Towards bridging the gap between deep learning and brains

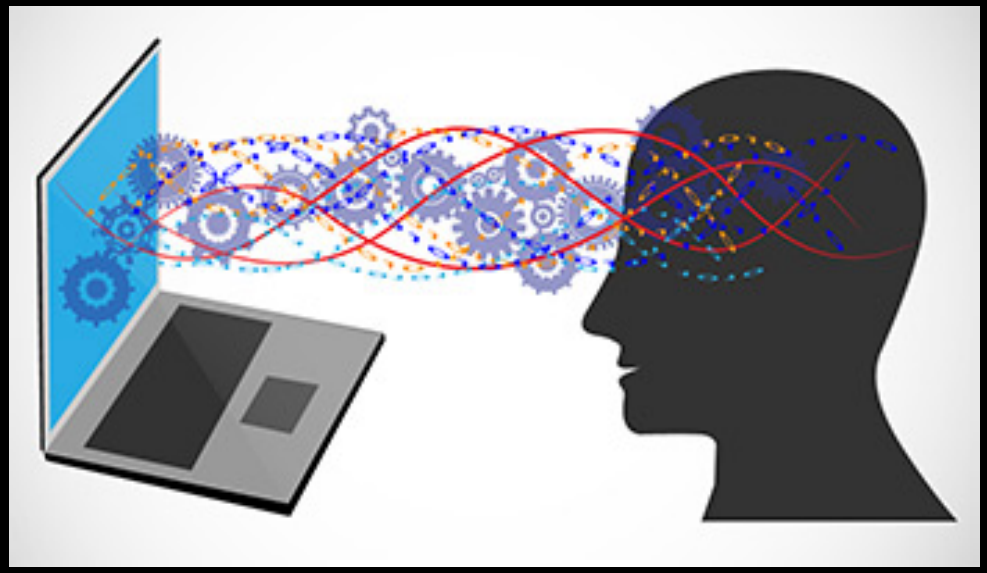
Yoshua Bengio

MIT

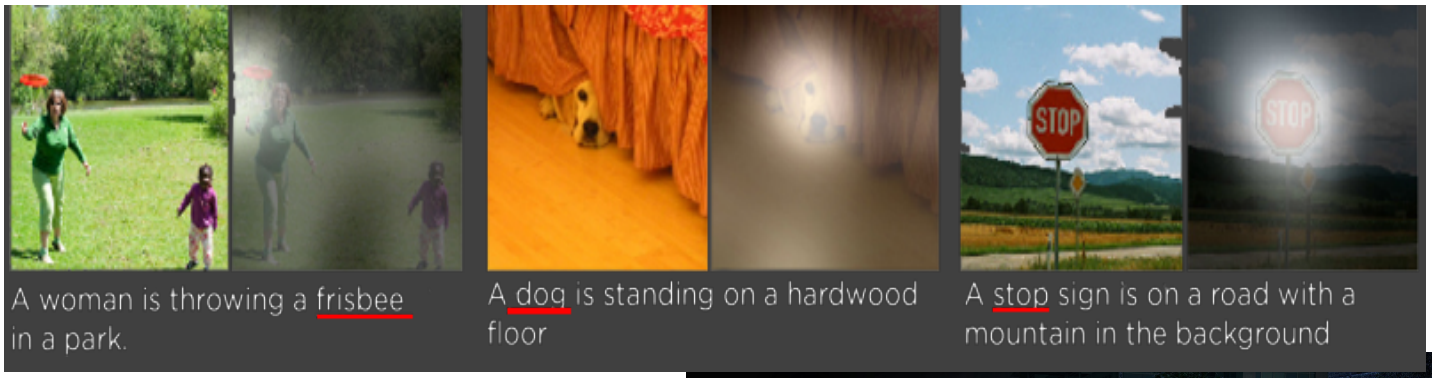
18 OCT 2018

AI & Knowledge

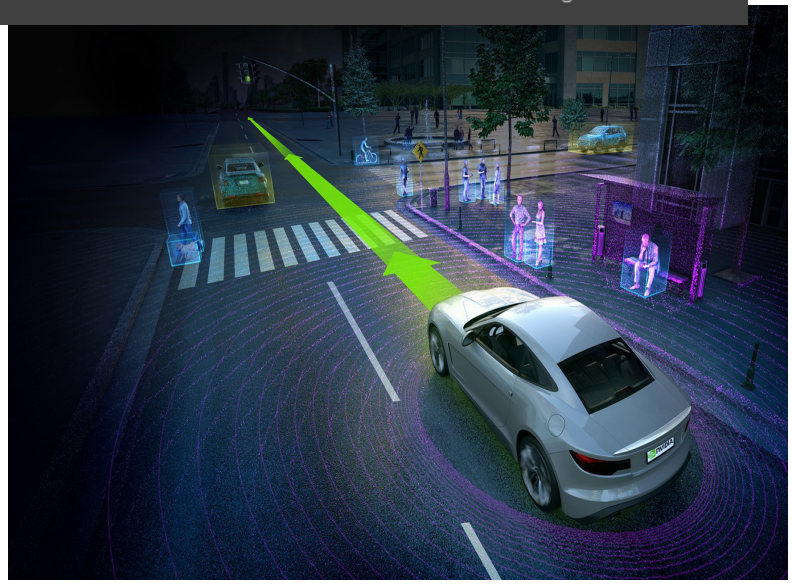
- Putting knowledge into computers
- Much knowledge is intuitive, uncommunicable



Deep Learning → AI Breakthroughs



Computers have made huge strides in **perception**, and to a lesser extent, in manipulating language, reasoning, ...

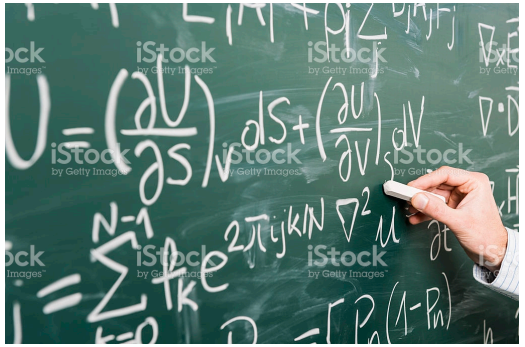


Drawing inspiration for AI from living intelligence

- 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

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.



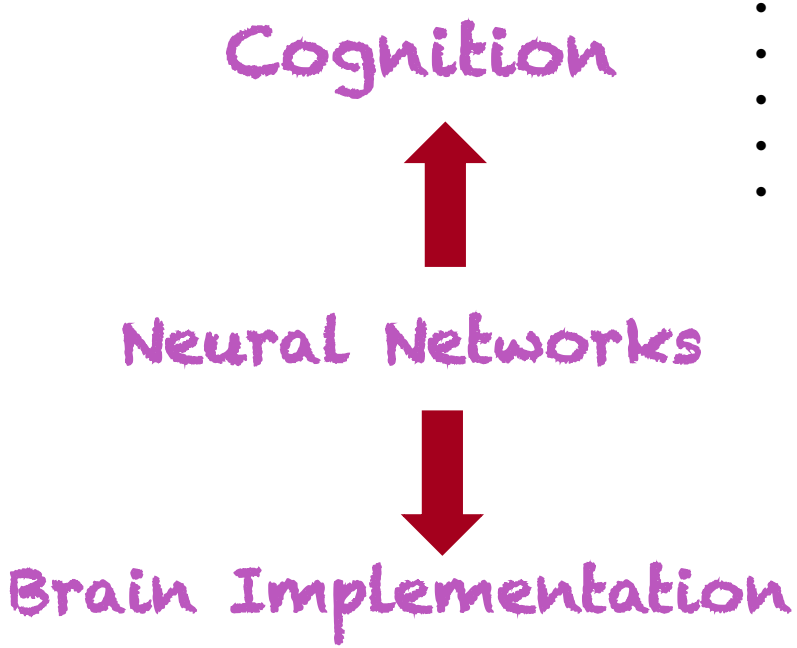
The Learning Mechanism is a Compact and Abstract Explanation of the Brain

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
- semantics & language
- agency & causality
- consciousness

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

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 ⟷ Forward Pass

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

$$\beta = 0$$

-read prediction at the outputs

Weakly Clamped Phase ⟷ Backward Pass

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

$$\beta \gtrapprox 0$$

-compare prediction/target
-compute error derivatives

requires:
-special computational circuit
-special kind of computation

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

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

Loss fn

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**

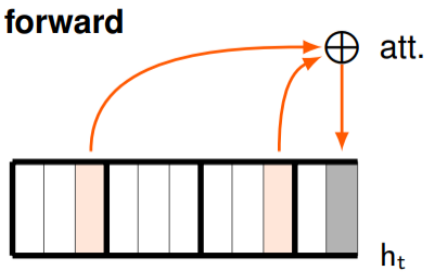
Mila

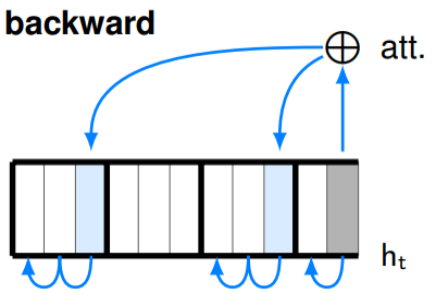
Sparse Attentive Backtracking

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

NIPS 2018

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





Mila

Mila

Still Far from Human-Level AI

- Industrial successes mostly based on **supervised** learning

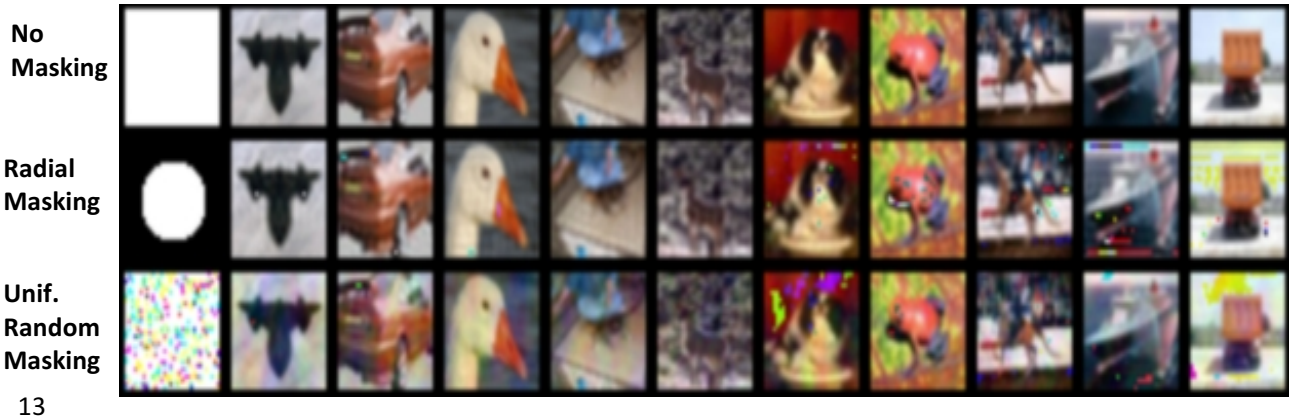
- Learning superficial clues, not generalizing well enough outside of training contexts, easy to fool trained networks:
  - Current models cheat by picking on surface regularities

Measuring the Tendency of CNNs to Learn Surface Statistical Regularities

Jason Jo and Yoshua Bengio 2017, arXiv:1711.11561

- **Hypothesis**: *Deep CNNs have a tendency to learn superficial statistical regularities in the dataset rather than high level abstract concepts*.

- From the perspective of learning high level abstractions, Fourier image statistics can be *superficial* regularities, not changing object category, but changing them leads CNNs to make mistakes

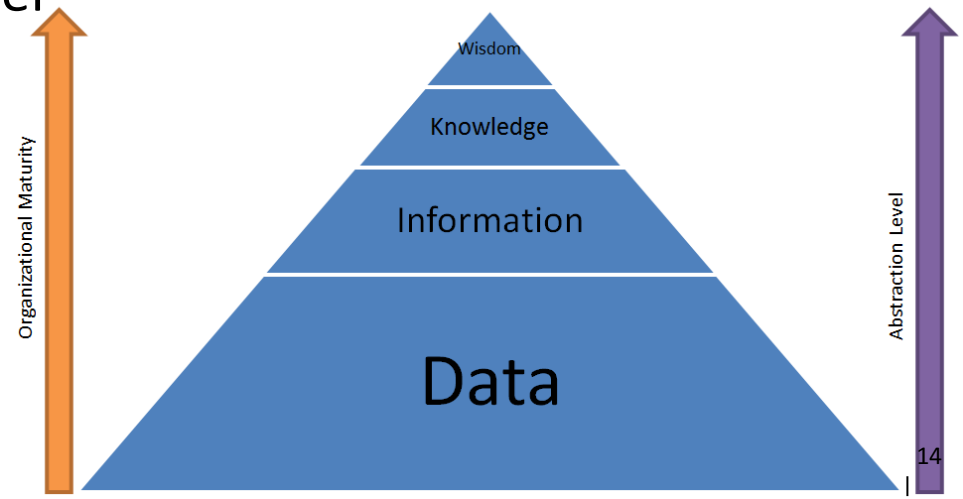


Learning Multiple Levels of Abstraction

(Bengio & LeCun 2007)

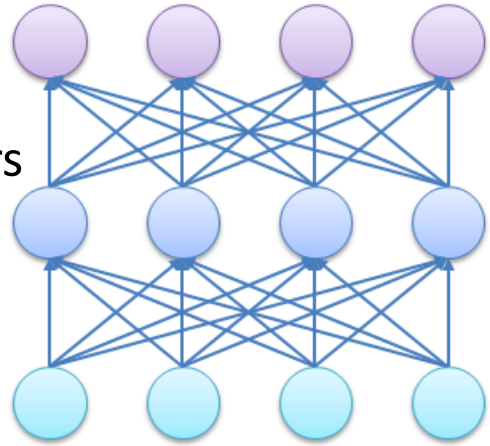
- The big payoff of deep learning is to allow learning higher levels of abstraction
- Higher-level abstractions **disentangle the factors of variation**, which allows much easier generalization and transfer

New concern:
Also disentangle the computation (modules) and the hypothesized causal mechanisms



How to Discover Good Disentangled Representations

- How to discover abstractions?
- What is a good representation? ***(Bengio et al 2013)***
- ***Dependencies are simple in the right representation***
- Need clues (= priors) to help **disentangle** the underlying factors, 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**
  - intuitive, fast heuristic, UNCONSCIOUS, non-linguistic
  - *what current **deep learning** does quite well*
- **System 2**
  - slow, logical, sequential, CONSCIOUS, linguistic, algorithmic
  - *what **classical symbolic AI** was trying to do*
- **Grounded language learning**: combine both language learning and world modeling

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

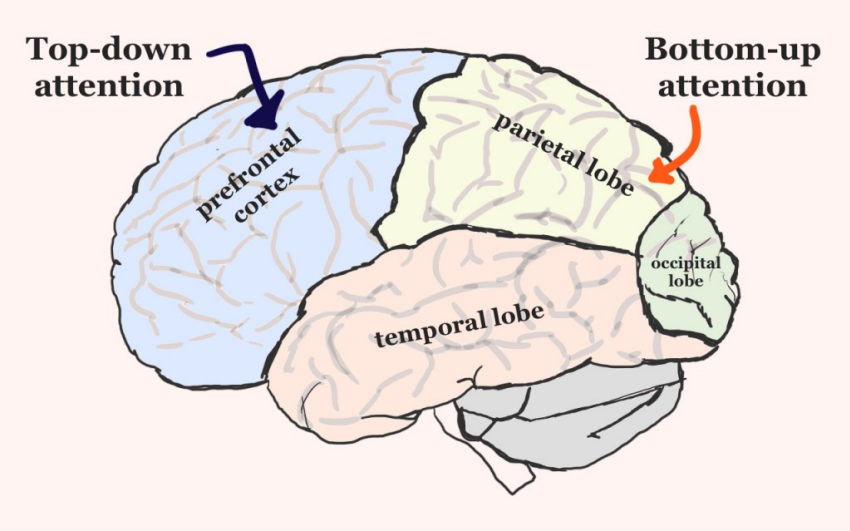
Mila

On the Relation between Abstraction and Attention

- Attention allows 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

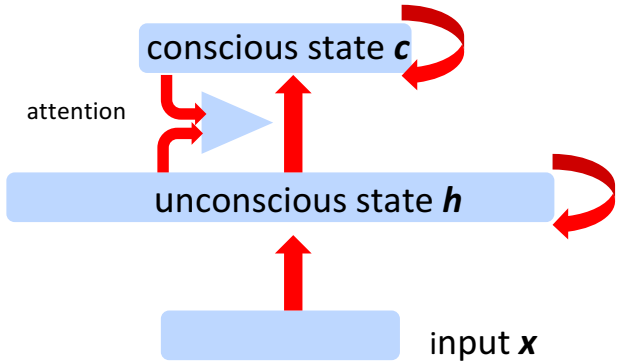
- Different from sparse auto-encoders: controller chooses focus, conditionally



The Consciousness Prior

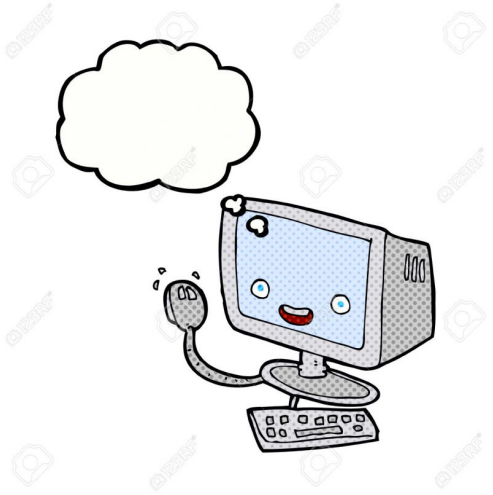
Bengio 2017, arXiv:1709.08568

- 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

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*

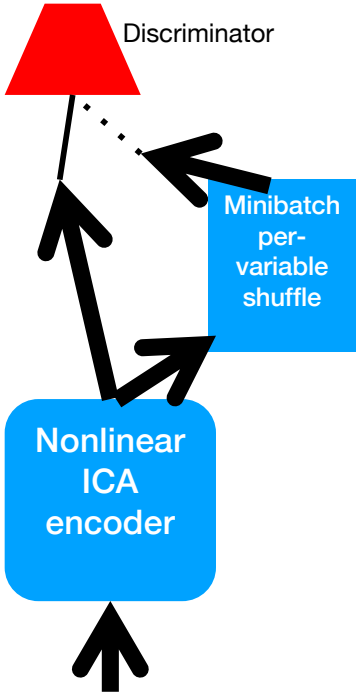
Using a discriminator to optimize **independence**, mutual information or entropy

Brakel & Bengio ArXiv:1710.05050

- Train a discriminator to separate between pairs (A,B) coming from P(A,B) and pairs coming from P(A) P(B)

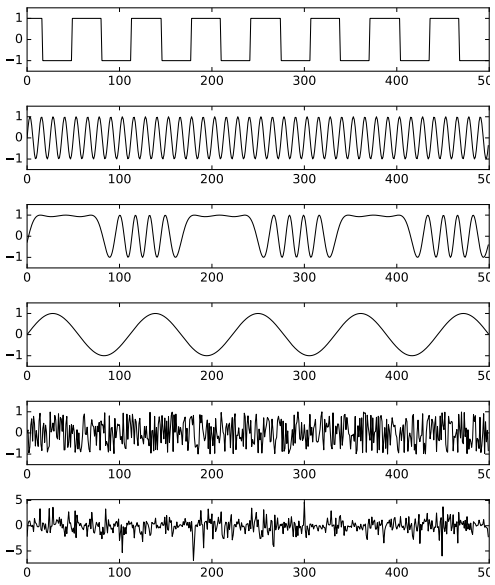
- Generalize this to measuring **independence** of all the outputs of a representation function (encoder). Maximize independence by backpropagating the independence score into the encoder

→ NON-LINEAR ICA.

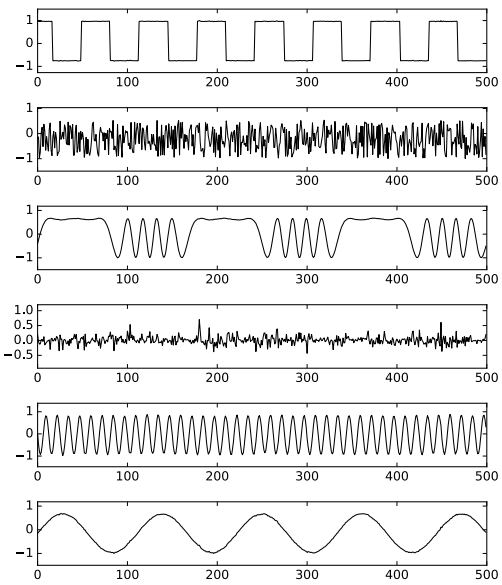


Non-Linear Independent Component Analysis Results

- Sources were either mixed linearly or non-linearly, independent components recovered in both cases

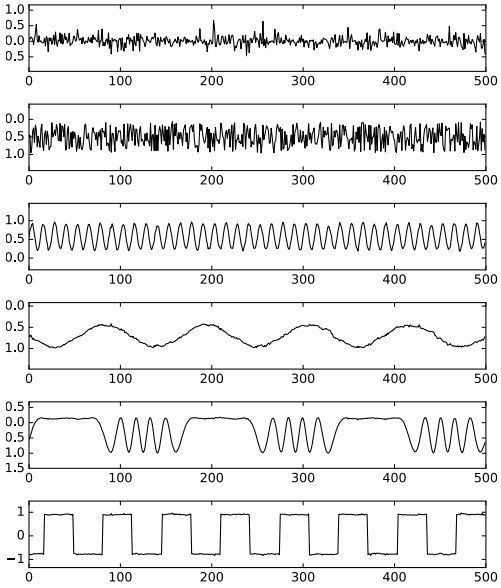


(a) Source signals.



(b) Anica reconstructions $\rho_{\max} = .997$.

Linearly mixed



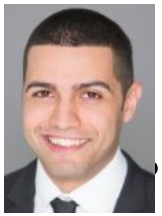
(a) Anica PNL reconstructions $\rho_{\max} = .997$.

Nonlinearly mixed

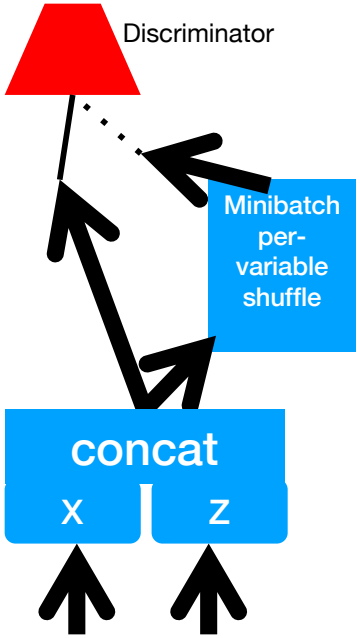
Using a discriminator to optimize independence, mutual information or entropy

MINE: Mutual Information Neural Estimator

Belghazi et al ArXiv:1801.04062



Same architecture, but with a twist in the training objective which provides an asymptotically consistent estimator of mutual independence



Mutual information, KL divergence and Donsker-Varadhan Representation [Belghazi et. al., 2018]

Mutual information: measure of dependence btwn 2 variables

$$I(X;Z) = \mathcal{D}_{KL}(\mathbb{P}_{X,Z} || \mathbb{P}_X \otimes \mathbb{P}_Z) = \mathbb{E}_{\mathbb{P}_{X,Z}} \left[\log \left(\frac{p(x,z)}{p(x)p(z)} \right) \right]$$

$$I(X;Z) = H(X) + H(Z) - H(X,Z) = D_{KL}(\mathbb{P}_{XZ} \,||\, \mathbb{P}_X \otimes \mathbb{P}_Z)$$

(Donsker & Varadhan, 1983):

$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) = \sup_{T:\Omega \to \mathbb{R}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T])$$

Optimal T:

$$T^* = \log \frac{d\mathbb{P}}{d\mathbb{Q}} + C$$

With suboptimal T:

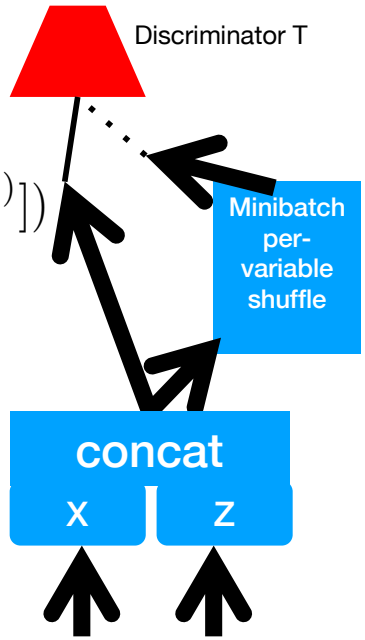
$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) \geq \sup_{T \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T])$$

MINE: Estimator of MI

Given two r.v. X & Z and samples of their joint & marginals:

$$\widehat{I(X;Z)}_n = \mathbb{E}_{\hat{\mathbb{P}}^{(n)}_{XZ}}[T_{\hat{\theta}_n}(x,z)] - \log(\mathbb{E}_{\hat{\mathbb{P}}^{(n)}_{X} \otimes \hat{\mathbb{P}}^{(n)}_{Z}}[e^{T_{\hat{\theta}_n}(x,z)}])$$

where discriminator T is optimized to maximize the rhs

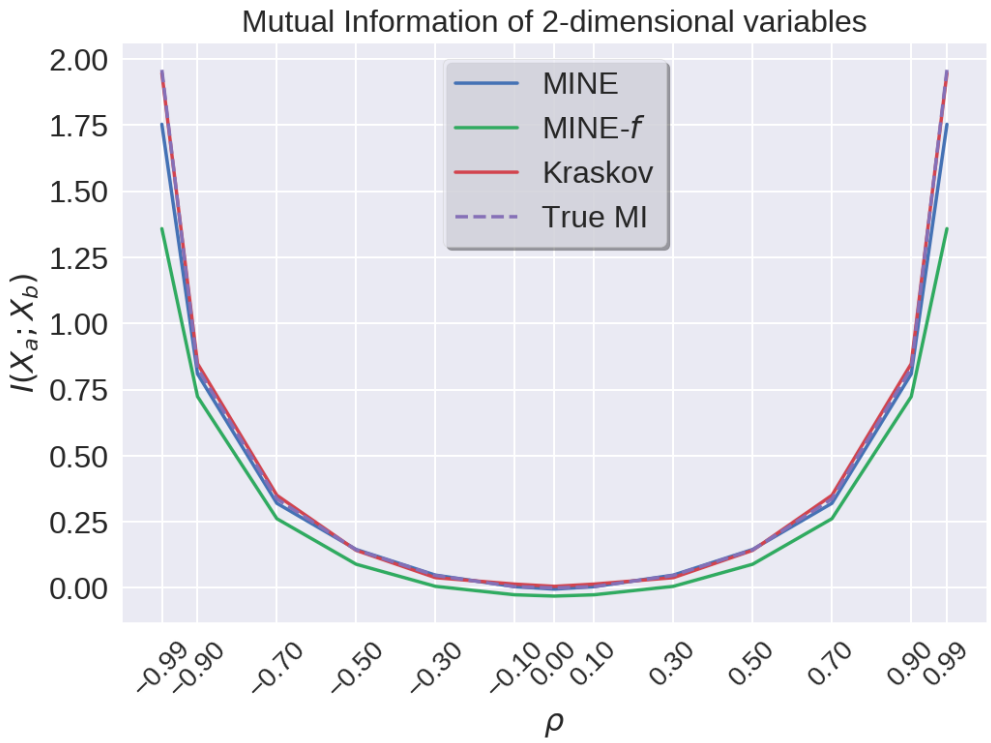


MINE: Consistency

Theorem: there exists a neural net architecture such that for all $\epsilon > 0$ there exists an integer *N* s.t.

$$\forall n \geq N, \quad |I(X,Z) - \widehat{I(X;Z)}_n| \leq \epsilon \text{ with probability one}$$

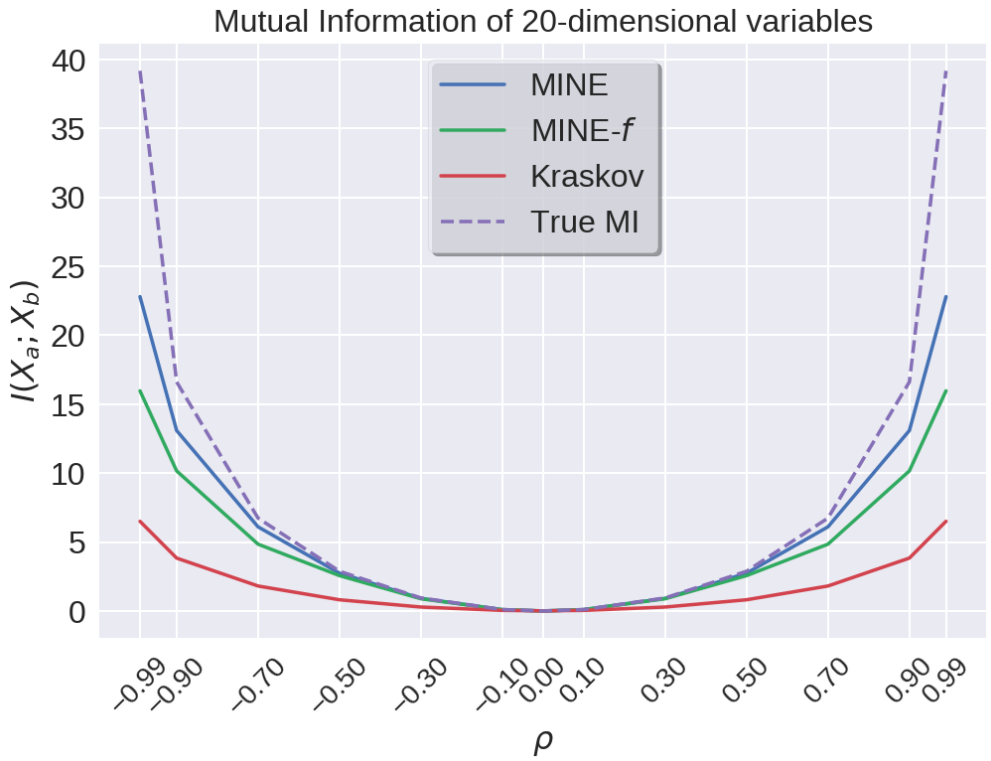
Demonstration of estimation



MI between 2 Gaussians

[Belghazi et. al., 2018]

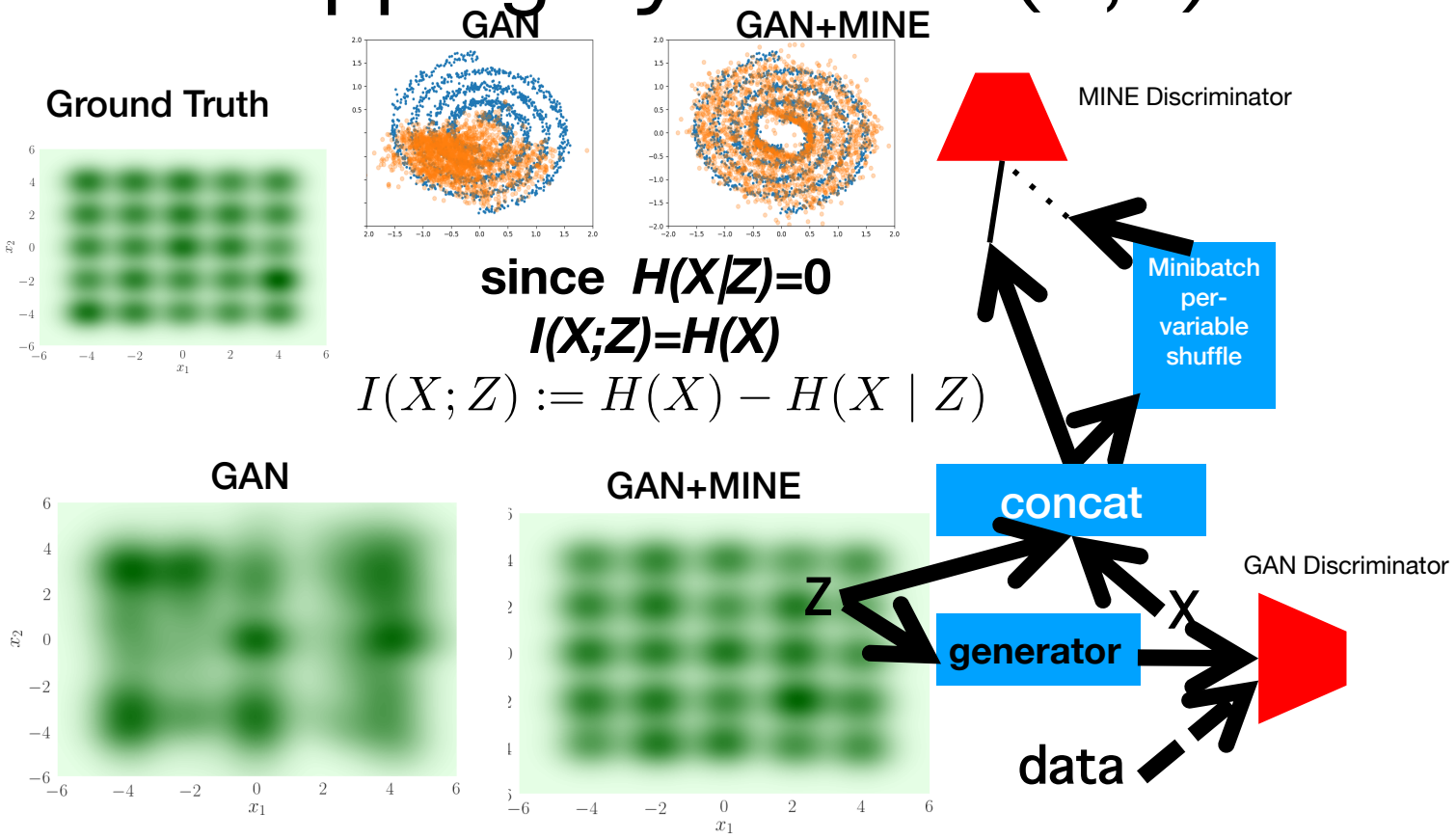
Demonstration of estimation



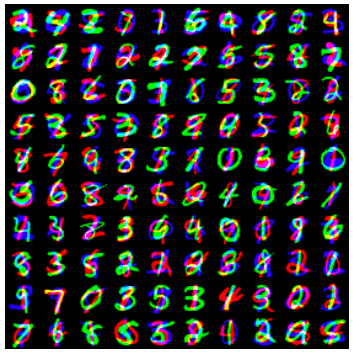
MI between 2 Gaussians

[Belghazi et. al., 2018]

Maximizing ENTROPY: avoid GAN mode dropping by max MI(X,Z)



Maximizing entropy at the output of a neural net (stacked MNIST)



	Modes (max 1000)	$\mathcal{D}_{KL}(\mathbb{P}_Y \| \mathbb{Q}_Y)$
DCGAN	99	3,4
ALI	16	5,4
Unrolled GAN	48,7	4,32
VEEGAN	150	2,96
PacGAN	1000	0,6
DCGAN+MINE	1000	0,5

[Belghazi et. al., 2018]

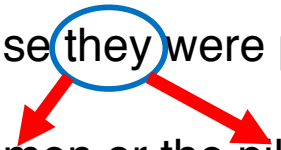
Back to the consciousness prior: joining system 1 and system 2

Most statistical NLP uses only natural language corpora & annotations

- Most NLP tasks currently dealt with using only text + labels
  - Speech recognition, language modeling, text compression, machine translation
  - Parsing
  - Question Answering, reading comprehension
  - Document classification
  - Disambiguation
  - Dialogue, chatbots, personal assistants

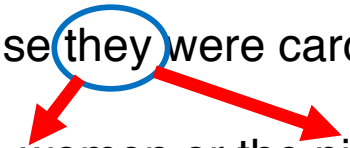
Common Sense & Winograd Schemas

The women stopped taking pills because they were pregnant.



Which entities were pregnant? The women or the pills?

The women stopped taking pills because they were carcinogenic.



Which entities were carcinogenic? The women or the pills?

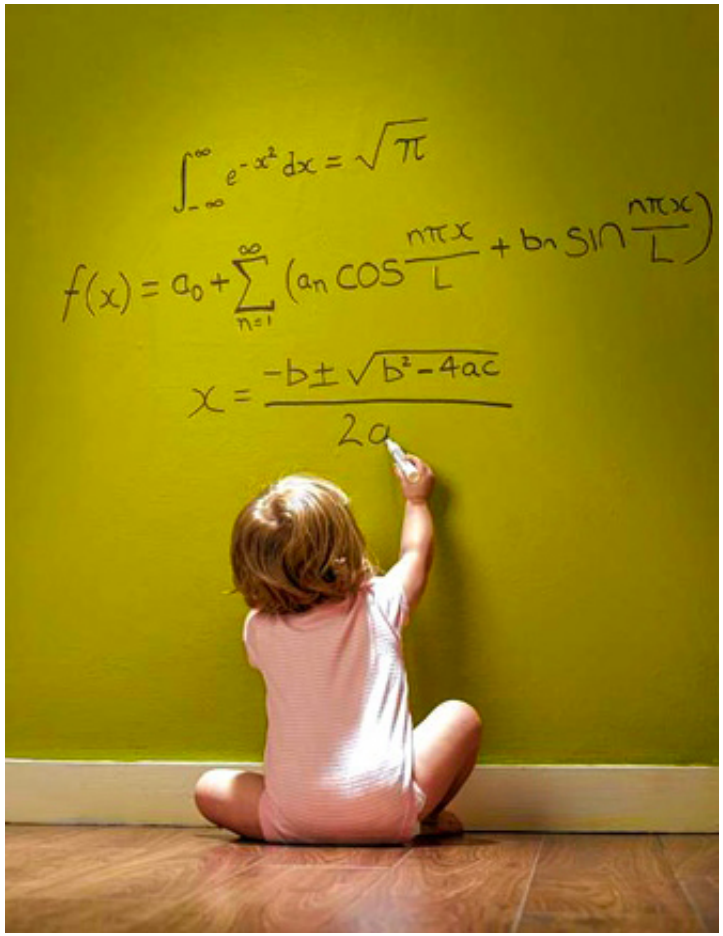
Humans: 100% accurate

SOTA systems: 56% accurate

Chance: 50% accuracy

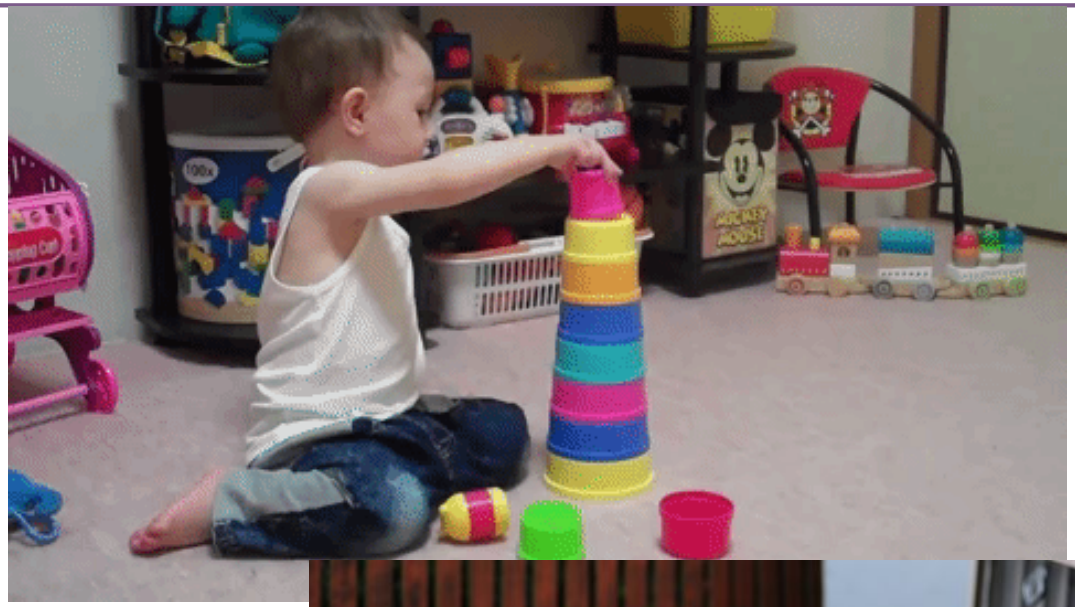
Humans outperform machines at unsupervised learning

- Humans are very good at unsupervised learning, e.g. a 2 year old knows intuitive physics
- Babies construct an approximate but sufficiently reliable model of physics, how do they manage that? Note that they interact with the world, not just observe it.



Mila

Intuitive Psychology and Intuitive Physics



Informal 'common sense' knowledge.

Still lacking in our best AIs

Mila

Knowledge!

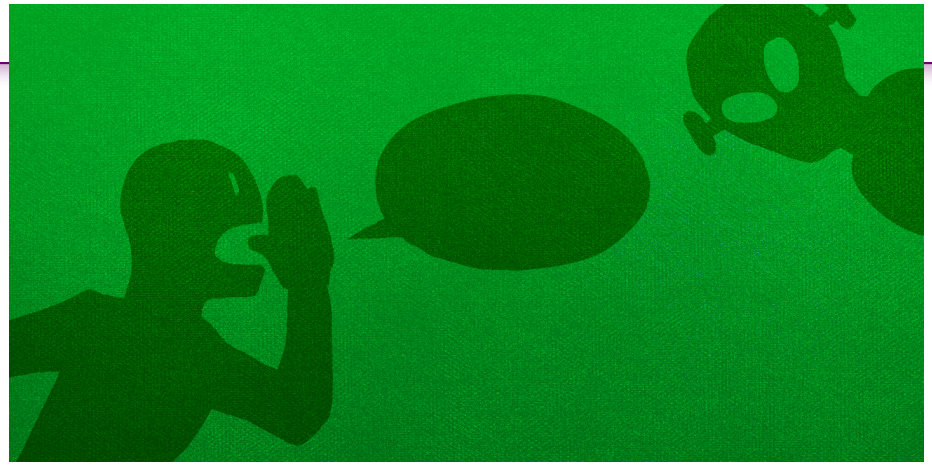
- What does it mean for a machine to understand a question, a document?
- What kind of knowledge would be required to do that?
- How is that knowledge to be acquired by the computer?

Alien Language Understanding: a Thought Experiment

- Imagine yourself approaching another planet and observing the bits of information exchanged by aliens communicating with each other

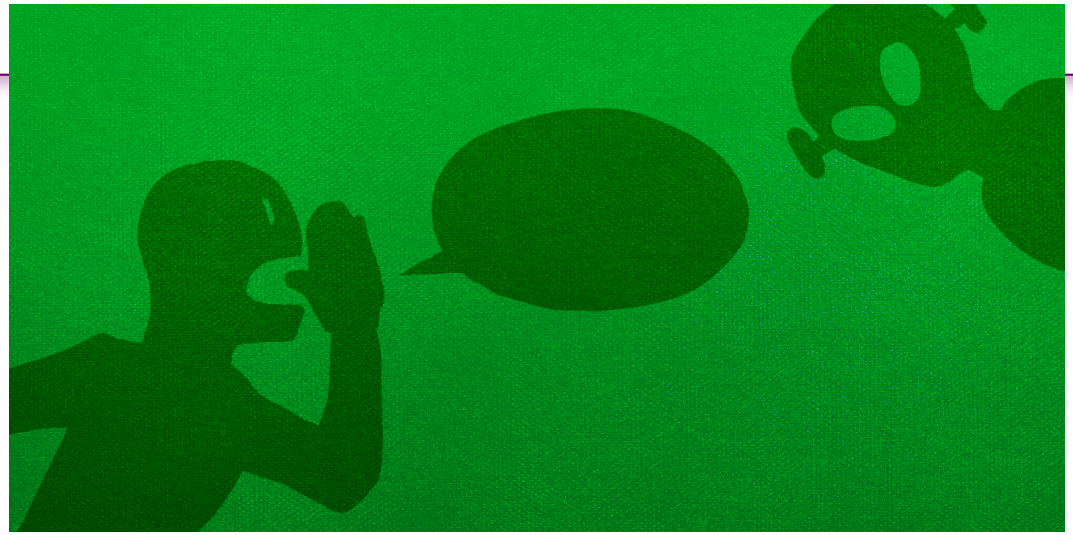
- Unlike on Earth, their communication channel is noisy, but like on Earth, bandwidth is expensive → the best way to communicate is to maximally compress the messages, which leads to sequences of random bits being actually exchanged.

- If we only observe the compressed messages, there is no way we can ever understand the alien language



Alien Language Understanding: a Thought Experiment

- How can we learn to understand the alien language?
- We need to do grounde language learning: we need to observe what the aliens are doing jointly with their messages, to try to decipher their intentions, context, etc.
- For this we need to build an 'Alien World Model' which captures the causal structure of their behaviors and resulting changes in their environment.



Jointly Learning Natural Language and a World Model

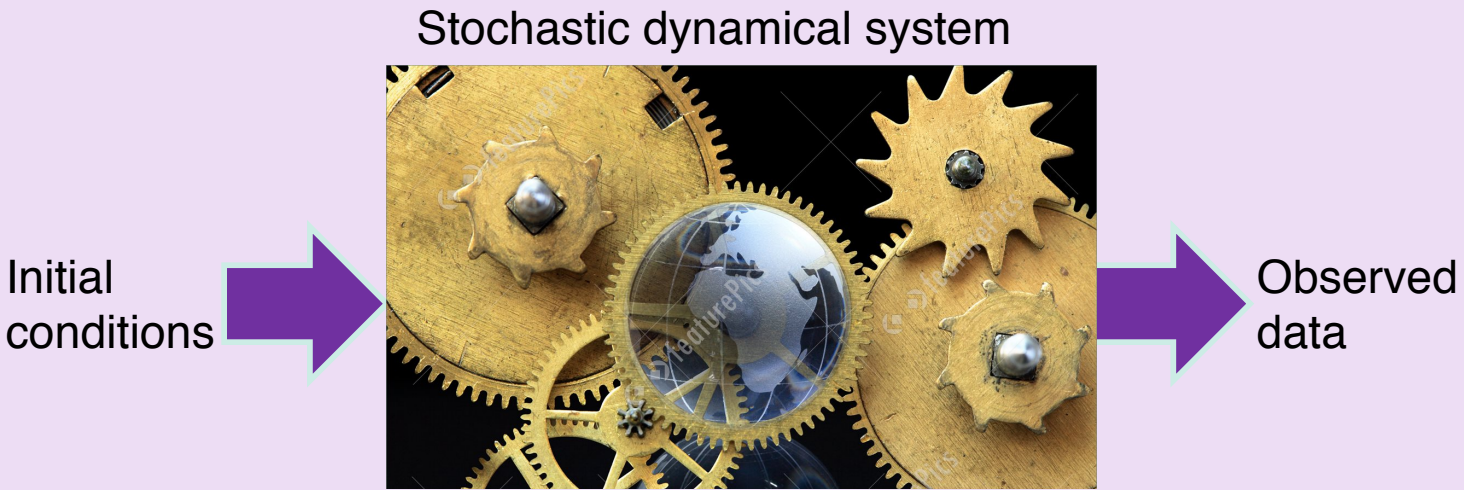
- Should we first learn a world model and then a natural language description of it?
- Or should agents jointly learn about language and about the world?
- I lean towards the latter.
- Consider top-level representations from supervised ImageNet classifiers. They tend to be much better and easier to learn 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.

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

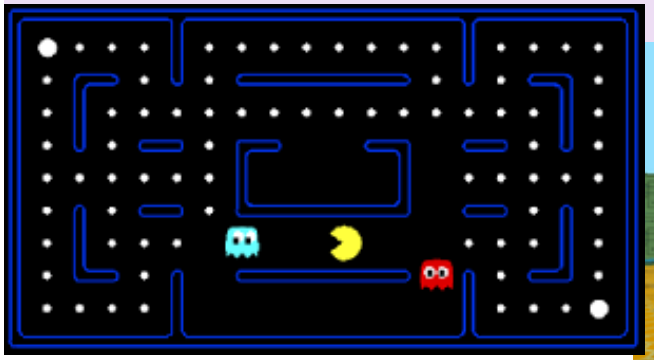
Beyond the iid assumption

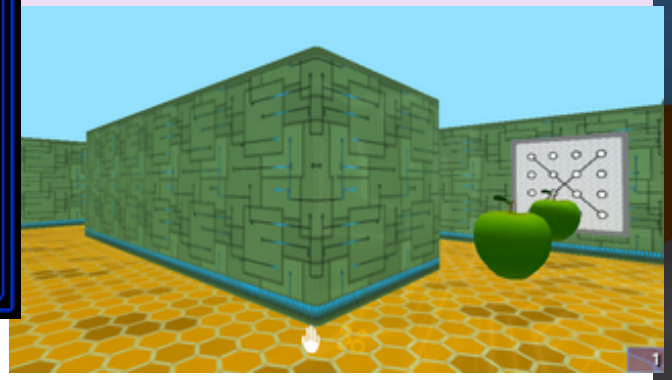
- The assumption that the test data is from the same distribution as the training data is too strong, and it is often violated in practice, leading to poor out-of-distribution generalization.
- I propose to consider relaxed assumptions: the test data was generated under the same causal dynamics, but from different initial conditions (which may be unlikely under the training distribution).

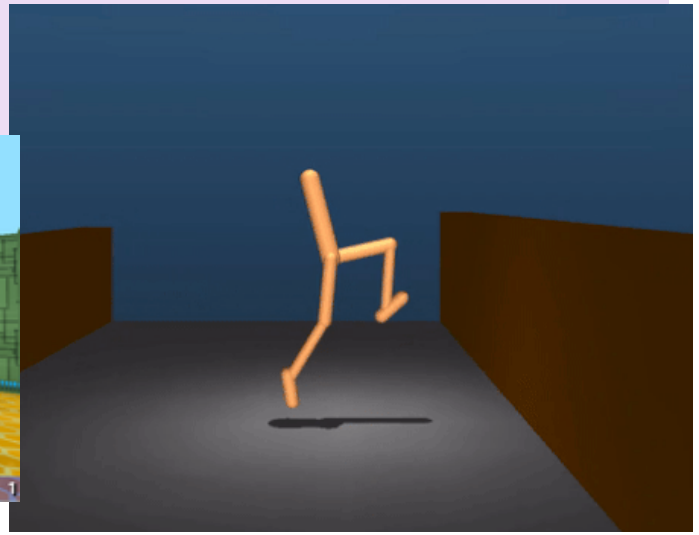


Develop learning procedures which figure out how their small-scale environment works

- Outcome of ML **research** = learning framework, not a trained learner
- Solve simple environments before human-level understanding of our world
- Current trained models are very poor at understanding our world. Working on a simpler virtual environment leads to a faster research cycle.
- We can gradually make the environments more realistic and complex as our learning methods and our computational capabilities improve





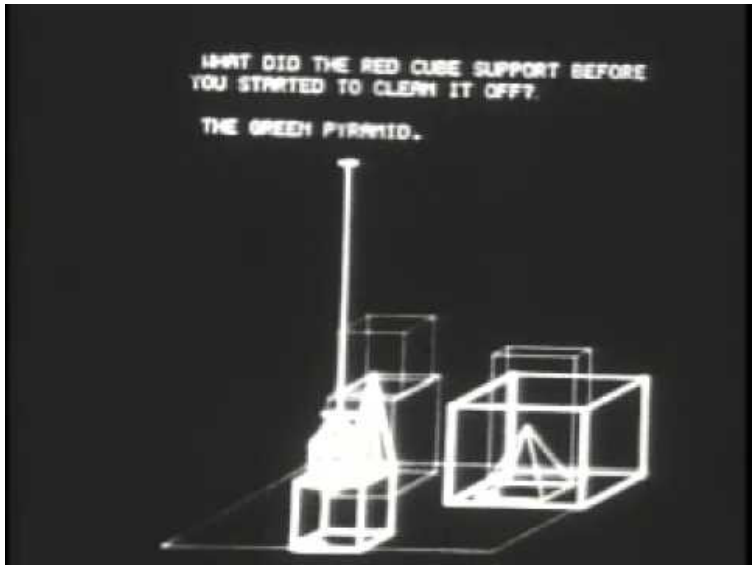


How to Teach Agent to Understand Language?

Hard-code everything?

```
Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE
HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS
TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
```

- **doesn't scale**
- **likely a dead-end**



SHRDLU by Winograd (1971)

***BabyAI**: First Steps Towards Grounded Language Learning With a Human In the Loop; submitted to ICLR 2018, Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, Yoshua Bengio*

Motivation for BabyAI Platform

Let's bite the bullet - language learning with a human in the loop would be great!

- a human can gradually increase the complexity of the task (curriculum learning)
- a human can choose the task based on what the agent has already learned (interactive learning)
- you could tune your intelligent helper / household robot / R2D2 to understand you better!

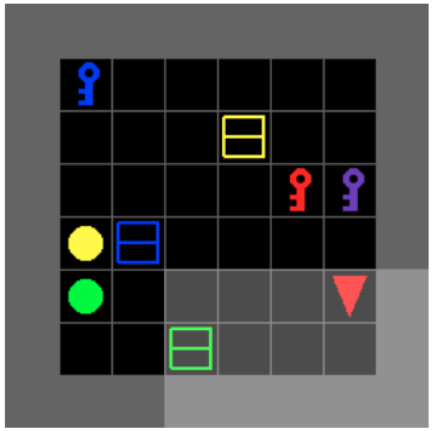
But are we there in terms of data efficiency?

BabyAI Platform

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

Comprises:

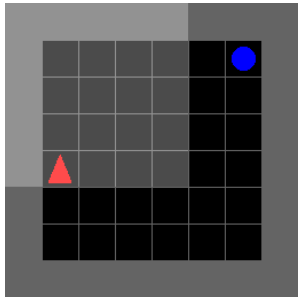
- a gridworld with partial observability (Minigrid)
- a compositional natural-looking Baby language with over 10^19 instructions
- 19 levels of increasing difficulty
- a heuristic stack-based expert that can solve all levels



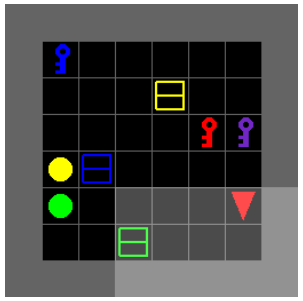
(b) PutNextLocal: "put the blue key next to the green ball"

github.com/mila-udem/babyai

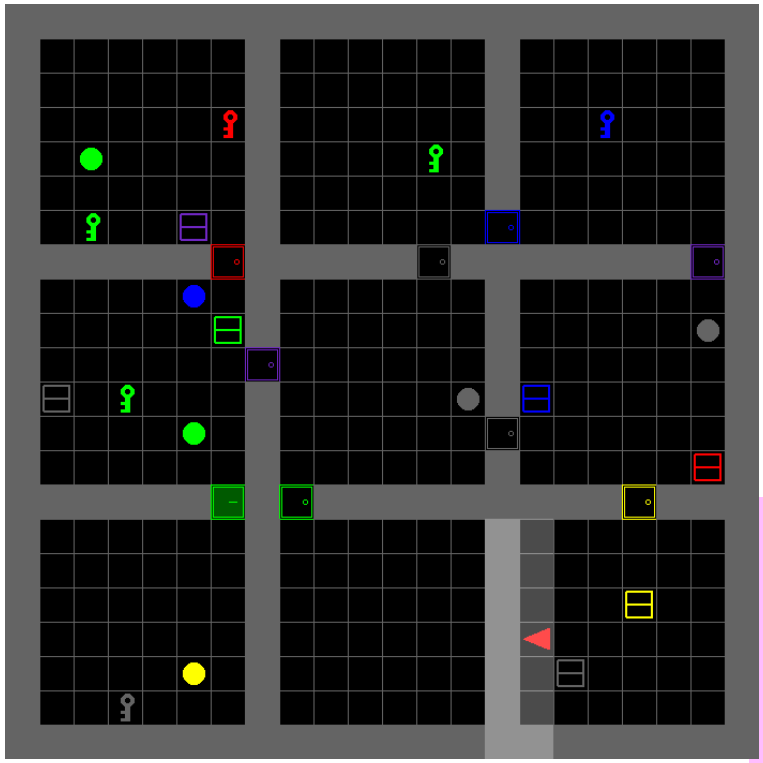
Early Steps in the Baby AI Game Project



(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

BabyAI Competencies

- we distinguish 13 competencies, e.g.
  - MAZE = 3x3 maze navigation
  - UNLOCK = find a key to unlock the door
  - LOC = understand “in front of”, “behind”, ...
- each level is defined by the set of required competencies

LEVEL BOSS

open a door and pick up the green box, then pick up the green key and put a blue ball next to a grey ball !!!!!!

	ROOM	DISTR-BOX	DISTR	MAZE	UNBLOCK	UNLOCK	IMP-UNLOCK	GOTO	OPEN	PICKUP	PUT	LOC	SEQ
GoToObj	x												
GoToRedBallGrey	x	x											
GoToRedBall	x	x	x										
GoToLocal	x	x	x					x					
PutNextLocal	x	x	x								x		
PickUpLoc	x	x	x							x		x	
GoToObjMaze	x			x									
GoTo	x	x	x	x				x					
Pickup	x	x	x	x						x			
UnblockPickup	x	x	x	x	x					x			
Open	x	x	x	x					x				
Unlock	x	x	x	x		x			x				
PutNext	x	x	x	x							x		
Synth	x	x	x	x	x	x		x	x	x	x		
SynthLoc	x	x	x	x	x	x		x	x	x	x	x	
GoToSeq	x	x	x	x				x					x
SynthSeq	x	x	x	x	x	x		x	x	x	x	x	x
~~GoToImpUnlock~~	~~x~~	~~x~~	~~x~~	~~x~~			~~x~~	~~x~~					
BossLevel	x	x	x	x	x	x	x	x	x	x	x	x	x

And How Is Data Efficiency?

- we measure the number of demos/episodes needed to get 99% success rate
- because who cares about 80% accurate agents???

all numbers are thousands!!!

Level	IL from Bot	RL
GoToRedBallGrey	5.7 - 8	377 - 379
GoToRedBall	44.2 - 62.5	453 - 470
GoToLocal	125.2 - 177	1167 - 1320
PickupLoc	250 - 354	2591 - 2608
PutNextLocal	354 - 500	1875 - 2587
GoTo	250 - 354	1057 - 2177

Results of 1st benchmark: data efficiency needs work!

- hundreds of thousands of demonstrations are needed for very simple tasks
- it takes 3 times as much data to get from 95% to 99%
- **a lot of progress is needed before putting a human in the loop!**
- use BabyAI for your data efficiency studies!
- … but don’t try too hard (e.g. semantic parsing) cause it’s a gridworld

What Next? Abstract Word Models

- Current ML and RL tends to model dependencies in data space
- Current ML and RL tends to model temporal sequences via the unfolding of one-step predictions

$$P(\text{next frame} \mid \text{previous frames})$$

- Humans' plans are very different:
  - We project ourselves at arbitrary points into the future or the past
  - A plan is a sequence of events which are not at regularly spaced intervals
  - A future event in a plan does not need to be specified at a particular time, e.g. "tomorrow I will…"
  - We imagine not the a future state but only very specific and abstract aspects of it (see 'The Consciousness Prior')