Yoshua Bengio
(DL from system 1 to system 2)

23 June 2020, 10:00 NY

- Thinking fast & slow – Daniel Kahneman

System 1	System 2
- Intuitive, fast, habitual unconscious, non-linguistic	- Slow, logical, sequential, algorithmic conscious, planning, reasoning
- Implicit knowledge	- Explicit knowledge
- Current DL	- DL 2.0

- Current DL
  - Mostly supervised
  - Adversarial Attacks

IMPLICIT VS VERBALIZABLE KNOWLEDGE: UNDERLYING ASSUMPTIONS BEHIND VERBALIZABLE KNOWLEDGE

- Most knowledge in our brain is implicit and **not verbalizable** (hence the explainability challenge, even for humans)
- Some of our knowledge however is verbalizable and we can reason and plan explicitly with it
- The concepts manipulated in this way are those we can name with language
- The joint distribution between these concepts and the way that distribution can change over time satisfies special assumptions, exploited in system 2 tasks and conscious processing
- We want to clarify these assumptions as priors to be able to embed them in ML architectures and training frameworks → ie. we need priors which belong in system 2 category for DL 2.0

Mila

INDUCTIVE PRIORS WHICH COULD GO IN DEEP LEARNING 2.0

- *Sparse factor graph in space of high-level semantic variables*
- *Semantic variables are causal: agents, intentions, controllable objects*
- Simple mapping between high-level semantic variables / thoughts and words / sentences
- Shared 'rules' across instance tuples (as arguments), requiring variables & indirection
- *Distributional changes due to localized causal interventions (in semantic space)*
- Meaning (e.g. grounded by an encoder) is stable & robust wrt changes in distribution
- Credit assignment is only over short causal chains

Mila

AGENT LEARNING NEEDS OOD GENERALIZATION

Agents face non-stationarities

Changes in distribution due to

- their actions
- ***ESPECIALLY:***
  actions of other agents
- different places, times, sensors, actuators, goals, policies, etc.
- ***What is the "objective"?***

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

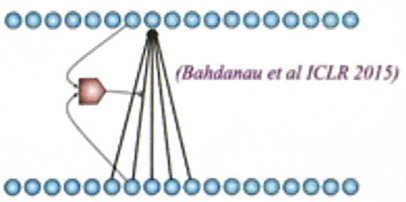
Mila

→ Not easy / possible with DL right now

- Current DL doesn't work well on tasks which involves recombining concepts we already know but has zero probability under training distribution

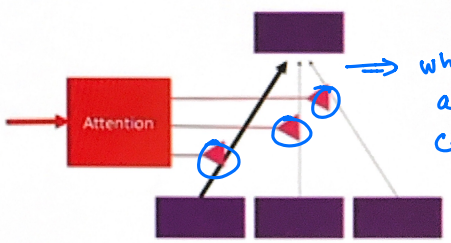
System 2 Basics : Attention & Conscious Processing

CORE INGREDIENT FOR CONSCIOUS PROCESSING: ATTENTION



(Bahdanau et al ICLR 2015)

- **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)*
- Operating on unordered SETS of (key, value) pairs
- SOTA in NLP

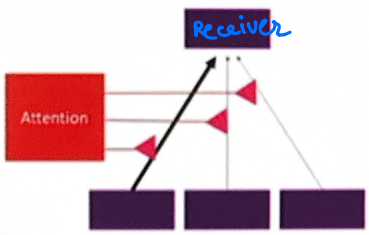


Mila

↓
Attention is like working on sets

⇒ when we have attention, we can know from where the input is coming

FROM ATTENTION TO **INDIRECTION**



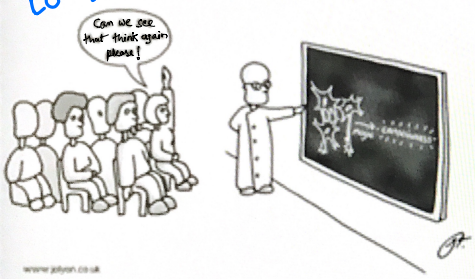
- Attention = dynamic connection
- Receiver gets the selected value
- Value of what? From where?
  → Also send 'name' (or key) of sender
- Keep track of 'named' objects: indirection
- Manipulate sets of objects (transformers)

P.S. contrary to convnets doing object recognition, sequential tasks involving memory and attention typically involve a more difficult optimization problem, and fighting underfitting (including the issue of long-term dependencies)

Mila

ML FOR CONSCIOUSNESS & CONSCIOUSNESS FOR ML

lol!



- 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

Mila

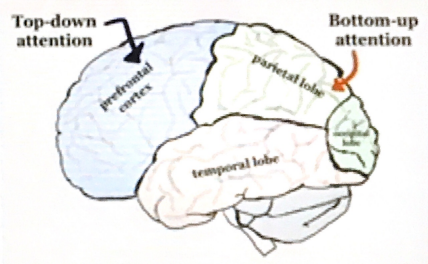
FROM ATTENTION TO **CONSCIOUSNESS**

C-word not taboo anymore in cognitive neuroscience

Global Workspace Theory

(Baars 1988++, Dehaene 2003++)

- Bottleneck of conscious processing
  - *WHY A BOTTLENECK?*
- Selected item is broadcast, stored in short-term memory, conditions perception and action
- System 2-like sequential processing, conscious reasoning & planning & imagination
- Can only run 1 simulation at a time, unlike a movie, only few abstract concepts involved at each step



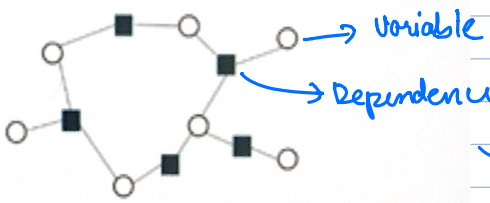
Mila

- Grounded language learning : BabyAI (Chevalier-Boisvert et al, ICLR 2019)

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
- Inference involves few variables at a time, selected by **attention mechanism** and memory retrieval

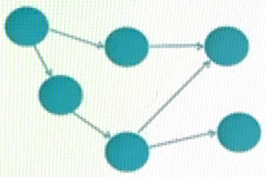


Mila

Deep Learning Objective: discover causal representation

→ THINK!

- What are the right representations? Causal variables explaining the data
- How to discover them?
- How to discover their causal relationship, the causal graph?



Mila

(Read Granger Causality told by Avinash Kori)