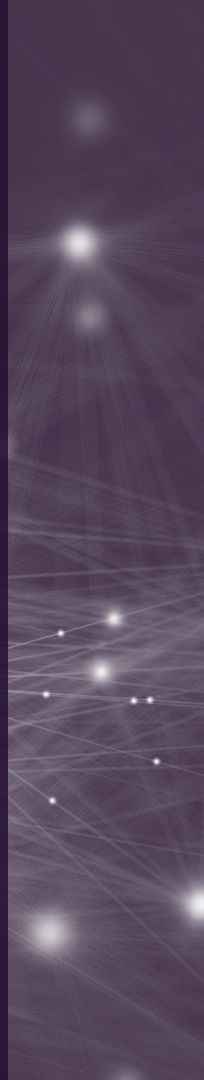
Mila

Inspiration from Brains for Deep Learning and Inspiration from Deep Learning for Brains

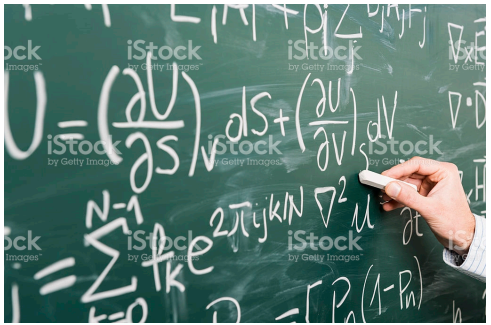
Yoshua Bengio

GRSNC NEURODISCUSSIONS
2 AVRIL 2019



Underlying Assumption

- There are principles giving rise to intelligence (machine, human or animal) via learning, simple enough that they can be described compactly, similarly to the laws of physics, i.e., our intelligence is not just the result of a huge bag of tricks and pieces of knowledge, but of general mechanisms to acquire knowledge.



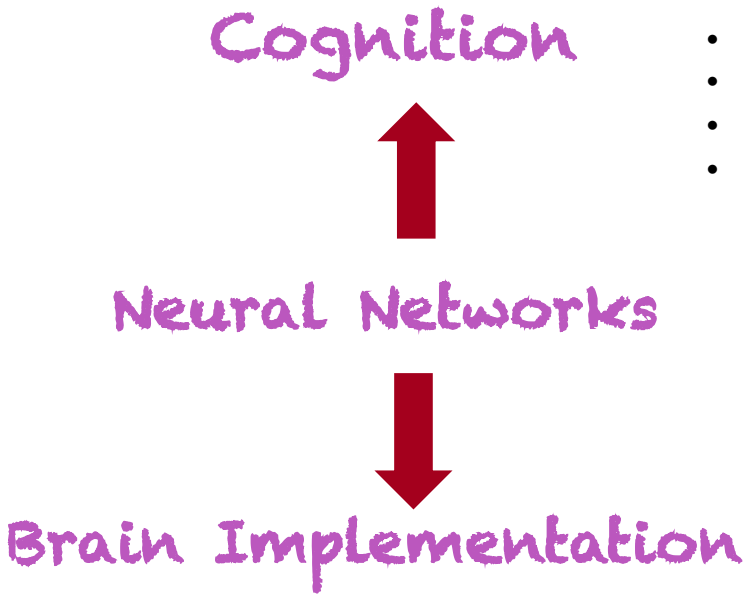
The Learning Mechanism is a Compact and Abstract Explanation of the Brain

Similar to the laws of physics: e.g. we consider **understanding** the physical world, mostly by having figured out the laws of physics, not just by describing its consequences (the immense complexity of describing the physical world)

Successful learning framework (e.g. architecture, optimizer, objective) is a compact abstract explanation, much more so than the actual detailed neuron-by-neuron functions performed by a trained brain

ML validation: can learn complex tasks

Neuroscience validation: matches biology at some level



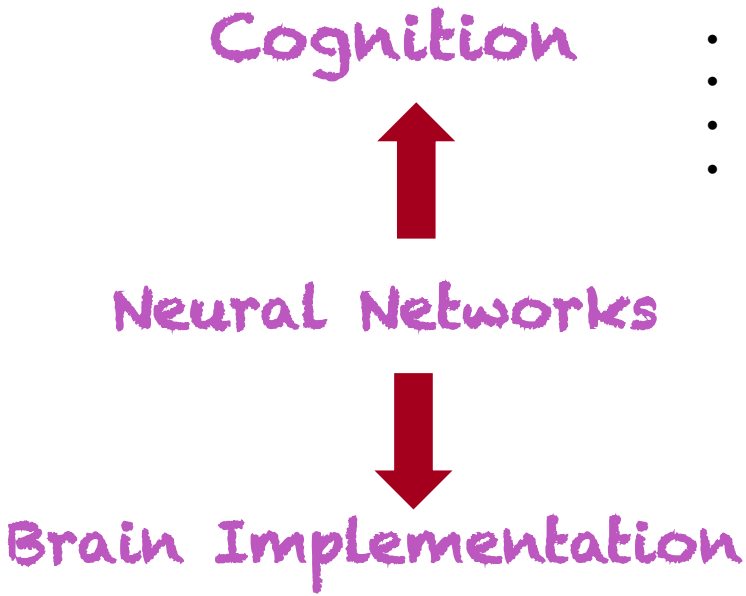
- Attention: vectors → data structures
- memory access, one-shot memorization
- reasoning & planning
- semantics & language
- agency & causality
- consciousness
- teaching & curriculum learning

- neural computation
- biological backprop
- dropout & spikes
- multi-module architecture

Brain Intelligence Inspiration for Deep Learning

Drawing inspiration for AI from living intelligence

- Neurons, networks, plasticity & learning
- Distributed representations
- Visual cortex, convnets & depth
- Neural nonlinearity & ReLUs
- Spikes: dropout & quantized activations
- Curriculum learning
- Cultural evolution & distributed training
- Affordances, options, exploration & controllable factors
- Attention
- Lateral connections, softmax, clustering & attractors
- Associative memories, hippocampus & episodic memory
- System 2, reasoning, planning & consciousness



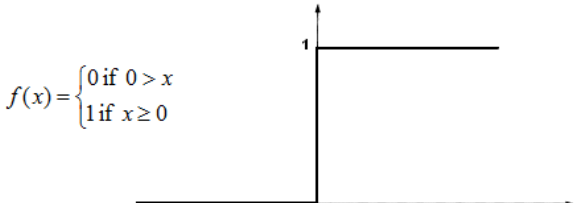
Attention:

- vectors → data structures
- memory access, one-shot memorization
- reasoning
- semantics & language
- agency & causality
- consciousness

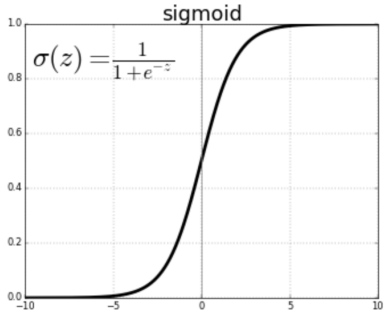
- biological backprop
- dropout & spikes
- multi-module architecture

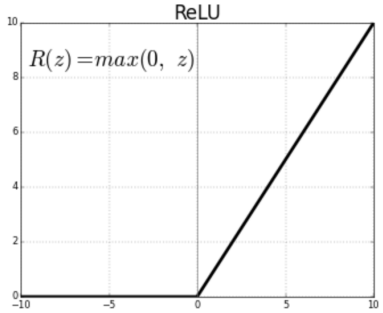
Neural Nonlinearity & ReLU

- First approximation: linear threshold units



- Second approximation: sigmoids

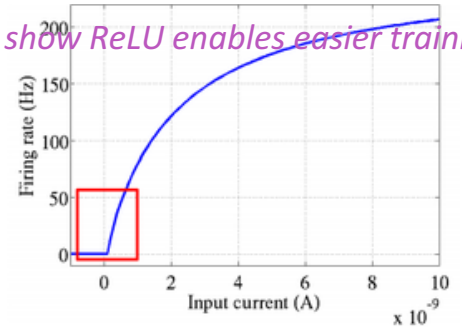




- Third approximations: piecewise-linear rectifiers (ReLU)

(Glorot & Bengio AISTATS 2011) show ReLU enables easier training of deep nets

- Still some way to go to approximate biological nonlinearity…

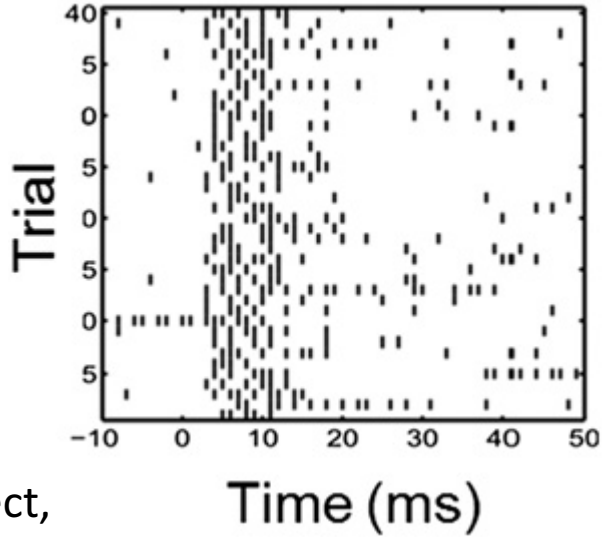


Spikes, Dropout & Quantization

- Real neurons send low-precision pseudo-noisy signals: spikes

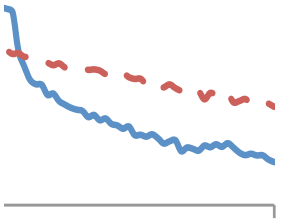
- Inspiration for dropout & other noise injection regularizers (Hinton et al 2012)

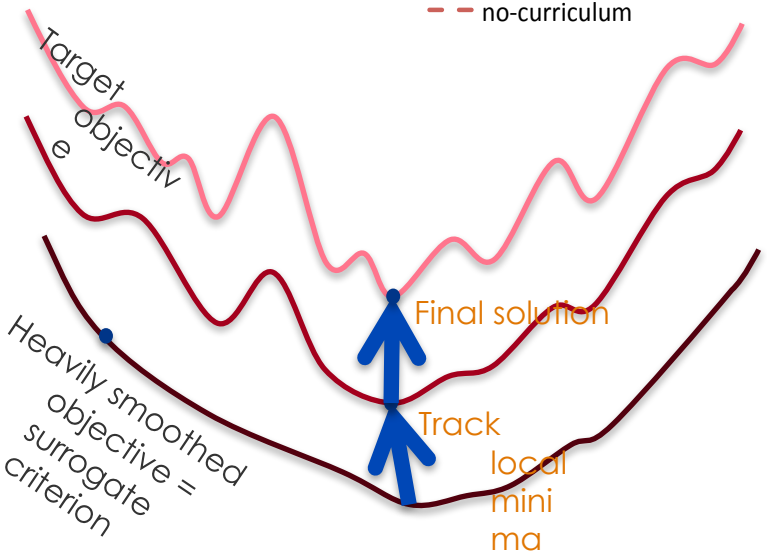
- Inspiration for low-precision (stochastically) quantized activations (Courbariaux & Bengio, Binary Connect, NIPS 2015)



Curriculum Learning

- Start from easier tasks and gradually build hierarchy of competences
- Shown to help training of deep nets (Bengio et al ICML 2009), acting as a form of continuation method





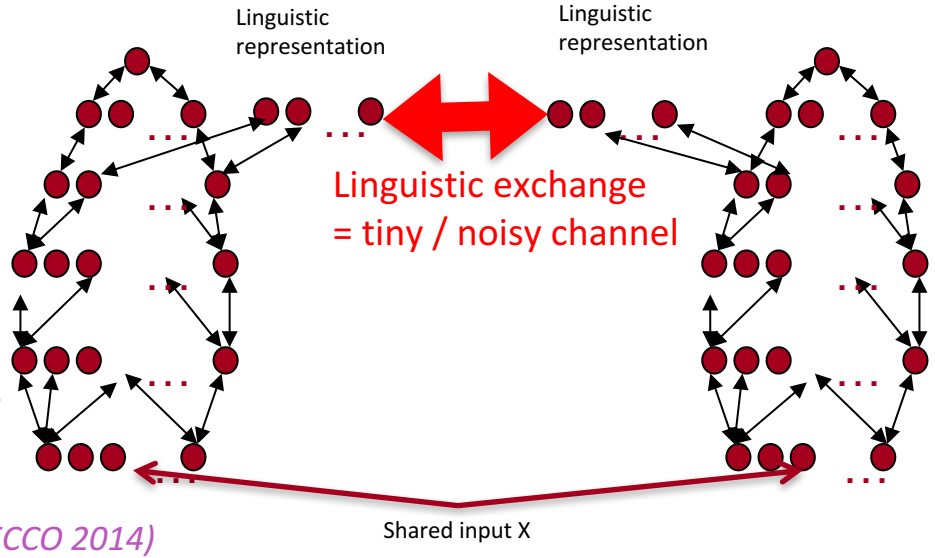
Cultural Evolution & Memes

- Memes: transmittable & evolving nuggets of cultural information
- Genetic & memetic evolution:
  - more powerful than random search
  - exponential advantage: combining sub-solutions
- Memes: more efficient than genes
  - more appropriate level of abstraction

(Bengio, GECCO 2014, Deep learning & cultural evolution)

How is one brain transferring abstractions to another brain?

Two individuals sharing a similar visual input, the teacher gives hints to the student about high-level abstractions

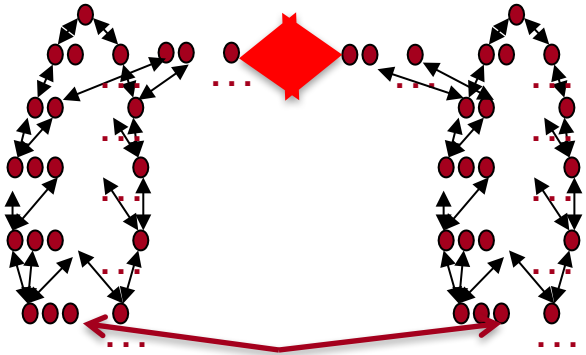


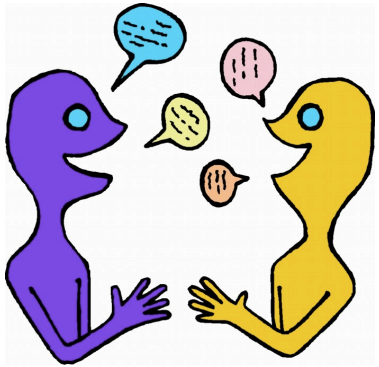
The linguistic output of one individual is modeled by the other one, jointly with X.

(Bengio, GECCO 2014)

Cultural Evolution, Distillation & Distributed Training

- P-A Manzagol & D Erhan worked on this with me in 2009 (unpublished)
- Instead of sharing weights, different networks can share activity for a shared input *(Bengio, GECCO 2014)*
  - Fitnets *(Romero et al & Bengio, ICLR 2015)*
  - Distillation *(Hinton, Vinyals & Dean 2015)*
  - Co-distillation *(Anil et al & Hinton ICLR 2018)*

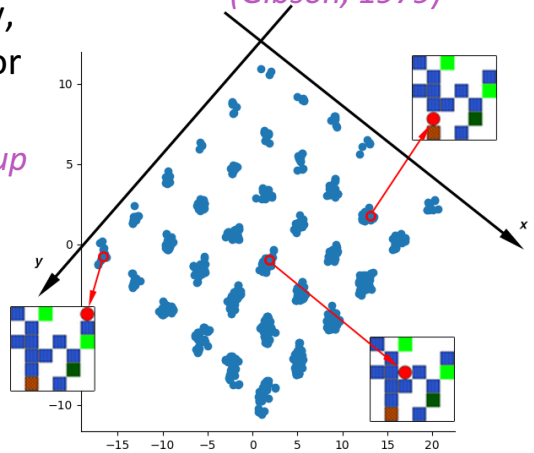




Affordances, options, exploration & controllable factors

- Affordances: concepts / aspects of the environment which can be changed by the agent
- Temporal abstractions: options, super-actions, macros or procedures, which can be composed to form more complex procedures *(Sutton, Precup & Singh 1999)*
- Controllable factors: jointly learn a set of (policy, factor) such that the policy can control the factor and maximize mutual information between policies and factors *(Bengio, Thomas, Pineau, Precup & Bengio 2017)*
- Intrinsic & exploration rewards: unsupervised aspect of reinforcement learning

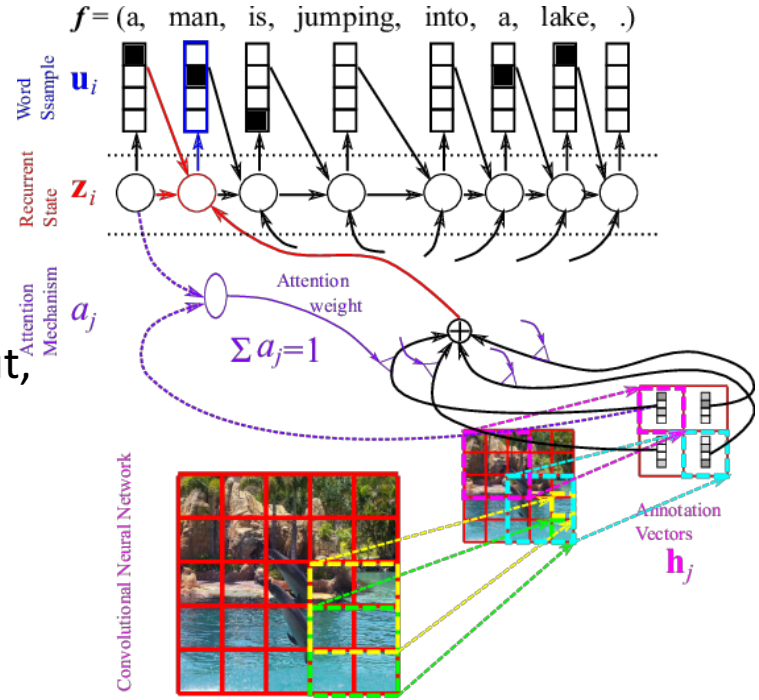
(Gibson, 1979)

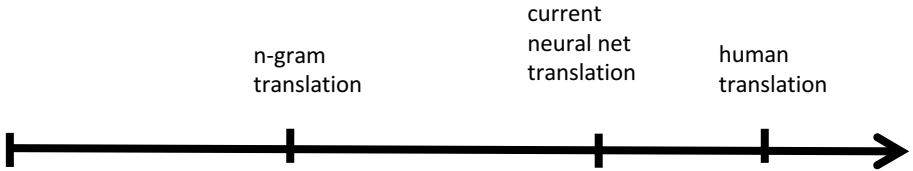


Attention!

(Bahdanau et al & Bengio 2014)

- Iterative computation
- Each step focuses on a few elements out of a larger set
- Attention can be on raw input, memory elements, or representations
- Content-based attention: condition on all available information
- Soft attention: backprop-trainable attention mechanism

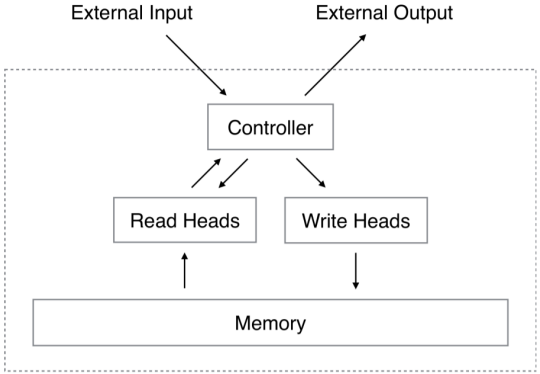




Associative Memories, Hippocampus & Episodic Memory

- Auto-encoders: encode-decode cycle implements one form of associative memory; can be iterated to converge to a manifold (or a set of manifolds, corresponding to different memories, as in Hopfield networks)
- Hippocampus stores episodic memories, has been an inspiration for memory augmented neural networks

Neural Turing Machines
Graves et al 2014



Reminding and Credit Assignment

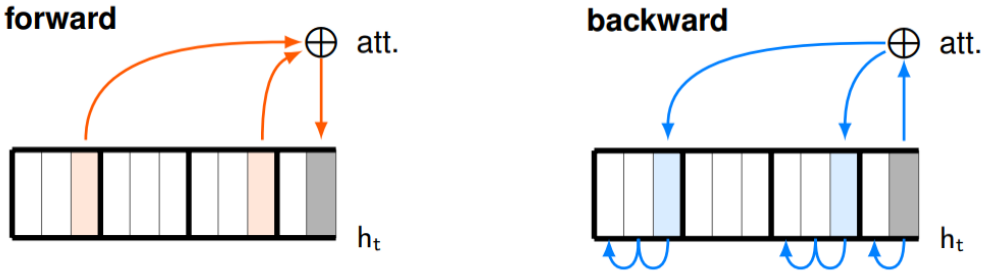
- Humans selectively recall memories that are relevant to the current behavior.
- Automatic reminding:
  - Triggered by contextual features.
  - Can serve a useful computational role in ongoing cognition.
  - Can be used for credit assignment to past events?
- Assign credit through only a few states, instead of all states:
  - Sparse, local credit assignment.
  - How to pick the states to assign credit to?

Sparse Attentive Backtracking

Rosemary Ke, Anirudh Goyal, Olexa Bilaniuk, Jonathan Binas, Mike Mozer, Yoshua Bengio,

NeurIPS 2018

The attention mechanism of the associative memory picks up past memories which match (associate with) the current state.

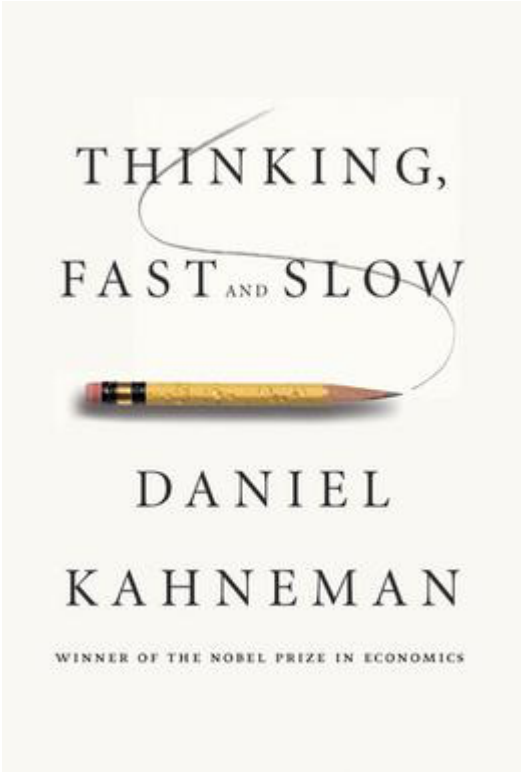


Mila

System 1 & System 2 cognitive processing

(Kahneman 2011)

- System 1: intuitive, fast, automatic, anchored in perception
  - What current deep learning is very good at
- System 2: rational, sequential, slow, logical, conscious, expressible with language
  - What future deep learning needs to do better



Deep Learning Inspiration for Neuroscience

Deep Learning & Neuroscience: Still a Large Gap

- **Backprop** and the ability to jointly train multiple layers is the workhorse of current deep learning successes. END-TO-END TRAINING OF DEEP COMPUTATIONS ROCKS. **Backprop is the building block behind modern unsupervised (generative) learning and RL**.

- But has been deemed not biologically plausible.

  - How to **efficiently** train a stochastic continuous-time dynamical system wrt a **global** objective?
    - *Random perturbation-based methods do not scale, BP does beautifully*

Equilibrium Propagation

(Scellier & Bengio 2017, Frontiers in Neuroscience)

Backpropagation

Free Phase

-network relaxes to fixed point
-read prediction at the outputs

$$\beta = 0$$

Forward Pass

-read prediction at the outputs

Weakly Clamped Phase

-nudge outputs towards targets
-error signals (back)propagate
-network relaxes to new nearby fixed point

$$\beta \gtrapprox 0$$

$$F(\theta, \beta, s) = E(\theta, s) + \beta C(s)$$

$$\frac{ds}{dt} = -\frac{\partial F}{\partial s}$$

Loss fn

Backward Pass

-compare prediction/target
-compute error derivatives

requires:
-special computational circuit
-special kind of computation

Mila

Equilibrium Propagation Theorem

(Scellier & Bengio, Bridging the Gap Between Energy-Based Models and Backpropagation, *Frontiers in Neuroscience, 2017)*

- Gradient on the objective function (cost at equilibriu can be estimated by a ONE-DIMENSIONAL finite-difference

$$\frac{dJ}{d\theta} = \lim_{\beta \to 0} \frac{1}{\beta} \left(\frac{\partial F(\theta, \beta, s)}{\partial \theta} - \frac{\partial F(\theta, 0, s)}{\partial \theta} \right)$$

Small nudging

after nudging

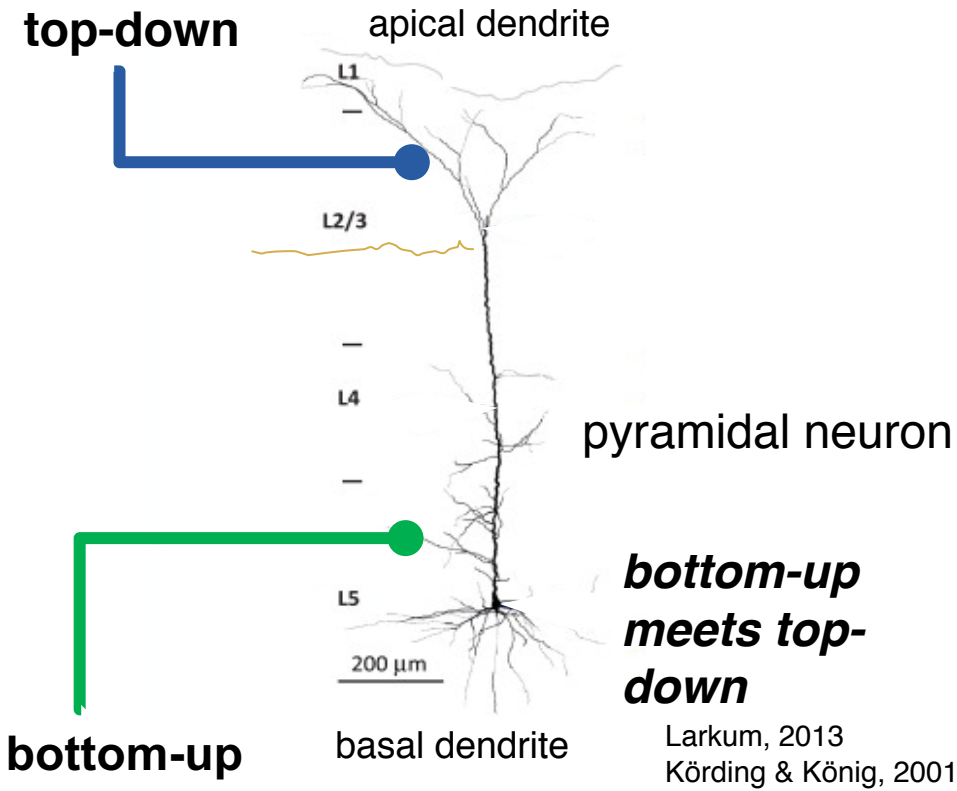
before nudging

There is a stochastic version too

→ **Gives rise to Hebbian / anti-Hebbian updates with Hopfield net energy fn**

→ **Theory is not limited to point neurons, any set of variables with dynamics, could be used for analog circuits or for adapting within-neuron dynamics**

A cortical circuit for error coding



pyramidal neuron

bottom-up meets top-down

Larkum, 2013
Körding & König, 2001
Guerguiev et al., 2017

Mila

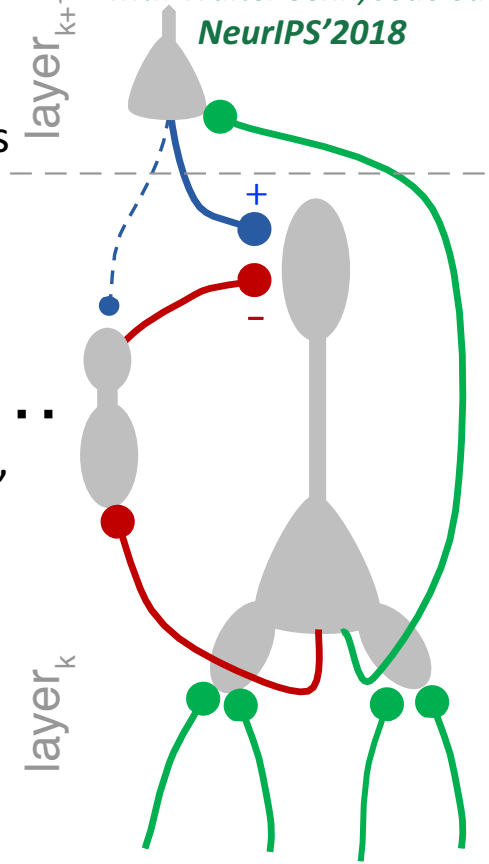
Dendritic cortical microcircuits approximate backpropagation

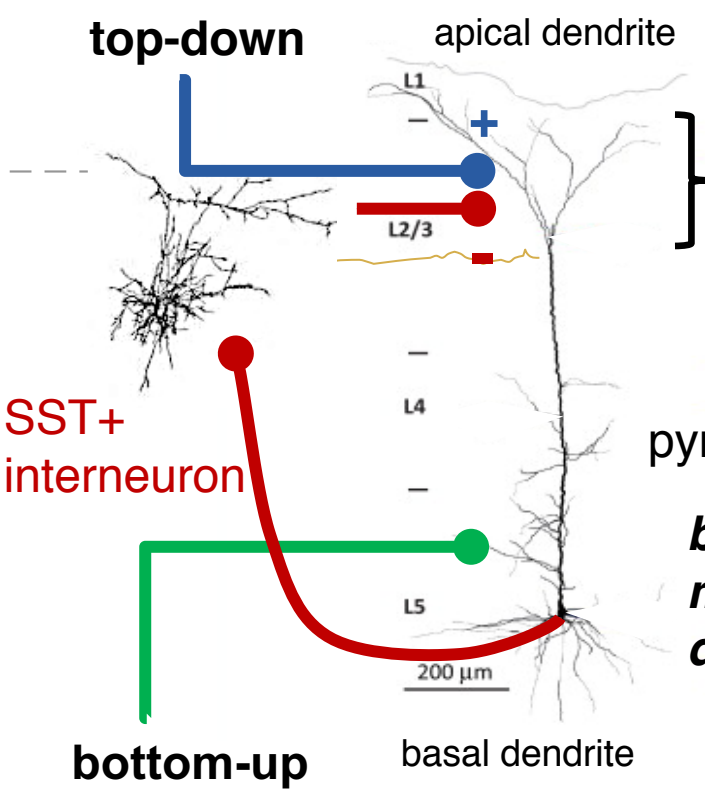
with Walter Senn, Joao Sacramento & Rui Ponte Costa

NeurIPS'2018

With no nudging, cancellation is perfect because next layer is predictable.

With nudging, difference = backprop error signal.





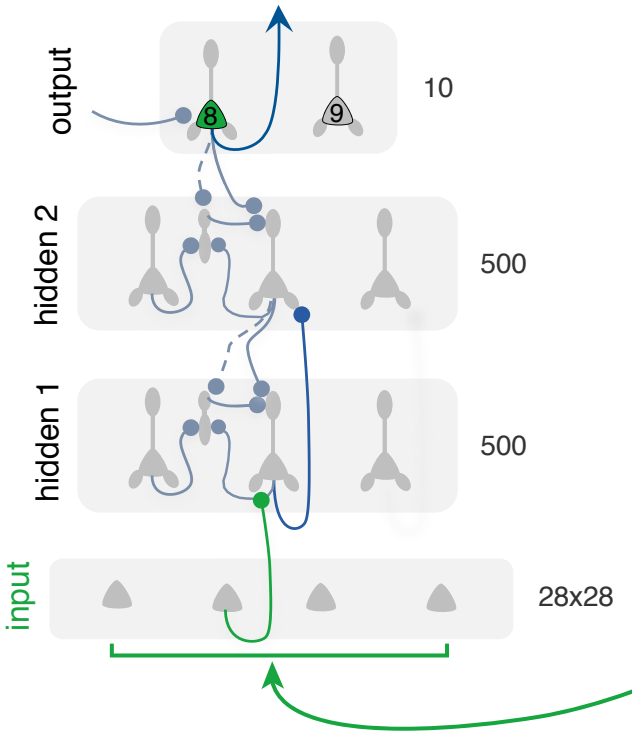
pyramidal neuron

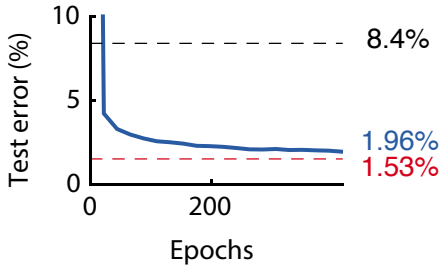
bottom-up meets top-down

Larkum, 2013
Körding & König, 2001
Guerguiev et al., 2017

Mila

Learning to classify MNIST digits

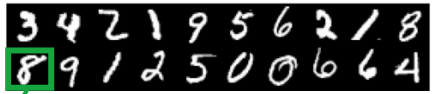




single-layer

500+500+10 microcircuit

500+300+10 backprop

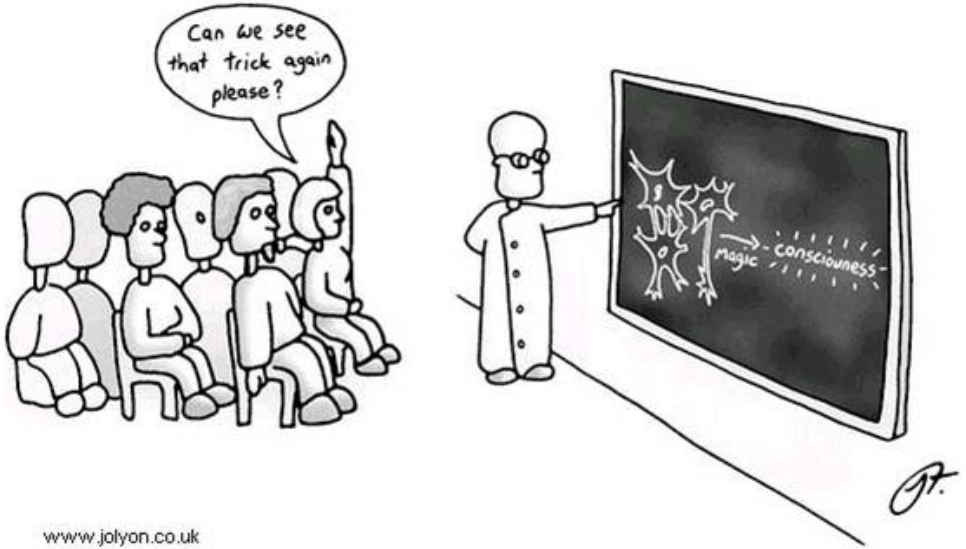


MNIST handwritten digit images

Mila

Taking the Magic out of Consciousness

- Brains are complex machines, probably stochastic
- What we commonly call consciousness should be associated with various computational mechanisms and properties and contrasted / linked with intelligence



The ML View on Consciousness

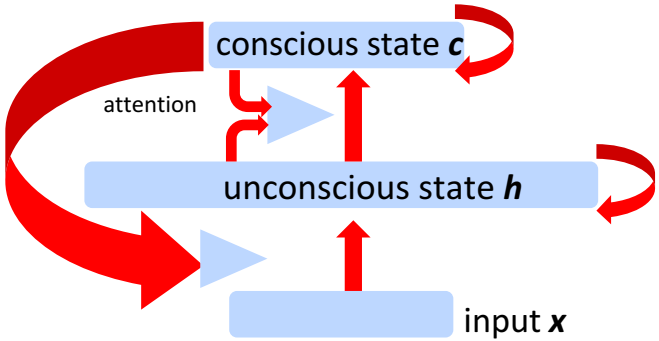
3 computational aspects of consciousness:

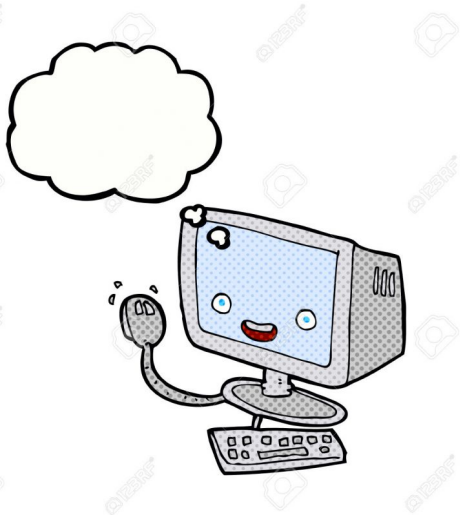
- Self-consciousness
  - Notion of self as part of the agent's state, which conditions the agent's decisions
- Access consciousness, conscious attention
  - While conscious, focus at each time step on a few attended elements which condition action/planning/imagination
- Qualia, subjective perception
  - The focus of conscious attention is mostly in a high-level abstract space in which perception is context-dependent and depends on the agent's history, goals, emotions, etc.

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

The Consciousness Prior

Bengio 2017, arXiv:1709.08568

- Focus on **representation learning** and one aspect of consciousness:
- Conscious thoughts are very low-dimensional objects compared to the full state of the (unconscious) brain = analogous to a sentence or a rule in rule-based systems
- Yet they have unexpected predictive value or usefulness

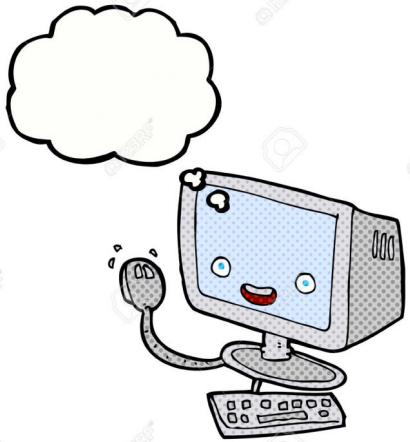
  → strong constraint or prior on the underlying representation

- **Thought**: composition of few selected factors / concepts at the highest level of abstraction of our brain

- Richer than but closely associated with short verbal expression such as a **sentence** or phrase, a **rule** or **fact** (link to classical symbolic AI & knowledge representation)

Need to disentangle both

- Variables in rule ⇔ features in representation space
- Rules ⇔ causal mechanisms



Mila

What Training Objective?

- How to train the attention mechanism which selects which variables to predict?
  - Representation learning without reconstruction:
    - Maximize entropy of code
    - **Maximize mutual information between past and future representations** (*Becker & Hinton 1992*), **between intentions (policies) and changes in representations** (affordances, independently controllable factors)
  - *Objective function completely in abstract space, higher-level parameters model dependencies in abstract space*
  - *Usefulness of thoughts: as conditioning information for action, i.e., a particular form of planning for RL*

Deep Objective: discover causal representation

- What are the right representations? Causal variables explaining the data
- How to disentangle them?
- How to discover their causal relationship, the causal graph?
- How does the brain represent such high-level concepts (expressed linguistically) and their relations?