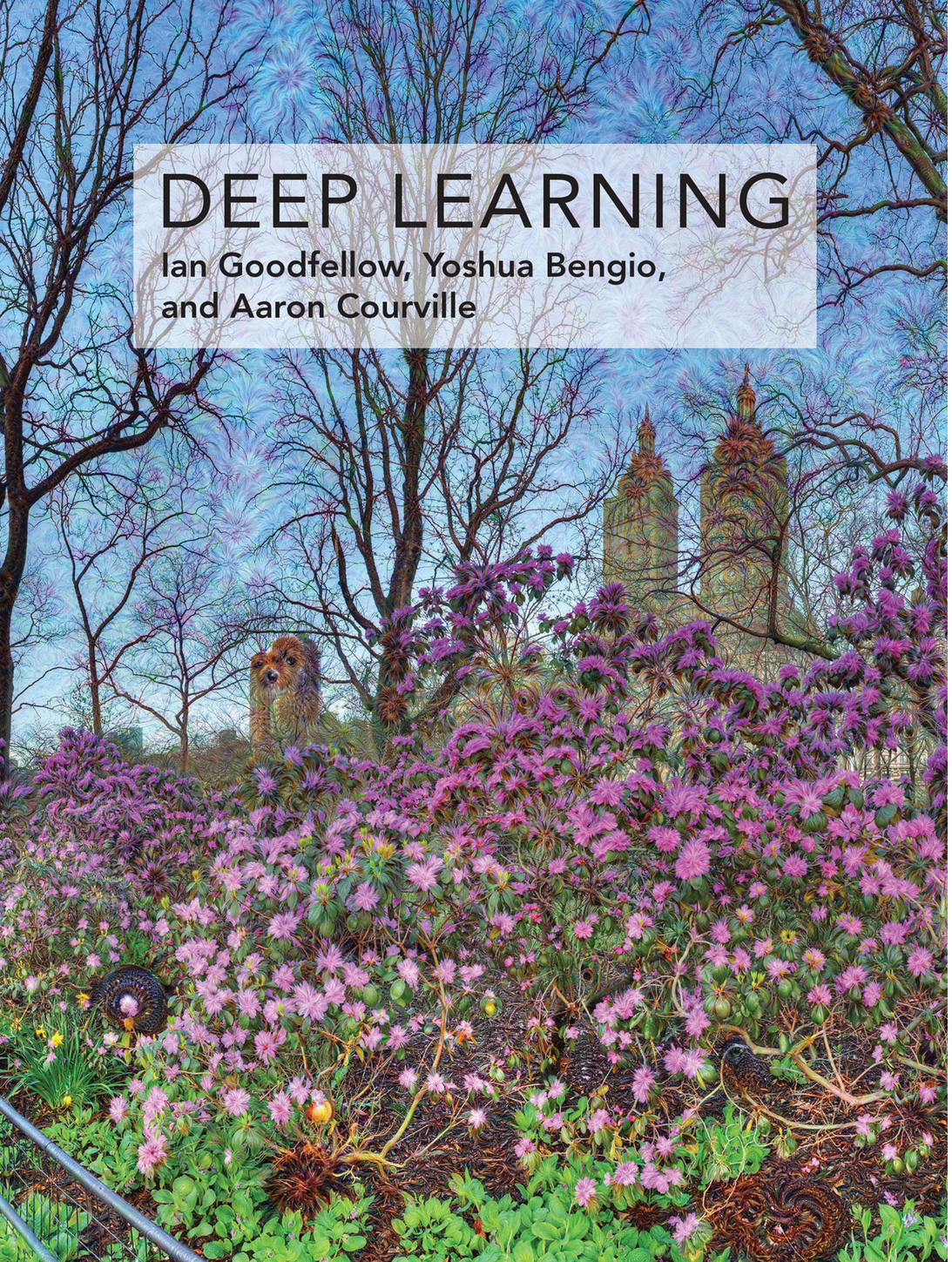


DEBATE WITH GARY MARCUS

YOSHUA BENGIO

December 23rd, 2019, Mila, Montreal

Mila

Université de Montréal

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

ICRA
INSTITUT CANADIEN DE RECHERCHES AVANCÉES

MAIN POINTS

- Core challenge: OOD generalization
- Deep learning has made good progress on system 1, what extensions are needed for system 2?
- Attention mechanisms: key ingredient, opens door to dynamic recombination, systematic generalization, causal factorization of knowledge
- How is that different from sticking GOFAI algorithms on top of deep perception?

ON THE TERM DEEP LEARNING

- Deep learning is not a fixed architecture, training methodology
- It's not MLPs, Convnets, RNNs or backprop
- It is an evolving approach to build intelligent learning & generalizing machines inspired by the brain
- Gradually building a corpus of principles (priors) guiding the design of tens of thousands of papers, e.g.
    - Brains provide powerful architectural priors (e.g. neural architecture, neural nonlinearities, local connectivity, spatial representations, etc)
    - Learning as optimization (or multiple optimizations in game-theoretical setup, as in GANs), i.e. coordinated learning of multiple parts of the system (e.g. end-to-end, actor-critic, …)
    - Gradient-based optimization (especially SGD) is extremely successful, especially for generalization
    - Distributed representations and depth provide powerful combinatorial priors
    - Sharing computation and representations across tasks, environments, etc enables multi-task learning, transfer learning, and learning to learn
    - Reasoning/search/inference can be implemented by energy-minimization and also approximated by deep nets

See impact on neuroscience: Richards et al 2019, Nature Neurosc. A deep learning framework for neuros

AGENT LEARNING NEEDS **OOD GENERALIZATION**

Agents face non-stationarities

Changes in distribution due to

- their actions
- actions of other agents
- different places, times, sensors, actuators, goals, policies, etc.

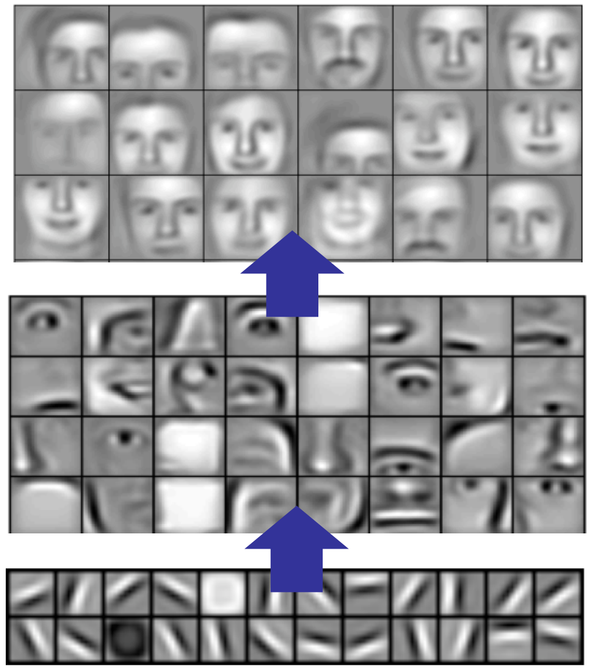
Multi-agent systems: many changes in distribution
Ood generalization needed for continual learning

Mila

COMPOSITIONALITY HELPS IID AND OOD GENERALIZATION

Different forms of compositionality
each with different exponential advantages

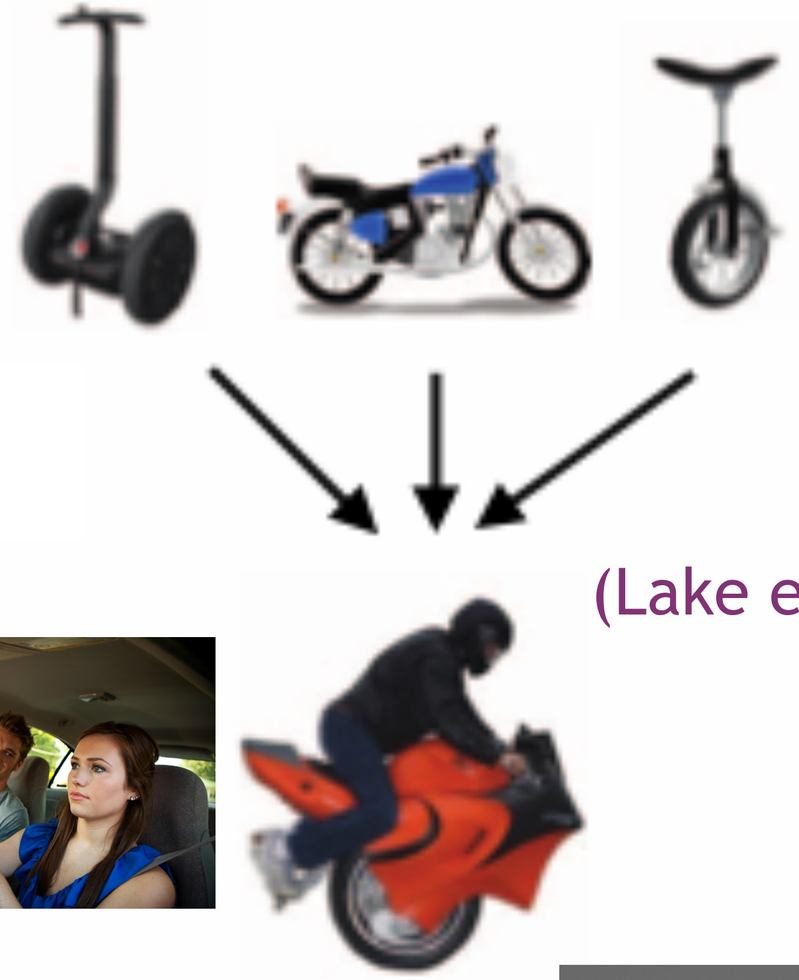
- Distributed representations
  (Pascanu et al ICLR 2014)
- Composition of layers in deep nets
  (Montufar et al NeurIPS 2014)
- **Systematic generalization in language, analogies, abstract reasoning? TBD**



(Lee, Grosse, Ranganath & Ng, ICML 2009)

SYSTEMATIC GENERALIZATION

- Studied in linguistics
- **Dynamically recombine existing concepts**

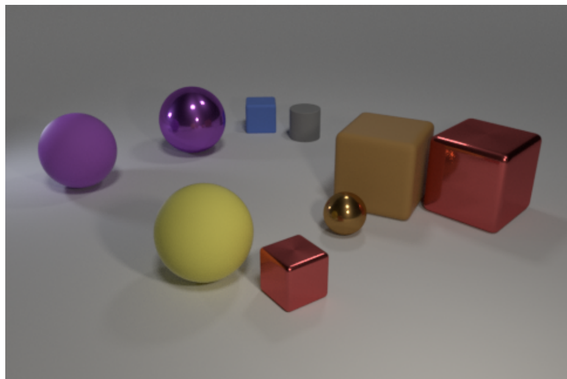


(Lake et al 2015)

- Even when new combinations have 0 probability under training distribution
  - E.g. Science fiction scenarios
  - E.g. Driving in an unknown city

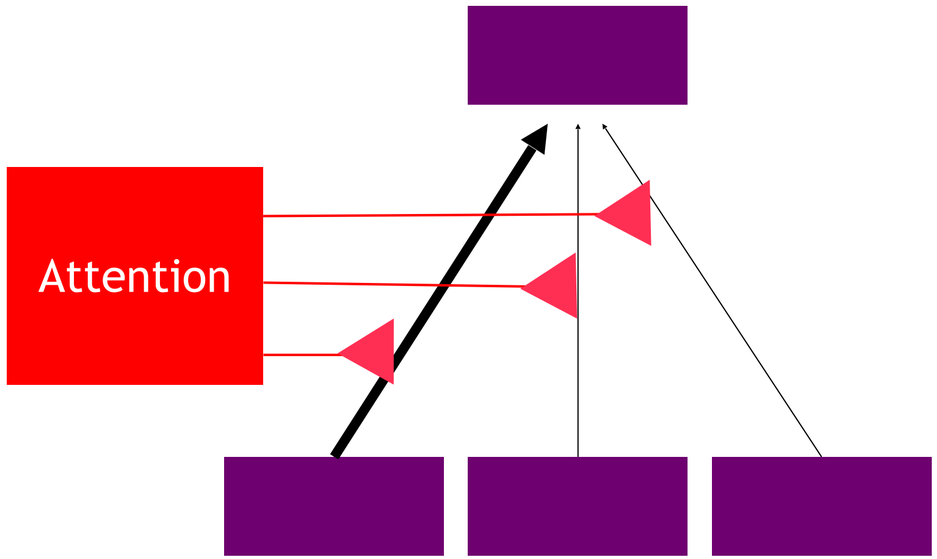
- Hot topic in DL research

(Lake & Baroni 2017)
(Bahdanau et al & Courville ICLR 2019)
CLOSURE: ongoing work by Bahdanau et al & Courville on CLEVR



Mila

FROM ATTENTION TO **INDIRECTION**



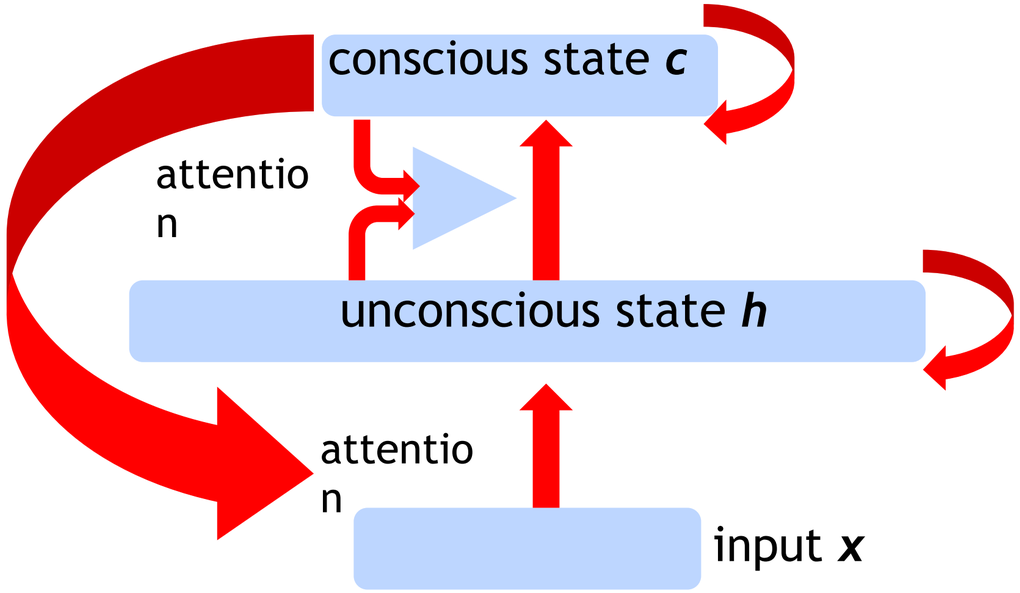
- Attention = dynamic connection
- Receiver gets the selected value
- Value of what? From where?

→ Also send ‘name’ (or key) of sender

- Keep track of 'named’ objects: indirection
- Manipulate sets of objects (transformers)

CONSCIOUSNESS PRIOR

Bengio 2017, arXiv:1709.08568



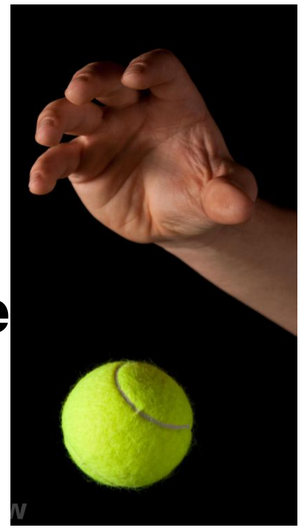
Different kinds of attention in the brain

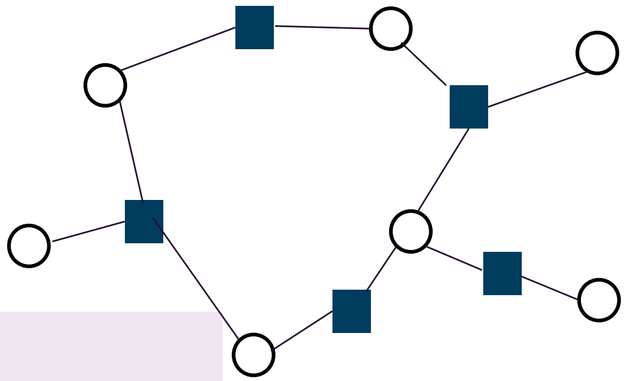
- **Attention: to form conscious state, thought**
- **A thought is a low-dimensional object**, few selected aspects of the unconscious state
- Need 2 high-level states:
  - Large unconscious state
  - Tiny conscious state

- Part of inference mechanism wrt joint distribution of high-level variables

CONSCIOUSNESS **PRIOR**
➔ **SPARSE FACTOR GRAPH**

Bengio 2017, arXiv:1709.08568



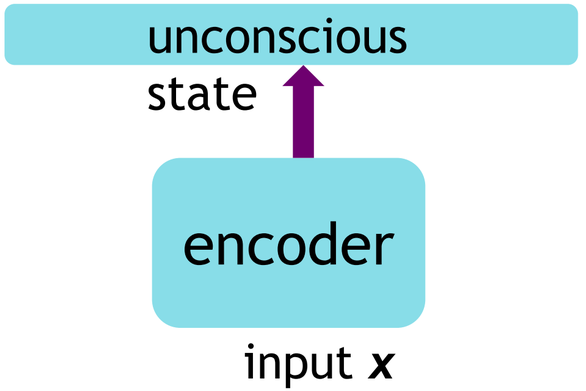


- Property of **high-level variables which we manipulate with language**:
  we can predict some given very few others
  - E.g. "if I drop the ball, it will fall on the ground”

Prior puts pressure on encoder

- **Disentangled factors** != marginally independent,
  e.g. ball & hand

- **Prior**: sparse factor graph joint distribution between high-level variables, consistent with inference mechanism which looks at just a few variables at a time.



WHAT **CAUSES** CHANGES IN DISTRIBUTION?

Underlying physics: actions are localized in space and time.

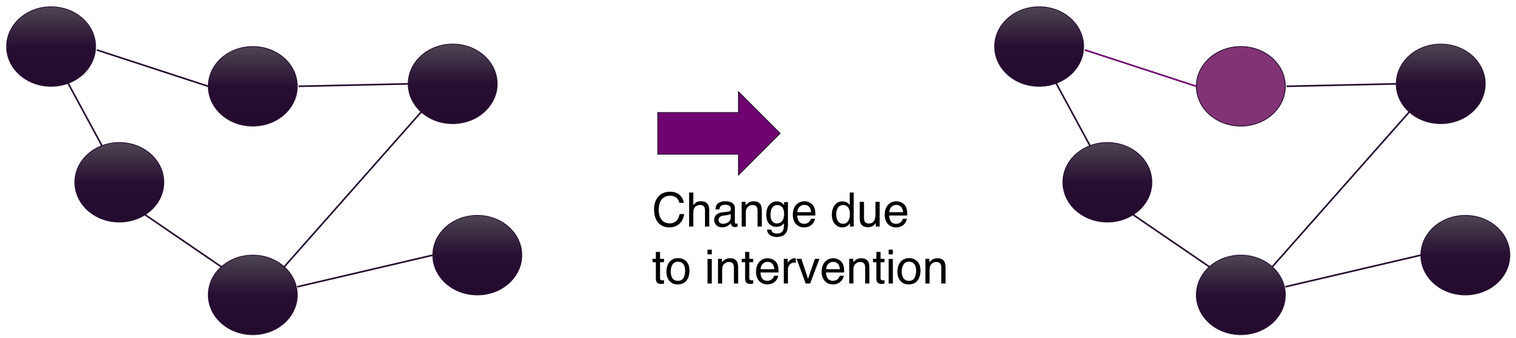
Hypothesis to replace iid assumption:

changes = consequence of an intervention on few causes or mechanisms (usually by an agent)

Extends the hypothesis of (informationally) Independent Mechanisms *(Scholkopf et al 2012)*

➔ local inference or adaptation in the right model

➔ good ood generalization/fast transfer/small ood sample complexity *(Bengio et al ICLR 2020)*

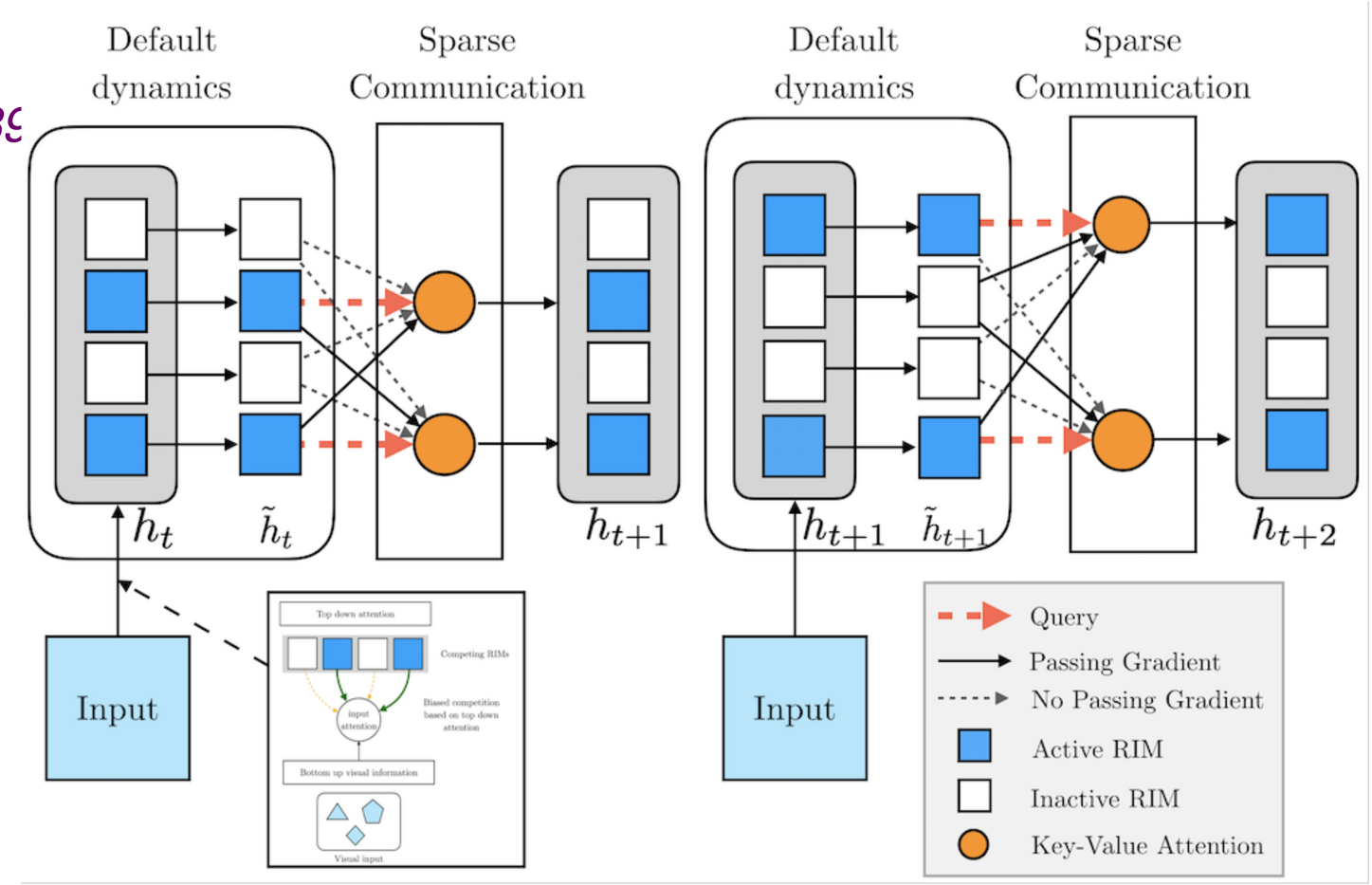


RIMS: MODULARIZE COMPUTATION AND OPERATE ON SETS OF NAMED AND TYPED OBJECTS

Recurrent Independent Mechanisms *Goyal et al 2019, arXiv:1909.108*

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

Results: better ood generalization



Builds on rich recent litterature on object-centric representations (mostly for images)

PRIORS FOR LEARNING HIGH-LEVEL SEMANTIC REPRESENTATIONS

- Consciousness prior: sparse factor graph
- Dependencies (rules/constraints) are shared (variables vs instances)
- HL variables tend to be causal
- HL variables tend to refer to agents, objects or actions
- Distributional changes arise from localized causal interventions (in semantic space)
- Different pieces of knowledge, w/ different stability/time scales

CONTRAST WITH **THE SYMBOLIC AI PROGRAM**

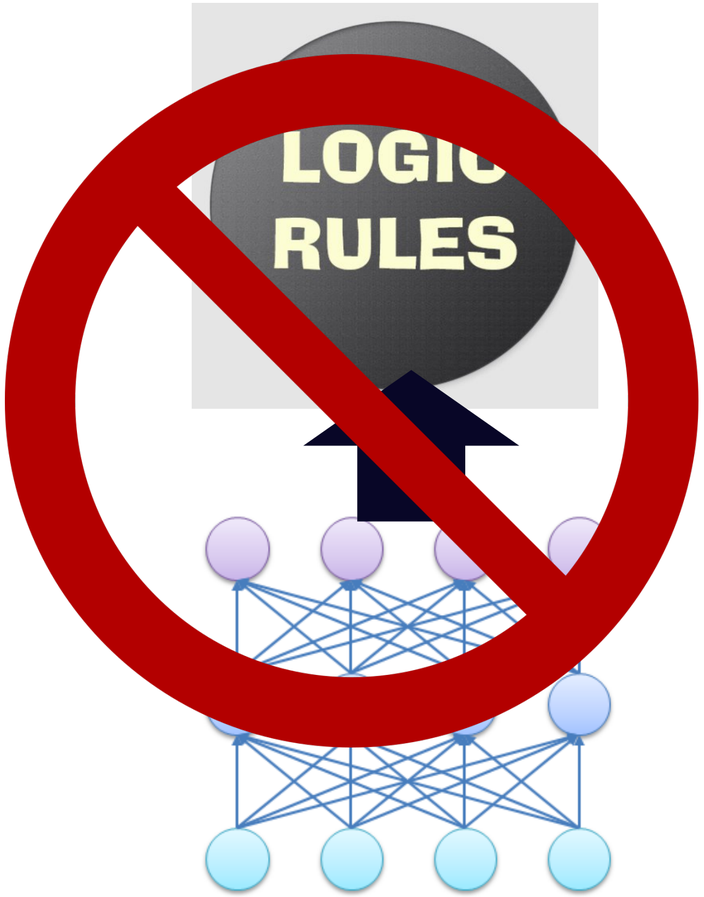
Avoid pitfalls of classical AI rule-based symbol-manipulation

- Need efficient & coordinated large-scale learning
- Need semantic grounding in system 1 and perception-action loop
- Need distributed representations for generalization
- Need efficient = trained search (also system 1)
- Need uncertainty handling

But want

- Systematic generalization
- Factorizing knowledge in small exchangeable pieces
- Manipulating variables, instances, references & indirection

MY BET: NOT A SIMPLE HYBRID OF GOFAI & DEEP NETS



Why? Many reasons:

(1) you need learning in the system 2 component as well as in the system 1 part,

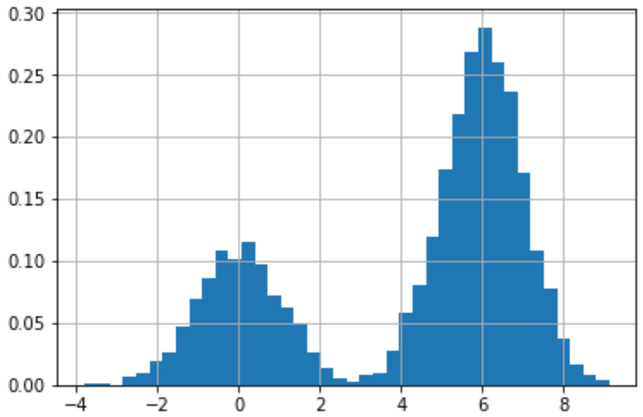
(2) high-level abstract concepts need to be grounded and have distributed representations to achieve generalization

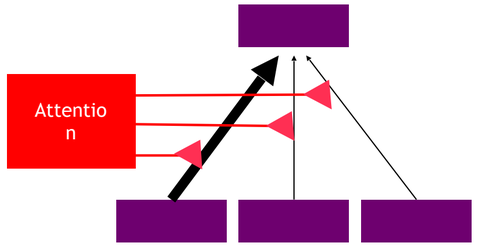
(3) you need to represent uncertainty there as well

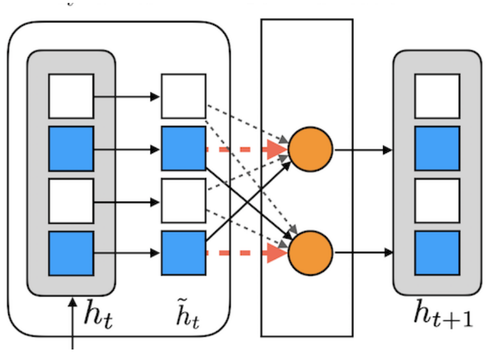
(4) brute-force search (the main inference tool of symbol-processing systems) does not scale, instead humans seem to use unconscious (system 1) processing to guide the search involved in reasoning, so system 1 and system 2 are very tightly integrated

(5) your brain is a neural net all the way

EXPLICIT OR IMPLICIT SYMBOLS?







My bet: DL implementing some of the functionalities of symbols

- Categories: multimodality of representation distribution, multiple manifolds, Gumbel softmax, inhibitory connections for competing attractors
- Indirection & variables: via attention & dynamic routing of information
- Recursion: via recurrent processing dynamically calling the same or different computational modules
- Context independence: only to some extent, using rich distributed representations of type and context, systematic generalization via dynamically activated combinations of mechanisms

Mila

Let's Debate!

Mila

Université de Montréal

McGill

Québec

CIFAR