Numerous analyses are possible with these data, and you can also try new methods. The analyses could ask, for example, (1) what information about character identity is present in the movie stimuli and (2) what information about their actions/interactions is present?

You can also compare the results derived by analyzing these data with PCA/ICA to the results derived from an analysis in which voxels are clustered according to their response across timepoints/stimuli. A central question here is whether these fMRI data are better fit by clustering (in which each voxel is assigned to a unique functional profile) or by linear analyses (in which the functional response of each voxel is fit by a linear weighted sum of components).

Additionally, you can also run a series of experiments on synthetic neural data and augmented real neural data to better understand the kinds of structure these "data-driven" methods discover. For example, what percentage of voxels need to exhibit a certain response profile in order to be discovered by clustering/ICA? How robust is ICA to correlations between different response profiles and different patterns of voxel weights? The results of these analyses can help inform what structure our current analyses are missing, as well as extensions to existing data-driven techniques that may be more sensitive to certain aspects of structure in neural responses.

Resources

The Kanwisher Lab website has resources related to the Group-Constrained Subject-Specific (GSS) method to algorithmically discover functional regions of interest (fROIs) in fMRI data that are activated systematically across subjects.

5. Theories for Intelligence

Project 5.1 – Neuronal implementation of probabilistic computation

Instructors: Haim Sompolinsky, SueYeon Chung, Yasmine Meroz

In many situations, cognitive tasks with uncertainty are thought to be well described by Bayesian models of the brain. If the brain is like a neural network, what kind of network structures (e.g. types of nonlinearities, number of layers and number of units each layer) can implement these optimal probabilistic computations? Perceptual invariance is one example where Bayesian optimal computation provides a desired robustness to trial-to-trial variability provided by the stimuli.

- Consider the situation where a network has to classify M different concepts out of P examples. The network is provided the templates for all P concepts, while in each trial, it is not provided with the template for the specific trial. The input for the network is the receptive fields' noisy realization of the example. What kind of network structure can achieve close to Bayesian optimal performance? Is there a learning algorithm that would learn the Bayesian optimal performance?

- Suppose now the source of noise is not from random neural variability, but from a dependence of neuron's tuning on smoothly varying latent variables. In this case, what kind of nonlinear transformations will make the classification more robust to the variability from the stimulus parameter? (e.g. quadratic/sigmoid nonlinearities? multiple layers?)

References

Bengio. Y. Learning deep architectures for AI. *Foundations and Trends in Machine Learning* 2, no. 1 (2009): 1-127.

Seung, H. S. and H. Sompolinsky. Simple models for reading neuronal population codes. *Proc. National Academy of Sciences* 90 (1993): 10749-10753.

Project 5.2 - Learning invariant visual representation from natural videos

Instructors: Tomaso Poggio, Fabio Anselmi, Georgios Evangelopoulos, Gemma Roig

Recent work has shown that convolutional neural networks (CNNs) can be trained not only using millions of labeled examples, but also through unlabeled videos and exploiting their temporal component as a weak form of supervision. This provides a way to build neural networks in a more biologically plausible manner, e.g by introducing frame similarity/feature slowness constraints, modifying the loss function of supervised networks or exploring autoencoder (AE)-type networks with reconstruction similarity constraints. In addition, similar ideas are prevalent in i-theory, where the predicted memory-based learning of invariances can be done through sampling transforming templates from videos depicting the same semantic content. Possible projects exploring these insights and their integration can include:

- Build an unsupervised network for invariance (i-theory network) using a dataset of natural videos and designing appropriate learning schemes, e.g. sampling/memory-based, supervised with (unsupervised) slowness constraints or unsupervised with reconstruction constraints. Try to implement networks, by either modifying existing CNN or AE architectures or sampling-based template selection for memory-based learning. You can use datasets of videos where a single object is undergoing natural transformations, for example, the CBMM face video dataset (YouTube clips of moving faces), the CBMM people, objects, actions and interactions dataset (full length commercial movies), or YouTube objects dataset (database of object videos from YouTube).

- Explore the pre-processing and visual meta-data that might be useful or employed for unsupervised training on videos. How can the sequence of frames in a video be used? Test the effects of motion filtering, cropping, detection/tracking, saliency detection on the training videos. Eye-tracking data and some object labels are available for the CBMM people, objects, actions, and interactions dataset. Optical flow and bounding boxes are provided for the YouTube objects dataset. https://data.vision.ee.ethz.ch/cvl/youtube-objects/

References

Agrawal, P., J. Carreira, and J. Malik. Learning to see by moving, *Proc. IEEE International Conference on Computer Vision (ICCV)* (2015).

Goroshin, R., J. Bruna, J. Tompson, D. Eigan, and Y. LeCun. Unsupervised learning of spatiotemporally coherent metrics. *Proc. IEEE Int. Conf. Computer Vision* (2015): 4086-4093.

Liao, Q, J. Z. Leibo, and T. Poggio. Unsupervised learning of clutter-resistant visual representations from natural videos. CBMM Memo No. 23 (2015).

Mobahi, H., R. Collobert, and J. Weston, Deep learning from temporal coherence in video (PDF). *26th Annual International Conference on Machine Learning (ICML)* (2009) (video)

Wang, X. and A. Gupta. Unsupervised learning of visual representations using videos. *Proc. IEEE Int. Conf. Computer Vision* (2015): 2794-2802.

Resources

See CBMM Code/Datasets page and Poggio Lab Code/Datasets page for additional resources.

Invariance and Selectivity in Representation Learning (CBMM Thrust 5 project page)

Code/Libraries

HMAX:

- CNS
- Color HMAX
- hmin: Minimal HMAX implementation

MatConvNet: CNNs for MATLAB

Torch, Keras (TensorFlow/Theano), Caff (or your flavor of a Deep Learning Toolkit, see: The Big List of Deep Learning Toolkits)

Project 5.3 - Comparing different architectures for low sample complexity representation learning in brains and machines

Instructors: Tomaso Poggio, Fabio Anselmi, Georgios Evangelopoulos, Gemma Roig

A recent computational theory about invariance and selectivity in data representation (i-theory) suggests hierarchies built of modules performing filtering and pooling, like the simple and complex cells described by Hubel and Wiesel in primary visual cortex and deep convolutional neural networks. Such networks can compute a representation for an object, starting from the image/pixel domain that is both invariant to typical class transformations (e.g., translation, scaling, rotation) and selective with respect to different class-specific features. In this project, you will test empirically the properties of such networks for invariance and the analogies to kernel machines or rbf neural networks.

- Multi-layer vs. one-layer i-theory networks: compare the invariance, selectivity and efficiency of single and multilayer i-theory networks. Define or use appropriate performance metrics for each one and study the dependency on the range of transformations, the number of classes, the depth vs. width specifications of the network, the size of the training set size, etc. Use the CBMM iCub dataset (first- person images collected from the iCub robot) and SUFR dataset (subtasks of unconstrained face recognition) for classification tasks.

- Systematically study and evaluate the role of nonlinearities for invariant representations and multilayer networks. Try different parametric families of nonlinearities.

- Radial Basis Function (RBF) networks and kernels: explore the connections between (a) one-layer, trained, RBF networks, (b) learning with Gaussian kernels, and (c) i-theory networks. What is the dependency on the number (and center) of the RBF functions? Is there a connection of optimal/learned RBF centers to i-theory templates?

References

Anselmi, F., J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio. Unsupervised learning of invariant representations. *Theoretical Computer Science* 633 (2015): 112-121. (arxiv preprint and CBMM Memo 01)

Anselmi, F., L. Rosasco, and T. Poggio. On invariance and selectivity in representation learning. Information and Inference, 2016 (arxiv preprint)

Anselmi, F., L. Rosasco, C. Tan, and T. Poggio. Deep convolutional networks are hierarchical kernel machines. CBMM Memo No. 35 (2015).

Poggio, T. and F. Girosi. Networks for approximation and learning. *Proc. of the IEEE* 78, no. 9 (1990): 1481-1497.

Resources

Invariance and Selectivity in Representation Learning (CBMM Research Thrust 5 project page)

CBMM Memos by T. Poggio (many related to i-theory)

Code

GURLS: a Least Squares Library for Supervised Learning http://lcsl.mit.edu/#/downloads/gurls

RB networks in MATLAB (NN toolbox): http://www.mathworks.com/help/nnet/ug/radial-basis-neural-networks.html

Resource: Brains, Minds and Machines Summer Course
Tomaso Poggio, and Gabriel Kreiman