AI and Neuroscience: Bridging the Gap

Irina Rish
AI Foundations - Learning
IBM Research AI
A common goal: **discover universal laws governing both biological and artificial intelligence**
Feature Engineering (prior knowledge)
- e.g. network properties
  [Rish et al, PLoS One 2013], [Cecchi et al, NIPS 2009]

Feature Extraction: Learning Representations
- dictionary learning, deep convnets learning
  [Rish et al, SfN 2011], [Rish et al, ICML 2008],
  [Bashivan et al, ICLR 2016], [Garg et al, submitted]

Feature Selection (sparsity)
[Rish et al, SPIE Med.Imaging 2012], [Honorio et al, AISTATS 2012],
[Rish et al, Brain Informatics 2010], [Carroll et al, Neuroimage 2009]
AI 4 NEURO: NEUROIMAGING DATA ANALYSIS

Mental states in videogames: sparse regression, 70-95%

Pain perception: sparse regression, 70-80% accuracy, “holographic” patterns

Cocaine addiction: sparse Markov net biomarkers; MPH effects analysis (“stimulant 4 stimulant”)

Cognitive load prediction: 91% w/ recurrent ConvNets

Schizophrenia classification: 74% to 93% accuracy

Symptom severity prediction

Nonlinear dynamical models of Cal and fMRI

[Carroll et al, Neuroimage 2009]
[Scheinberg & Rish, ECML 2010]

[Rish et al, Brain Informatics 2010]
[Rish et al, SPIE Med. Imaging 2012]
[Cecchi et al, PLOS Comp Bio 2012]

[Honorio et al, AISTATS 2012]
[Rish et al, SPIE Med. Imaging 2016]

[Bashivan et al, ICLR 2016]

[Cecchi et al, NIPS 2009]
[Rish et al, PLOS One, 2013]

[Abrevaya et al, 2018, submitted]
As predicted by the World Health Organization, by 2030 the amount of worldwide disability and life loss attributable to depression may become greater than for any other condition, including cancer, stroke, heart disease, accidents, and war.

However, many people do not receive an adequate treatment; one of the major factors here is limited availability of mental health professionals, compared to the number of potential patients.

Goal: easily accessible, round-the-clock therapeutic services provided by a conversational agent ("PsyBot"?)
“Language is a window into the brain” - M. Covington
INFOGAIN-DRIVEN DIALOGUE VIA HASHCODE REPRESENTATIONS

Outperforms deep net systems:

- Works great on small datasets where deep nets failed
- Learns orders of magnitude faster: hours instead of days

Patient

Therapist

\[ \argmax_{M_h} I(C_p : C_t; M_h) = H(C_t; M_h) - H(C_t|C_p; M_h) \]

Construct hash codes of responses

Optimize hashing model to maximize mutual information between patient and therapist

Learn a predictive model to infer therapist’s response to patient

**Dialog Response Inference: Human Evaluation**

Our models take more than 80% of the votes.

900 test instances.

Chose 2 best responses from each instance.

108 Amazon turkers (human annotators, not experts in psychiatry).

Diverse (%)

Appropriate (%)
I did lose three pounds, did you notice that?

That's great, I did notice that.

---------------------- hashing methods:

Yes, you did.

To help you with your diet?

Okay what's the explanation?

------------------------ deep network methods:

I don't want to be a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of a lot of the way

Yeah. You

Uh-huh. you
<patient> Which is stupid because like, ya know, (sniffles) I guess if it's not
what they like, then it's not what they like and they can like do something
else, but like (sniffles) ya know there's some psychiatrists who charge like $300
an hour. If you did that you would really have to work like...if you worked an
hour a day five days a week that would be fifteen hundred dollars a week.
Like...you wouldn't have to work at all...like barely. I mean, though, most
psychiatrists probably don't charge $300 an hour...yeah...it's just kinda
ridiculous that...

<therapist> I'm sorry you're having such trouble getting...getting access to
the care that you need (pause).

-------------------- hashing methods:

<RMM> You got to do.
<RkNN> Oh yeah, you never tried Lithium?
<RLSTM> And you ought not to just shove it away out of your life because you
will need it again.

-------------------- deep network methods:

<LSTM> Uh-huh. you
<HRED> Yeah. you
<VHRED> Uh-huh. you
A virtual AI assistant on a smartphone which implements the following four main steps: (1) **data collection**; (2) **mental state recognition**; (3) **taking action** to improve the mental state; (4) **receiving feedback** from a person to improve future actions.

**Roles:**
24/7 personal coach, assistant, therapist, caretaker, or just a “digital friend”
NEURO 4 AI: WHY DO WE NEED IT?

• Successful examples: reinforcement learning, deep learning

• Still, artificial brains are way behind the natural ones:
  
  • brains develop from a single cell via neurogenesis and plasticity, while artificial NNs (ANNs) are manually constructed
  
  • brains can easily adapt to very different environments and new tasks over lifetime, ANNs are still highly specialized and inflexible
  
  • Attention, memory, learning mechanism (backprop) in ANNs can be improved by more biologically plausible implementations
  
  • Brain is a dynamical system changing even without the input, in resting-state, while machine-learning models are mainly “static”
OUR QUESTIONS

• What are major limitations of modern AI?

• What can AI learn from neuroscience?

• What directions should we focus on first?
Deep Learning: limitations

- Network engineering is (still) mostly ad-hoc/manual
- Catastrophic forgetting + limited transfer in continual learning
- Requires huge amounts of data
- Vulnerable to adversarial attack
- Sensitive to hyperparameters
- Requires massive power during test and train (kW of power)

Biological brains: advantages

- Brains develop in embryo from a single cell: neurogenesis, plasticity
- Balances stability (memory) & plasticity (adaptation)
- Few shot learning is the norm
- Resilient to small input changes
- Robust/adaptive to internal changes
- Runs on ~20W

Why Neuro-AI?

- Sensitive to hyperparameters
- Robust/adaptive to internal changes
Important differences

Deep Learning:
- Real-valued activations propagate
- Learning rules are non-local, require floating point precision
- Synchronous computations
- Converges to a fixed model

Biological brains:
- Discrete “spikes”, precise timing
- Learning is local, weight updates possible with limited precision
- Asynchronous computation
- Dramatic diversity of cell types and connectivity patterns, compositionality (subsystems)
- Multi-scale dynamics
- Dynamical system: nonlinear, coupled, non-equilibrium; activity never stops, even without input
NEURO-AI: INFORMATION BOTTLENECK

X
Neuroscience
- cellular
- systems
- cognitive

M
(models/methods)

Y
AI challenges
- AutoML
- Efficiency: samples, time
- Language understanding, generation
- Lifelong learning
- RL

min $I(X;M)$
max $I(M;Y)$

Compress X into M
most-relevant to Y

An algorithm for interdisciplinary research:
1. Identify Y: problems AI can’t solve (well) today
2. Sample from (infinite?) X: recent (or not) discoveries?
3. Build M (piece of cake 😊)
4. Repeat
LONG-TERM GOAL

Next-generation **AI** based on better understanding of brain functioning including **plasticity, attention, memory, reward processing, motivation** and beyond, while approaching both brain and AI as **non-equilibrium stochastic dynamical systems** rather than fixed predictive models.
Current Focus

AI

Better Learning Algorithms and Neuronal Models

Automated AI
Lifelong, Continual Learning

Advancing RL

Real-Time Behavior
Neuromorphic Hardware

Language Understanding
Dialogue Generation

Neuro-AI

Beyond Backprop
Compartmental neuron models

Neuro-genesis (-evolution)
Stability/Plasticity Models

Reward and Attention Models

Spiking Networks (3rd-gen ANNs)
Nonlinear Dynamical Models

Modeling Psychology of Dialogue
Multicompartment models:

Segregated Dendrites (Bengio NIPS 2018; Lillicrap, Richards; etc)

“Capsules” [Hinton et al]

Neuron Models?

Simple binary neurons [McCullogh & Pitts, 1943]

Learning algorithms? (Credit assignment)

More bio-plausible algorithms:

Target Propagation variants [LeCun, 1986], [Bengio 2015], [Hinton 2018]

Krotov&Hopfield [2018], Chklovski [2018]

[Carreira-Perpinan 2014], [Taylor 2016]

[Our AltMin method – submitted to ICML-2019]
WHAT'S WRONG WITH BACKPROP?

Biologically implausibility:

- Error feedback does not influence neural activity, and hence does not conform to known biological feedback mechanisms underlying neural communication

- Weight transport problem: symmetric weight connectivity for feedforward and feedback directions

- Many other issues (precise clocking between feedforward and backprop phases, violation of Dale’s law, etc)

Computational Issues:

- Vanishing gradients (due to chain of derivatives)

- Difficulty handling non-differentiable nonlinearities (e.g., binary spikes)

- Lack of cross-layer weight update parallelism
ALTERNATIVE: TARGET PROPAGATION


Lee, Zhang, Fischer and Bengio. Difference target propagation. ECML-2015


Unfortunately, backprop still outperforms target prop on standard benchmarks
OUR APPROACH: ALTMIN


Beyond Backprop: Online Alternating Minimization with Auxiliary Variables

Anna Choromanska* 1  Benjamin Cowen* 1  Sadhana Kumaravel* 2  Ronny Luss* 2  Mattia Rigotti* 2
Irina Rish* 2  Brian Kingsbury 2  Paolo DiAchille 2  Viatcheslav Gurev 2  Ravi Tejwani 3  Djalal Bouneffouf 2

Breaking gradient chains with auxiliary activation variables:

• Explicit propagation of activations
• Noisy neuronal activity
• More local updates than backprop
• Parallel, distributed, asynchronous

\[
f(W, C) = \sum_{t=1}^{n} \mathcal{L}(y_t, \sigma_L(c_t^L), W^{L+1}) + \mu \sum_{t=1}^{n} \sum_{l=1}^{L} \| c_t^l - W^l \sigma_{l-1}(c_t^{l-1}) \|_2^2 + \lambda_W \| W^l \|_1 + \lambda_C \| C^l \|_1\]
PROMISING RESULTS

- AltMin matches Backprop performance, often converging faster on fully-connected nets (MNIST, CIFAR-10), RNNs (seqMNIST), LeNet5 (MNIST)

- Online AltMin GREATLY outperforms OFFLINE AltMin methods

Our method LEARNS FASTER than backprop in 1st epoch, then matches BP

---

Similar behavior on some RNNs and sequential data (seqMNIST), And LeNet5 (MNIST)
Adaptation at Different Time Scales

Plasticity: architectural changes

- Adult neurogenesis as an inspiration for hidden-layer adaptation (neuronal birth and death) – e.g., in sparse linear autoencoder (online dictionary learning)

- Next: “ensemble-level” changes: add/delete network “blocks”

Reward-driven Attention: input and architecture selection

- “External” attention: input selection (generalizing visual attention) driven by reward in online decision-making (e.g., contextual bandit setting)

- Next: “internal” attention as dynamic choice of subnetworks
Plasticity: Adult Neurogenesis

- Adult neurogenesis (AN): generation of new neurons in adult brains throughout life (balanced by neuronal death)

- In dentate gyrus of the hippocampus (in humans)

- Increased AN is associated with better adaptation to new environments. But why is it necessary, besides the usual synaptic plasticity (i.e. learning weights in neural nets)?

- Can a computational model of AN support this observation? Can it lead to an adaptive representation learning algorithm?
Baseline: Sparse Autoencoder

- Current neuroscience theories suggest that the hippocampus functions as an autoencoder to create and evoke memories.

- A simple autoencoder model: single-hidden-layer sparse linear autoencoder (classical sparse coding of Olshausen & Field, 1996), also known as dictionary learning model:

- Solved via $l_1$-regularized optimization:

$$\min_{D, C} \|X - CD\|_2^2 + \lambda \sum_i \|C(i, :)\|_1$$
**Neurogenetic Online Sparse Autoencoder: Neuronal Birth and Death**

\[ \hat{f}_t(D) = \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \|x_i - D\alpha_i\|_2^2 + \lambda_c \|\alpha_i\|_1 + \lambda_g \sum_j \|d_j\|_2 + \sum_j \lambda_j \|d_j\|_1 \]

- Reconstruction error
- Sparsity on codings
- \(L_1/L_2\) group sparsity
- Sparse elements

**Reconstruction error too high on new samples?**

- **yes**
  - Neuronal birth (new random elements)

- **no**
  - Memory Update
  - Dictionary Update + Neuronal Death via group sparsity
  - Encode new samples
Experiments in Non-Stationary Environments: Switching Between Different Domains

Images: from urban ("Oxford") to nature (flowers, dogs, cats)

NODL improves reconstruction accuracy of ODL on both old data and learns more compact representations

NODL adapts to change, while not forgetting the past (‘memory’ matrices)
**Reward-Driven Attention: External and Internal**

**“External” attention:** input selection [IJCAI 2017]

- Inspiration: visual attention (foveation, sequence of glimpses)
- Recent neuroscience literature suggests that attention is a reward-driven mechanism which in turn drives the future reward

**“Internal” attention as dynamic execution / network path selection** [AAMAS 2018]

Decision: e.g., image class, type of medication, ad on webpage etc.
very high dimensionality: ~100,000 time series (neurons), even in a collapsed 2D version (average) of the original 3D data

high temporal (0.8Hz) and high spatial (few voxels/neuron) resolution (unlike fMRI/EEG)

oscillatory behavior

Can we learn a model capturing the underlying dynamics of this system?
Can this model predict the temporal evolution of brain activity?
Can it be interpretable – i.e., relate to prior neuroscientific knowledge?
Why use van der Pol?

1. Oscillatory behavior is common in all neural systems; van der Pol is simplest nonlinear oscillator with rich enough behavior.

2. Voltage-like $x_1$ (activity), and recovery-like $x_2$ (excitability), similarly to neuro literature [Izhikevich 2007]

How to best estimate oscillator parameters from data?
Not a mainstream problem - neither in machine learning, nor in physics!
Parameter fitting:
stochastic search + variable-projection optimization [Aravkin et al, 2016]

0.7-0.8 correlation between the model and the actual data

Interpretability: learned $W$ gives Interaction strength among brain areas
PREDICTING FUTURE BRAIN ACTIVITY WITH VAN DER POL AND LSTM

- Issue: van der Pol prediction is so-so; also, LSTM suffers from small data problem
- Solution: hybrid approach outperforms both methods (and baseline VAR model) - data-augmented LSTM with fitted van der Pol simulating more data

<table>
<thead>
<tr>
<th>Calculation</th>
<th>Calcium Imaging (zebrafish)</th>
<th>Functional MRI (people)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE (root mean square error)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hybrid outperforms LSTM, van der Pol and VAR on both calcium and fMRI data
Next-generation AI based on better understanding of brain functioning including plasticity, attention, memory, reward processing, motivation and beyond, while approaching both brain and AI as non-equilibrium stochastic dynamical systems rather than fixed predictive models.
**Cellular neuroscience:** what inspiration can we derive from cellular level mechanisms in the neurons, beyond the classical McCullough and Pitts neuron? Can we derive inspiration for new learning rules and computational mechanisms from the biophysics and cellular machinery of neurons?

**Systems neuroscience:** how does the local network architecture of the brain enable robust, efficient learning and reasoning? Are there clues from the systems-and computational-neuroscience literature for new kinds of architectures for deep learning?

**Cognitive neuroscience:** what findings from psychology and cognitive and behavioral science can inspire us to think differently about problem formulations and potential mechanisms for learning and reasoning?

**Beyond Backprop:** more bio-plausible (and better!) learning models & methods

**Neural Dynamics:** temporal evolution of neural network activity, synaptic strength and structure; spiking networks (3rd-gen ANNs); oscillatory dynamics; continual/lifelong learning

**Behavior Dynamics:** Bio-RL -- adding more bio-plausible decision-making mechanisms to reinforcement learning (attention-reward models; positive vs negative reward processing biases); psychology of dialogue (infogain-driven collaborative dialogue models)
Recent Focus (NIPS, ICLR): Two Main Directions

- **Beyond-backprop: bio-plausible error-propagation and neuron models**
  - Assessing the Scalability of Biologically-Motivated Deep Learning Algorithms and Architectures
  - Initialized Equilibrium Propagation for Backprop-Free Training
  - Dendritic cortical microcircuits approximate the backpropagation algorithm
  - Improved Expressivity Through Dendritic Neural Networks

- **Dynamical Systems Approaches: spiking networks and beyond**
  - Gradient Descent for Spiking Neural Networks
  - Long short-term memory and Learning-to-learn in networks of spiking neurons
  - Deep Rewiring: Training very sparse deep networks
  - **NeurIPS Best Paper Award:** Neural Ordinary Differential Equations
    - Non-temporal! Dynamics over continuous (rather than discrete) layers, not
Surya Ganguli on Future Neuro Inspirations for AI blog post 5/12/18

- Biologically plausible credit assignment
- Incorporating synaptic complexity
- Taking cues from systems-level modular brain architecture
- Unsupervised learning, transfer learning and curriculum design
- Building world models for understanding, planning, and active causal learning
- Achieving energy-efficient computation in a post Moore’s law world
- Seeking universal laws governing both biological and artificial intelligence
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
**POPULAR TOPIC:**

**THE RISE OF NEURO-AI?**

- **Increasing number of conferences, workshops, papers (NIPS, etc) on the topic**

- **Surveys:**
  - Steve M. Potter (Georgia Tech): *What Can AI Get from Neuroscience?* 2007
  - DeepMind: *Neuroscience-inspired AI*, Hassabis et al, 2017
  - Stanford (S. Ganguli): *The intertwined quest for understanding biological intelligence and creating artificial intelligence*. blog post 5/12/18

Is there enough momentum for a breakthrough beyond deep nets? We believe so. But someone has to make it 😊

It could be us – if we really set this as a priority