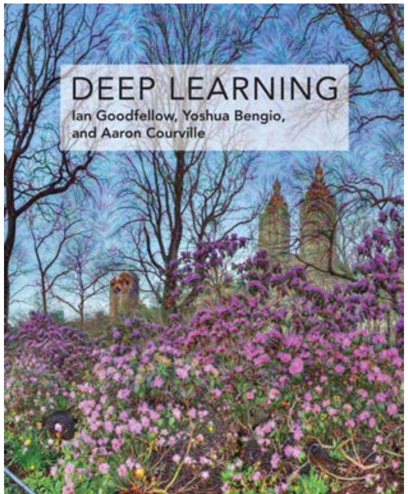
Towards compositional understanding of the world by deep learning

Yoshua Bengio



Peking University
Distinguished Lecture

October 8th, 2019,
(remote talk)

Mila

CIFAR | ICRA
CANADIAN INSTITUTE FOR ADVANCED RESEARCH | INSTITUT CANADIEN DE RECHERCHES AVANCÉES

Université de Montréal

Current AI is far from Human-Level AI

- Sample complexity is high for supervised learning, even more for RL
  - Real-world actions can be lethal, experience is limited & costly
  - We don’t have a good simulator of the real world (esp. involving humans)

- High-level concepts provided by human designers or labelers

- Errors made by trained systems reveal that their ‘understanding’ is very shallow and superficial

- The dream of deep learning discovering and disentangling high-level explanatory variables is far from achieved

Mila

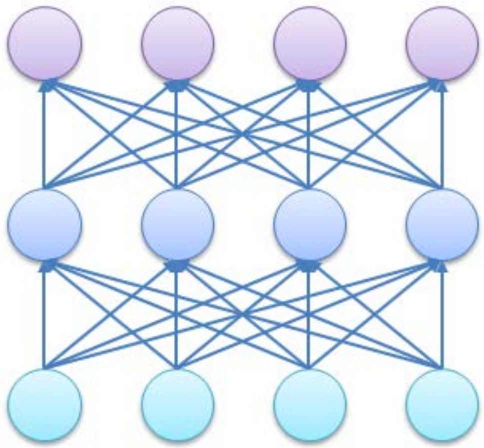
Learning Multiple Levels of Abstraction

(Bengio & LeCun 2007)

- The big payoff of deep learning is to allow learning higher levels of abstraction
- Higher-level abstractions would **disentangle the factors of variation**, which allows much easier generalization, transfer, reasoning, and language understanding
- These factors are composed to form observed data

How to Discover Good Disentangled Representations

- How to discover abstractions?
- What is a good representation? ***(Bengio et al 2013)***
- Need clues (= priors) to help **disentangle** the underlying factors (**not necessarily statistically independent**), such as
  - Spatial & temporal scales
  - Marginal independence
  - Simple dependencies between factors
    - *Consciousness prior*
  - Causal / mechanism independence
    - *Controllable factors*



System 1 vs System 2 Cognition

Two systems (and categories of cognitive tasks):

- **System 1**
  - Cortex-like (state controller and representations)
  - intuitive, fast heuristic, UNCONSCIOUS, non-linguistic
  - what current DL does quite well
- **System 2**
  - Hippocampus (memory) + prefrontal cortex
  - slow, logical, sequential, CONSCIOUS, linguistic, algorithmic
  - what classical symbolic AI was trying to do
- **Grounded language learning**: combine both systems

Manipulates high-level / semantic concepts, which can be recombined combinatorially

Mila

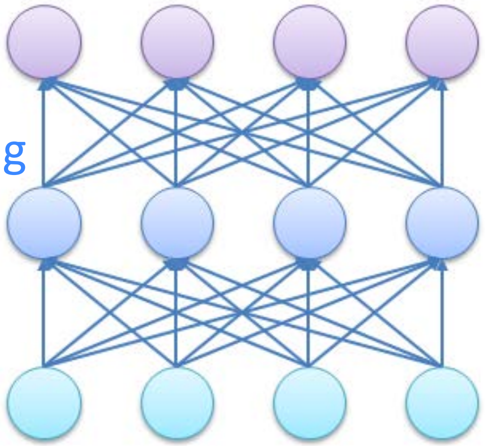
Compositionality to bypass the curse of dimensionality

We need to build compositionality into our ML models

- Just as human languages exploit compositionality to give representations and meanings to complex ideas

Exploiting compositionality can give an **exponential** gain in representational power

- Distributed representations / embeddings: feature learning
- Deep architecture: multiple levels of feature learning



Prior assumption: compositionality is useful to describe the world around us efficiently

Missing from Current ML: Understanding & Generalization Beyond the Training Distribution

- Learning theory only deals with generalization within the same distribution
- Models learn but do not generalize well (or have high sample complexity when adapting) to modified distributions, non-stationarities, etc.
- Poor reuse, poor modularization of knowledge: humans are good at systematic generalization (e.g., combining known words in new ways unlikely under the training distribution)

Mila

The Need for Meta-Learning

Meta-Learning / Learning to Learn

- Generalize the idea of hyper-parameter optimization
  - Inner loop optimization (normal training), a fn of meta-params

$$\theta_t(\omega) = \mathrm{approxmin}_\theta C(\theta, \omega, \mathcal{D}^t_{train})$$

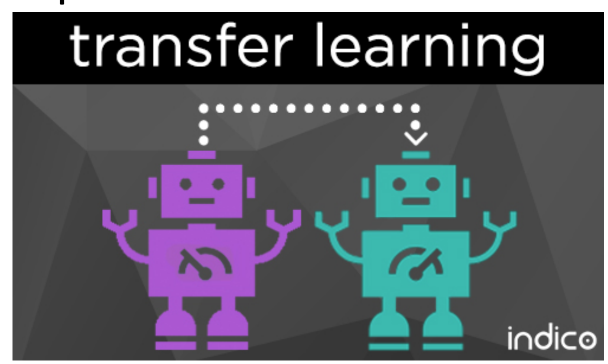
  - Outer loop optimization (meta-training), optimize meta-params

$$\omega = \mathrm{approxmin}_\omega \sum_t L(\theta_t(\omega), \omega, \mathcal{D}^t_{test})$$

- Meta-parameters can be the learning rule itself *(Bengio et al 1991; Schmidhuber 1992)*, learn 2 optimize
- Meta-learn an objective or reward function, or a shared encoder
- Meta-learning can be used to learn to generalize or transfer
- Can backprop through θ_t, use RL, evolution, or other tricks

Learning to Generalize and Adapt End-to-End

- We can optimize through the sequence
  - see regular training data (and learn from it)
  - see (a few) out-of-distribution examples (and optionally fine-tune / adapt to them)
- if these steps involve some meta-parameters which can be tuned so that we optimize the generalization performance in the second step
  - 0-shot generalization = out-of-distribution generalization
  - k-shot generalization: the learner is allowed to use a few examples of the modified distribution, we are doing **transfer learning**



Mila

Beyond iid: Hypotheses about how the environment changes, Independent Mechanisms and the Small Change Hypothesis

- Independent mechanisms:
  - changing one mechanism does not change the others *(Peters, Janzig & Scholkopf 2017)*
- Small change:
  - Non-stationarities, changes in distribution, involve few mechanisms at a time (e.g. the result of a single-variable intervention)
- How can we discover these independent mechanisms, i.e., factor knowledge?

Mila

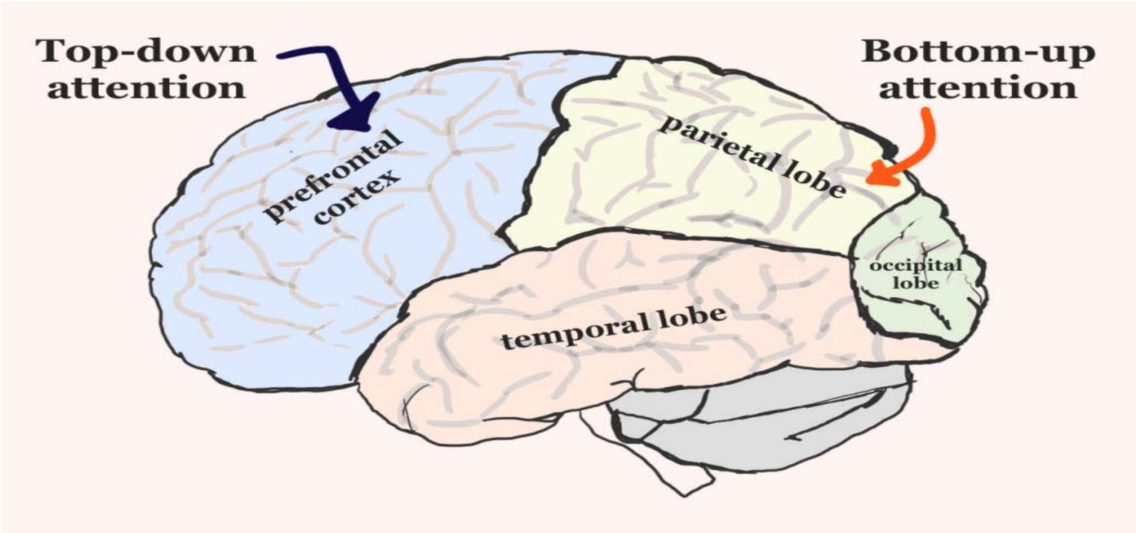
The Need for Sparsely Interacting Modules

Mila

On the Relation between Abstraction, Thought and Attention

- **A thought is a low-dimensional object**, few aspects of the state
- Attention allows us to focus on a few elements out of a large set
- Soft-attention allows this process to be trainable with gradient-based optimization and backprop

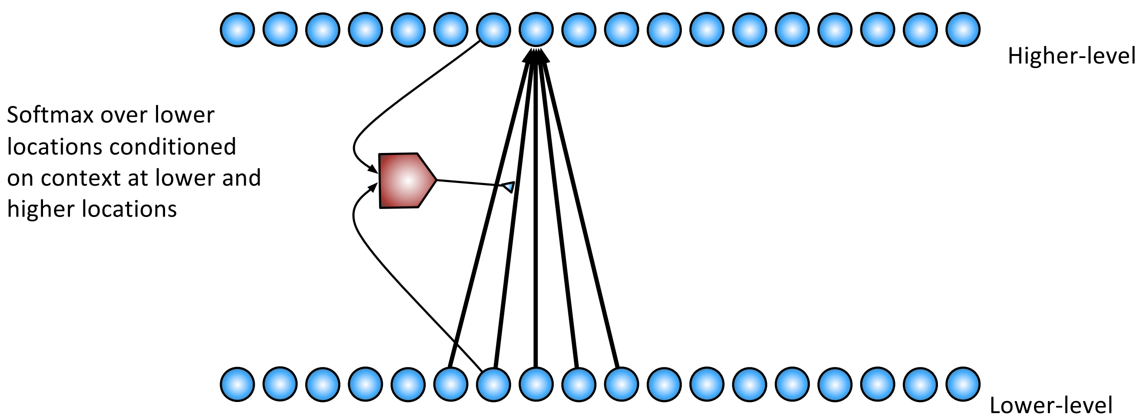
Attention focuses on a few appropriate abstract or concrete elements of mental representation



The Attention Revolution in Deep Learning

- **Attention mechanisms exploit GATING units,** have unlocked a breakthrough in machine translation:

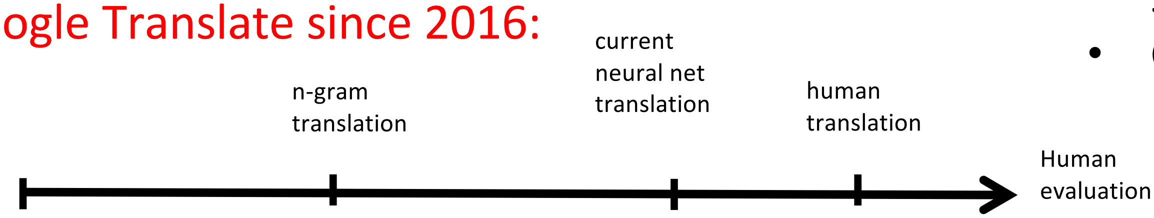
Neural Machine Translation (ICLR'2015)



Attention enables:

- Differentiable memory access
- Operating on sets
- Long-term dependencies
- Self-attention, transformers, SOTA NLP
- **Consciousness**

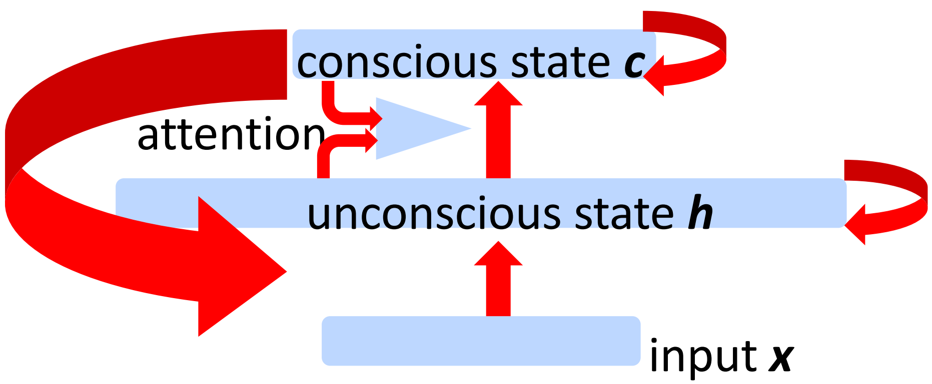
- In Google Translate since 2016:

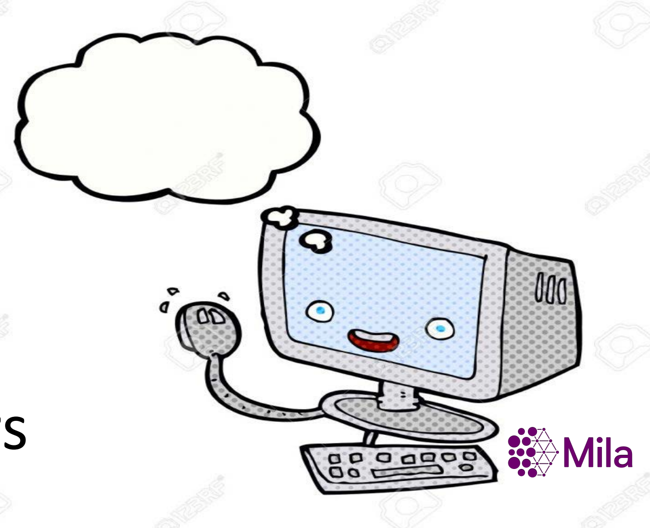


The Consciousness Prior

Bengio 2017, arXiv:1709.08568

- 2 levels of representation:
  - High-dimensional abstract representation space (all known concepts and factors) ***h***
  - Low-dimensional conscious thought ***c***, extracted from ***h***





- ***c*** includes names (keys) and values of factors

Mila

Why do I call it a Prior?

- There is something very special about **the kind of high-level variables which we manipulate with language**:
  - we can predict some given very few others
    - E.g. "if I drop the ball, it will fall on the ground"
  - corresponds to a sparse **factor graph**
  - **Each factor captures an independent piece of knowledge**
    - Strong interactions between few variables

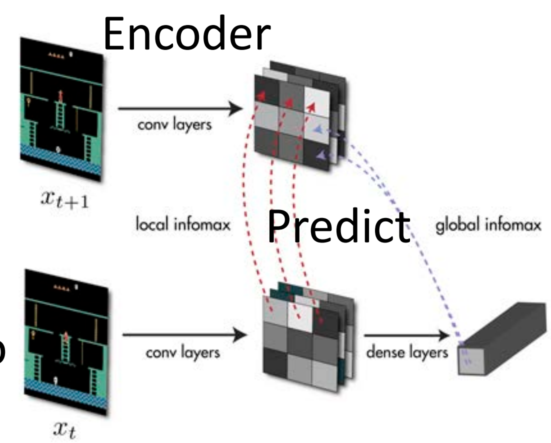
$$P(V) \propto \prod_k \phi(V_{s_k})$$

where V_{s_k} is the subset of V with indices s_k

Mila

Learn Generative Models in Latent Space, not Pixel Space

- For human-like brains, generative models are useful for planning (**model-based RL**), imagination, counterfactuals, inference over causes and explanations, high-level credit assignment
  - NONE OF THIS REQUIRES WORKING IN PIXEL SPACE
- Current generative models are trained wrt pixel-space objectives, how to train purely in the space of abstract representations? We want the encoder mapping pixel space to abstract space to be trained wrt the high-level goals too.
- There is an issue of possible collapse of representations if we maximize predictability (e.g. max likelihood) in latent space



Deep InfoMax or DIM
(Hjelm et al & Bengio ICLR 2019)

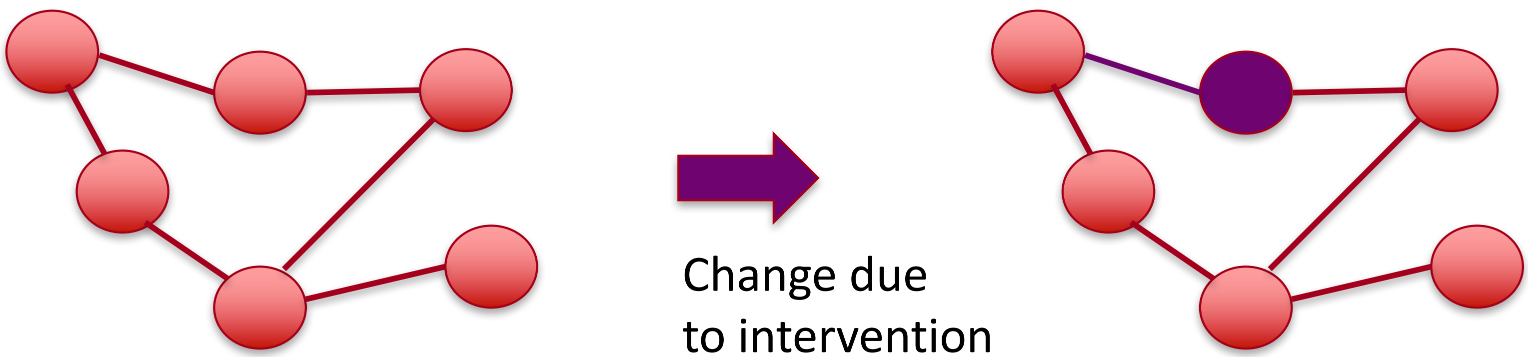
Mila

Integrating System 1 and System 2

- System 2 model is very coarse and imperfect, unlike system 1
- System 2 abstract concepts need to be grounded via system 1
- System 2 thinking allows counterfactual reasoning, i.e., imagining scenarios which did not and will not happen, as an exercise (e.g. for credit assignment, if I had done that…), allows generalization far from the training data, imagine dangerous scenarios without having to take the actual risks
- System 2 is too slow and inefficient, compile to system 1 into habits and intuitive behavior

Separating Knowledge in Small Re-Usable Pieces

- Pieces which can be re-used combinatorially
- Pieces which are stable vs nonstationary, subject to interventions



Change due
to intervention

Mila

Wrong Knowledge Factorization Leads to Poor Transfer

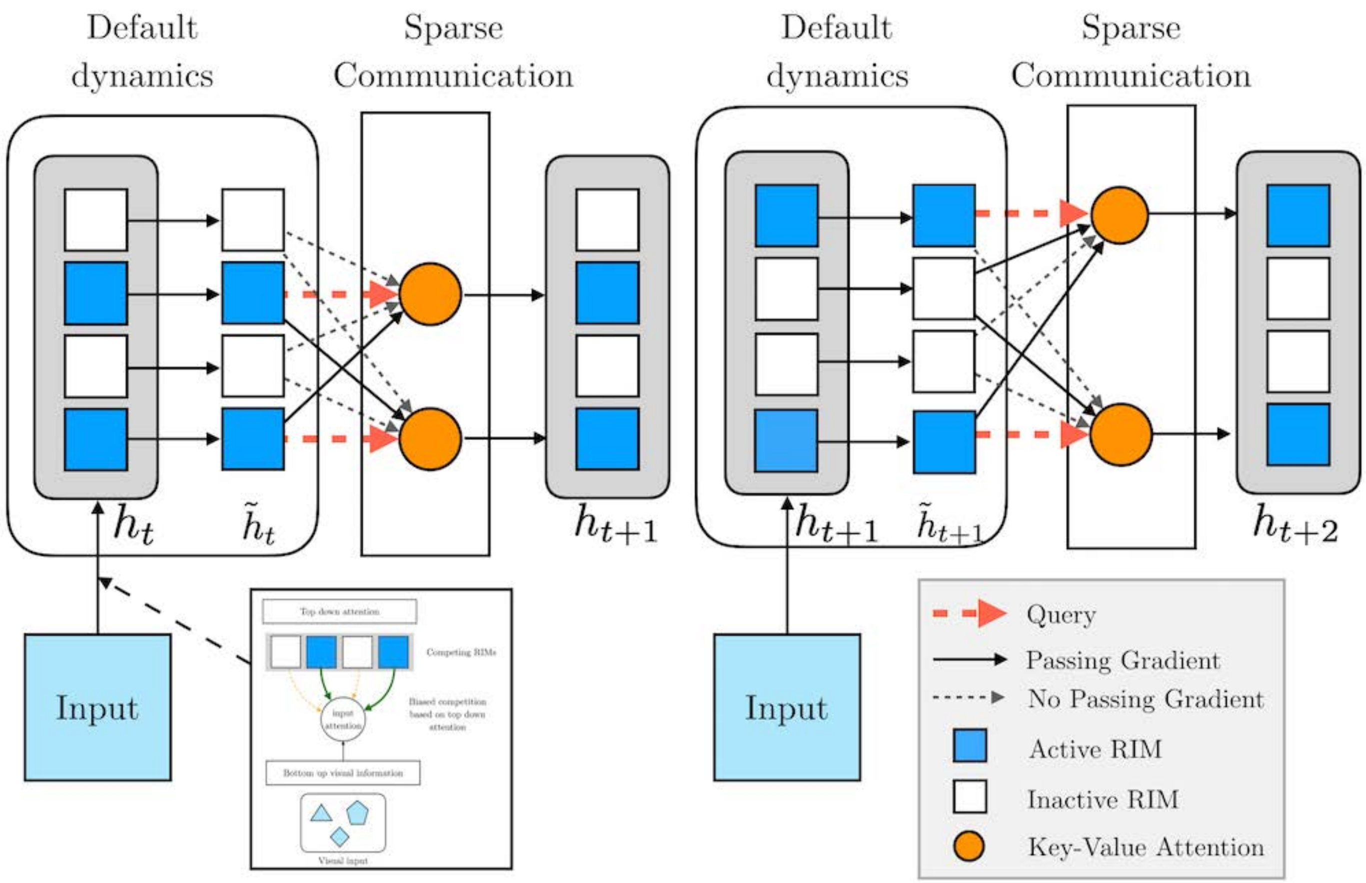
- With the wrong factorization P(B) P(A|B), a change in ground truth P(A) influences both modules, all the parameters
  - poor transfer: all the parameters must be adapted
- This is the normal situation with standard neural nets: every parameter participates to every relationship between all the variables
  - this causes ***catastrophic forgetting, poor transfer, difficulties with continual learning or domain adaptation, etc***

Mila

Recurrent Independent Mechanisms

Goyal et al, arXiv:1909.10893

Multiple recurrent sparsely interacting modules, each with their own dynamics, with object (key/value pairs) input/outputs selected by multi-head attention



Mila

Recurrent Independent Mechanisms

Goyal et al, arXiv:1909.10893

Copying	k_{T}	k_{A}	h_{size}	Train(50) CE	Test(200) CE
RIMs	6	5	600	0.01	3.5
	6	4	600	**0.00**	**0.00**
	6	3	600	**0.00**	**0.00**
	6	2	600	**0.00**	**0.00**
	5	3	500	**0.00**	**0.00**
LSTM	-	-	300	0.00	2.28
	-	-	600	0.00	3.56
NTM	-	-	-	0.00	2.54
RMC	-	-	-	0.00	0.13
Transformers	-		-	0.00	0.54

Sequential MNIST	k_{T}	k_{A}	h_{size}	16 x 16 Accuracy	19 x 19 Accuracy	24 x 24 Accuracy
RIMs	6	6	600	85.5	56.2	30.9
	6	5	600	88.3	43.1	22.1
	6	4	600	**90.0**	**73.4**	**38.1**
LSTM	-	-	300	86.8	42.3	25.2
	-	-	600	84.5	52.2	21.9
EntNet	-	-	-	89.2	52.4	23.5
RMC	-	-	-	89.58	54.23	27.75
DNC	-	-	-	87.2	44.1	19.8
Transformers	-		-	**91.2**	51.6	22.9

RIMs generalize better than SOTA methods for sequential learning to out-of-distribution data (longer sequences, larger images).

Mila

The Need for Causal Understanding

Learning « How the world ticks »

- So long as our machine learning models « cheat » by relying only on superficial statistical regularities, they remain vulnerable to out-of-distribution examples
- Humans generalize better than other animals thanks to a more accurate internal model of the **underlying causal relationships**
- To predict future situations (e.g., the effect of planned actions) far from anything seen before while involving known concepts, an essential component of reasoning, intelligence and science

Deep Learning Objective: discover high-level representation capturing cause and effect variables

- What are the right representations?
  - Causal variables explaining the data
  - Pixels are not causal variables
- How to discover them? (learn the mythical encoder)
- How to discover their causal relationship, the causal graph?

Mila

Turning a Hindrance into a Useful Signal

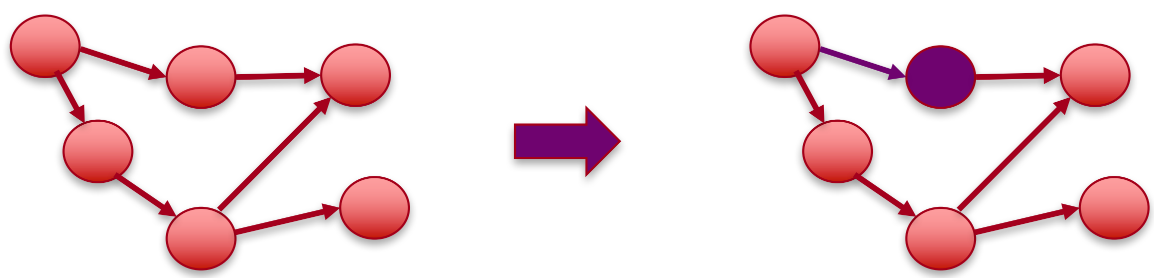
ArXiv paper, Bengio et al 2019: *A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms*

- Changes in distribution (nonstationarities in agent learning, transfer scenarios, etc) are seen as a bug in ML, a challenge
- Turn them into a feature, an asset, to help discover causal structure, or more generally to help **factorize knowledge:**
- **Tune knowledge factorization (e.g. causal structure) to maximize fast transfer**
- *"Nature does not shuffle environments, we shouldn't"*
  L. Bottou

Mila

Small Change →
Small Sample Complexity

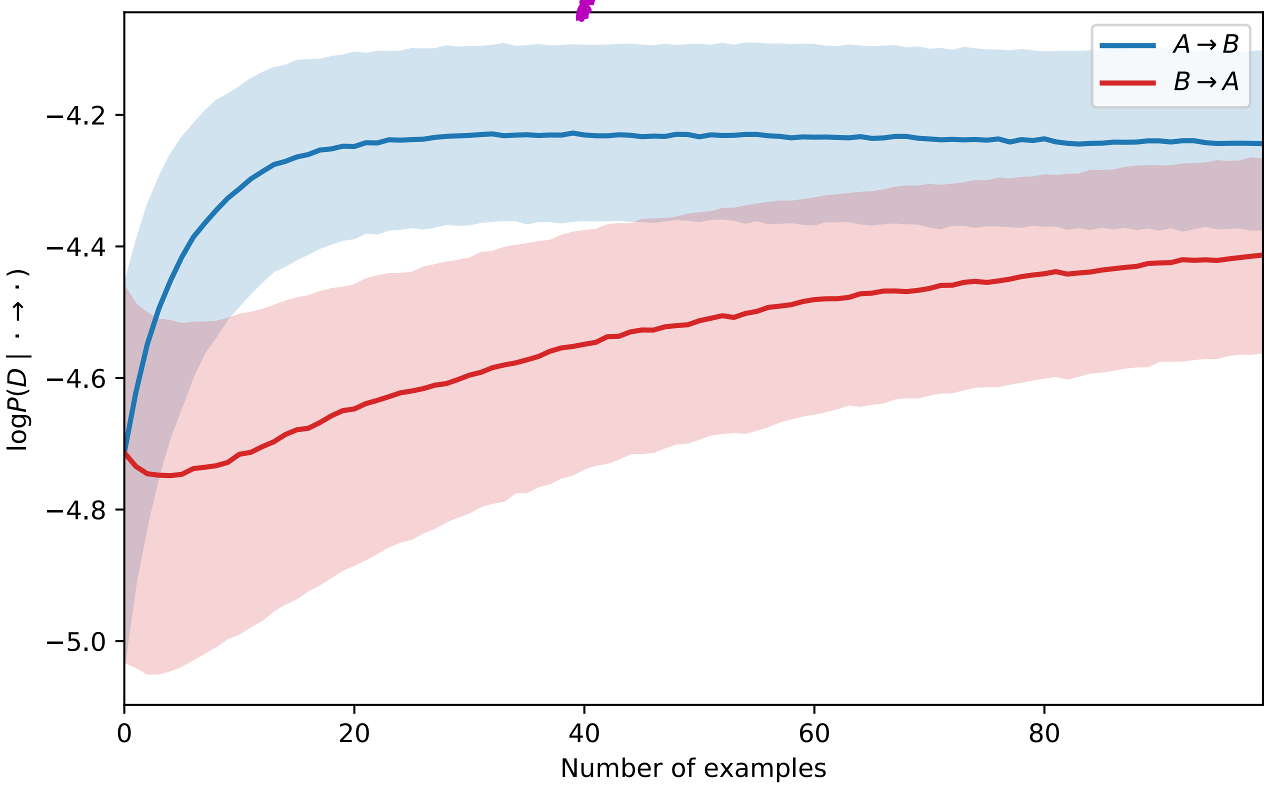
Few parameters need to change → small L2 change → few examples needed to recover from the change



Under the right parametrization → fast adaptation to interventions

Mila

Empirical Confirmation: Correct Causal Structure Leads to Faster Adaptation



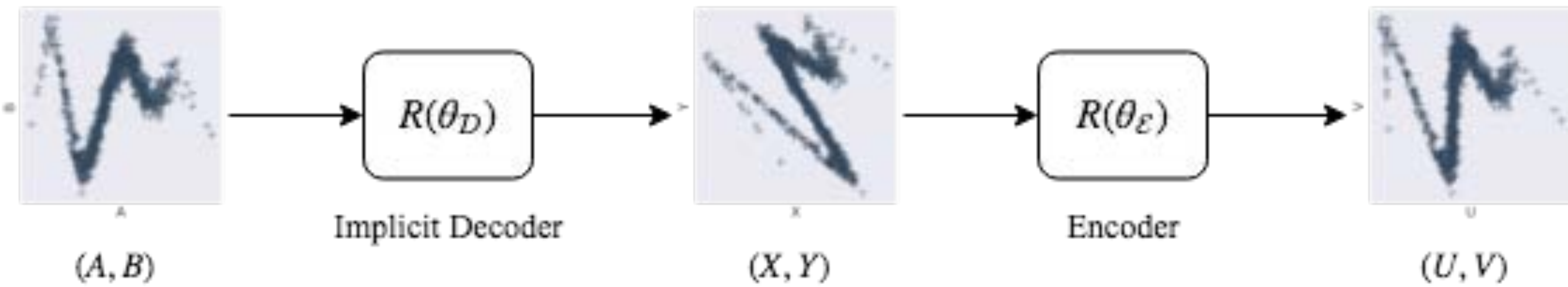
A→B is the correct causal structure: faster online adaptation to modified distribution = lower NLL regret

Mila

A Novel Approach to Causality: Disentangling the Causes

Bengio et al 2019: *A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms*

- Realistic settings: causal variables are not directly observed
- Need to learn an encoder which maps raw data to causal space
- Consider both the encoder parameters and the causal graph structural parameters as meta-parameters trained together wrt proposed meta-transfer objective



Mila

Experiments successful in 2-D with simple linear mappings, Bengio et al 2019.

Learning Neural Causal Models from Unknown Interventions:

Avoiding *Super-exponential Search*

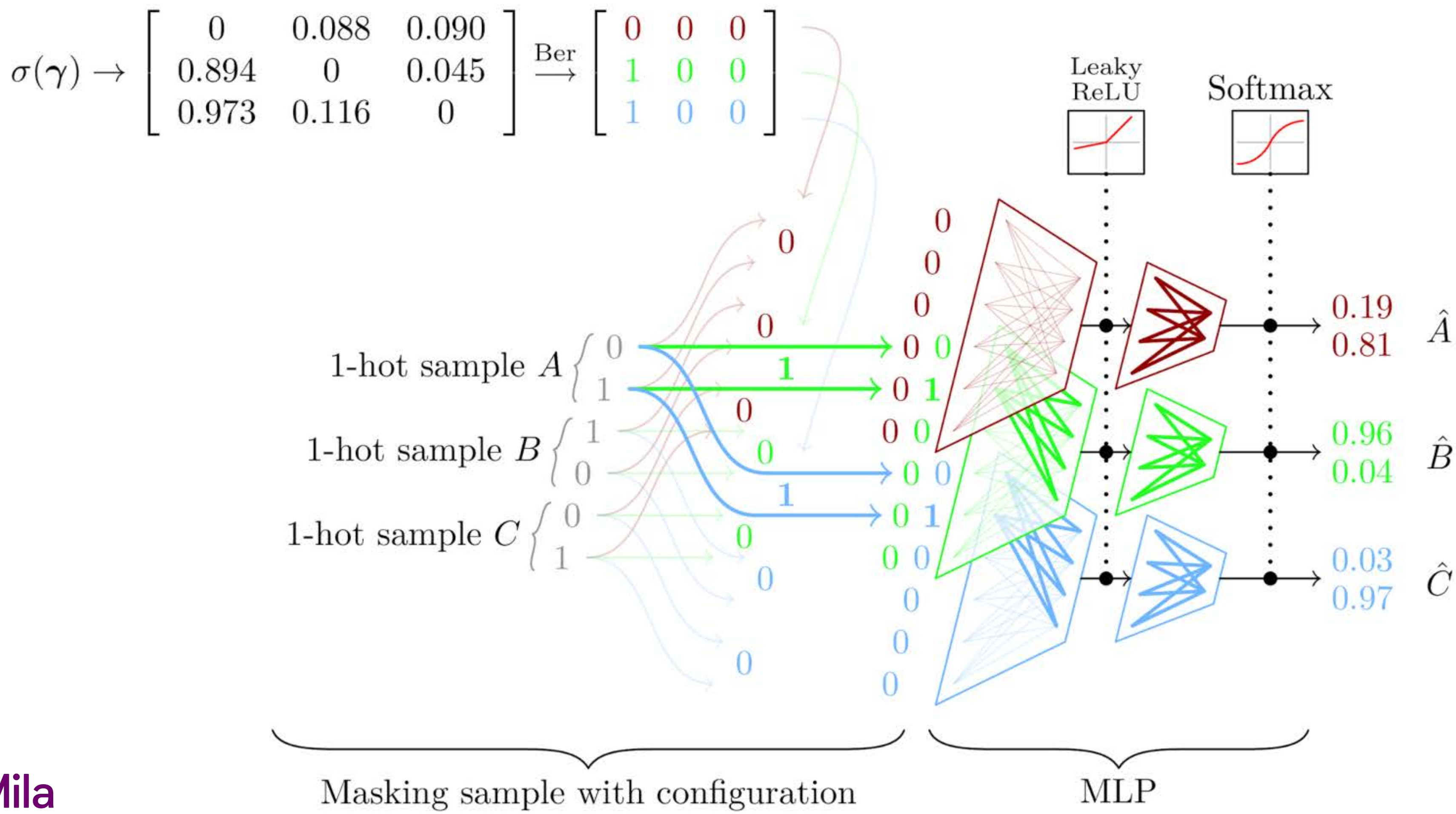
Ke et al & Bengio arXiv:1910.01075

- Most causal induction methods search over super-exponential number of possible graphs
  - Difficult to scale to larger graphs

- How to bypass the super-exponential search?
  - Learn ensemble of structured causal models (SCM)
  - More efficient, does not have to search through super-exponential set of possible DAGs.

Mila

Multivariate Categorical MLP Conditionals



Comparative Results

Asia graph, CE on ground truth edges, comparison against other causal induction methods

Our method	(Eaton & Murphy, 2007a)	(Peters et al., 2016)	(Zheng et al., 2018)
0.0	0.0	10.7	3.1

Evaluating the consequences of a previously unseen intervention

	fork3	chain3	confounder3	collider3
Our Model	-0.4502	-0.3801	-0.2819	-0.4677
Baseline	-0.5036	-0.4562	-0.3628	-0.5082

Ke et al & Bengio arXiv:1910.01075
Learning Neural Causal Models from Unknown Interventions

Mila

Observing Other Agents

- Can infants figure out causal structure in spite of being almost passive observers?
- Yes, if they exploit and infer the interventions made by other agents
- Our approach does not require the learner to know what the action/intervention was (but it could do inference over interventions)
- But more efficient learning if you can experiment and thus test hypotheses about cause & effect

Mila

The Need for the Agent Perspective in Deep Learning

Mila

The Agent Perspective for Deep Learning

- Classical deep learning and ML only considered a fixed data distribution
- Agents can modify their environment through their actions
- There may be multiple agents, also leading to non-stationarities, changing distribution
- Difficult to generalize out-of-distribution
- Need for the agents to "really understand" their environment
- Acting purposely can help to gather knowledge, discover good representations

Jointly Learning Natural Language and a World Model

- Should we first learn a world model and then a natural language description of it?
- Or should agents jointly learn about language and about the world?
- I lean towards the latter.
- Consider top-level representations from supervised ImageNet classifiers. They tend to be much better and easier to learn than those learned by unsupervised learning. Why?
- Because language (here object categories) provides to the learner clues about relevant semantic high-level factors from which it is easier to generalize.
- See my earlier paper on cultural evolution, which posits that culture can help a learner escape from poor optimization, guide (through curricula) the learner to better explanations about the world.

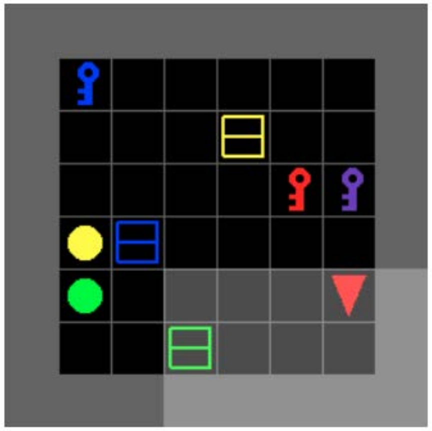
Grounded Language Learning

BabyAI Platform *Chevalier-Boisvert et al & Bengio ICLR 2019*

Purpose: simulate language learning from a human and study data efficiency

Comprises:

- a gridworld with partial observability (Minigrid)
- a compositional natural-looking Baby language with over 10^19 instructions
- 19 levels of increasing difficulty
- a heuristic stack-based expert that can solve all levels

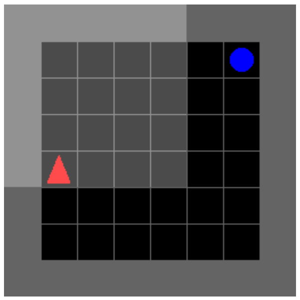


(b) PutNextLocal: "put the blue key next to the green ball"

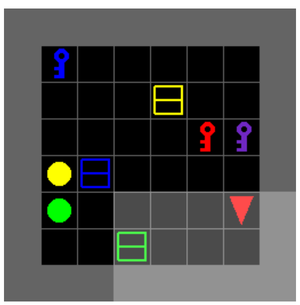
github.com/mila-udem/babyai

Mila

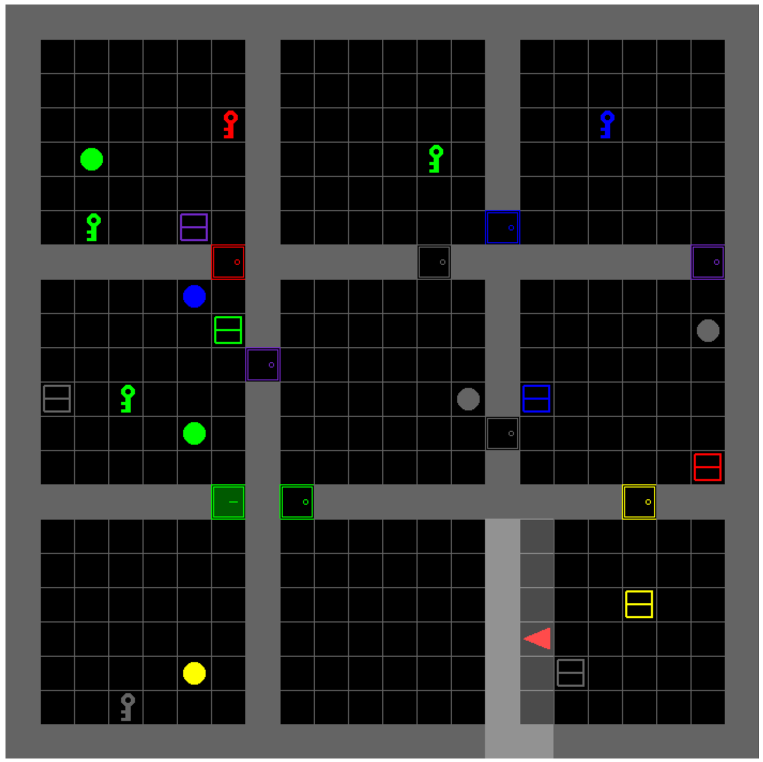
Early Steps in Baby AI Project



(a) GoToObj: "go to the blue ball"



(b) PutNextLocal: "put the blue key next to the green ball"



(c) BossLevel: "pick up the grey box behind you, then go to the grey key and open a door". Note that the green door near the bottom left needs to be unlocked with a green key, but this is not explicitly stated in the instruction.

- Designing and training experts for each level, which can serve as teachers and evaluators for the Baby AI learners
- Partially observable, 2-D grid, instructions about objects, locations, actions

go to the red ball

open the door on your left

put a ball next to the blue door

open the yellow door and go to the key behind you

put a ball next to a purple door after you put a blue box next to a grey box and pick up the purple box

Acting to Guide Representation Learning & Disentangling

(E. Bengio et al, 2017; V. Thomas et al, 2017; more recently see Warde-Farley et al ICLR 2019, Kim et al ICML 2019)

- **Some factors (e.g. objects) correspond to 'independently controllable' aspects of the world**
  - Corresponds to maximizing mutual information between intentions (goal-conditioned policies) and changes in the state (trajectories), conditioned on the current state.
- *Can only be discovered by acting in the world*
  - *Control linked to notion of objects & agents*
  - *Causal but agent-specific & subjective: affordances*

Mila

Four Tools for More Compositional Deep Learning

1. Meta-Learning (to adapt quickly to changes in distribution)
2. Sparsely interacting mechanisms at the top level (consciousness prior)
3. High-level variables are causal and their dependencies are represented in a modular way
4. System 1 and system 2 together actively acquire a world model and corresponding semantic concepts (grounded language learning), can be composed for reasoning and planning, and representations of actions and state are linked (affordances)

Looking Forward

- Build a world model which meta-learns causal effects in abstract space of causal variables, able to quickly adapt to changes in the world and generalize out-of-distribution
- Acting to acquire that knowledge (exploratory behavior)
- Bridging the gap between system 1 and system 2, old neural nets and conscious reasoning, all neural