Fakultät für Ingenieurwesen
Facoltà di Ingegneria
Faculty of Engineering

Recognizing Human Actions and Goals in an Open Environment – A Brain-Inspired Approach

03/07/2024

Franz Alexander Van-Horenbeke Echevarria

Supervisor: Angelika Peer
Second Supervisor: Tamim Asfour

Ph.D. in Advanced-Systems Engineering 35th cycle

A Introduction

B Methods and Results

B.1 The NILRNN

B.2 The HLRNN

B.3 The U-LRNN

C Conclusion

* part of the content (e.g., problem formalization, model behavior analysis, etc.) has been left out due to lack of time

A Introduction

B Methods and Results

B.1 The NILRNN

B.2 The HLRNN

B.3 The U-LRNN

C Conclusion

Action and Goal Recognition

Humans

Innate ability

Allows us to **understand** the state and **predict** the behavior of others

Outstanding performance

Related skills: **understand** and **learn new** actions/goals, **adapt** to execution changes, etc.

Machines

Challenging: uncertainty, variability, incomplete knowledge, missed events, etc.






Common Approaches

Hybrid logic-probabilistic

mainly for plan/goal recognition

✓
- Highly structured
- Highly expressive
- Generative

✗
- Require much manual work
- Rigid
- Bad at generalizing

Deep learning

mainly for action recognition

✓
- Very flexible
- Deal well with sensory input
- Deal well with uncertainty
- Hierarchical

✗
- Require much labeled data
- Hard to interpret
- Bad at dealing with unknown actions

Foundation model-based

✓
- Zero-shot
- Contain much general knowledge

✗
- Require much computation
- Bad to learn online

In general

✗

Dynamic open environments

Introduction

Motivation

Our brain

Very good at dynamic open environments

Brain-inspired systems

- Typically very application-specific

Approach

Work with general mechanisms and models of regions of the brain that apply to this and other problems

Introduction

Objective

Develop an **action and goal recognition system** for real unconstrained environments

Develop a **new unsupervised cognitive framework** inspired by known mechanisms from the brain

Develop a system able to **recognize known actions and goals** based on this cognitive framework

Adapt the system to other **fully unsupervised tasks** such as action prediction or selection

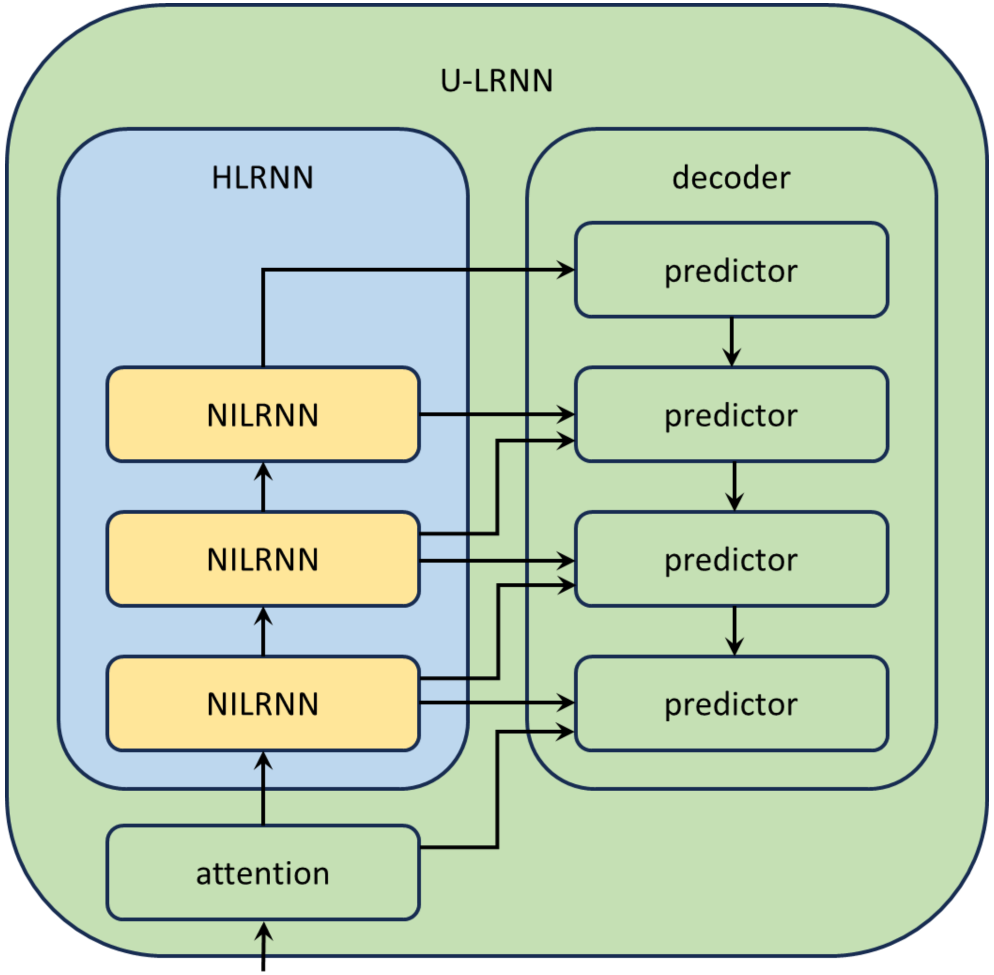
Introduction

Outline

NILRNN
Neocortex inspired locally recurrent neural network [1,2]
- **Shallow** self-supervised representation learning system for temporal data
- **Model** of the primary visual cortex

HLRNN
Hierarchical locally recurrent neural network [3]
- **Deep** stack of NILRNNs

U-LRNN
U-shaped locally recurrent neural network
- **Encoder-decoder** architecture with HLRNN as encoder
- For action and goal **prediction** and **selection**



[1] Van-Horenbeke, Franz A., and Angelika Peer. "NILRNN: a neocortex-inspired locally recurrent neural network for unsupervised feature learning in sequential data." *Cognitive Computation* 15.5 (2023): 1549-1565.
[2] Van-Horenbeke, Franz A., and Angelika Peer. "The Neocortex-Inspired Locally Recurrent Neural Network (NILRNN) as a Model of the Primary Visual Cortex." *IFIP AIAI*. Cham: Springer International Publishing, 2022.
[3] Van-Horenbeke, Franz A., and Angelika Peer. “HLRNN: building a hierarchy of locally recurrent neural networks for self-supervised representation learning in temporal data.“ Manuscript submitted for publication.

A Introduction

B Methods and Results

B.1 The NILRNN

B.2 The HLRNN

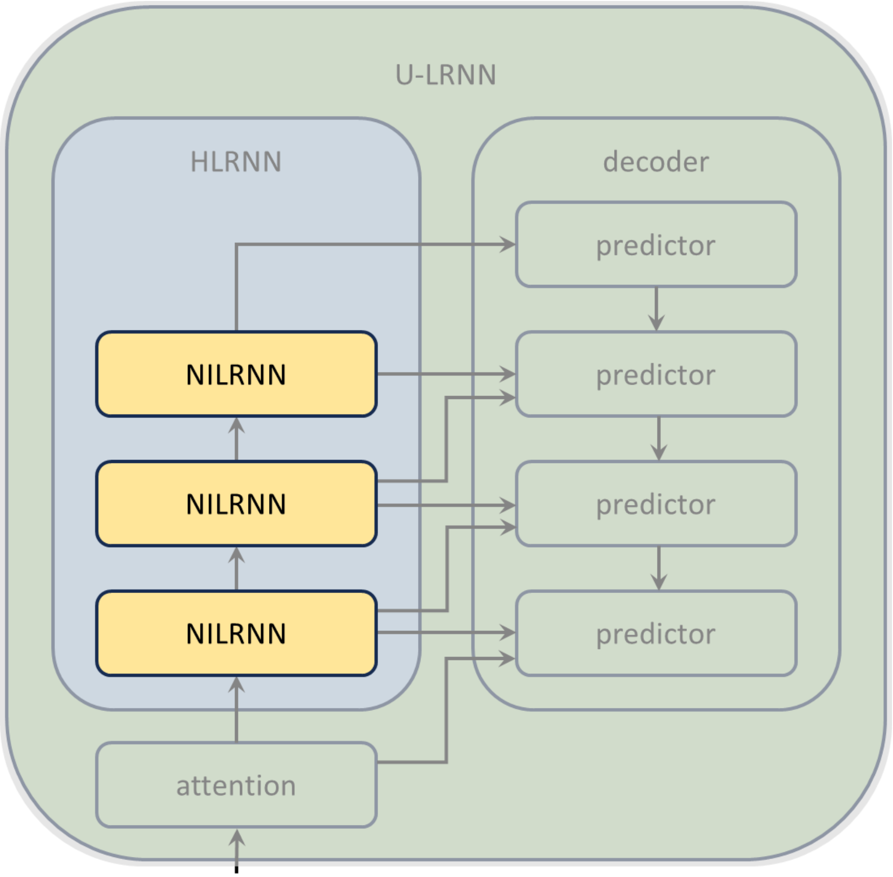
B.3 The U-LRNN

C Conclusion

The NILRNN

The Neocortex-Inspired Locally Recurrent Neural Network

- Main elementary **block**
- **Shallow** self-supervised representation learning system
- Inspired by areas of the **neocortex**
- Learns structure from **temporal** data
- Tested on data from different **domains**
- **Outperforms** other shallow systems
- Shows **analogous** behavior to the primary visual cortex



The NILRNN

The Neocortex

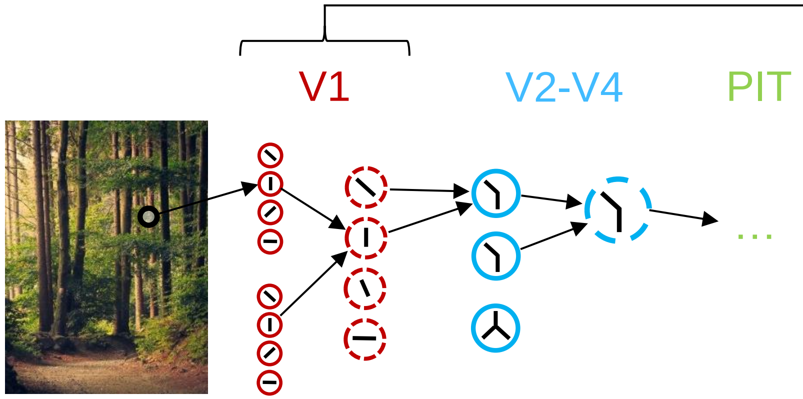
Involved in **high-level** cognitive **tasks**

Distributed in **areas**

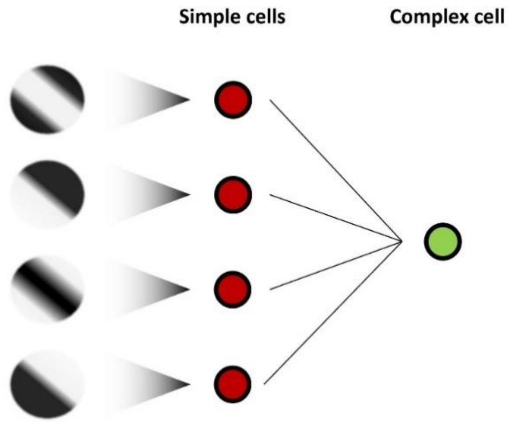
Organized **hierarchically**

Quite **uniform**

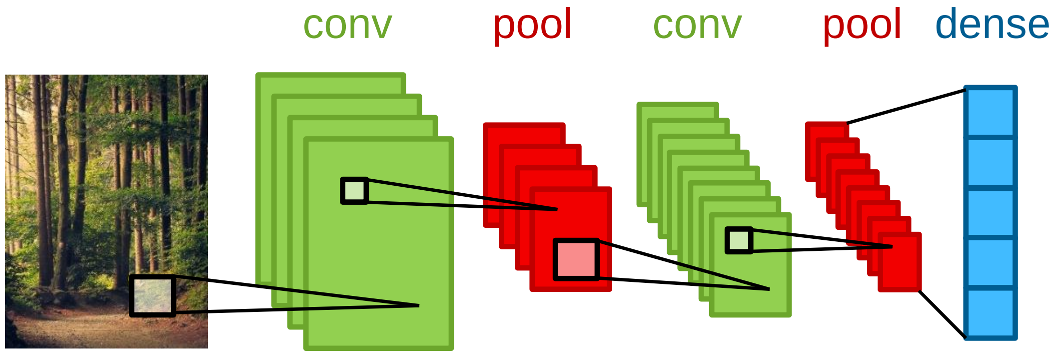
CNNs as Models of the Visual Cortex



Model of the visual cortex



Model of the primary visual cortex



CNN

Spatial pooling

- Presence of pattern: relevant
- Exact position: irrelevant low-level information

Models of the Primary Visual Cortex

Model by Antolík and Bednar (2011)[1]

Achieves **orientation order** and **phase disorder**

Uses **realistic** patterns of **connectivity**

Relies on **shifted patterns** occurring **close in time**

This pooling

- Presence of sequence of patterns: relevant
- Exact pattern: irrelevant low-level information

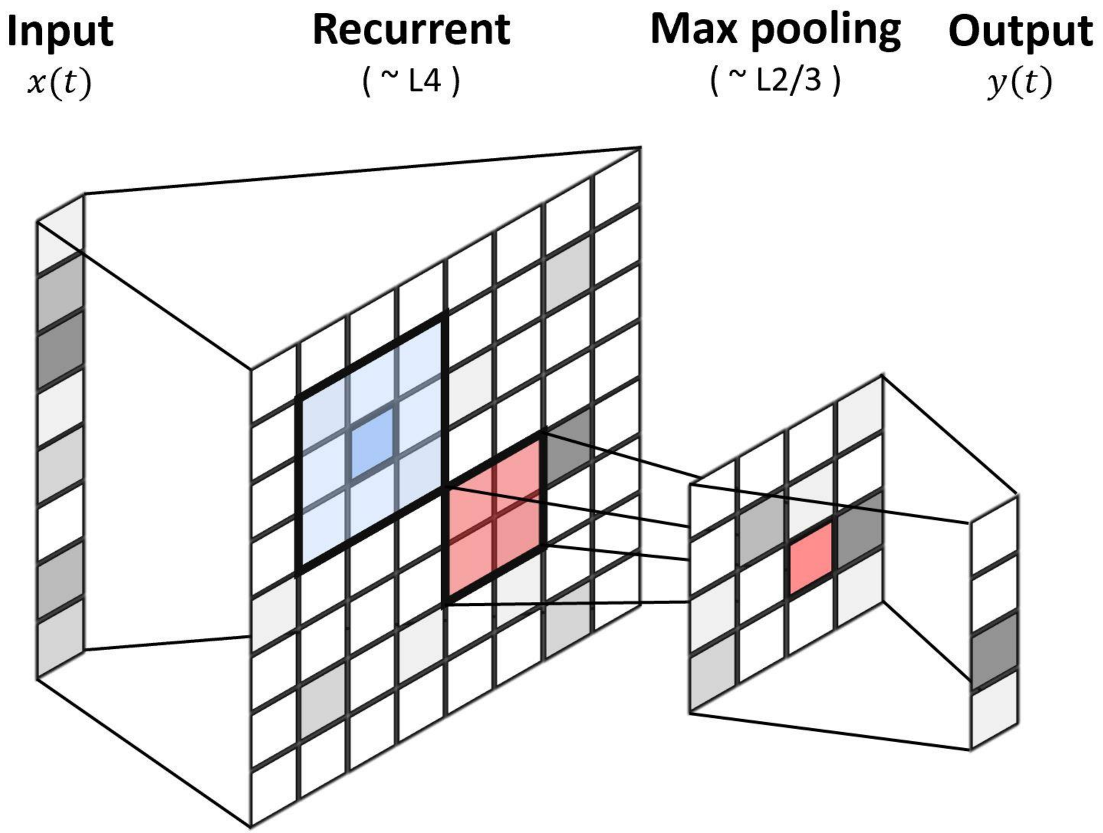
Different from temporal pooling

- **Generalization** of spatial pooling
- Potential mechanism describing other **neocortical areas**

- **Unsupervised representation** learning system
- **Sparse** representations
- **Semantic** order

[1] Antolik, Jan, and James A. Bednar. "Development of maps of simple and complex cells in the primary visual cortex." Frontiers in computational neuroscience 5 (2011): 17.

The Feature Extraction System



Fully connected input

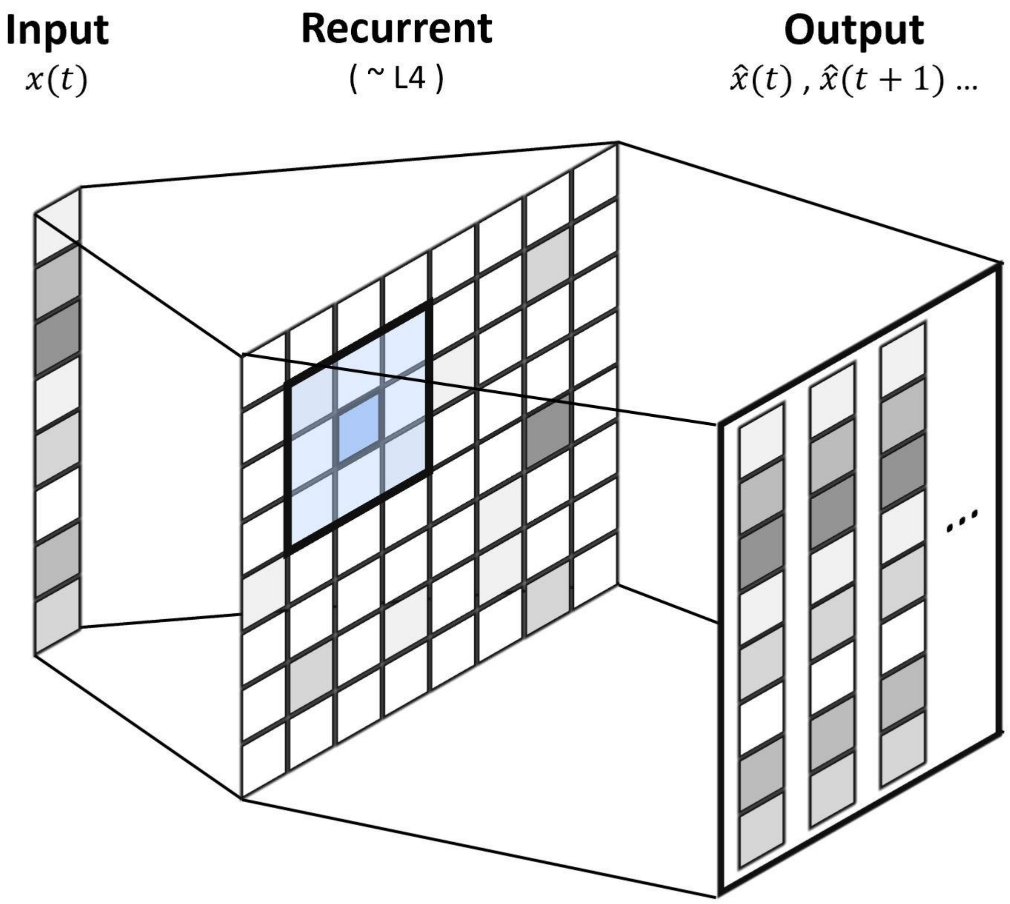
2D locally connected recurrent layer

Circular shape kernels

Sigmoid activation functions

Designed to get sparse inputs

The Self-supervised Learning System



Self-supervised learning through input reconstruction and prediction

Loss function:

$$J(W,b) = J_{error} + \lambda \cdot J_{regularization} + \beta \cdot J_{sparse}$$

$$J_{error} = \frac{1}{2m} \sum_{i=1}^{m} \| \sqrt{w_{\hat{x}}} \circ (h_{W,b}(x_i) - y_i) \|_2^2$$

$$J_{regularization} = \frac{1}{2} \|W\|_2^2$$

$$J_{sparse} = \sum_{i=1}^{s_{hidden}} D_{KL}(\rho || \hat{\rho}_i)$$

The NILRNN

Data Inputs

Comparison with other systems

Dataset	Type	Preproc.	Sparse	Sample size	# samples	# classes
WARD	actions (inertial)	no	no	25	565,755	13
FSDD	speech	spectrogram	yes	40	126,750	10
Synth. actions	actions	grid + att.	yes	55	$\sim\infty$	4

Comparison against the primary visual cortex

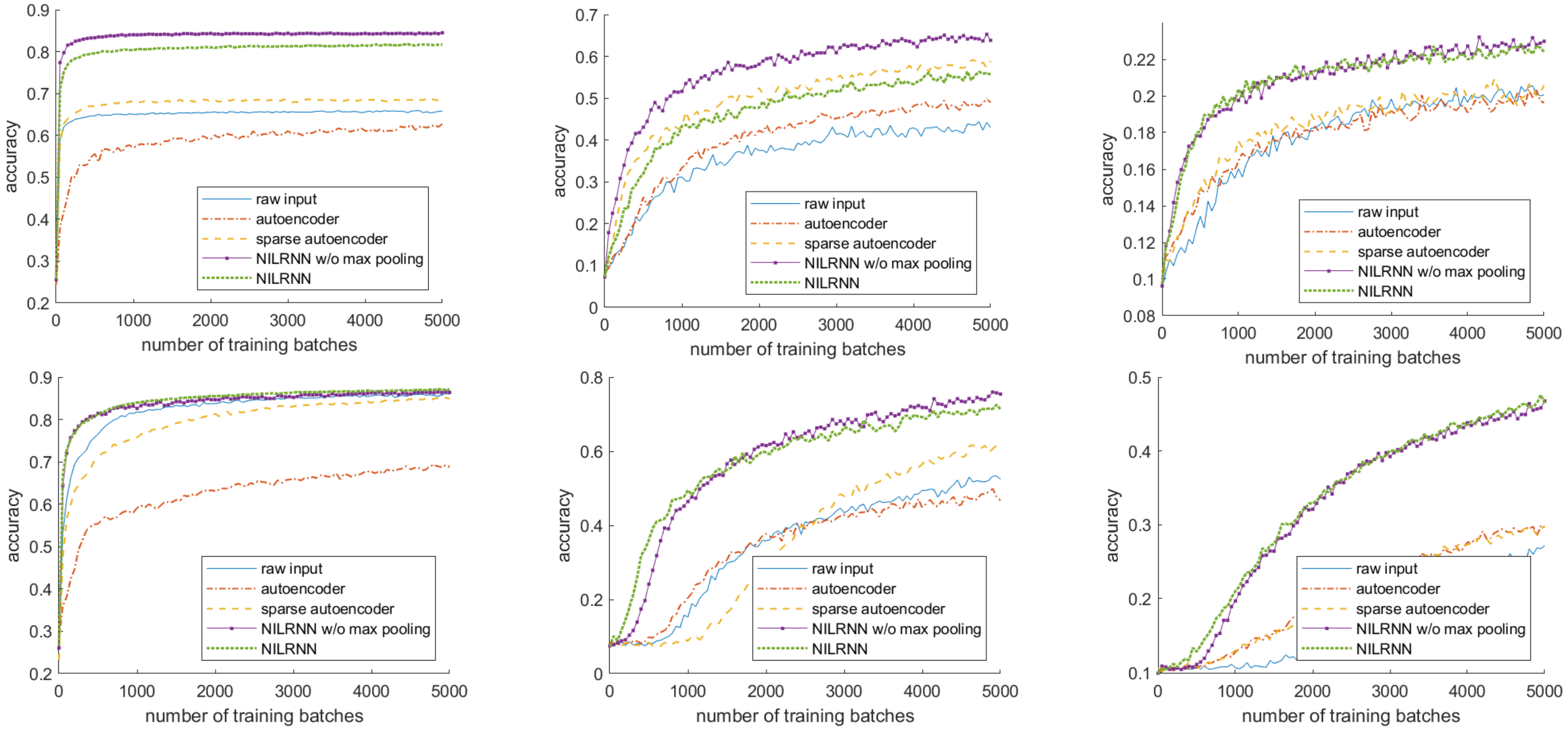
Sequences of 16×16 shifting patches of whitened natural images

Comparison with Other Systems

Hyperparameters chosen using genetic algorithm

Linear

RNN



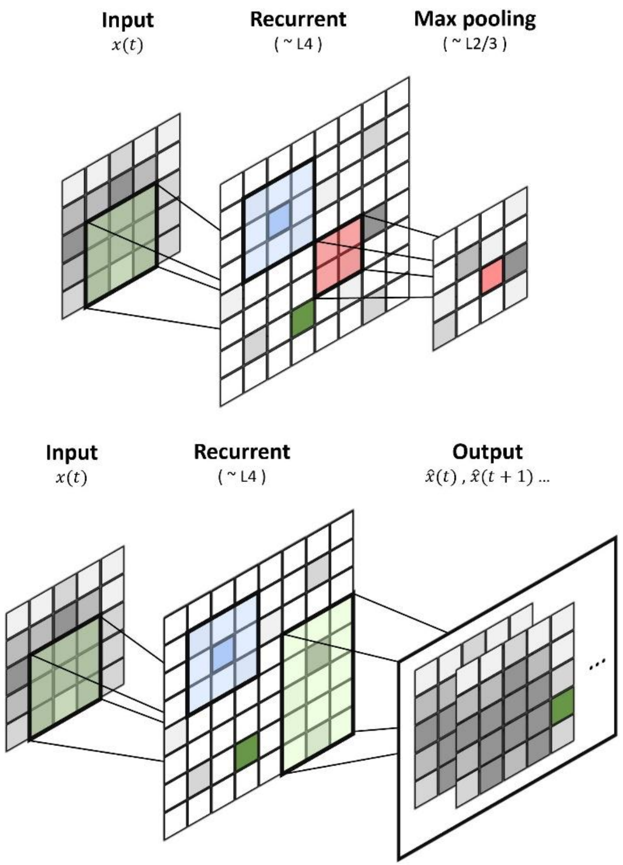
Synthetic action input

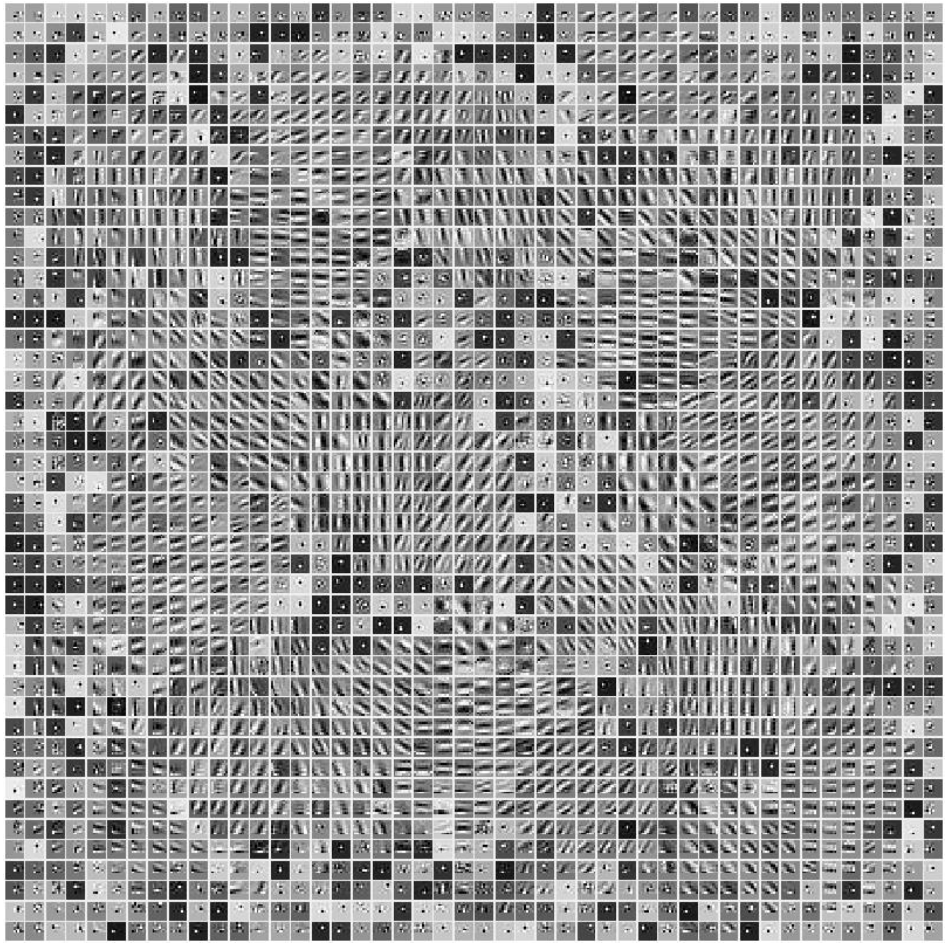
WARD (inertial) dataset

FSDD (speech) dataset

Our system **outperforms** all other systems

Comparison Against the Primary Visual Cortex





Normalized learned input weights

Our system learns **edges** with the expected **order**

Conclusion

NILRNN: neocortex-inspired shallow self-supervised representation learning system for temporal data

Images

Behavior analogous to the primary visual cortex

- Desired behavior
- Valid model of it

Other data

Outperforms other shallow self-supervised learning systems

- Probably desired behavior
- Potential model of other neocortical areas

Further steps

Further analysis

- Max pooling layer
- Non-sparse input
- Modifications
- Neocortex comparison
- …

Build hierarchy

A Introduction

B Methods and Results

B.1 The NILRNN

B.2 The HLRNN

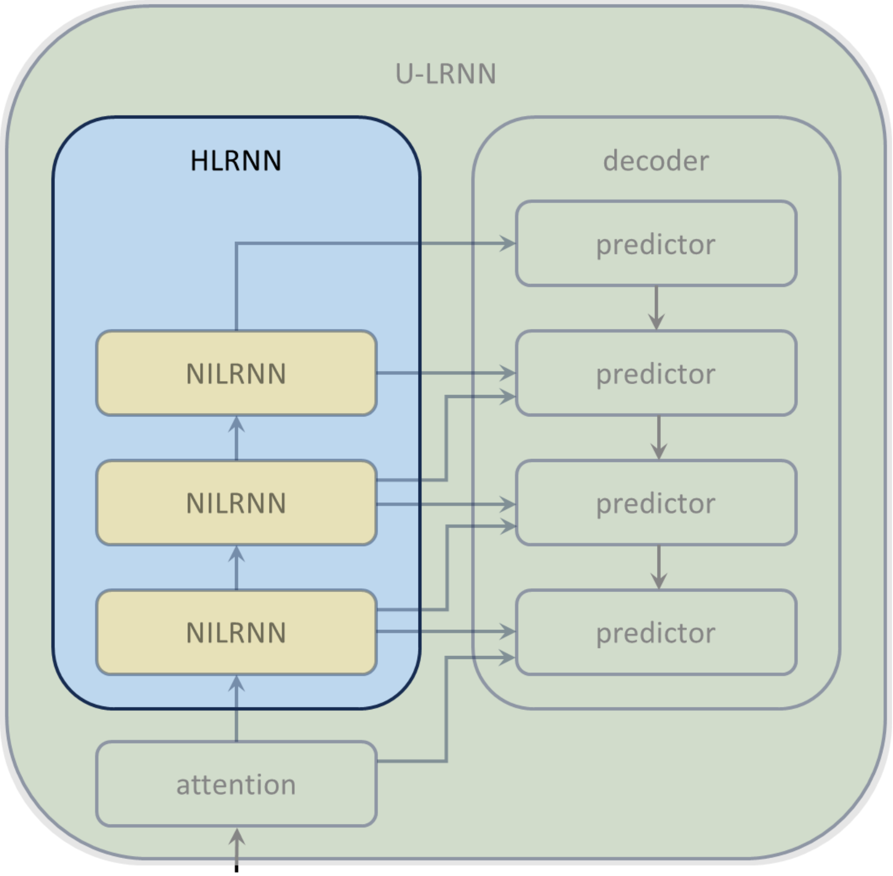
B.3 The U-LRNN

C Conclusion

The HLRNN

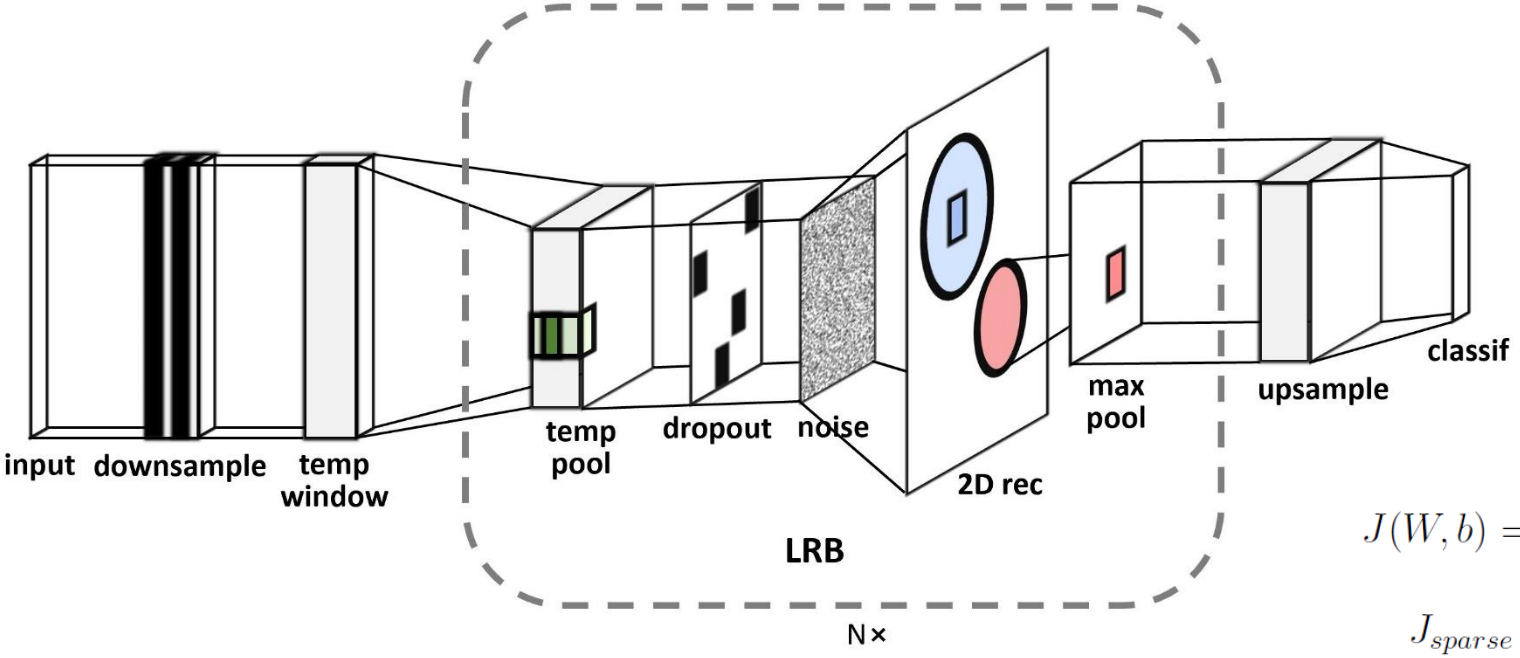
The Hierarchical Locally Recurrent Neural Network

- **Hierarchical** self-supervised representation learning system
- **Stack** of enhanced NILRNNs
- **Mimics** feedforward circuits of hierarchies of the neocortex
- Tested on data from different **domains**
- **Outperforms** other SotA systems
- Shows **expected** hierarchical behavior



The HLRNN

The Architecture



Stack of LRBs (robust downsampling version of NILRNN)

Trained in a greedy way

Deep LRB variant for dense input

Loss function:

$$J(W,b) = J_{error} + \lambda \cdot J_{regularization} + \beta \cdot J_{sparse} + \gamma \cdot J_{slowness}$$

$$J_{sparse} = \frac{1}{m} \sum^{m} \|a_i^{(r)}\|_1$$

$$J_{slowness} = \frac{1}{2 \cdot \delta \cdot (m-\delta)} \sum_{i=1}^{m-\delta} \sum_{j=1}^{\delta} \|a_i^{(p)} - a_{i+j}^{(p)}\|_2^2$$

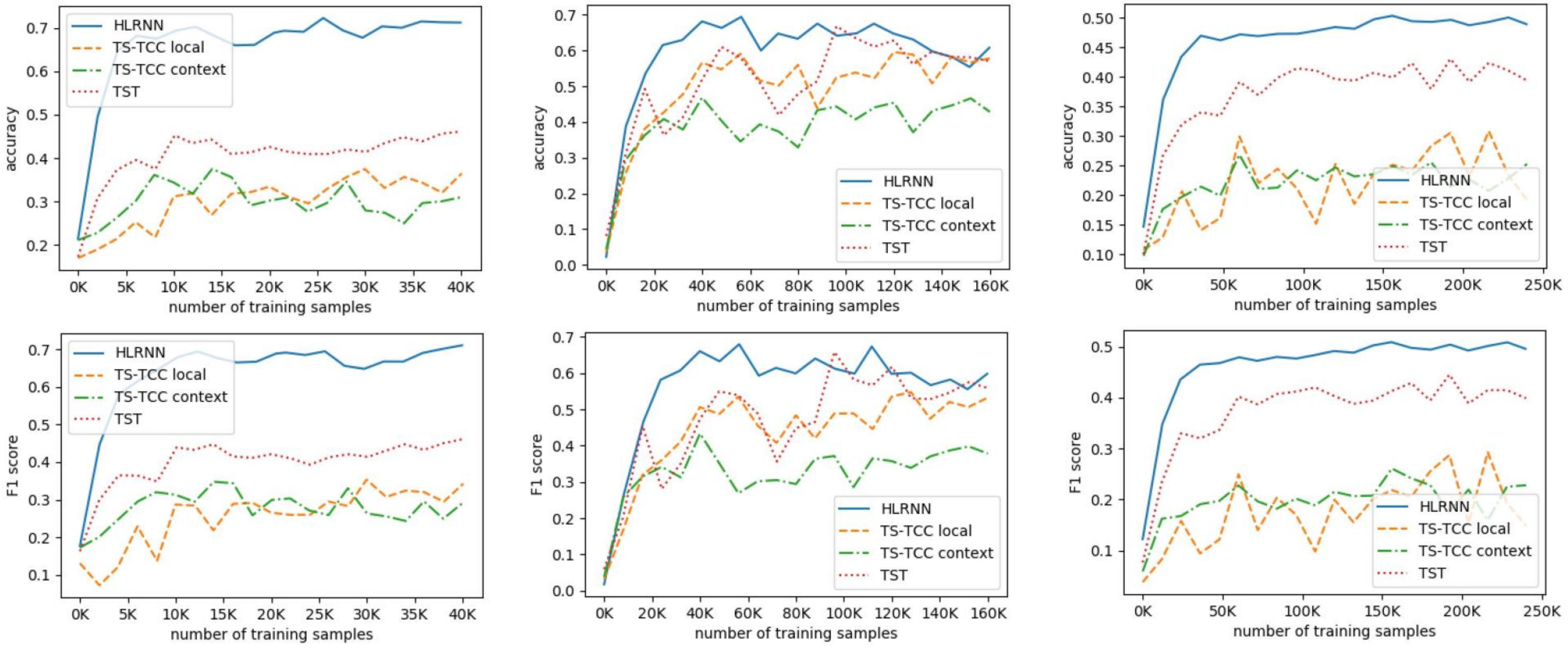
The HLRNN

Comparison with Other Systems

Hyperparameters chosen using Bayesian optimization

Accuracy

F_1 score



Synthetic plan input | WARD (inertial) dataset | FSDD (speech) dataset

Our system **outperforms** all other systems

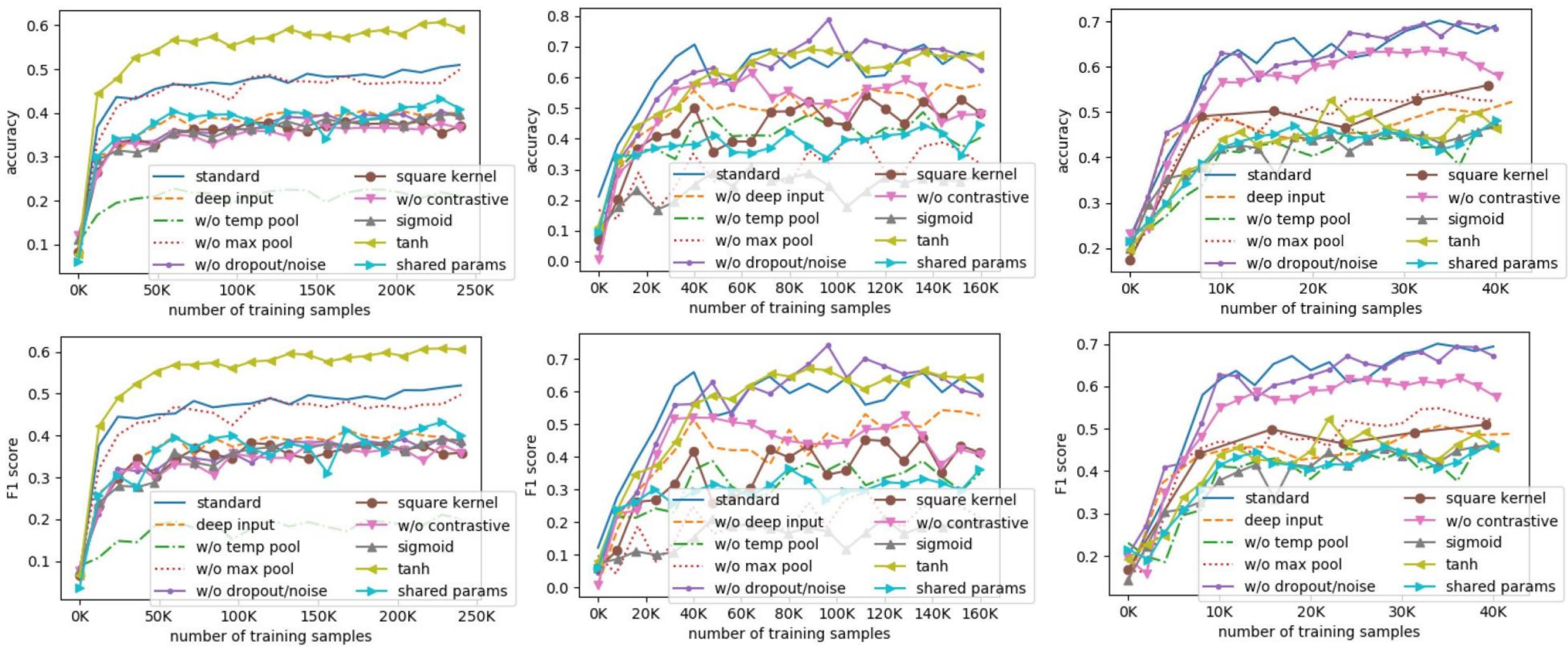
The HLRNN

Ablation Study

Hyperparameters chosen using Bayesian optimization

Accuracy

F_1 score



Synthetic plan input

WARD (inertial) dataset

FSDD (speech) dataset

tanh variant reaches performances **similar** to ReLU

The HLRNN

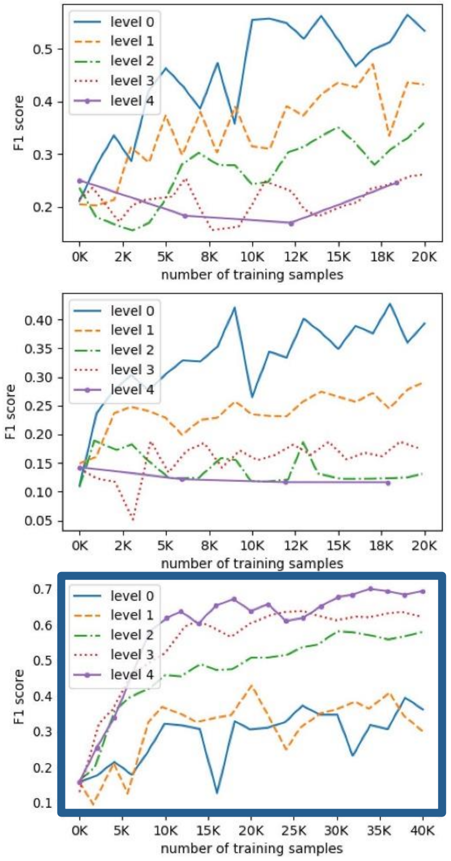
Hierarchy Analysis

Hyperparameters chosen using Bayesian optimization

Primitives

Actions

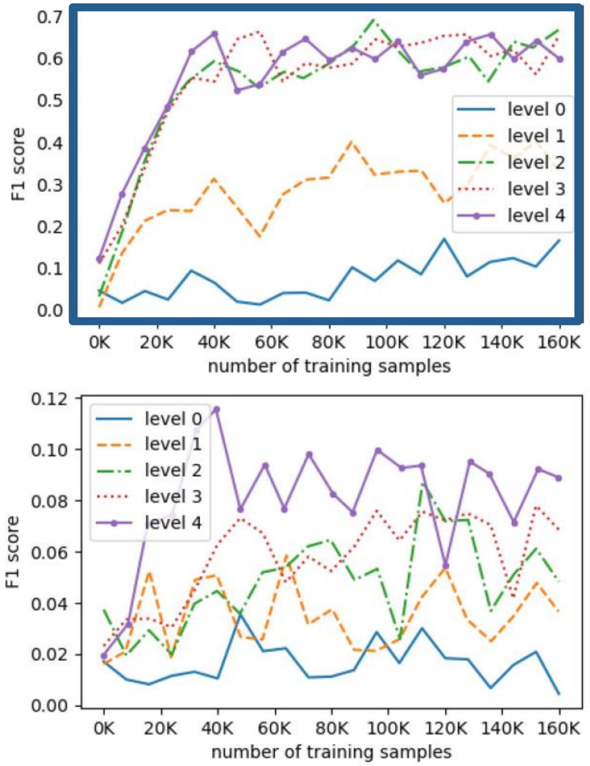
Goals



Synthetic plan input

Actions

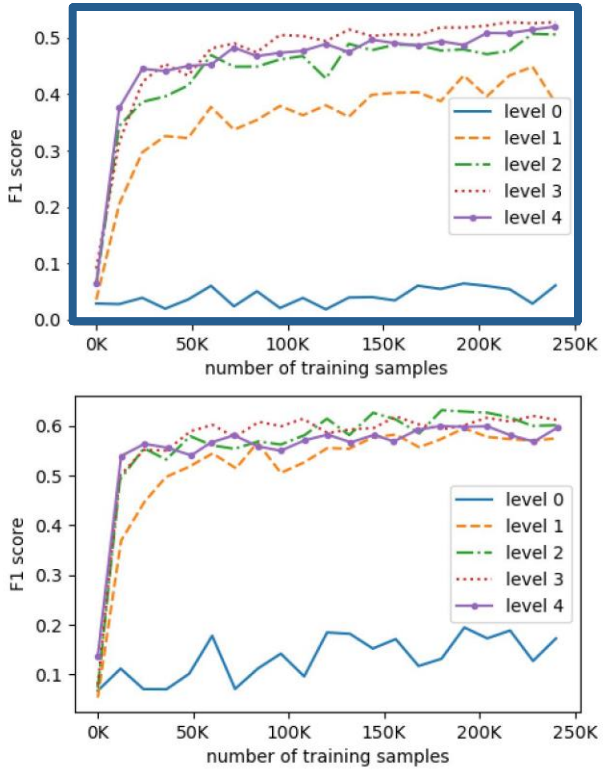
Subjects



WARD (inertial) dataset

Digits

Subjects



FSDD (speech) dataset

For the right configuration, the **hierarchy** works as **desired**

Conclusion

HLRNN: hierarchical self-supervised representation learning system for temporal data

HLRNN

- Outperforms other SotA self-supervised learning systems on different domains
- Potential model of neocortical hierarchies

LRB

- Works at different levels
- Successful improvement of NILRNN

Further steps

Further analysis
- ReLU vs. tanh
- General-purpose representations
- …

Extend functionality
- Encoder-decoder

A Introduction

B Methods and Results

B.1 The NILRNN

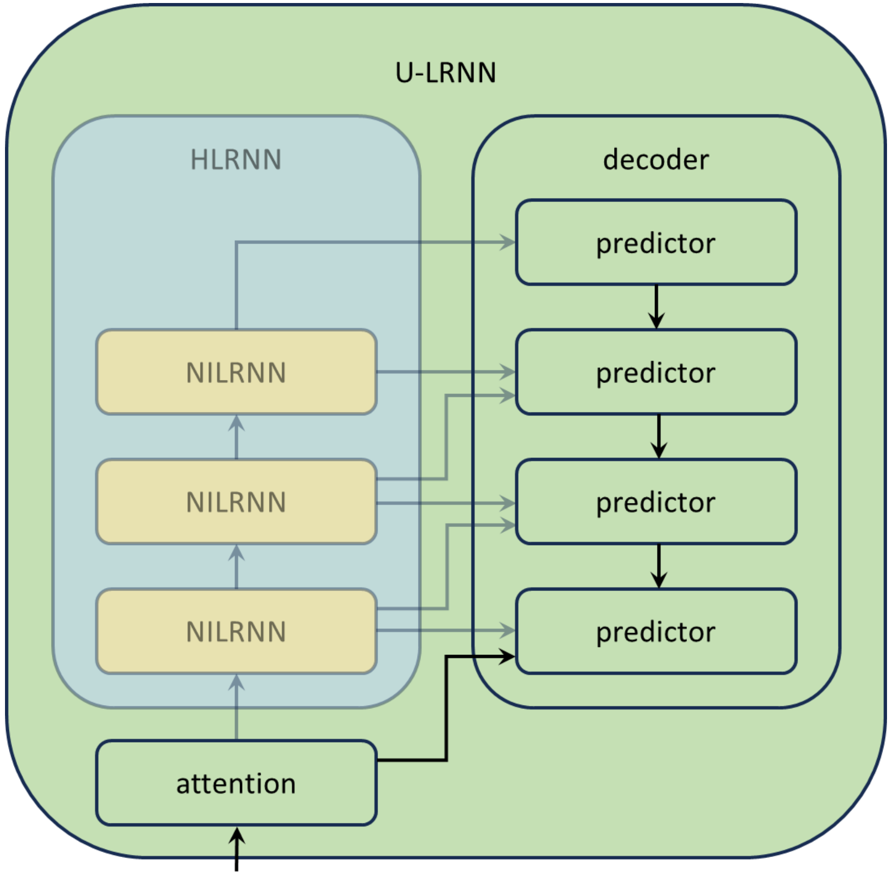
B.2 The HLRNN

B.3 The U-LRNN

C Conclusion

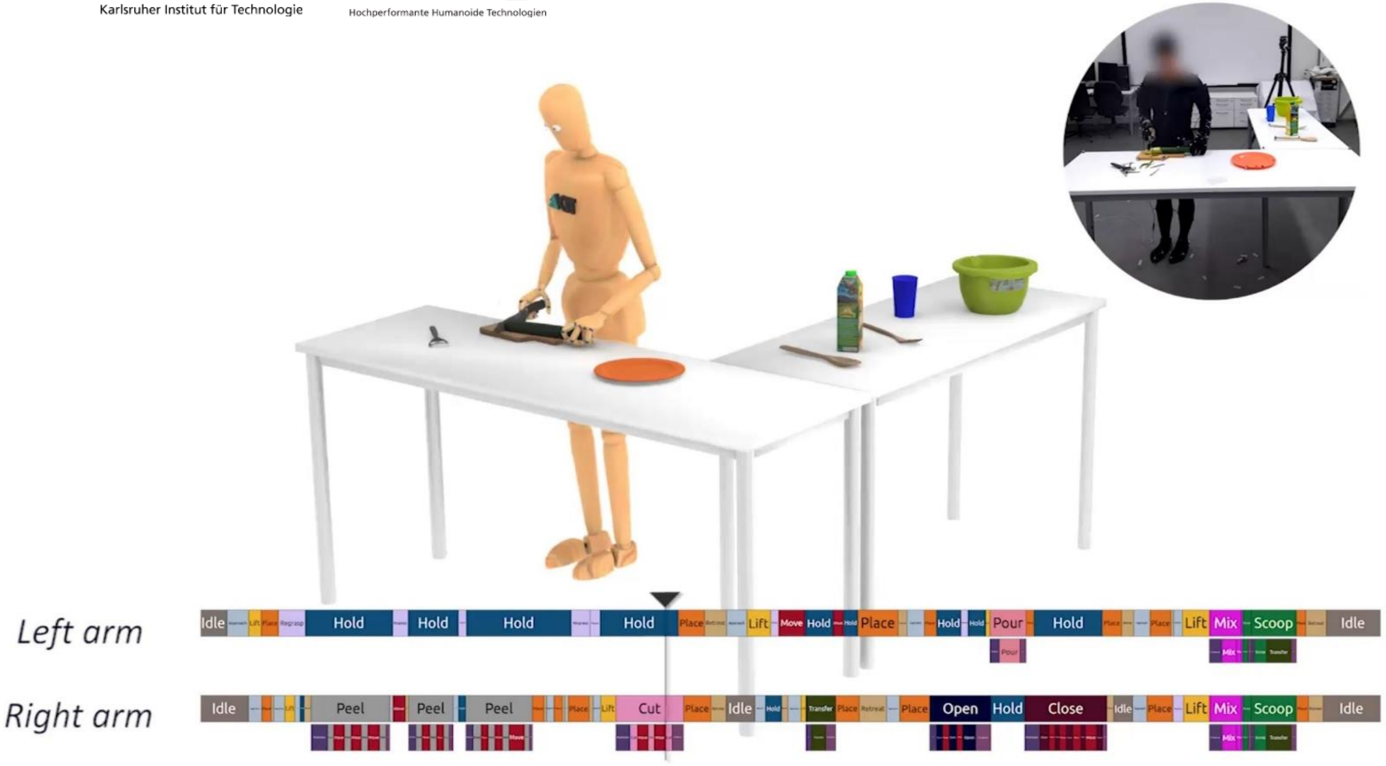
The U-shaped Locally Recurrent Neural Network

- Self-supervised **encoder-decoder** architecture
- HLRNN as **encoder**
- Multi-horizon probabilistic **predictive** decoder
- Includes input self-supervised **attention** learning block
- For action **prediction** and **selection**
- **Mimics** feedforward and feedback circuits of hierarchies of the neocortex



The U-LRNN

The Extended KIT Bimanual Manipulation Dataset



Contains recordings of subjects performing kitchen **actions** and **plans**

Multi-modal

Segmented and **labeled** at different **levels** of abstraction

Designed for tasks such as **imitation** learning and human motion **analysis**

Limitation: too simple classification

Enhancing the KIT Dataset

Only **class-specific** objects present

- **Add** objects dynamically

Most subjects **right-handed**

- Randomly **mirror**

Classes very **different** from each other

- Define **new** classes
- Perform new **recordings** (in collaboration with H2T)

New Recordings

Data Augmentations

Sparse Data Representation

Designed to easily integrate **new objects**

Expressed in an **egocentric** reference frame

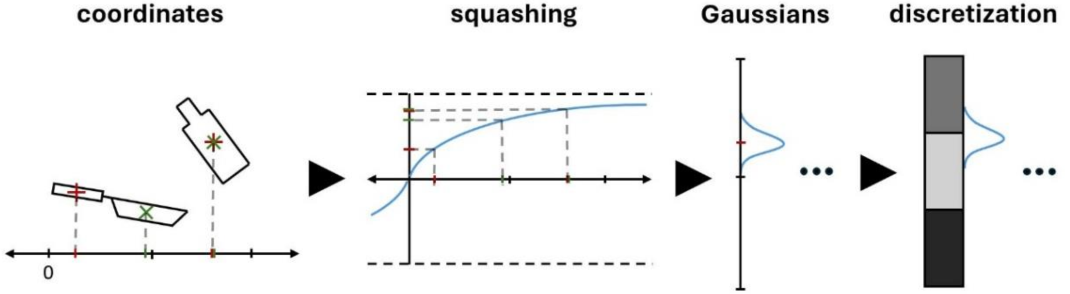
Admits **symmetry** invariant representations

fixed					
torso		head		hand (x2)	
pos*	yaw*	pos	rot (/2)	pos	rot

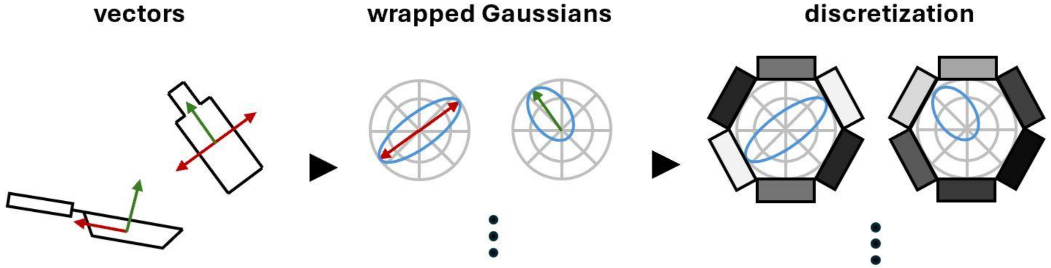
variable			
object1			object2…
id	pos (x2)	rot	…

geom.	long.	empty	open	cont.	handle	sharp	mater.

Identifier representation



Position representation



Orientation representation

The Self-supervised Attention Learning System

Multi-head attention system for sparse data

Loss function: $J(W,b) = J_{error} + \lambda \cdot J_{regularization} + \psi \cdot J_{focus}$

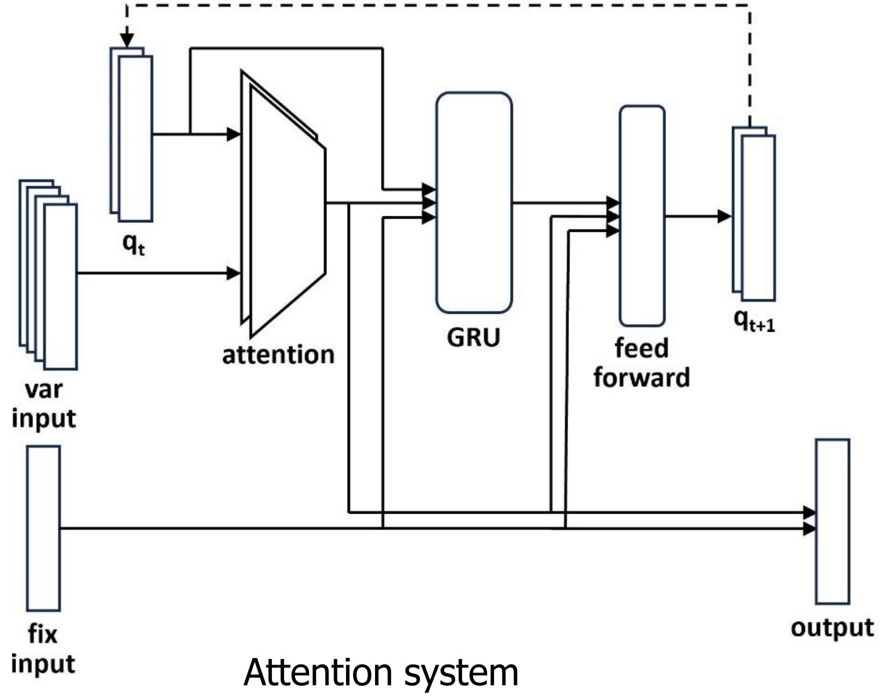
$$J_{focus} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{h}(\|w_{i,j}\|_1 - \max(w_{i,j}))$$

Results

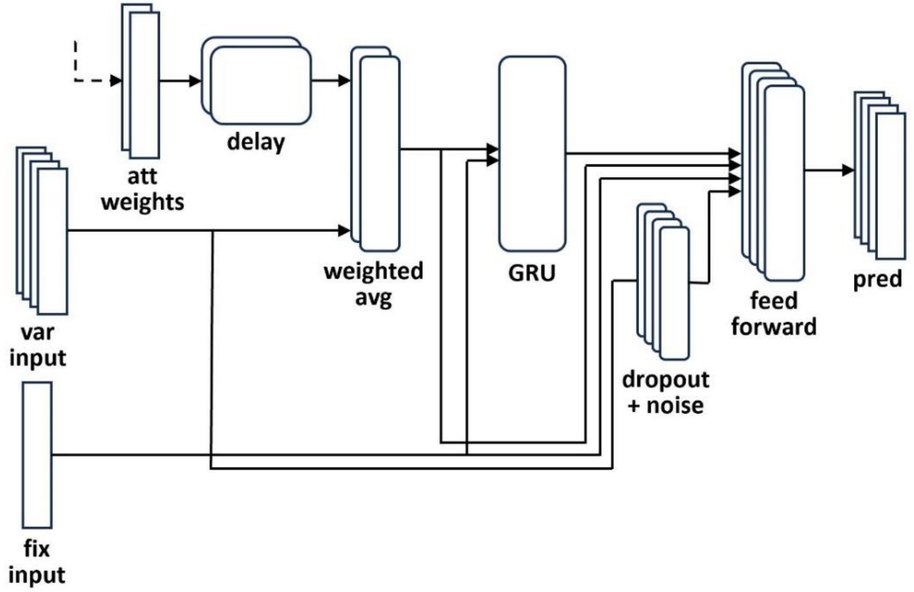
Good general observed behavior

68.2% of time focused on main object

Average max weight of **0.972**

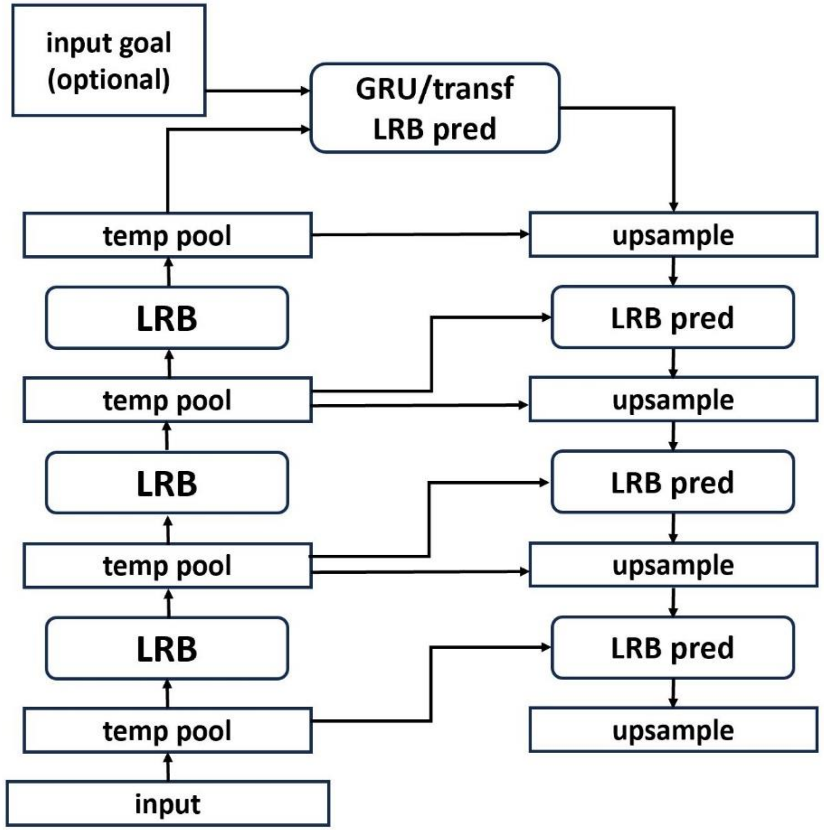


Attention system



Additional training circuit

The Architecture



Multi-level one-step-ahead predictive decoder

Mixture of rectified Gaussian distribution predictions

Multi-horizon through sampling and refeeding

Predictions rely on current and context information

Multi-purpose: action and goal recognition, prediction, and selection

Loss function: $J(W,b) = J_{NLL} + \lambda \cdot J_{regularization}$

$$J_{NLL} = -\frac{1}{m}\sum_{i=1}^{m} log\left(\sum_{j=1}^{n} \pi_{i,j} \cdot \prod_{k=1}^{s} f_{NR}(y_{i,k};\mu_{i,j,k},\sigma^2_{i,j,k})\right)$$

Conclusion

U-LRNN: neocortex-inspired self-supervised encoder-decoder for action and goal recognition, prediction, and selection

U-LRNN

- Multi-level
- Multi-purpose
- Multi-horizon
- Probabilistic
- Flexible/extendable
- Potential model of neocortical hierarchies

Input

- KIT dataset extension
- Augmentations
- Sparse representation
- Self-supervised attention system

Further steps

- Further analysis
- Extensions/adaptations
- Implementation in autonomous agent/robot
- Brain-like modifications

A Introduction

B Methods and Results

B.1 The NILRNN

B.2 The HLRNN

B.3 The U-LRNN

C Conclusion

Conclusion

Summary of Contributions

Multi-purpose flexible and adaptable self-supervised learning brain-like architecture for action and goal recognition, prediction, and selection in real dynamic open environments

Other

- **SotA** analysis
- Problem **formalization**
- **Synthetic** actions and plans input + simulation environment
- NILRNN **behavior** analysis

NILRNN

Shallow self-supervised representation learning system for temporal data outperforming others of its kind

- **Model** of the primary visual cortex
- Novel **semantic pooling** mechanism

HLRNN

Self-supervised representation learning system for temporal data outperforming SotA systems

- Learns representations at different **levels**
- **Analogous** to neocortical feedforward circuits
- NILRNN improvements (**LRB**)
- NILRNN as building **block**
- Novel **slowness** loss term

U-LRNN

Self-supervised encoder-decoder for action and goal recognition, multi-horizon probabilistic prediction, and selection

- **Analogous** to neocortical hierarchies
- **Extendable** to other applications and domains
- Self-supervised **attention** learning system for temporal data
- KIT **dataset extensions** for action recognition
- **Symmetry**-invariant motion **sparse** representation

Conclusion

Future Directions

Design

Further analysis
- Internal behavior
- Neocortex comparison
- Testing on different domains

Improvements

Extension

- High-level reasoning
- Cognitive attention
- Reinforcement learning
- Multimodality
- Developmental
- Human-robot interaction

More Brain-like

Architecture
- Merge encoder and decoder

Mechanisms
- Hebbian learning
- Spiking neural network

This improvements may lead to a better performing and more brain-like system and to an advancement in AI and cognitive neuroscience

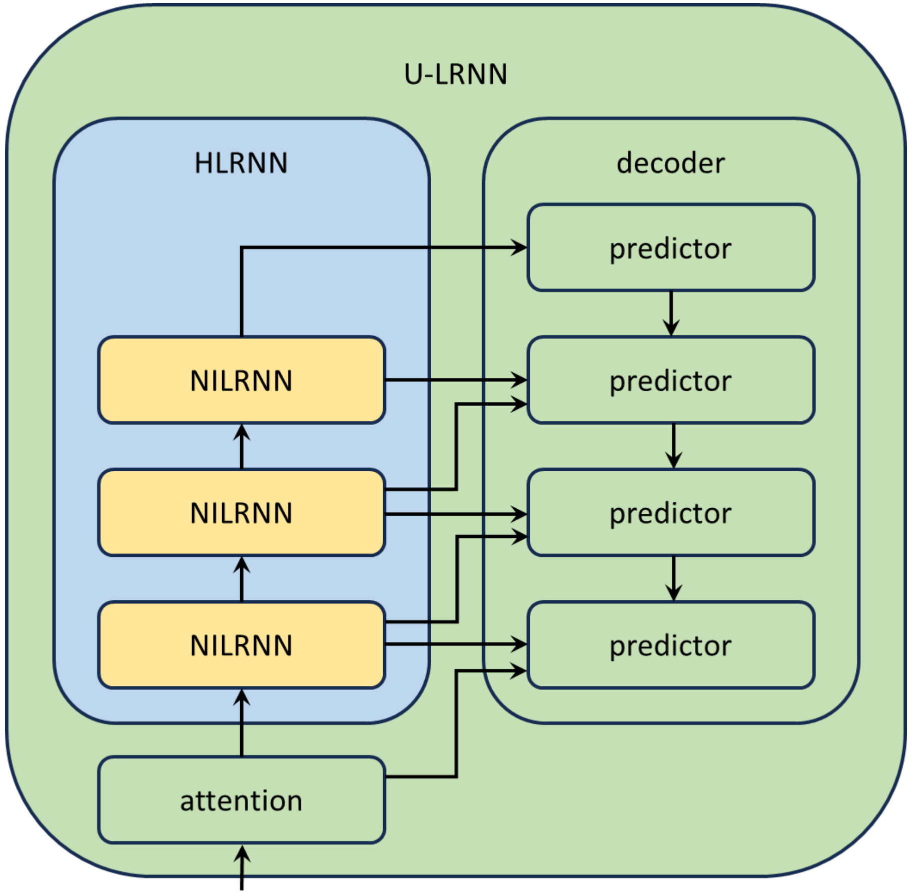
Conclusion

Neocortex-inspired self-supervised representation learning system for **action** and **goal recognition**, **prediction** and **selection**

Flexible and versatile:

- Good performance on **different domains** with temporal data
- Adaptable to **real world online** applications
- Extendable to **multiple tasks**

Its **analogous** behavior to the **neocortex** makes it a valid model of it



Conclusion

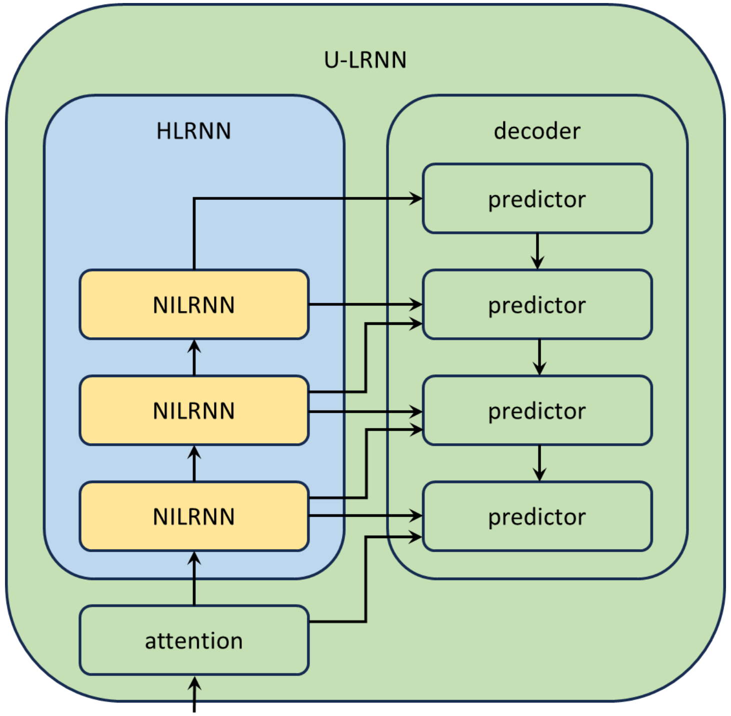
Publications

Journal papers

Name	Journal
Activity, Plan, and Goal Recognition: A **Review**	Frontiers in Robotics and AI
NILRNN: A Neocortex-Inspired Locally Recurrent Neural Network for Unsupervised Feature Learning in Sequential Data	Cognitive Computation
HLRNN: Building a Hierarchy of Locally Recurrent Neural Networks for Self-Supervised Representation Learning in Temporal Data	(Submitted)

Conference papers

The Neocortex-Inspired Locally Recurrent Neural Network (**NILRNN**) as a Model of the Primary Visual Cortex	AIAI 2022



Conclusion

Courses

Name	CFU
Theory of Scientific Method	3,00
Advanced Scientific English	3,00
Advanced Statistics	3,00
Machine Learning	6,00
Decision Making and Support Systems	6,00
Series of Lectures	2,00
Total:	**23,00**

Fakultät für Ingenieurwesen
Facoltà di Ingegneria
Faculty of Engineering

Recognizing Human Actions and Goals in an Open Environment – A Brain-Inspired Approach

03/07/2024

Franz Alexander Van-Horenbeke Echevarria

Supervisor: Angelika Peer
Second Supervisor: Tamim Asfour

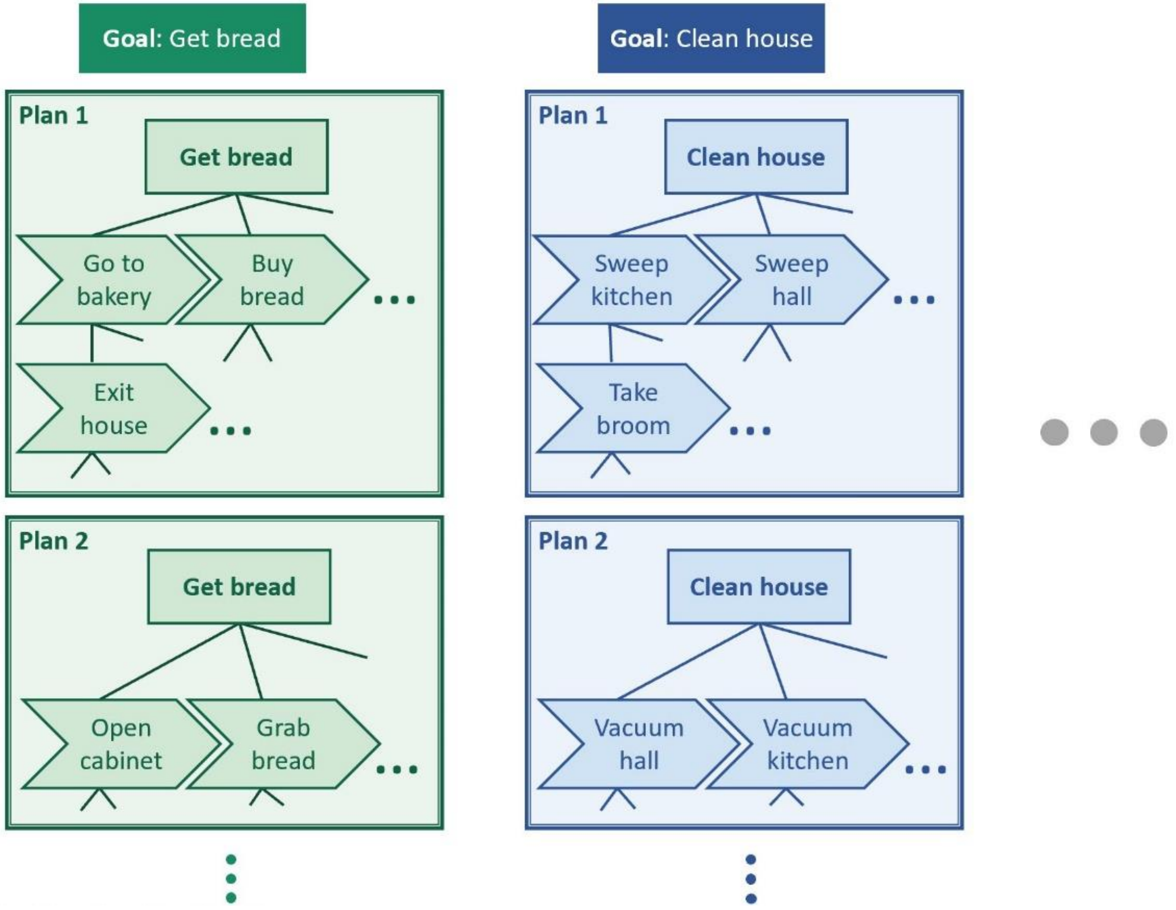
Ph.D. in Advanced-Systems Engineering 35th cycle

Conclusion

Funding Sources

This research was supported by the Euregio project OLIVER (Open-Ended Learning for Interactive Robots) with grant agreement IPN86, funded by the EGTC Europaregion Tirol-Südtirol-Trentino within the framework of the third call for projects in the field of basic research.

Actions, Plans and Goals



Signals vs. labels

Structured vs. non-structured

…

Problem Classification

Observer

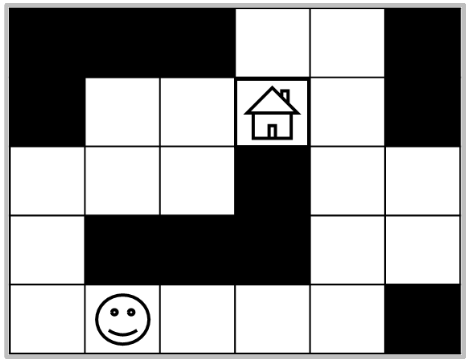
Intervention	Recognition	Knowledge
none	offline	complete
offline	online	partial
online		

Actor

Intentionality	# agents
agnostic	single
adversarial	multiple
intended	

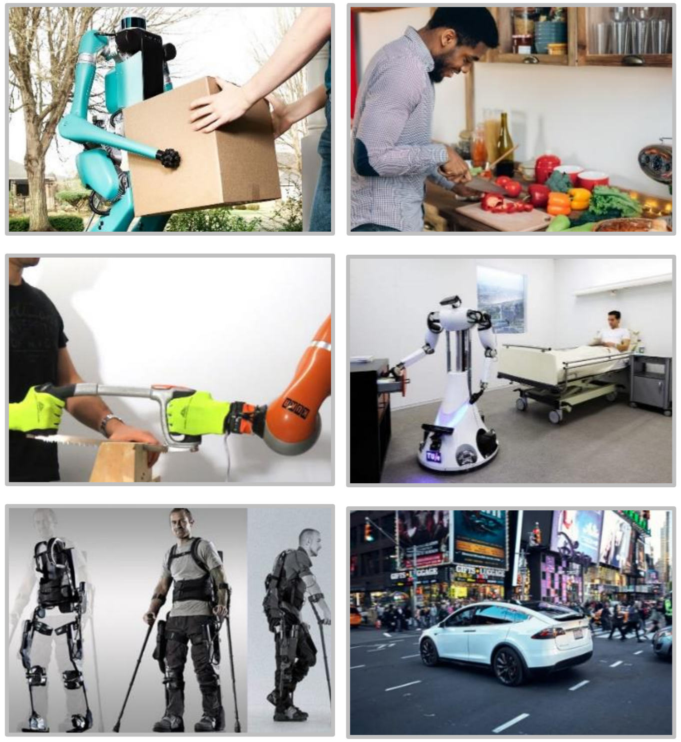
Environment

Observability	Predictability	Continuity
full	deterministic	discrete
partial	stochastic	continuous

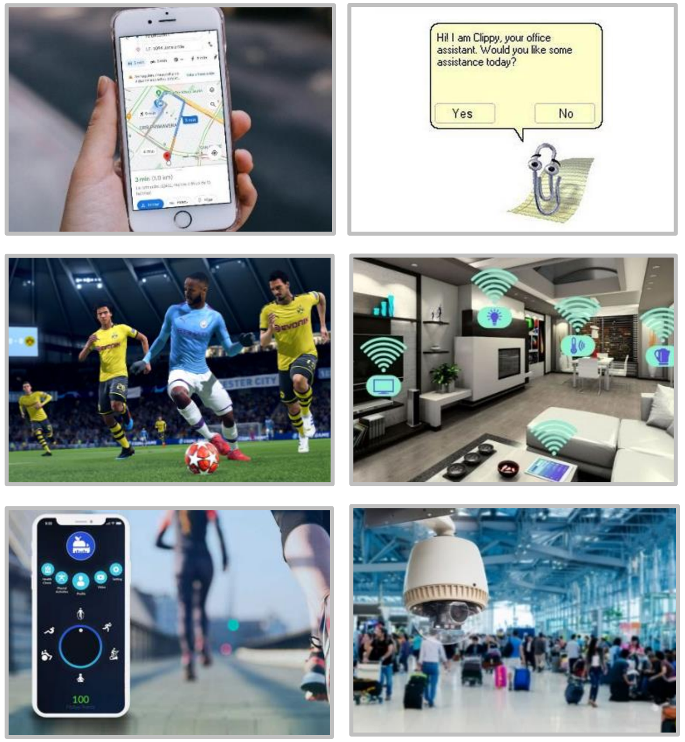


Literature Review

Applications



Human-robot interaction



Others

Challenges

Things to deal with

- Uncertainty
- Variability
- Incomplete knowledge
- Unknown transitions
- Interleaved plans
- Interrupted plans
- Actions with multiple goals
- Plans developed by multiple agents
- Irrelevant actions

Relevant information

- Body movements
- Context
- Objects/agents interacting with
- Previously observed actions
- Effects of actions
- Observed agent characteristics
- Temporal order of events

System characteristics

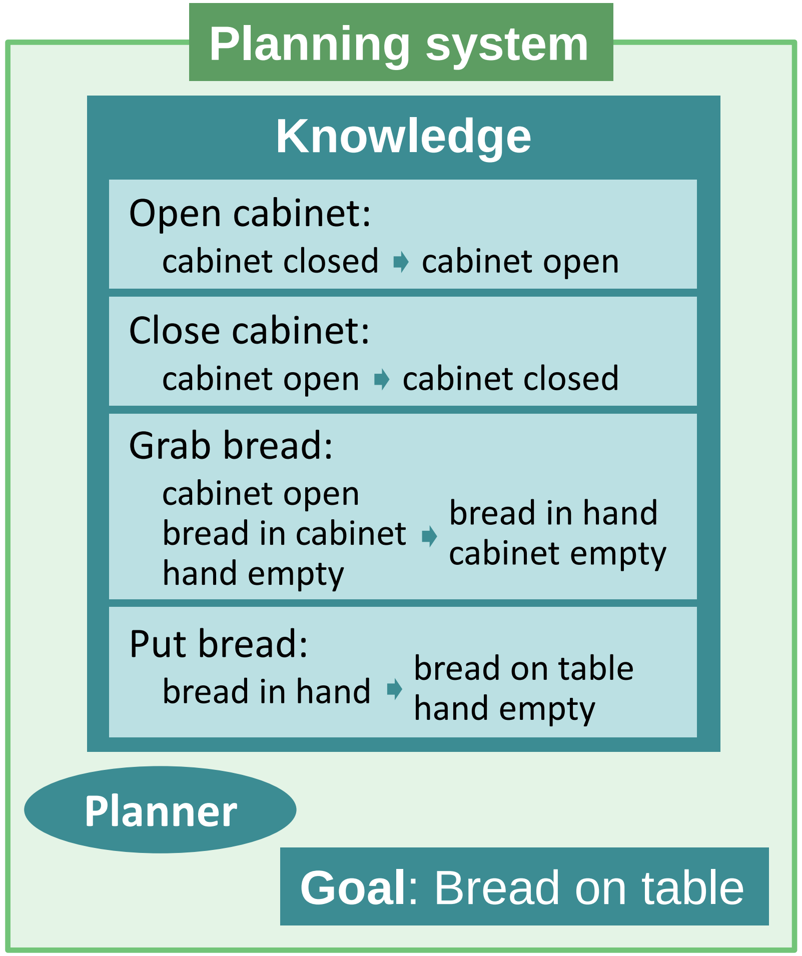
- Predictive
- Expressive
- Scalable
- Adaptable

Plan Recognition as Planning

- **Planning systems** generate candidate plans
- Candidate plans are evaluated **probabilistically** based on **observations**

Strengths
- Highly structured
- Highly expressive
- Generative

Weaknesses
- Require much manual work
- Rigid
- Bad at generalizing

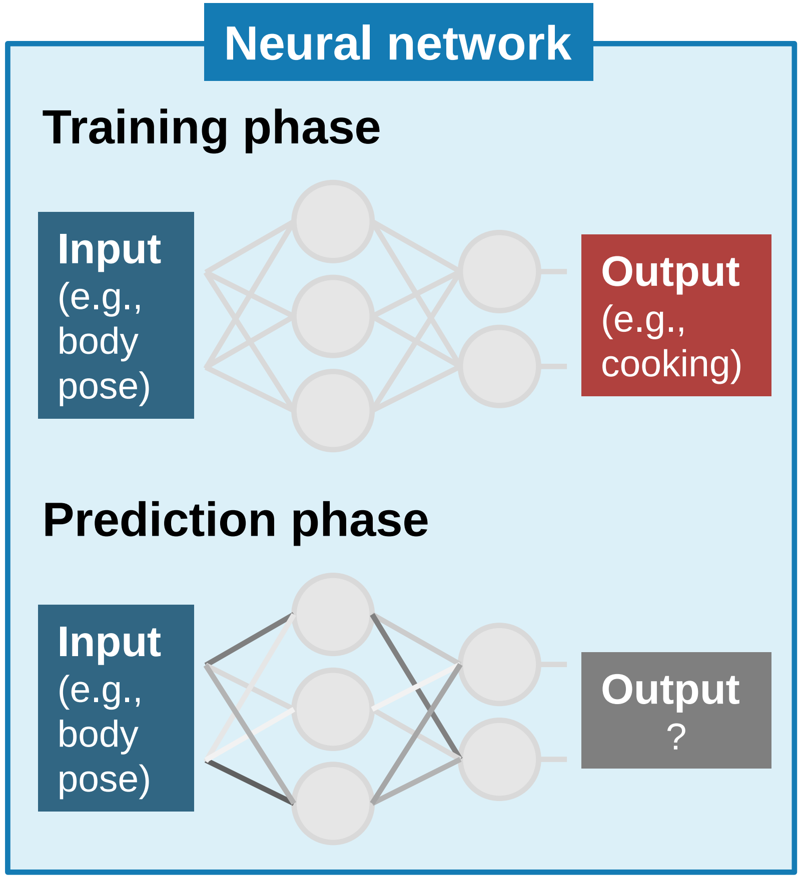


Action Recognition through Neural Networks

- The network is shown many **labeled examples** of actions
- It **learns** to predict the label and **generalize** to unseen examples

Strengths
- Very flexible
- Deal well with sensory input
- Deal well with uncertainty
- Hierarchical

Weaknesses
- Require much labeled data
- Hard to interpret
- Bad at dealing with unknown actions



Hybrid Action and Plan Recognition

- **Action recognition** from sensor data using **neural network**
- Recognized actions used as input for **plan recognition as planning**

Strengths

- Deal well with sensory input
- Deal well with uncertainty
- Highly structured
- Highly expressive

Weaknesses

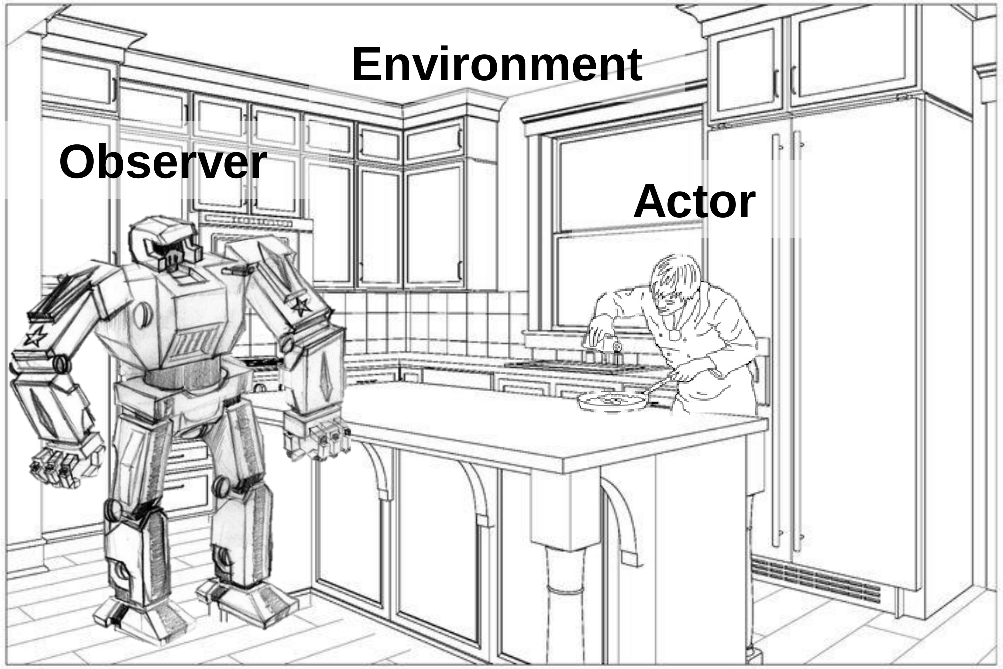
- Require much manual work
- Bad at dealing with unknown actions
- Bad at generalizing

Literature Review

Comparison

	PRAP	NN	Hybrid	Ours
Structure	✓✓	○	✓✓	○
Expressivity	✓✓	✗	✓✓	✗
Uncertainty	○	✓✓	✓	✓✓
Flexibility	✗✗	✓✓	○	✓✓
Sensory input	✗✗	✓✓	✓✓	✓✓
Human effort	✗✗	✗	✗✗	✓✓
Scalability	✗	✓✓	✓	✓✓
Open environment	✗	○	○	✓

Formalization



Environment: $(S, S_0, A^{over}, O^{from}, T, E)$

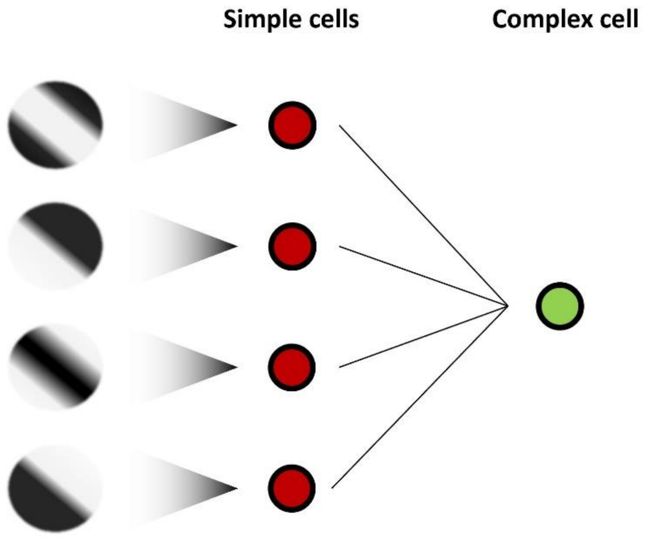
Agent: $(S, S_0, A^{by}, A^{over}, O^{by}, O^{from}, T, E, M, \pi)$

S : State space
- S' : Substate space
- K: Knowledge space
- G : Goal space

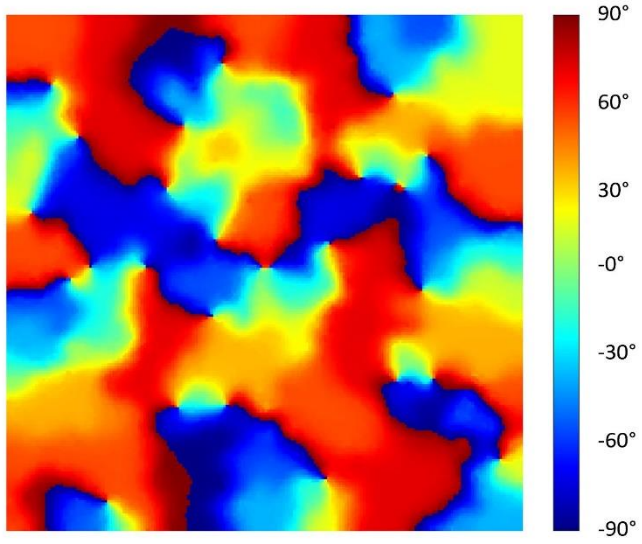
S_0 : Initial state space
A^{by} : Action space
A^{over} : Affordance space
O^{by} : Observation space
O^{from}: Observable state space
T: Transition function
E: Emission function
M: Sensor model
π: Policy

Problem: $(K_{obs,0}, S_{act,rec}, A^{by}_{obs}, O^{by}_{obs}, F_{sys}, g_{rec})$

Models of the Primary Visual Cortex



V1 neuron model



Orientation map

Model by Antolík and Bednar (2011)[1]

Achieves **orientation order** and **phase disorder**

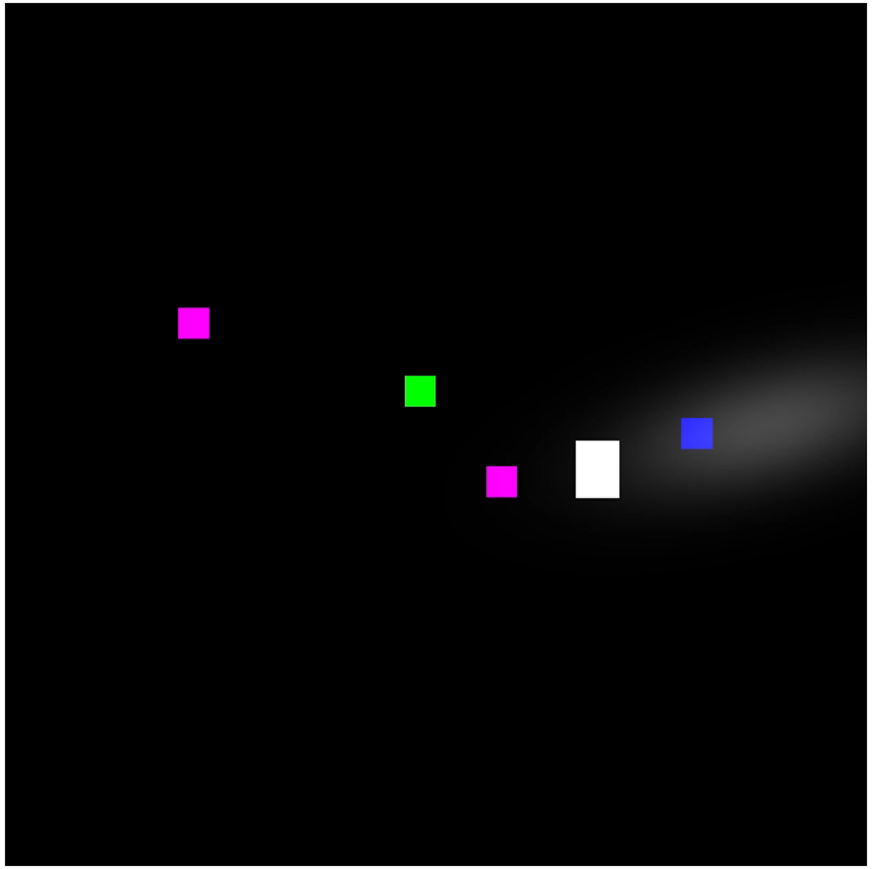
Uses **realistic** patterns of **connectivity**

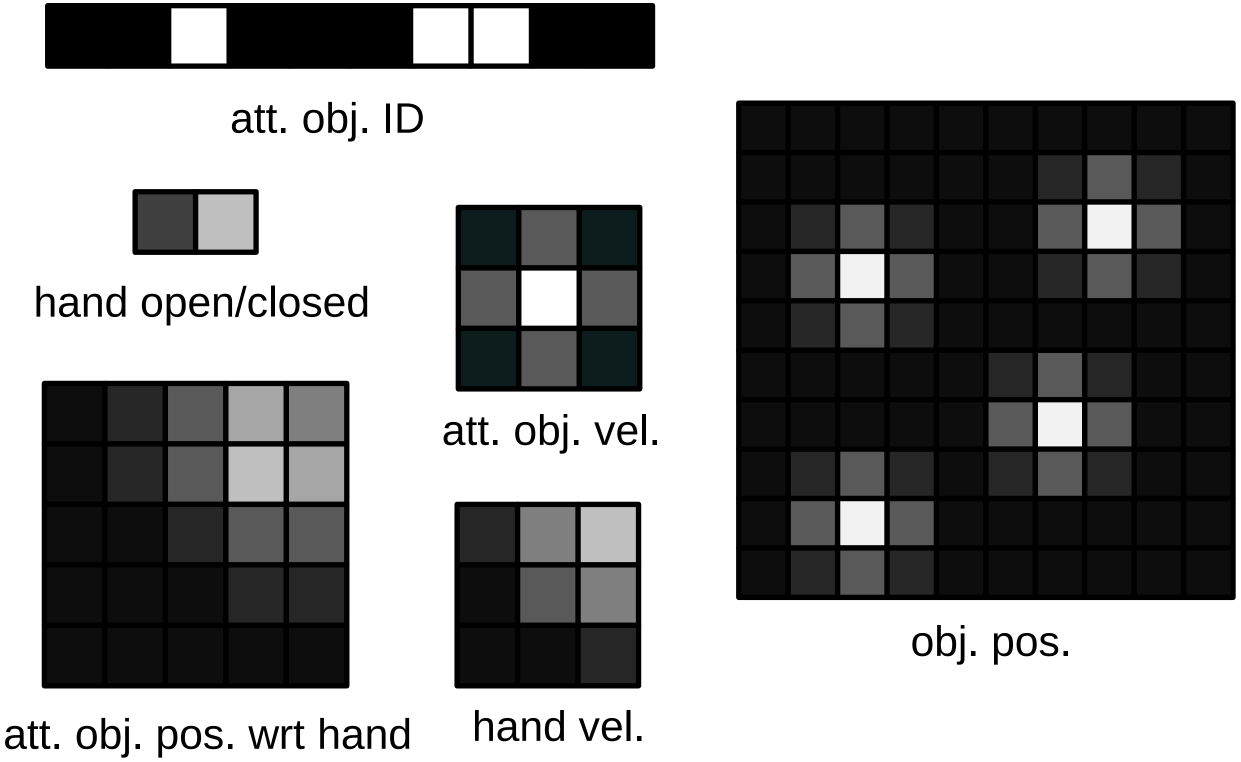
Relies on **shifted patterns** occurring **close in time**

[1] Antolik, Jan, and James A. Bednar. "Development of maps of simple and complex cells in the primary visual cortex." Frontiers in computational neuroscience 5 (2011): 17.

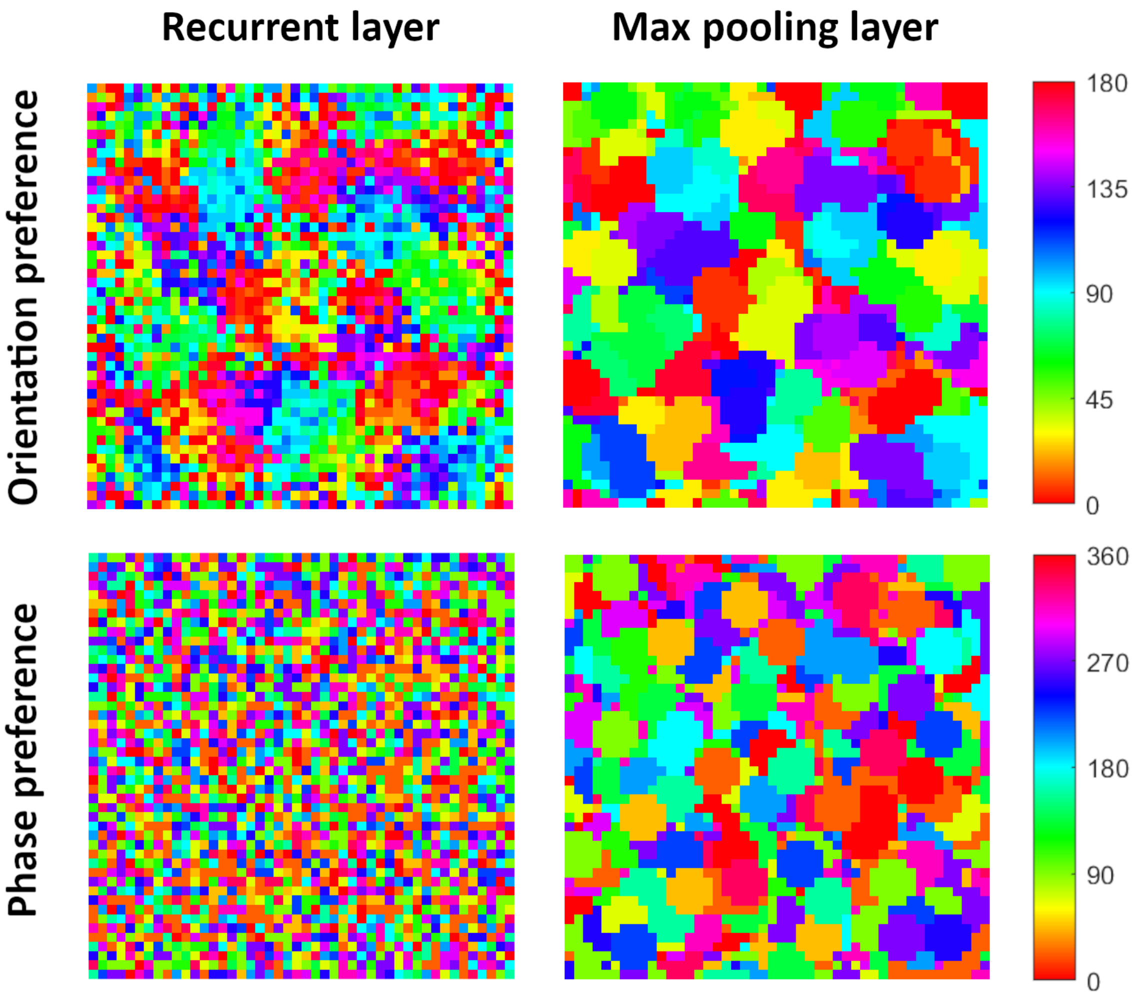
Data Inputs

Synthetic Input





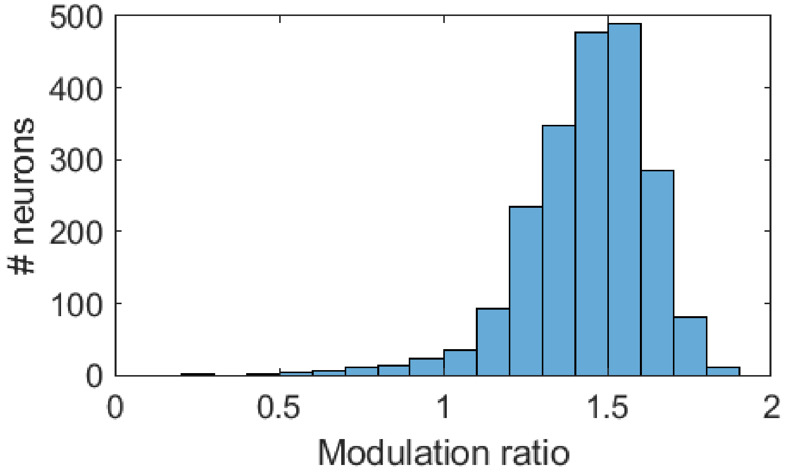
The NILRNN



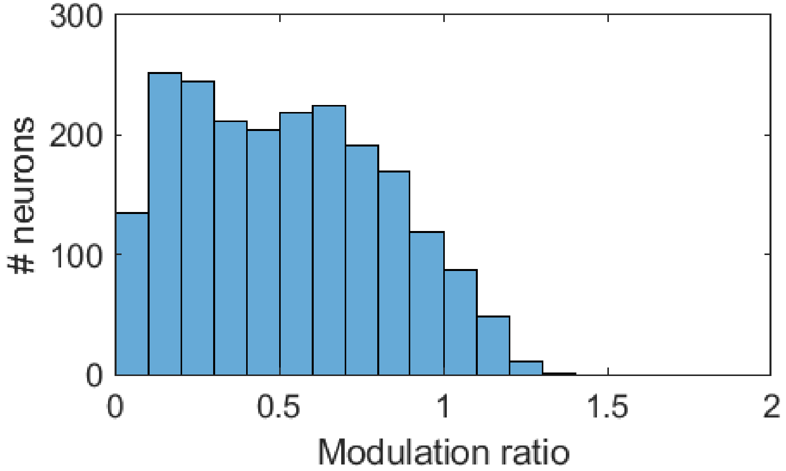
Orientation and Phase Maps

The NILRNN

Modulation Ratios



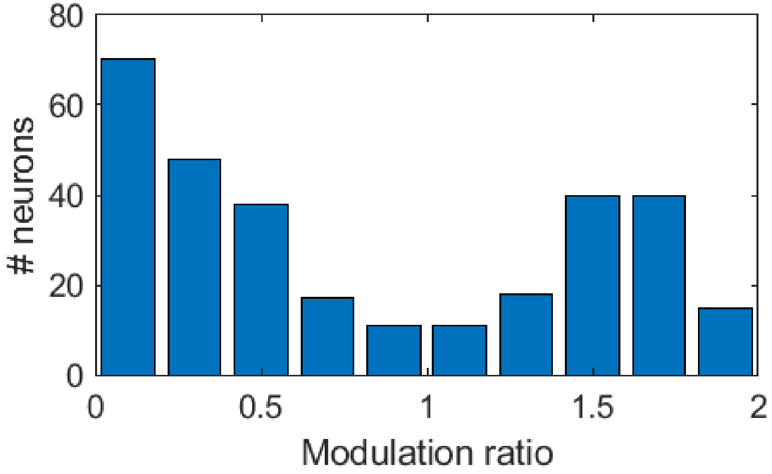
Recurrent layer



Max pooling layer

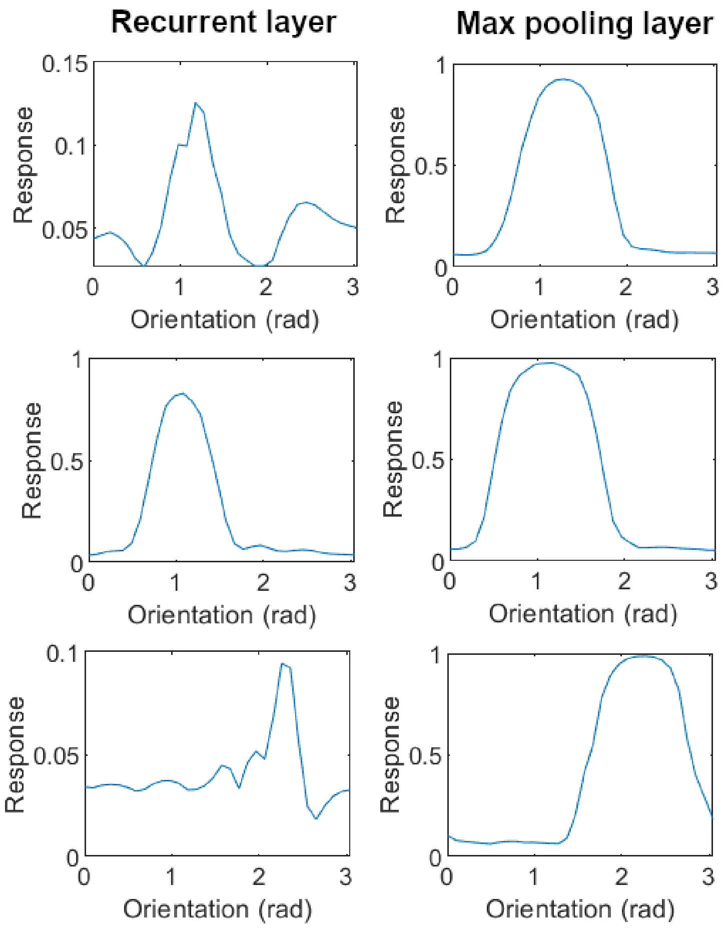
Modulation ratios in a macaque monkey

(Ringach et al., 2002)

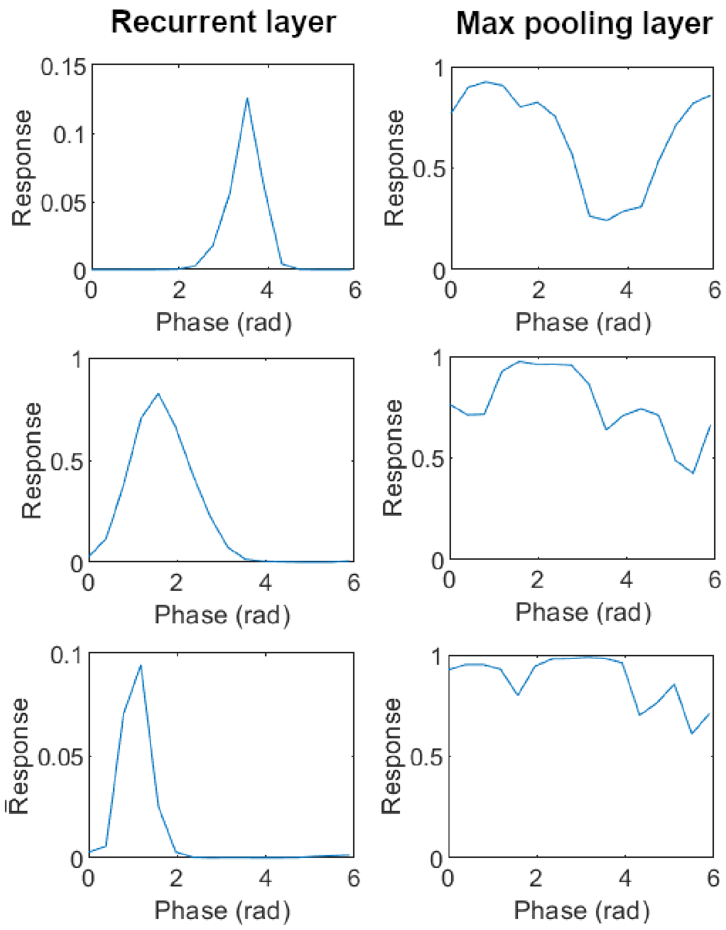


The NILRNN

Orientation Tuning Curves



Phase Responses



The U-LRNN

Other Possible Extensions

High-level reasoning

Model of PFC + hippocampus

Knowledge-based system

LLM

Cognitive attention

Focus on representation regions

Top-down

Similar to feedback circuits

Reinforcement learning

Learn/fine-tune actions

Active perception + attention

Basal ganglia function

Multimodality

Sensor-specific preprocessing

Association areas-like fusion

Developmental

Incremental set up + training

Similar to neocortical maturation

…

Conclusion

Further Future Directions

Structure & expressivity vs. flexibility & human effort & open environment

Make our system hybrid
(would bring other limitations)

Make our system predictive
(would express a single plan)

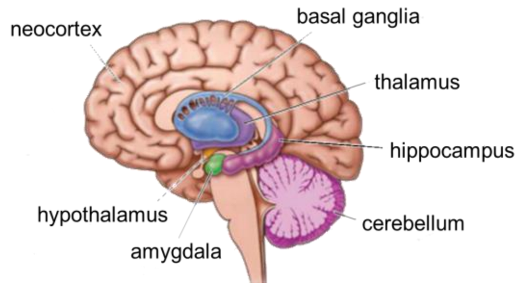
Mimic the hippocampus
(learn patterns + predictive)

Hippocampus + PRAP
(unsupervised learning of knowledge)

Faster at learning but still slower than humans

Mimic the amygdala
(faster learning but also forgetting)

Incremental/few-shot learning
(+ hippocampus patterns)



Cannot deal with unknown unlabeled actions

Anomaly detection
(supervised and unsupervised)

Zero-shot learning
(meaningful label representations)

Integrate other inputs
(e.g., verbal feedback)

Mimic the basal ganglia
(reinforcement learning)