Report from Dagstuhl Seminar 22422

Developmental Machine Learning: From Human Learning to Machines and Back

James M. Rehg[*1], **Pierre-Yves Oudeyer**[*2], **Linda B. Smith**[*3], **Sho Tsuji**[*4], **Stefan Stojanov**[†5], **and Ngoc Anh Thai**[†6]

1 **Georgia Institute of Technology – Atlanta, US.** rehg@gatech.edu
2 **INRIA – Bordeaux, FR.** pierre-yves.oudeyer@inria.fr
3 **Indiana University – Bloomington, US.** smith4@indiana.edu
4 **University of Tokyo, JP.** tsujish@gmail.com
5 **Georgia Institute of Technology – Atlanta, US.** sstojanov@gatech.edu
6 **Georgia Institute of Technology – Atlanta, US.** athai6@gatech.edu

Abstract

This interdisciplinary seminar brought together 18 academic and industry computer science researchers in artificial intelligence, computer vision and machine learning with 19 researchers from developmental psychology, neuroscience and linguistics. The objective was to catalyze connections between these communities, through discussions on both how the use of developmental insights can spur advances in machine learning, and how computational models and data-driven learning can lead to novel tools and insights for studying child development. The seminar consisted of tutorials, working groups, and a series of talks and discussion sessions. The main outcomes of this seminar were 1) The founding of DevelopmentalAI (http://www.developmentalai.com), an online research community to serve as a venue for communication and collaboration between develpomental and machine learning researchers, as well as a place collect and organize relevant research papers and talks; 2) Working group outputs – summaries of in-depth discussions on research questions at the intersection of developmental and machine learning, including the role of information bottlenecks and multimodality, as well as proposals for novel developmentally motivated benchmarks.

Seminar October 16–21, 2022 – http://www.dagstuhl.de/22422
2012 ACM Subject Classification Computing methodologies → Artificial intelligence
Keywords and phrases developmental psychology, human learning, machine learning, computer vision, language learning
Digital Object Identifier 10.4230/DagRep.12.10.143

1 Executive Summary

Pierre-Yves Oudeyer
James M. Rehg
Linda B. Smith
Sho Tsuji

Recent advances in artificial intelligence, enabled by large-scale datasets and simulation environments, have resulted in breakthrough improvements in areas like object and speech recognition, 3D navigation, and machine translation. In spite of these advances, modern

[*] Editor / Organizer
[†] Editorial Assistant / Collector

artificial learning systems still pale in comparison to the competencies of young human learners. The differences between human learning and the paradigms that currently guide machine learning are striking. For example, children actively identify both the concepts to be learned and the data items used for learning, they infer the labels for learning from ambiguous perceptual data, and they learn from continuous streams of percepts without storing and curating large datasets. Artificial intelligence researchers are increasingly looking to developmental science for ideas and inspiration to improve machine learning, while developmental scientists are adopting tools from data science and machine learning to analyze large datasets and gain insights into developmental processes.

This seminar created new connections between the developmental and machine learning research communities by bringing together researchers in linguistics, psychology, cognitive science and neuroscience with investigators working in computer vision, machine learning and robotics. The seminar focused on three research questions:

1. What are the key computational problems and challenges that need to be addressed in creating a developmentally-inspired machine learner? Existing machine learning methods are built on a set of canonical problem formulations such as supervised learning and reinforcement learning. At the same time, decades of research in developmental science have produced an increasingly detailed characterization of learning in children. How can we leverage these insights to create new and more powerful machine learners and revise standard ML problem formulations?
2. What criteria are necessary for agent-based simulation models of development to advance machine learning and provide useful tests of developmental hypotheses? Advances in computer graphics and physics simulation have made it possible to create synthetic environments for training reinforcement learning agents to perform developmentally-relevant cognitive tasks such as navigating 3D space and manipulating objects. Can such computational experiments serve as useful tests of developmental hypotheses?
3. How can data-driven computational models be used to advance developmental science? It is increasingly feasible to collect dense sensor data that captures the perceptual inputs children receive (e.g. via wearable cameras and eye trackers), their behaviors during naturalistic interactions, and a variety of contextual variables relevant to cognitive tasks. These rich datasets, in conjunction with advances in deep learning have created the opportunity to create machine learning models which can "solve" certain developmental tasks such as object recognition. Given that such deep models do not speak directly to mechanisms of human learning, how can such research advance developmental science?

Through a seminar program consisting of tutorials, talks, working group meetings, and an early career mentorship sessions, we gained interdisciplinary insights into these core research questions. Attendees discussed the potential research directions that different research disciplines can benefit from each other, as well as collaboration opportunities and future development of the community. As the initial step, we aim to connect interested researchers online through social media and provide a common repository for relevant literature.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

3.1 Studying visual object learning with egocentric computer vision

David J. Crandall (Indiana University – Bloomington, US)

While early work in computer vision was inspired by studies of human perception, most recent work has focused on techniques that work well in practice but probably have little biological basis. But low-cost, lightweight wearable cameras and gaze trackers can now record people's actual fields of view as they go about their everyday lives. Such first-person, "egocentric" video contains rich information about how people see and interact with the world around them, potentially helping us better understand human perception and behavior while also yielding insights that could improve computer vision. I'll describe a recent interdisciplinary project (with Chen Yu and Linda Smith) in which we used computer vision to try to characterize the properties of childrens' egocentric views as they interact with objects – the "training data" of the child's learning system – and then showed that injecting similar properties into the training data of computer vision algorithms could improve the algorithms' accuracies as well.

3.2 Using machine learning in early language acquisition research: Examples from long-form audio-recordings

Alejandrina Cristia (LSCP – Paris, FR)

In 2022, we may not have hoverboards, but we have seen artificial intelligences beat humans at go, write in the style of Shakespeare, and generate novel continuations to incomplete spoken sentences. Feats like these have, in part, been due to the rise of self-supervision machine learning techniques, in which systems are trained with vast amounts of unlabeled data. In this talk, I argue that such techniques are useful to infant researchers working in under-described languages and cultures in two key ways: First, to create classifiers that describe and annotate the vast amounts of infant-centered data we can now easily collect; and second, to build systems that potentially learn like infants do. I draw from recent work using audio recordings collected with wearables to illustrate these two avenues of work in the description of children's spoken language environment, while highlighting both opportunities and challenges, including saliently ethical and legal ones.

3.3 Can Machine Learning Inform the Science of Infant Development, and Vice-Versa?

Rhodri Cusack (Trinity College Dublin, IE)

It can be difficult for psychologists and neuroscientists to conceptualise how cognitive functions emerge in the infant brain. Computational models may help, by providing a way to quantify the statistics of the environment, and to test the efficacy of proposed learning objectives and inductive biases. I will describe our ongoing work using deep neural networks to model the development of diverse aspects of the visual system, and the neuroimaging experiments we are running to evaluate them. Finally, I will discuss the potential for translation in the opposite direction, by reviewing how what we have learned about the development of infant cognition might inform the next generation of unsupervised machine learning.

3.4 Towards embracing complexity to understand atypical development: the case of Down syndrome

Hana D'Souza (Cardiff University, GB)

Development is a complex process, involving interactions between various domains across levels of description. Yet, many of our traditional developmental paradigms aim to isolate domains. The domains measured from various tasks are then correlated in order to understand how they are connected. However, everyday experiences emerge through the complex interactions of various domains – such as motor ability, attention allocation, and the actions of other social agents. Thus, in order to understand typical and atypical development, it is crucial to embrace complexity by putting these interactions at the very core of our research. Findings from studies using this approach have been challenging fundamental assumptions about typical development. I will introduce some of our initial steps in applying this approach to atypical development (Down syndrome) and explain why it has the potential to reconceptualise our understanding of neurodevelopmental disorders.

3.5 Simulating early language acquisition using self-supervised learning

Emmanuel Dupoux (LSCP – Paris, FR)

Recent progress in self supervised learning opens up the way to learn probabilistic models of language from raw audio signals. We propose to use these models as a proof of feasibility of the 'statistical learning hypothesis', which states that infants bootstrap into language primarily by extracting regularities from the speech input. Similarities and differences between the developmental curves in the models and infants are presented and discussed.

3.6 “ML as a tool” vs. “ML as a model” for the study of child development in the wild

Abdellah Fourtassi (Aix-Marseille University, FR)

Recent improvements in Machine Learning (ML) promise to transform research in developmental psychology by allowing the quantitative study of children’s behavior outside the lab. ML can help achieve this goal in two steps: 1) automatic annotation of a target behavior from naturalistic data, and 2) quantitative prediction of this behavior from complex (possibly causal) factors. These two steps are obviously related but they diverge in the nature of the ML they call upon. In the first, ML is a “tool” whose purpose is to overcome the limitations of manual labor. In the second, ML is considered a “model” whose purpose is to mimic the child’s behavior given a similarly rich input/stimuli. In this brief talk, I will illustrate – based on ongoing research in our team about children’s early conversational development – how “ML as a tool” and “ML as a model” can be articulated to help build quantitative theories of child development in the wild.

3.7 Predictive models of early language learning

Michael C. Frank (Stanford University, US)

How can we create mechanistic models of children’s early language learning? One key problem is the availability of data to train and evaluate such models. I’ll present our approach to combining data from large numbers of children – inputs from CHILDES, outcomes from Wordbank – to model early vocabulary acquisition across languages. A simple regression approach allows us to combine both descriptive and model-based predictors, holding the promise of more integrative, data driven theories.

3.8 Visual affordances from video: learning to interact by watching people

Kristen Grauman (University of Texas – Austin, US)

First-person or “egocentric” vision requires understanding the video that streams to a wearable camera. It offers a special window into the camera wearer’s attention, goals, and interactions, making it an exciting avenue for perception in augmented reality and robot learning. I will present our recent work using passive observations of human activity to inform active robot behaviors – such as learning object affordances from video to shape dexterous robot manipulation, transforming video into a human-centric topological map of the physical space and the activities it supports, or discovering compatible objects to shortcut visual semantic planning. We show how reinforcement learning agents that prefer human-like interactions can successfully accelerate their task learning and generalization. Finally, I will overview Ego4D, a massive new egocentric video dataset and benchmark built by a multi-institution collaboration that offers a glimpse of daily life activity around the world.

3.9 The First 1,000 Days Project

Uri Hasson, Casey Lew-Williams (Princeton University, US)

How do natural, everyday statistics in infants' environments give rise to learning? We will introduce a big-data project, the First 1,000 Days Project at Princeton University, inspired by prior video corpora, including the Human Speechome Project and the SAYCam corpus. Our dataset is designed to video-record 20 families for 1,000 days, beginning when the family returns home after birth. Each house is wired with eight cameras and four microphones that will record for 12 hours per day. Our team is deploying (and developing) machine learning tools for automated analysis of objects, people, space, proximity, and language, including a 'baby detector' and a pipeline that can analyze 300+ years of raw video and audio data. We have completed the development and automation of the research pipeline, and data collection has started with eight families in New Jersey and eastern Pennsylvania, with five additional families waiting to start once their babies are born. Our goal is to recruit a final sample of 20 families that represents the diversity of U.S. demographics.

3.10 How language can help machines to acquire general intelligence?

Felix Hill (Google DeepMind – London, GB)

Having and using language makes humans as a species better learners and better able to solve hard problems. I'll present three results that demonstrate how this can also be the case for artificial models of general intelligence. First, I'll show that agents with access to visual and linguistic semantic knowledge explore their environment more effectively than non-linguistic agents, enabling them to learn more about the world around them. Second, I'll demonstrate how an agent embodied in a simulated 3D world can be enhanced by learning from explanations – answers to the question "why?" expressed in language. Agents that learn from both classical reinforcement and explanations solve harder cognitive challenges than those trained from RL alone. Finally, I'll present evidence that the skewed and bursty distribution of natural language may explain how large language models can be prompted to rapidly acquire new skills or behaviours. This suggests how modelling language can make a neural network better able to acquire new cognitive capacities quickly, even when those capacities are not necessarily explicitly linguistic.

3.11 The Impact of Dataset Bias on Model Learning

Judy Hoffman (Georgia Institute of Technology – Atlanta, US)

Computer vision relies on learning from collections of data. The mechanisms used for collecting, curating, and annotating visual data results in datasets with distinct forms of bias. In turn, models that are trained using biased data, then perpetuate that bias into their learned representations. As the world changes the particular visual appearance bias of the initial data collection may not well represent the appearances the model is expected to operate on. This discrepancy leads to reduced performance and reliability of the learned model. In strong contrast, people are able to experience a biased sample of the world yet generalize (under certain conditions) to alternative world views, like a child who can recognize an elephant at the zoo after being shown cartoon drawings of an elephant. This talk will discuss two key challenges towards producing generalizable visual learning: 1) how can we leverage the learning process to help us identify bias in our data and 2) how can we mitigate bias through modified learning protocols or by adapting to new observations as they appear?

3.12 Truth, lies, and misinformation during cognitive development

Celeste Kidd (University of California – Berkeley, US)

I will talk about our lab's current work-in-progress exploring interventions designed to give children a greater ability to discern truth from falsity. I will discuss some of the foundational empirical studies in progress on two types of interventions designed to facilitate children's ability to discern fact from fiction. The first set of interventions target factors external to the child relating to the information ecosystems in which they are making judgements. The second set of interventions involve investigating internal mechanisms children may have available for helping them detect misinformed opinions. Both sets of work build off the lab's previous behavioral experiments and computational models about how children sample subsets of information from the world based on their uncertainty in order to form their beliefs and guide their subsequent sampling decisions. I will briefly provide some background on how our new work is building off of our prior papers.

3.13 Enhancement of cues and the oddball effect in child-directed speech

Eon-Suk Ko (Chosun University, KR)

People adapt their way of speaking when addressing children, and this speech register called Child-Directed Speech (CDS) is considered to provide features beneficial for infants' language learning. I present some of these features based on Korean mothers' interaction with their children. I then raise the question about the mechanism of how such features might benefit infants' learning given their small proportions provided in the input. I suggest that infants' novelty-driven learning and the oddball effect might help us understand aspects of such a mechanism.

3.14 Studying infant-like visual category generalization using the Toybox dataset

Maithilee Kunda (Vanderbilt University, US)

Infants can generalize from a small number of object instances within a category to novel instances. For computer vision, this problem can be posed as a domain adaptation problem, i.e., where the distribution of data in the training dataset differs from the distribution seen at test time. However, current domain adaptation tasks and datasets do not target learning across this particular type of distribution shift. We have used our lab's Toybox dataset of handheld object manipulation videos to create a new task that mimics this learning scenario, and I will present initial work on examining how existing domain adaptation models perform on this challenging new task. I will also briefly describe two other projects that investigate how agents might learn spatial reasoning skills and theory of mind reasoning skills.

3.15 Learning Vision for Walking

Jitendra Malik (University of California – Berkeley, US)

As a child interacts with the world around her, there is a barrage of sensory information – proprioception, tactile, audition, vision – together with knowledge of her own commanded actions via the efference copy. In AI and robotics, the cross-modal supervision that this would enable has been quite under-exploited. In recent work, `https://arxiv.org/abs/2211.03785` (to appear at ICRA 2023), we showed a concrete example of how this might work by learning a visual walking policy for a quadruped legged robot. We train a visual module in the real world to predict the upcoming terrain with our proposed algorithm Cross-Modal Supervision (CMS). CMS uses time-shifted proprioception to supervise vision and allows the policy to continually improve with more real-world experience. We evaluate our vision-based walking policy over a diverse set of terrains including stairs (up to 19cm high), slippery slopes (inclination of 35 degrees), curbs and tall steps (up to 20cm), and complex discrete terrains. We achieve this performance with less than 30 minutes of real-world data. Finally, we show that our policy can adapt to shifts in the visual field with a limited amount of real-world experience.

3.16 Does Affective communication increase the relation between children with ASD and their mothers?

Atsushi Nakazawa (Kyoto University, JP)

Affective communication has the function of facilitating smooth communication. Our group have been studying a French-originated affective communication method "Humanitude" which was originally developed for the nursing of dementia care. The Humanitude consists of face-to face communication (eye contact and facial expressions), touching, and talking, but there have been no studies quantifying the elements. Using computational behavioral science methods, our group have detected and analyzed the skill elements including eye contact, face-to-face communication using image recognition from first and third person video, developed and used the state-of-the-arts whole-body tactile sensor for the touch communication analysis, and developed a novel mobile facial myoelectric for facial expression recognition. As the result, our group revealed the skill elements of the methodology. Moreover, we developed the training system of the Humanitude using Augmented Reality (AR) technology which outperformed the existing communication trainings method. We will also introduce our efforts to apply this technique to improve parent-child relationships in ASD. While the experiment is preliminary, their eye contact and physical communication significantly increased after the intervention.

3.17 Language and Culture Internalization for Autotelic Human-Like AI

Pierre-Yves Oudeyer (INRIA – Bordeaux, FR)

Building autonomous artificial agents able to grow open-ended repertoires of skills is one of the fundamental goals of AI. To that end, a promising developmental approach recommends the design of intrinsically motivated agents that learn new skills by generating and pursuing their own goals – autotelic agents. However, existing algorithms still show serious limitations in terms of goal diversity, exploration, generalization or skill composition. This perspective calls for the immersion of autotelic agents into rich socio-cultural worlds. We focus on language especially, and how its structure and content may support the development of new cognitive functions in artificial agents, just like it does in humans. Indeed, most of our skills could not be learned in isolation. Formal education teaches us to reason systematically, books teach us history, and YouTube might teach us how to cook. Crucially, our values, traditions, norms and most of our goals are cultural in essence. This knowledge, and some argue, some of our cognitive functions such as abstraction, compositional imagination or relational thinking, are formed through linguistic and cultural interactions. Inspired by the work of Vygotsky, we suggest the design of Vygotskian autotelic agents able to interact with others and, more importantly, able to internalize these interactions to transform them into cognitive tools supporting the development of new cognitive functions. This perspective paper proposes a new AI paradigm in the quest for artificial lifelong skill discovery. It justifies the approach by uncovering examples of new artificial cognitive functions emerging from interactions between language and embodiment in recent works at the intersection of deep reinforcement learning and natural language processing. Looking forward, it highlights future opportunities and challenges for Vygotskian Autotelic AI research. This presentation will be an overview of some of the ideas in this paper: `https://arxiv.org/pdf/2206.01134.pdf`.

3.18 The Never-Ending VIsual classification Stream (NEVIS) 1.0

Marc'Aurelio Ranzato (DeepMind – London, GB)

Intelligent agents need to constantly adapt to change; for instance they need to adapt to change in the environment or change in the computation versus accuracy trade-off . Even modern large-scale models such as large vision and language models need to constantly adapt. They not only need to adapt to the current task but also use that experience to better learn future tasks. Unfortunately, there does not exist any benchmark today which is useful to investigate the question of how to efficiently adapt and consolidate knowledge over time and at scale. In this talk, I will provide an overview of NEVIS, a new benchmark which consists of a stream of very challenging and diverse visual classification tasks. I will then discuss the preliminary results we obtained using a variety of baseline approaches. NEVIS will be released in about a month, and it is meant to motivate researchers working in continual learning, meta-learning and auto-ml to join forces and to make strides together towards the development of robust systems that can become more apt and efficient over time.

3.19 Connecting 3D Shape Learning and Object Categorization

James M. Rehg (Georgia Institute of Technology – Atlanta, US)

A classical topic in computer vision and psychology is the link between knowledge of 3D object shape and the ability to categorize objects. In this talk we revisit this link in two machine learning contexts that are connected to development: few-shot learning and continual learning. We show that learning a representation of 3D shape in the form of dense local descriptors provides a surprisingly powerful cue for rapid object categorization. Our shape-based approach to low-shot learning outperforms state-of-the-art models trained on category labels. We also present the first investigation of continual learning of 3D shape and demonstrate significant differences relative to continual category learning, finding that 3D shape learning does not suffer from catastrophic forgetting.

3.20 Human infants' brains are specialized for social functions

Rebecca Saxe (MIT – Cambridge, US)

In this talk, I will argue that human infants have distinct social representations and motivations. Infants' learning about, and representations of, other people are not just a downstream consequence of generic processes that promote learning in the nonsocial environment, nor are they built by gradual, bottom-up adjustment to the statistics of visual experience. On the contrary, infants' attention to people depends on specific inferences about their social relevance; and is related to activity in distinctively social brain regions.

3.21 Towards Teachable Autonomous Agents: How can developmental psychology help?

Olivier Sigaud (Sorbonne University – Paris, FR)

As a developmental AI researcher, I will outline a research program where we try to endow autotelic agents (agents who learn to represent, pursue and reach their own goals) with a teachability property, so that we can influence their goals through social interactions. With such agents, we can mimic guided play interactions with children, where they learn both on their own and from the guidance of a tutor or caregiver. Then I will show that such a research program faces the language grounding problem and that a central issue is the acquisition of language-sensitive sensorimotor representations. I will question existing lines of AI research related to this challenge and conclude by showing that developmental psychology research can bring a lot to address it, by providing relevant concepts, models and experimental data about it.

3.22 Why self-generated behavior has more radical consequences than you might originally think

Linda B. Smith (Indiana University – Bloomington, US)

Humans, including toddlers, are adept at taking knowledge from past experiences and using it in compelling new ways. Learning and generalization depend on both the learning machinery and the training data on which the machinery operates. This talk will highlight findings from studies of toddler's self-generated experiences . The main point is that everyday experiences occur in time-extended episodes. Each unique episode is characterized by a suite of coherence statistics. I propose that these statistics are the secret ingredient to innovative intelligence. Moreover, they provide novel insights into the internal processes that learn, generalize and innovate.

3.23 Rethinking the developmental pathway of early infant language learning

Daniel Swingley (University of Pennsylvania, US)

Prominent empirical results of the 1980s and 1990s in which infants were revealed to have learned aspects of their language's system of phonetic categories (consonants and vowels) contributed to a standard theoretical model in which infants first learn to perceive speech sounds, then aggregate these into possible words, and then seek to identify meanings for those words while grasping at regularities caused by grammar. Modeling approaches that are based on this pathway have shown how simple statistical heuristics computed over phoneme sequences could help point infants to the early vocabulary. I will argue that this pathway is probably wrong and that current quantitative psychological models of infant word-form discovery are misguided. I will show that infant-directed speech is too variable and too unclear for such models to be plausible characterizations, and will sketch what an alternative looks like.

3.24 SCALa: A blueprint for computational models of language acquisition in social context

Sho Tsuji (University of Tokyo, JP)

Different views on language acquisition suggest a range of cues are used, from structure found in the linguistic signal, to information gleaned from the environmental context or through social interaction. Technological advances make it now possible to collect large quantities of ecologically valid data from young children's environment, but we still lack frameworks to extract and integrate such different kinds of cues from the input. SCALa (Socio-Computational Architecture of Language Acquisition) proposes a blueprint for computational models that makes explicit the connection between the kinds of information available to the social early language learner and the computational mechanisms required to extract language-relevant information and learn from it. SCALa further allows us to make precise recommendations for future large-scale empirical research.

3.25 Visual attention development in infancy

Ingmar Visser (University of Amsterdam, NL)

Eye-movements are a valuable source of information, next to responses and response times, for inferring cognitive states and processes. Infant research depends on eye-movements to a large extent as other behavioral response modalities are hard to use in this population. Eye-movement data comes with many challenges, many basic properties are not well known or understood. Optimal methods for defining fixations and saccades are still under much discussion. Free viewing presents a good way to study infant visual attention development and provides robust developmental trends for a number of phenomena that together form an interesting target for computational modeling.

3.26 Temporal patterns in vocal even sequences produced by human infants and computational vocal learning models

Anne Warlaumont (UCLA, US)

In recent years, my collaborators and I have analyzed the timings of when over the course of a day human infants produce vocalizations. These patterns tend to have a somewhat fractal structure, wherein vocalizations occur in clusters within clusters within clusters in time. More recently we have begun to identify relationships between how close two consecutive infant vocal events are in time and how similar they are acoustically. And we are finding that infant vocalizations also tend to be more likely to occur in quick succession in the aftermath of hearing vocalizations produced by adults. We are developing some hypotheses for why these patterns may be important for infant vocal learning. An increase in infant vocalization rate following a reward (either social or intrinsic) may be a mechanism through which human infants can gain additional practice making specific sound types, capitalizing on the current state of the relevant neural and vocal apparatus. In other words, vocalization rate is potentially a pathway to achieving acoustically targeted vocal exploration. This pathway may be particularly useful given that infants' voluntary vocal control is limited; it may be a mechanism for bootstrapping vocal motor learning. Most computational models of vocal learning do not concern themselves with when vocalization occurs in the first place, and also don't consider vocalization-to-vocalization patterns. I expect that some modeling approaches will be better suited than others to addressing these aspects of human vocal learning. These temporal patterns may provide a useful dimension for comparing models to human data, and prioritizing a fit along this dimension may turn out to favor more biologically realistic architectures.

3.27 Curiosity in infants and computational models

Gert Westermann (Lancaster University, GB)

Much of what we know about infants' cognitive development comes from studies in which infants are passive recipients of information presented to them on a computer screen in an order and duration determined by the experimenter. While this body of work has provided us with many insights about infants' learning and their cognitive abilities, these methods ignore a fundamental aspect of real-life learning: outside the lab, infants are actively involved in their learning through exploring their environment and engaging with information in the order and duration they choose. In our lab we investigate infants' information seeking using behavioural, eye tracking, EEG and computational modelling methods. I will give a very brief overview of the methods and studies currently going on in my lab, and then describe a simple auto-encoder neural network model used to simulate intrinsically-motivated exploration that is based on maximizing in-the-moment learning progress. This model learns a stimulus set used in seminal studies of infant category learning as well as a non-curious model embedded in an optimally structured external environment.

3.28 Magnifying Time and Space: New Ways of Studying Early Development and Learning from the Infant's Point of View

Chen Yu (University of Texas – Austin, US)

The three primary research goals in my lab are 1) to quantify the statistical regularities in the real world; 2) to examine the underlying learning mechanisms operated on the statistical data; and 3) to discover developmental pathways in a complex and multi-causal system. Toward the first goal, we have collected a corpus of infant-perspective visual scenes and infant gaze data as they play with their parent in a home-like environment. We have analyzed visual properties of infant-perspective scenes and quantified the ambiguity/transparency of individual parent naming events using infant gaze. We have also fed egocentric video to deep learning models to examine the quantity and quality of the statistical data that lead to successful learning. Toward the second goal, we have used the corpus of scenes that co-occur with parent naming to construct lab experiments which are composed of different mixes of high and low ambiguity naming events. Infants were trained and tested in multiple experimental conditions, varying in terms of the ambiguity of training trials and also in the composition and order of those trials to test specific hypotheses about statistical learning mechanisms. Toward the third goal, we have examined the social effects of joint attention in the development of the infant's own sustained attention and identified the potentially malleable pathway through which social interactions influence the self-regulation of sustained attention. I will conclude my talk by discussing developmental dependencies among motor development, visual perception, sustained attention, joint attention, and language learning.

3.29 Audio-visual self-supervised learning

Andrew Zisserman (University of Oxford, GB)

Lesson 1 from the classic paper "The Development of Embodied Cognition: Six Lessons from Babies" is 'Be Multimodal'. This talks explores how recent work in the computer vision literature on audio-visual self-supervised learning addresses this challenge. The aim is to learn audio and visual representations and tasks directly from the audio-visual data stream of a video (without providing any manual supervision of the data) – much as an infant could learn from the correspondence and synchronization between what they see and hear. It is shown that a neural network that simply learns to synchronize audio and visual streams is able to localize the faces that are speaking (active speaker detection). It is shown that a network that simply learns from the correspondence of faces and voice is able to cluster speakers according to their identity, and so be able to recognize the person from their face or voice.

4 Working Groups

Overview

We split the participants into five working groups to explore three main open-ended research questions related to developmental and machine learning as detailed below:

1. What are the connections between current computational research in self-supervised, weakly-supervised, and continual machine learning, and analogous developmental learning processes in humans and animals?
2. What is the role of computational models of learning (e.g., object recognition, machine perception, and reinforcement learning) in advancing developmental science? Can computational tools enable new developmental research questions? What kinds of data should developmental scientists produce that would be the most useful for computational approaches?
3. What is the role of multimodal learning (learning from diverse signal types such as visual, audio, touch, force, etc.) in development? What are the challenges and opportunities in multimodal machine learning?

Specifically, there are two groups investigated question 1 (Group 1.1 and 1.2), two groups did question 2 (Group 2.1 and 2.2) and one group examined question 3 (Group 3).

4.1 Group 1.1: The power of informational processing bottlenecks

Rhodri Cusack, Uri Hasson, Celeste Kidd, Marc'Aurelio Ranzato, Stefan Stojanov, Anh Thai

This working group explored the relationships between information processing bottlenecks in biological systems and machine learning techniques. Attention bottlenecks are pervasive in biological systems, for example in humans: 1) visual input streams are actively sampled by fixating only on one area of the field of view at a time, often in a context dependent way; 2) working memory is constrained to a few items and features; 3) humans are embodied and physically constrained to only perform one task at a time.

Information bottlenecks can be regarded both as a bug, because they force information to be thrown away, limit parallel processing, or turn an inherently multi-modal task into a unimodal one, or a feature, because they force abstraction, encourage generalization and reduce computational cost. Attention mechanisms have been developed in machine learning, in the form of transformers, LSTM networks. Further, the idea of core sets is concerned with removing data that is informative for learning by finding semantic redundancies. Last, specialised bottlenecked representations have been proposed e.g. variational autoencoders and sparse coding. This group identified that working to obtain high accuracy machine learning systems under constraints such as wall clock time, instantaneous compute, memory, hardware time, bandwidth, is potentially key to artificial general intelligence, in addition to the current trend of scaling model capacity and dataset size.

References

1 Cartwright-Finch, Ula, and Nilli Lavie. "The role of perceptual load in inattentional blindness." Cognition 102.3 (2007): 321-340

2 Alvarez, George A., and Steven L. Franconeri. "How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism." Journal of vision 7.13 (2007): 14-14.

3 Asplund, Christopher L., et al. "Surprise-induced blindness: a stimulus-driven attentional limit to conscious perception." Journal of Experimental Psychology: Human Perception and Performance 36.6 (2010): 1372.

4 Feigenson, Lisa, and Justin Halberda. "Conceptual knowledge increases infants' memory capacity." Proceedings of the National Academy of Sciences 105.29 (2008): 9926-9930.

4.2 Group 1.2: Developmental AI Benchmark

Emmanuel Dupoux, James M. Rehg, Daniel Swingley, Anne Warlaumont, Gert Westermann, Chen Yu

This group conceptualized a new developmental AI benchmark focusing on speech articulation. The outcome of successfully accomplishing this challenge would be an artificial speech system that faithfully reproduces children's speech production learning trajectories. The challenge is separated into three rounds.

The first round consists of learning a control model to articulate speech. More specifically, given an articulatory tract model [1](a model that simulates controllable muscles and vocal tract physics that can synthesize vocalizations including speech sounds), train a control system that can imitate heard speech, producing intelligible words. After training, the model will be tested on reproducing spoken English words and non-words. Such a system would be constrained by the physical properties of the human organs that are used to produce speech sounds, such that the control problem is dynamic and nonlinear, making it highly non-trivial.

The second round would consist of developing general learning and control algorithm such that a single model can operate multiple articulatory models provided (e.g. simulating individual and/or maturational differences among babies), requiring models to show the generality and adaptability characteristic of human learners.

The third round of this challenge is modeling child development trajectories. By providing realistic, recorded child input as a resource for model training, do similar child phonology phenomena appear in the trained models as they learn as in the children? Specific evaluation criteria may include matching children's speech errors and matching prelinguistic vocal milestone sequences.

The group members will work on creating this challenge, which will involve creating an articulatory speech synthesizer and an API, creating training and evaluation datasets, and organizing the challenge.

References

1 Boersma, Paul. Functional phonology. Netherlands Graduate School of Linguistics, 1998.

4.3 Group 2.1: Embodied Intention Prediction Challenge

Michael Frank, Naithilee Kunda, Marvin Lavechin, Pierre-Yves Oudeyer, Rebecca Saxe, Maureen de Seyssel, Ingmar Visser

The goal of this working group was to conceptualize a challenge that could be promoted to the AI community to facilitate the study of computational models that can learn about relationships between infants' attention and the language of their caregivers in naturalistic settings.

Investigating the relationship between language and attention is one way to operationalize larger problems around the role of multi-modal input in language development. For example, when a caregiver says "Look over there!", what is the response of the infant's visual attention, that is, where is the infant looking? Currently this problem is challenging to study because the mutual information between language and attention is hard to quantify and datasets are hard to annotate. Further, lab experiments on attention and intent may not generalize to naturalistic settings. We hoped that the challenge format would provide a focal point for bringing together annotated datasets and teams interested in bringing new models to bear.

This group proposed two complementary multi-modal prediction challenges. In both challenges, the model would have access to caregiver language, either in transcript or raw audio form, and video from an infant's head mounted camera. For the first challenge, the goal is to predict the infant's visual attention in a set of future frames, given the past caregiver language and infant's visual attention. For the second challenge, the goal is to predict the future caregiver language, from previous infant attention and caregiver language. The suggested input data length was 10 seconds, and then the language or attention prediction would be done over the next two seconds. The output for the visual attention challenge would be a vector indicating where the attention will move in the next 2 seconds, whereas for language it would be the words that the parent will say in the next 2 seconds.

Potential datasets that can be used for this challenge are the SAYCam [1] and SEEDLings [2] datasets. The models can be evaluated in two testing regimes, within-subject (training and testing done on data from only one child) or between-subject (training done on pooled data from multiple infants, and tested on data from other infants not seen during training).

A further approach to test such a model would be to use it in a closed-loop time-extended manner to drive attention of a learning agent in an environment where it would interact with a human, and test whether interaction is structured and coordinated in a way that reproduces high-level properties of similar child-caretaker interactions.

References

1 Sullivan, Jessica, et al. "SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective." Open mind 5 (2021): 20-29.

2 Bergelson, Elika, and Richard N. Aslin. "Nature and origins of the lexicon in 6-mo-olds." Proceedings of the National Academy of Sciences 114.49 (2017): 12916-12921.

4.4 Group 2.2: ML for Causal Theories of Child Development?

David Crandall, Alejandrina Cristia, Hana D'Souza, Abdellah Fourtassi, Clément Romac, Olivier Sigaud

This group studied using machine learning (ML) in understanding causal theories of child development. Machine learning can be used for studying child development in three main ways, based on the purpose of the study: ML for annotation, ML for modeling, and ML for simulation.

Machine learning tools can be used to aid automatic annotation for both observational and experimental data in various settings and for different purposes (e.g., object detection, social cues transcription). Data shareability is identified as a potential issue, including privacy concerns and implementation infrastructure needed to host and process data.

Machine learning models are further helpful for predicting child target behaviors or understanding learning mechanisms. However, current techniques to understand the mechanistic process of machine learning models remain superficial.

Another utility of machine learning tools is to study emerging behaviors in artificial contexts. Designing the context and initializing the agent's biases and properties should be carefully considered in order to obtain meaningful results from the simulation.

This group further proposed different ways in which machine learning and developmental learning communities can collaborate and provide helpful scientific insights for each other regarding computational tools and naturalistic data.

4.5 Group 3: Multimodality in babies and machines

Thomas Carta, Hiromichi Hagihara, Felix Hill, Judy Hoffman, Eon-Suk Ko, Casey Lew-Williams, Atsushi Nakazawa, Jelena Suvevic

This working group investigated the role of multimodal inputs in learning in both machines and babies. Multimodality is perceived differently by machine learning and developmental learning communities. For example, modality in human learning is a wide range of human senses from vision to social cues and motion awareness while the machine learning community mostly focuses on a much smaller set of modals such as vision and language or audio. One of the significant questions raised in the discussion concerns if we can bridge "the gap" between cross-disciplinary communities regarding learning with multimodal inputs.

Perceiving multimodal inputs is an inherent part of the human perception system [1, 2]. Different sensors can provide beneficial "redundant" information and create clearer learning moments that support one-shot learning where different sensors are complementary with each other, e.g. some certain toys might make some certain unique sounds. Another difference between human and machine is that multimodality learning cues in humans adapt depending on developmental needs [3, 4, 5]. For example, infants initially focus on faces when they are young but shift to hands which guide their attention to objects [4] when they are older. In contrast, in machines there is no such dynamic present in the integration of multimodal inputs.

Another distinction between learning in machines and babies lies in the processing efficiency of the underlying network structures (layered – machine vs dense – brain connection), synapses pruning and attention mechanisms.

An attempt to bridge the gap between human learning and machine learning is to construct data for learning from a wide-range of modalities for machines. Existing multimodal datasets such as SAYCam [6] and Databrary [7] provide more sensor data for machines than current standard datasets. This group proposed a means to obtain rich multimodality data for machines that is similar to play behavior observed in children.

References

1 Kuhl, Patricia K., and Andrew N. Meltzoff. "The bimodal perception of speech in infancy." Science 218.4577 (1982): 1138-1141.

2 Rosenblum, Lawrence D., Mark A. Schmuckler, and Jennifer A. Johnson. "The McGurk effect in infants." Perception & psychophysics 59.3 (1997): 347-357.

3 Mlincek, Miranda M., et al. "Posture matters: Object manipulation during the transition to arms-free sitting in infants at elevated vs. typical likelihood for autism spectrum disorder." Physical & Occupational Therapy In Pediatrics 42.4 (2022): 351-365.

4 Fausey, Caitlin M., Swapnaa Jayaraman, and Linda B. Smith. "From faces to hands: Changing visual input in the first two years." Cognition 152 (2016): 101-107.

5 Ko, E.-S., Abu-Zhaya, R., Kim, E.-S., Kim, T., On, K.-W., Kim, H., Zhang, B.-T., and Seidl, A. "Mothers' use of touch across infants' development and its implications for word learning: Evidence from Korean dyadic interactions", Infancy (2023). DOI: 10.1111/infa.12532

6 Sullivan, Jessica, et al. "SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective." Open mind 5 (2021): 20-29.

7 R. O. Gilmore, K. E. Adolph and D. S. Millman, "Curating identifiable data for sharing: The databrary project," 2016 New York Scientific Data Summit (NYSDS), New York, NY, USA, 2016, pp. 1-6, doi: 10.1109/NYSDS.2016.7747817.

Participants

- Thomas Carta
INRIA – Bordeaux, FR
- David J. Crandall
Indiana University – Bloomington, US
- Alejandrina Cristia
LSCP – Paris, FR
- Rhodri Cusack
Trinity College Dublin, IE
- Hana D'Souza
Cardiff University, GB
- Maureen de Seyssel
INRIA & ENS Paris, FR
- Emmanuel Dupoux
LSCP – Paris, FR
- Abdellah Fourtassi
Aix-Marseille University, FR
- Michael C. Frank
Stanford University, US
- Hiromichi Hagihara
University of Tokyo, JP
- Uri Hasson
Princeton University, US
- Felix Hill
Google DeepMind – London, GB
- Judy Hoffman
Georgia Institute of Technology – Atlanta, US
- Celeste Kidd
University of California – Berkeley, US
- Eon-Suk Ko
Chosun University, KR
- Maithilee Kunda
Vanderbilt University, US
- Marvin Lavechin
Meta AI – Paris, FR
- Casey Lew-Williams
Princeton University, US
- Atsushi Nakazawa
Kyoto University, JP
- Pierre-Yves Oudeyer
INRIA – Bordeaux, FR
- Marc'Aurelio Ranzato
DeepMind – London, GB
- James M. Rehg
Georgia Institute of Technology – Atlanta, US
- Clement Romac
INRIA – Bordeaux, FR
- Rebecca Saxe
MIT – Cambridge, US
- Olivier Sigaud
Sorbonne University – Paris, FR
- Stefan Stojanov
Georgia Institute of Technology – Atlanta, US
- Jelena Sucevic
University of Oxford, GB
- Daniel Swingley
University of Pennsylvania, US
- Ngoc Anh Thai
Georgia Institute of Technology – Atlanta, US
- Ingmar Visser
University of Amsterdam, NL
- Anne Warlaumont
UCLA, US
- Gert Westermann
Lancaster University, GB
- Chen Yu
University of Texas – Austin, US

Remote Participants

- Kristen Grauman
University of Texas – Austin, US
- Jitendra Malik
University of California – Berkeley, US
- Linda B. Smith
Indiana University – Bloomington, US
- Sho Tsuji
University of Tokyo, JP
- Andrew Zisserman
University of Oxford, GB