

FROM SYSTEM 1 DEEP LEARNING TO SYSTEM 2 DEEP LEARNING YOSHUA BENGIO

NeurIPS'2019 Posner Lecture December 11th, 2019, Vancouver BC







THE STATE OF DEEP LEARNING

Amazing progress in this century

• Is it enough to just grow datasets, model sizes, computer speed?

Still far from human-level AI!

- Narrow AI
- Sample efficiency
- Human-provided labels
- Robustness, stupid errors
- Next step completely different from deep learning?

Just get a bigger brain?





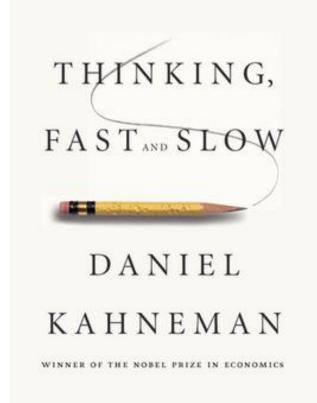
SYSTEM 1 VS. SYSTEM 2 COGNITION

2 systems (and categories of cognitive tasks):

System 1

- Intuitive, fast, UNCONSCIOUS, non-linguistic, habitual
- Current DL





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

System 2

- Slow, logical, sequential, **CONSCIOUS**, linguistic, algorithmic, planning, reasoning
- Future DL





MISSING TO EXTEND DEEP LEARNING TO REACH HUMAN-LEVEL AI

- Out-of-distribution generalization & transfer
- Higher-level cognition: system $1 \rightarrow \text{system 2}$
 - High-level semantic representations
 - Compositionality
 - Causality
- Agent perspective:
 - Better world models
 - Causality
 - Knowledge-seeking
- Connections between all 3 above!



CONSCIOUSNESS FUNCTIONALITIES: ROADMAP FOR PRIORS EMPOWERING SYSTEM 2

- 1. ML Goals: handle changes in distribution, necessary for agents
- 2. System 2 basics: attention & consciousness
- 3. Consciousness prior: sparse factor graph
- 4. Theoretical framework: meta-Learning, localized change hypothesis \rightarrow causal discovery
- 5. Compositional DL architectures: operating on sets of pointable objects with dynamically recombined modules

DEALING WITH CHANGES IN DISTRIBUTION

FROM IID TO OOD

Classical ML theory for iid data

Artificially shuffle the data to achieve that?

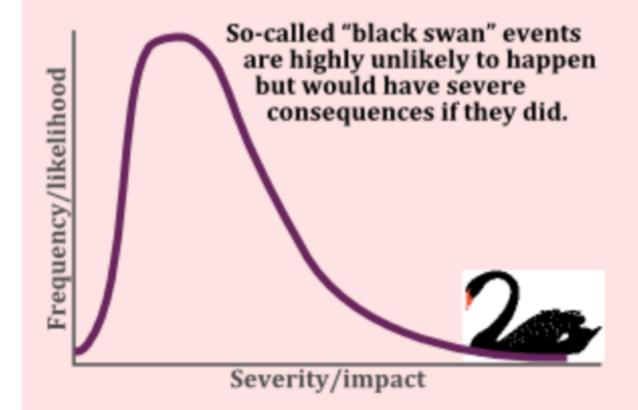
L. Bottou ICML 2019: Nature does not shuffle the data, we shouldn't

Out-of-distribution generalization & transfer

No free lunch: need new assumptions to replace iid assumption, for ood generalization



The Black Swan



Source: Chris Mandel, Sedgwick Inc.

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



COMPOSITIONALITY HELPS IID AND OOD GENERALIZATION

Different forms of compositionality each with different exponential advantages

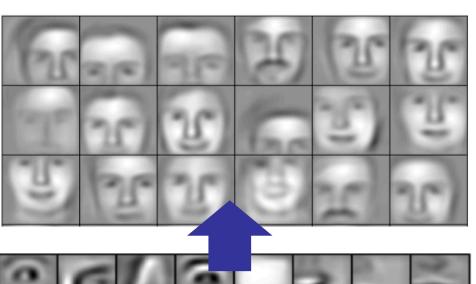
• Distributed representations

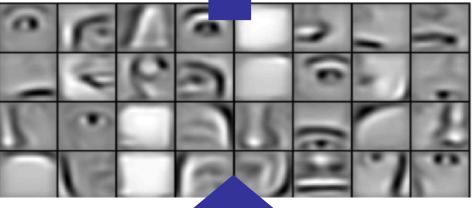
(Pascanu et al ICLR 2014)

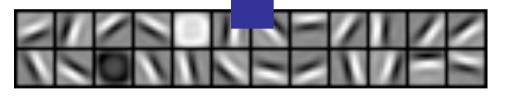
Composition of layers in deep nets

(Montufar et al NeurIPS 2014)

 Systematic generalization in language, analogies, abstract reasoning? TBD







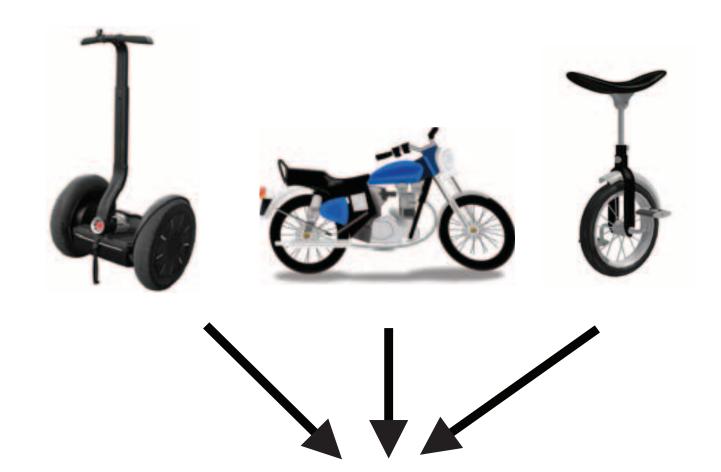
(Lee, Grosse, Ranganath & Ng, ICML 2009)



SYSTEMATIC GENERALIZATION

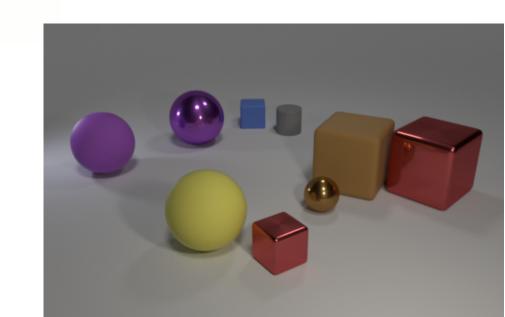
- Studied in linguistics
- Dynamically recombine existing concepts
- Even when new combinations have 0 probability under training distribution
 - E.g. Science fiction scenarios
 - E.g. Driving in an unknown city
- Not very successful with current DL

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



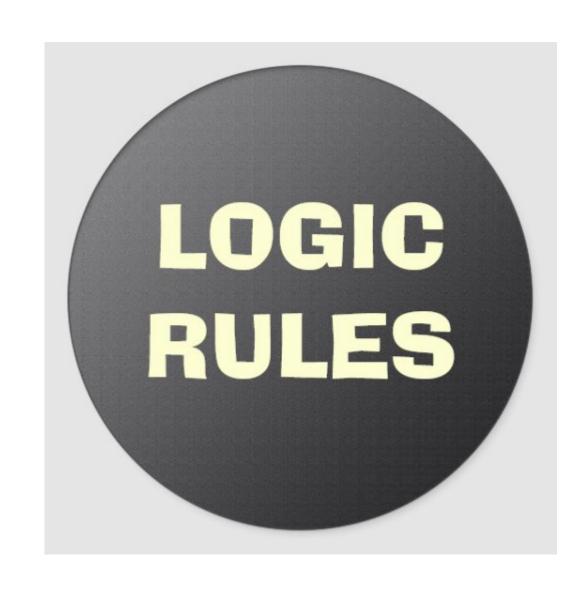








CONTRAST WITH THE SYMBOLIC AI PROGRAM



Avoid pitfalls of classical AI rule-based symbol-manipulation

- Need efficient large-scale learning
- Need semantic grounding in system 1
- 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



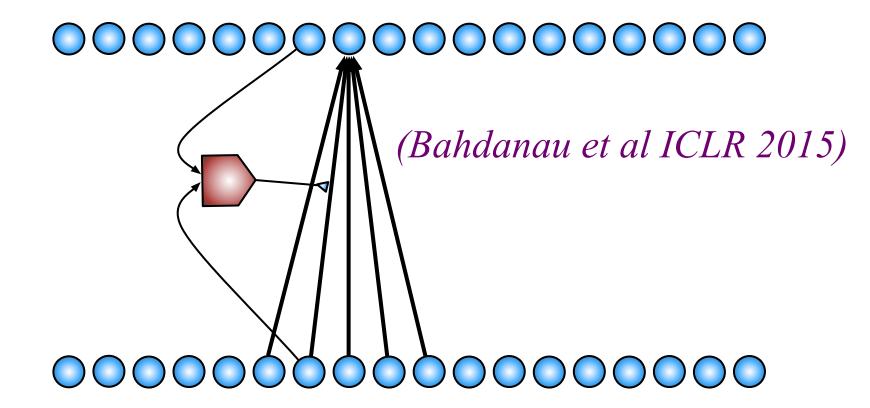
SYSTEM 2 BASICS: ATTENTION AND CONSCIOUSNESS

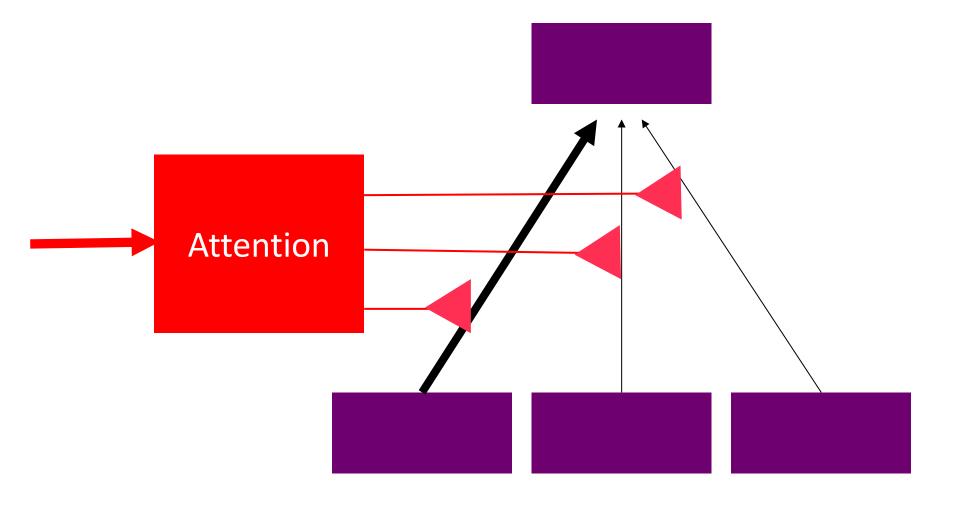
CORE INGREDIENT FOR CONSCIOUSNESS:

ATTENTION

• Focus on a one or a few elements at a time

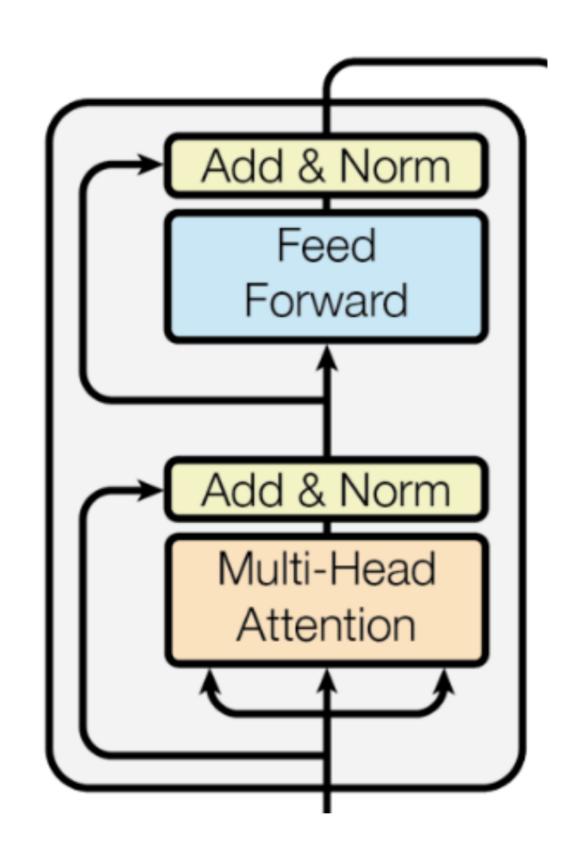
- Content-based soft attention is convenient, can backprop to *learn where to attend*
- Attention is an internal action, needs a learned attention policy (Egger et al 2019)







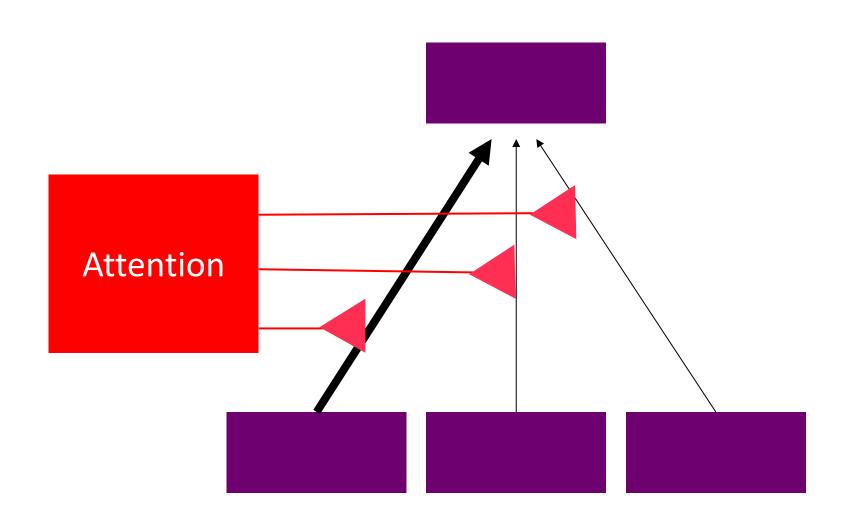
ATTENTION BENEFITS



- Neural Machine Translation revolution (Bahdanau et al ICLR 2015)
- SOTA in NLP (self-attention, transformers)
- Memory-extended neural nets
- Address vanishing gradients (Ke & al NeurIPS 2018)
- Operating on unordered SETS of (key, value) pairs



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)



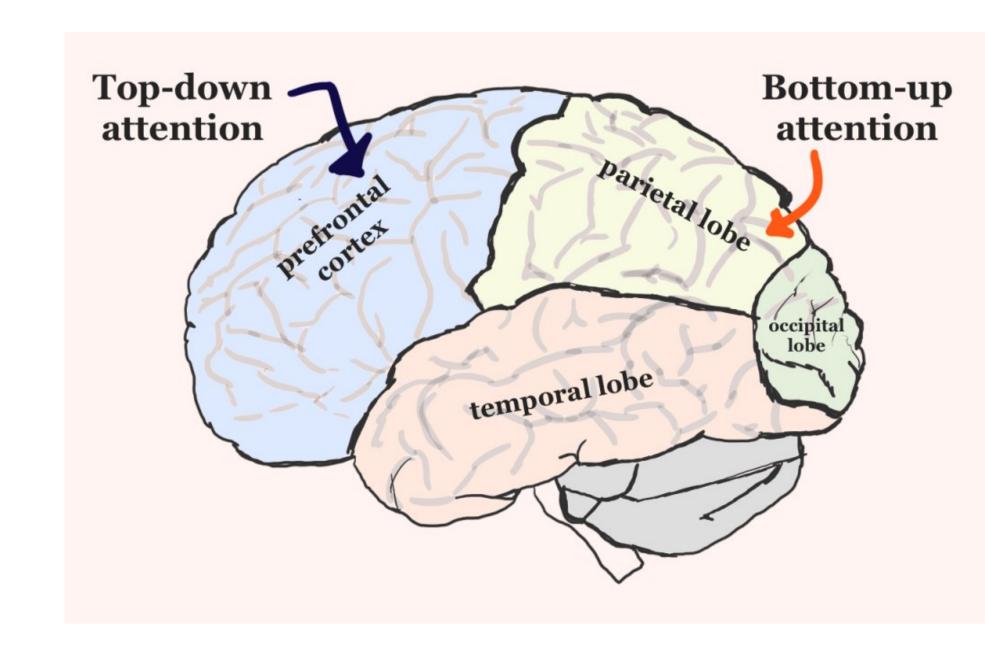
FROM ATTENTION TO CONSCIOUSNESS

C-word not taboo anymore in cognitive neuroscience

Global Workspace Theory

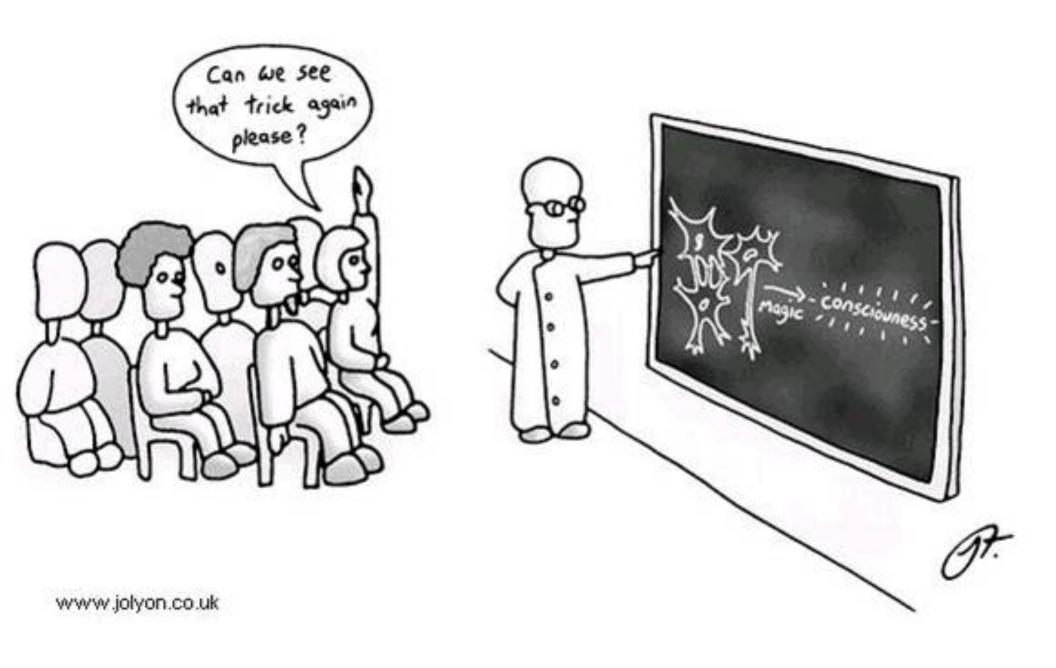
(Baars 1988++, Dehaene 2003++)

- Bottleneck of conscious processing
- Selected item is broadcast, stored in short-term memory, conditions perception and action
- System 2-like sequential processing, conscious reasoning & planning & imagination





ML FOR CONSCIOUSNESS & CONSCIOUSNESS FOR ML



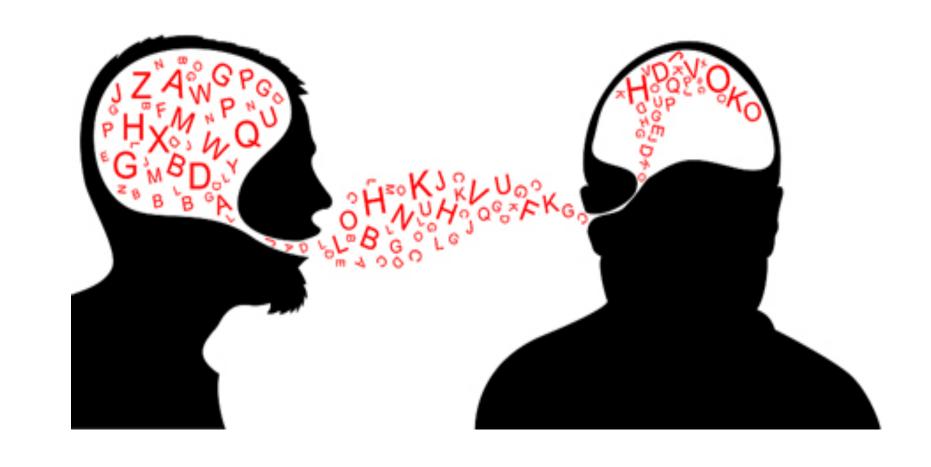
- Formalize and test specific hypothesized functionalities of consciousness
- Get the magic out of consciousness
- Understand evolutionary advantage of consciousness: computational and statistical (e.g. systematic generalization)
- Provide these advantages to learning agents



THOUGHTS, CONSCIOUSNESS, LANGUAGE

- Consciousness: from humans reporting
- High-level representations \iff language
- High-level concepts: meaning anchored in low-level perception and action → tie system 1 & 2
- Grounded high-level concepts
 - → better natural language understanding
 - Grounded language learning

e.g. BabyAI: (Chevalier-Boisvert and al ICLR 2019)

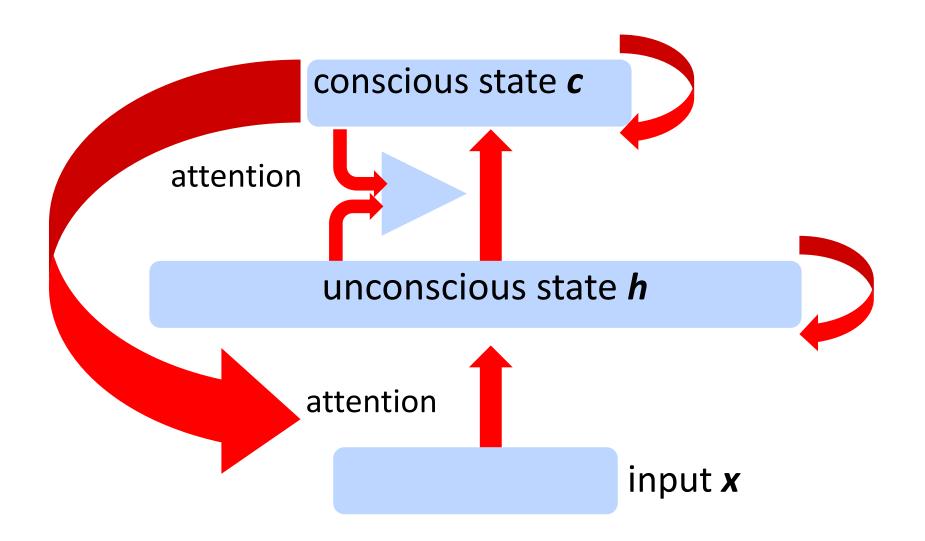




THE CONSCIOUSNESS PRIOR: SPARSE FACTOR GRAPH

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

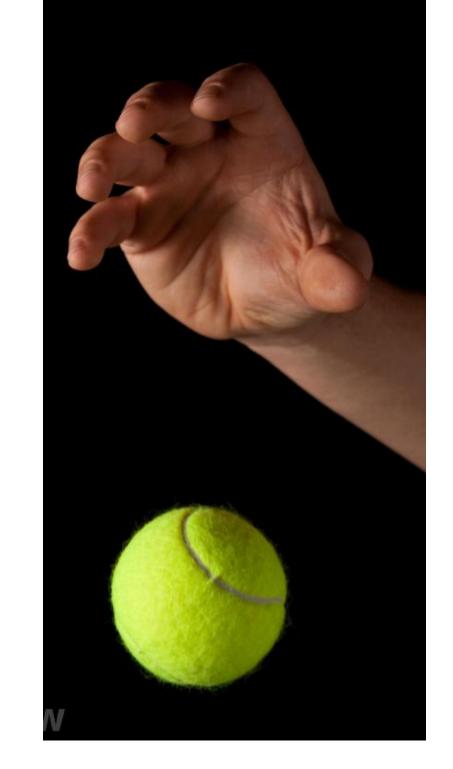
→ SPARSE FACTOR GRAPH

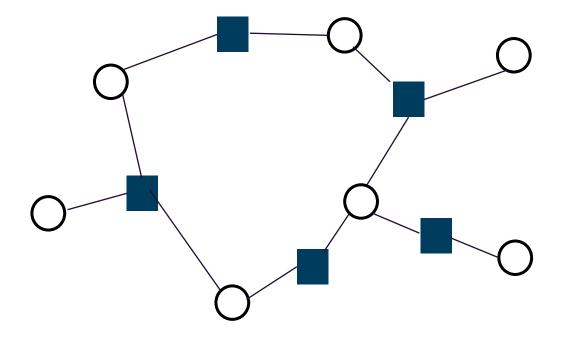
Bengio 2017, arXiv:1709.08568

 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"
- **Disentangled factors**!= marginally independent, e.g. ball & hand
- **Prior**: sparse factor graph joint distribution between high-level variables





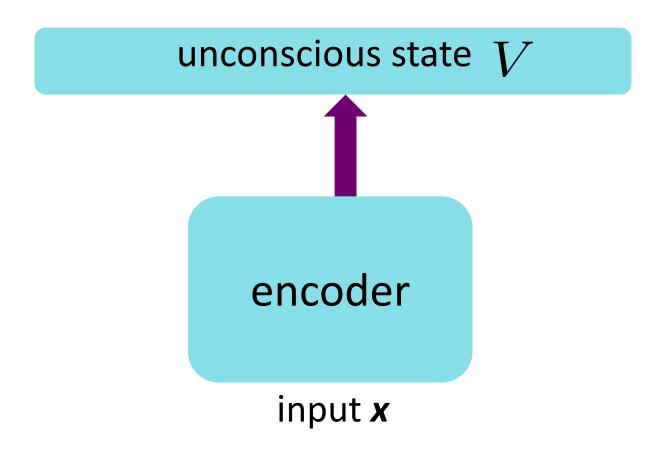


CONSCIOUSNESS PRIOR -> SPARSE FACTOR GRAPH

$$P(V) \propto \prod_{k} \phi_k(V_{s_k})$$

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

Prior puts pressure on encoder computing implicitly P(V|observations x)

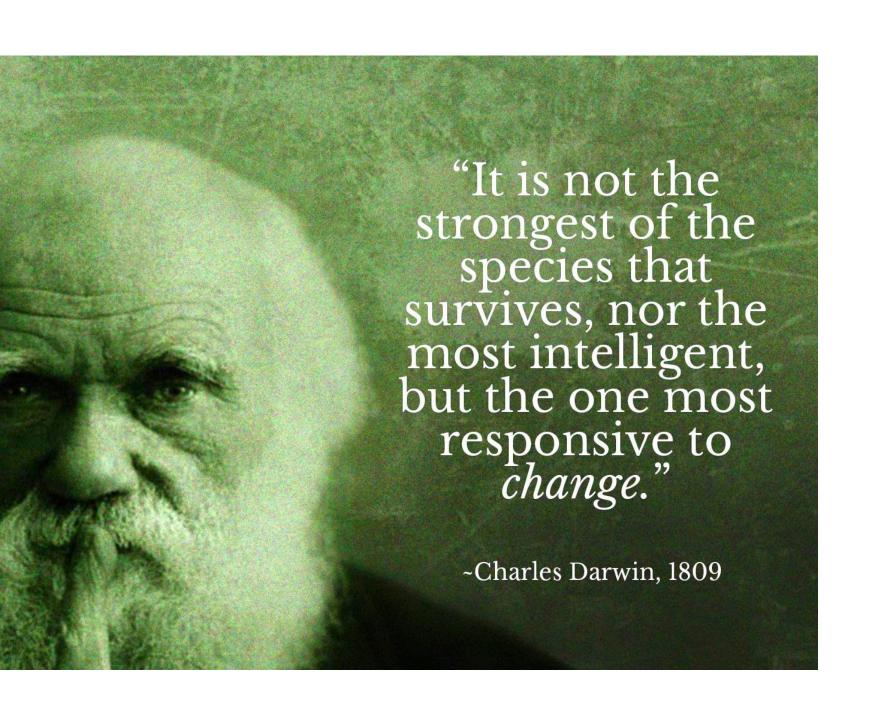


Bengio 2017, arXiv:1709.08568



META-LEARNING: END-TO-END OOD GENERALIZATION, LOCALIZED CHANGE HYPOTHESIS

META-LEARNING FOR TRAINING TOWARDS OOD GENERALIZATION



- Meta-learning or learning to learn
 - (Bengio et al 1991; Schmidhuber 1992)
 - Backprop through inner loop or REINFORCE-like estimators
- Bi-level optimization
 - Inner loop (may optimize something) → outer loss
 - Outer loop: optimizes E[outer loss] (over tasks, environments)
- E.g.
 - Evolution individual learning
 - Lifetime learning fast adaptation to new environments
- Multiple time-scales of learning
- End-to-end learning to generalize ood + fast transfer



WHAT CAUSES CHANGES IN DISTRIBUTION?

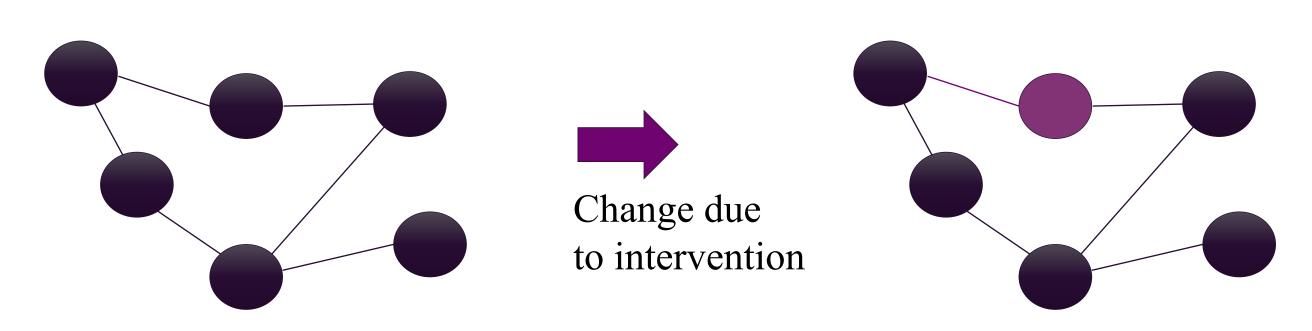
Hypothesis to replace iid assumption:

Underlying physics: actions are localized in space and time.

changes = consequence of an intervention on few causes or mechanisms

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

→ local inference or adaptation in the right model



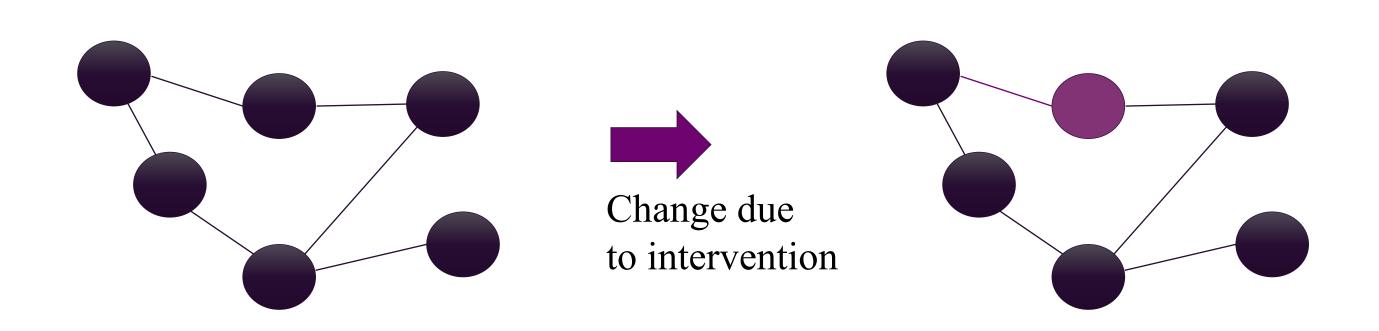




COUNTING ARGUMENT: LOCALIZED CHANGE→OOD TRANSFER

Good representation of variables and mechanisms + localized change hypothesis

- → few bits need to be accounted for (by inference or adaptation)
- → few observations (of modified distribution) are required
- → good ood generalization/fast transfer/small ood sample complexity







META-LEARNING KNOWLEDGE REPRESENTATION FOR GOOD OOD PERFORMANCE

- Use ood generalization as training objective
- Good decomposition / knowledge representation \rightarrow good ood performance
- Good ood performance = training signal for factorizing knowledge



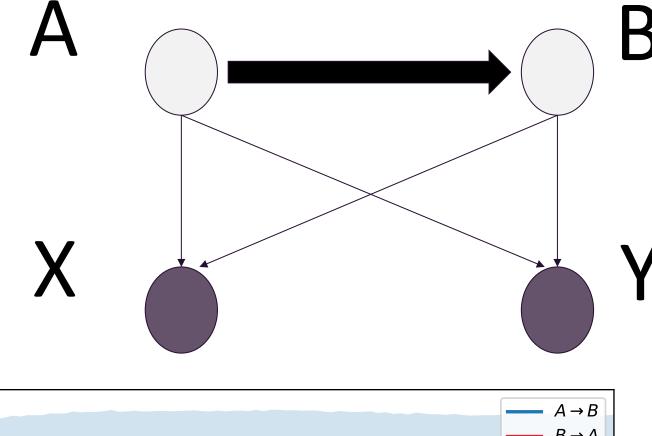


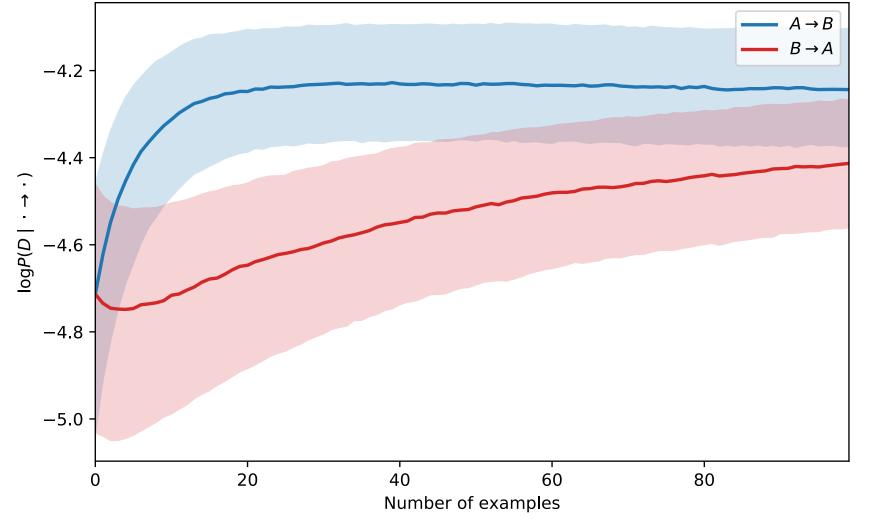
EXAMPLE: DISCOVERING CAUSE AND EFFECT = HOW TO FACTORIZE A JOINT DISTRIBUTION?

A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms

- Learning whether A causes B or vice-versa
- Learning to disentangle (A,B) from observed (X,Y)
- Exploit changes in distribution and speed of adaptation to guess causal direction

Bengio et al 2019 arXiv:1901.10912







EXAMPLE: DISCOVERING CAUSE AND EFFECT = HOW TO FACTORIZE A JOINT DISTRIBUTION?

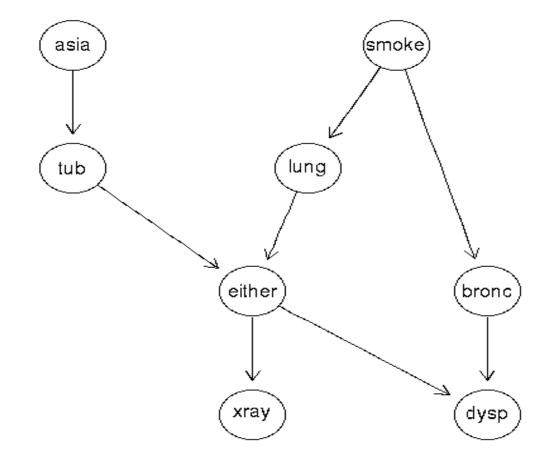
Learning Neural Causal Models from Unknown Interventions *Ke et al 2019 arXiv:1910.01075*

• Learning small causal graphs, avoid exponential explosion of # of graphs by parametrizing factorized distribution over graphs

• Inference over the intervention: faster causal discovery

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





OPERATING ON SETS OF POINTABLE OBJECTS WITH DYNAMICALLY RECOMBINED MODULES



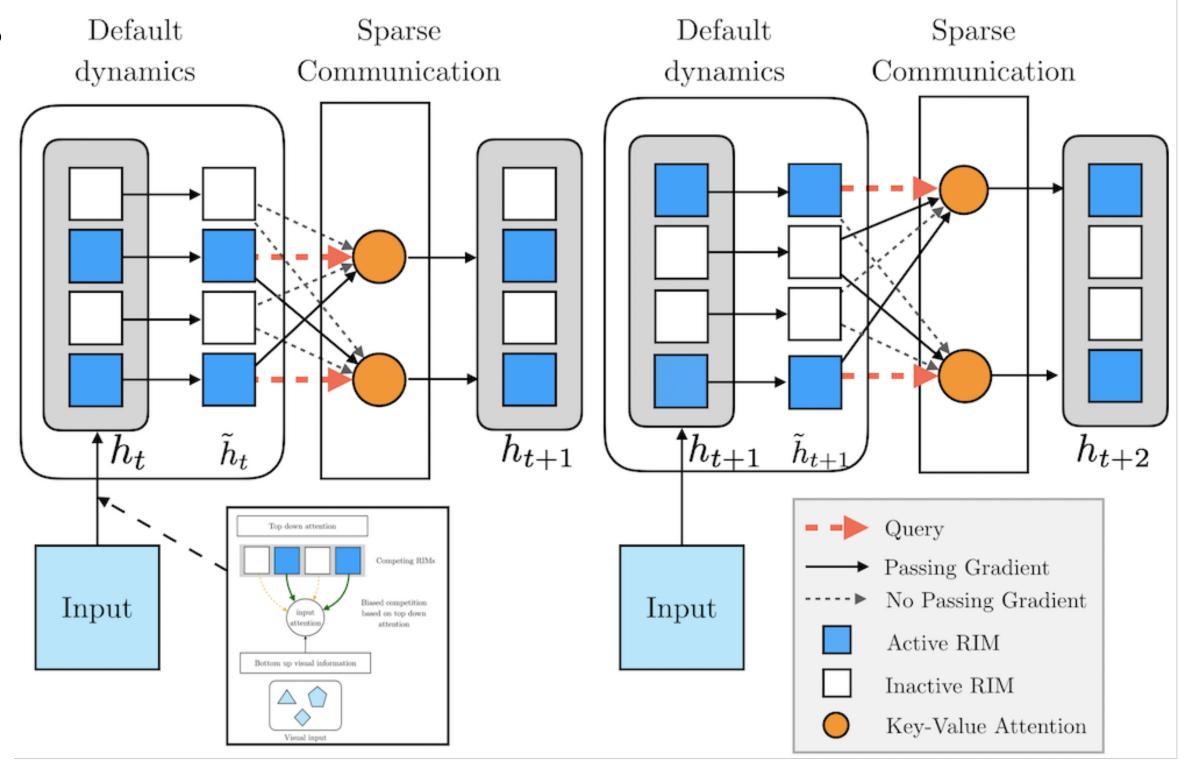
RIMS: MODULARIZE COMPUTATION AND OPERATE ON SETS OF NAMED AND TYPED OBJECTS

Recurrent Independent Mechanisms

Goyal et al 2019, 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

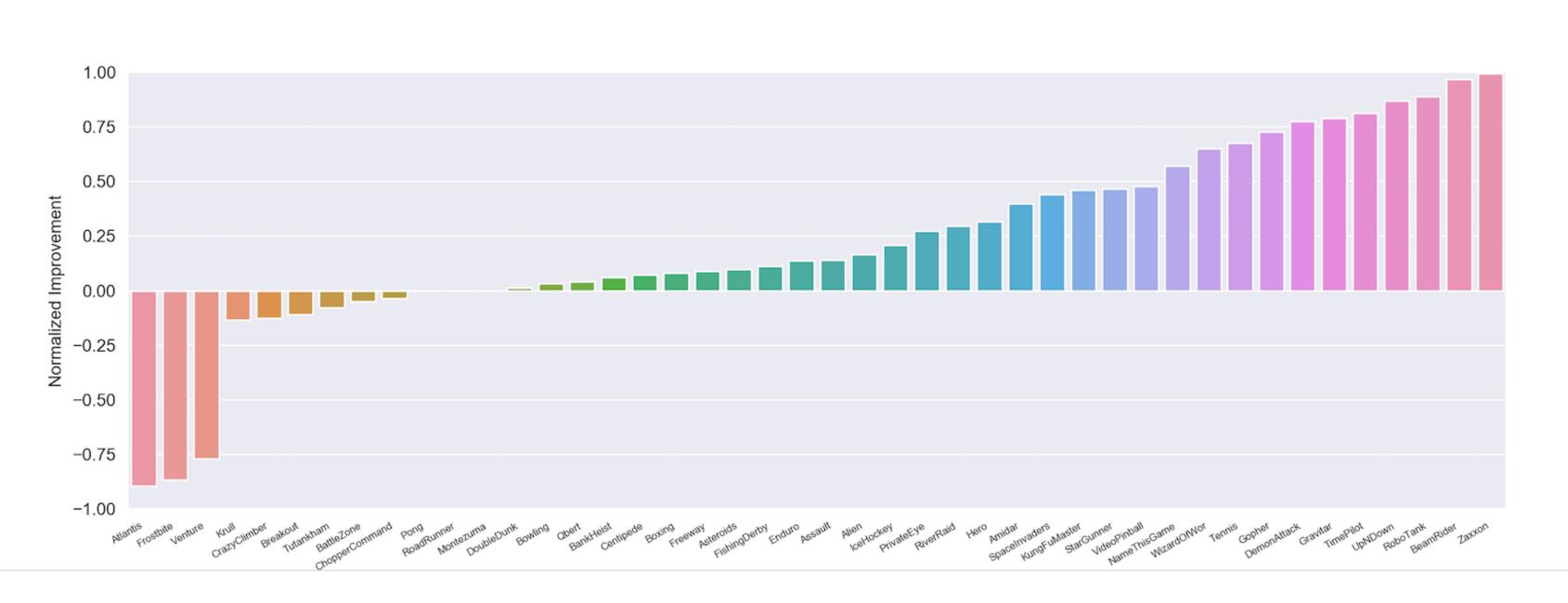
Results: better ood generalization





RESULTS WITH RECURRENT INDEPENDENT MECHANISMS

- RIMs drop-in replacement for LSTMs in PPO baseline over all Atari games.
- Above 0 (horizontal axis) = improvement over LSTM.





HYPOTHESES FOR CONSCIOUS PROCESSING BY AGENTS, SYSTEMATIC GENERALIZATION

- Sparse factor graph in space of high-level semantic variables
- Semantic variables are causal: agents, intentions, controllable objects
- Shared 'rules' across instance tuples (arguments)
- Distributional changes from localized causal interventions (in semantic space)
- Meaning (e.g. grounded by an encoder) stable & robust wrt changes in distribution



CONCLUSIONS

- After cog. neuroscience, time is ripe for ML to explore consciousness
- Could bring new priors to help systematic & ood generalization
- Could benefit cognitive neuroscience too
- Would allow to expand DL from system 1 to system 2



System 1

***Mila



System 2





