CIS 522: Lecture 2

Multilayer Perceptrons

Penn Engineering

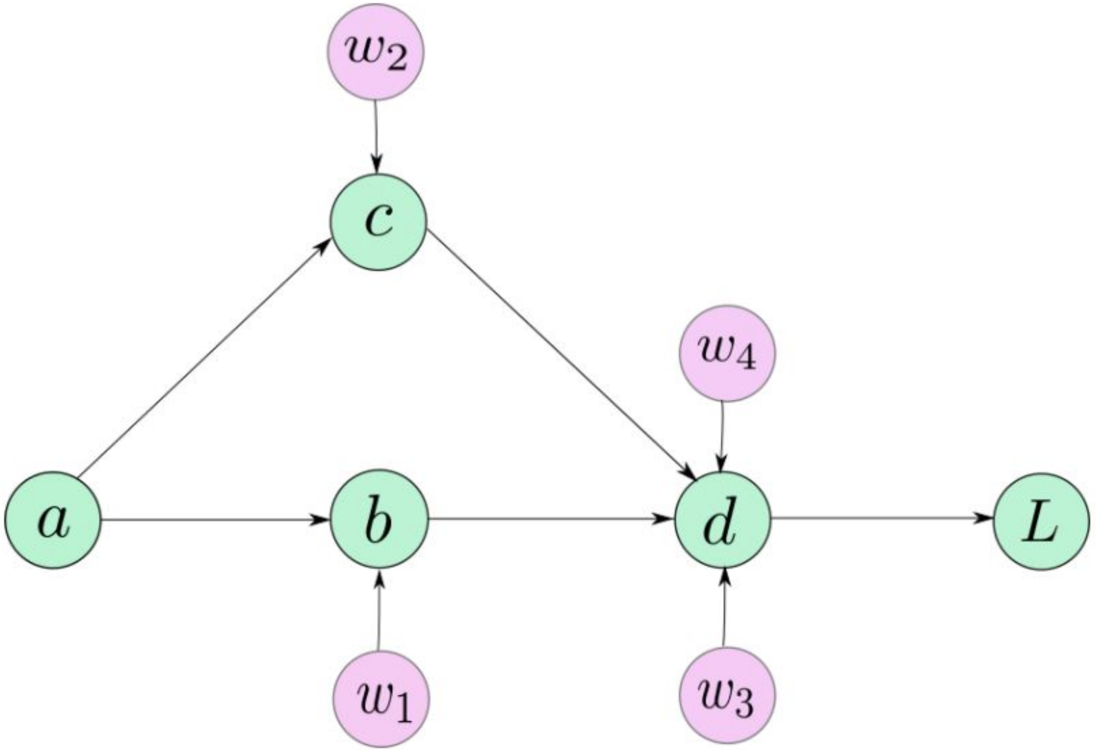
PR our tutorials

Find ways of improving our tutorials

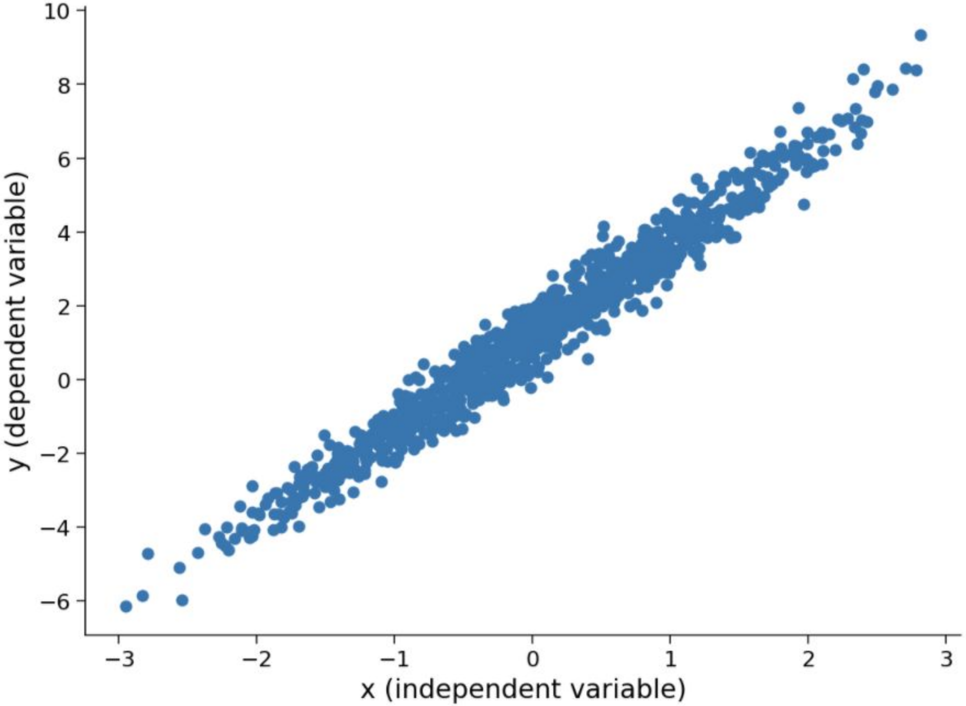
Put them in as a PR on github

Extra points!

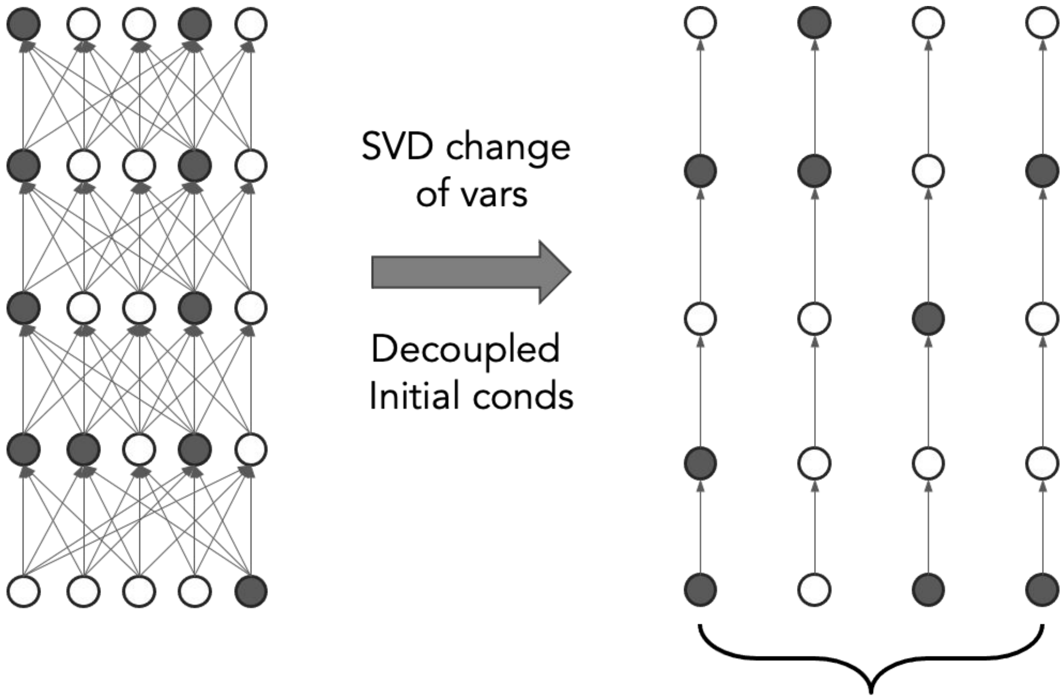
(2) Let us talk about this week’s notebooks



Linear Regression



Shallow vs deep linear networks



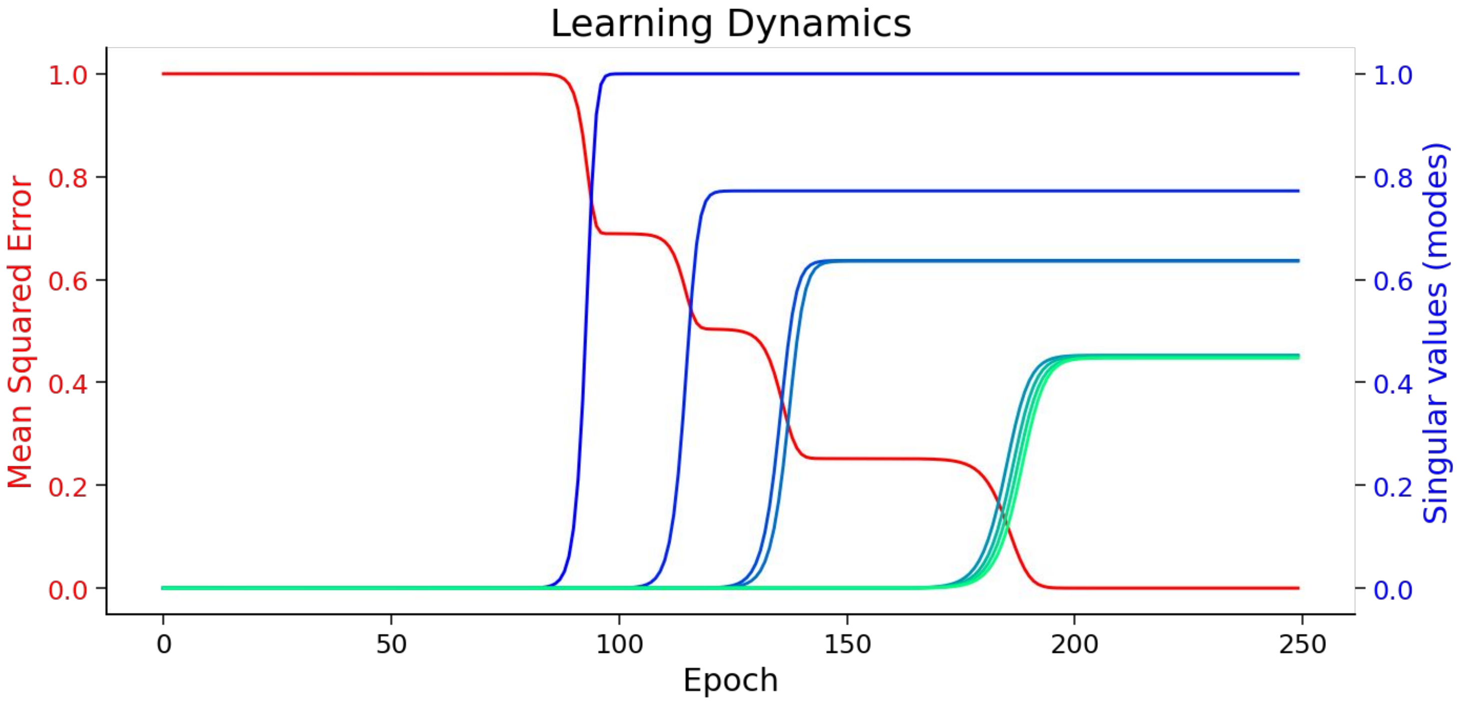
SVD change of vars

Decoupled Initial conds

N independent, scalar modes

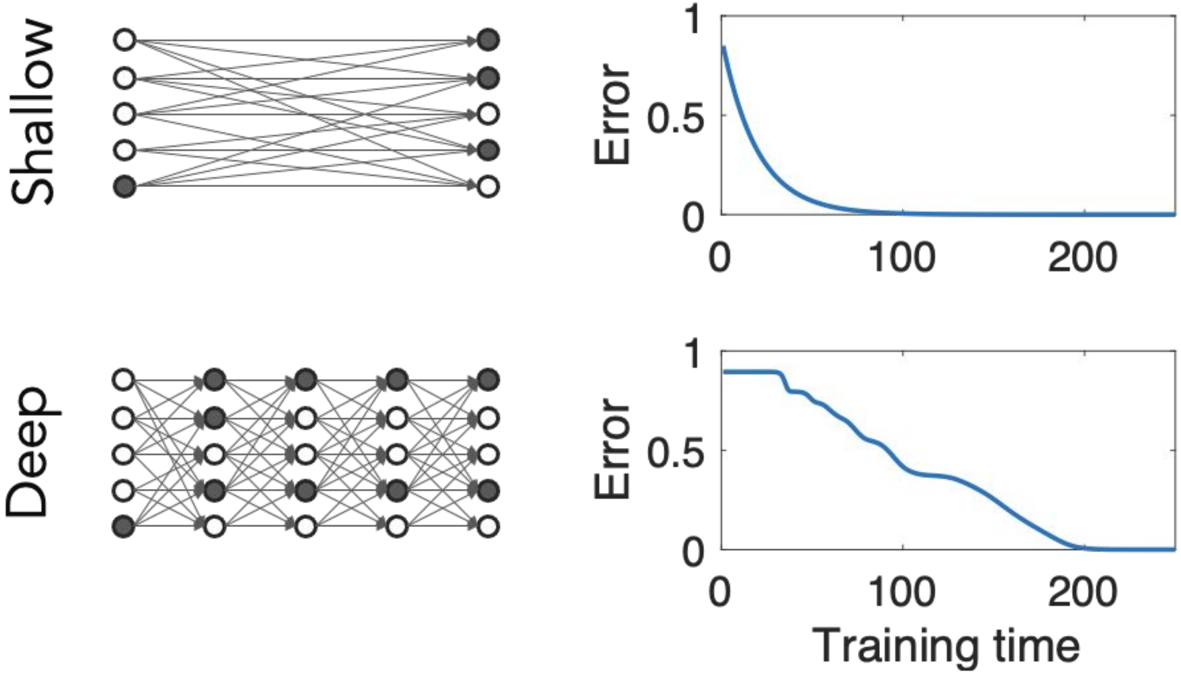
Graph by Saxe 4 NMA

Spent some time on race models



Graph by Saxe 4 NMA

The effect of depth on training



Graph by Saxe 4 NMA

Some cost function engineering

MSE

Mean absolute error (MAE)

% correct guesses

Whatever the problem calls for

The intricacies of network initialization

Xavier

He

The principles

Some high dimensional intuitions

Draw from N-dimensional isotropic Gaussian, N(0,sigma) per dimension, where N>>1

Angle between two vectors?

Proportion of mass within fixed delta skin of sphere

Variance?

Best estimator for future points?

Questions and Answers

Let us spend some time talking about them

Looking towards to the week ahead

Optimization

We will learn about

SGD - noisy approximation, and cheap

Batch normalization

Momentum

Rate scheduling

Adaptive Learning rates

Various problems

Deeper insights

Optimization is not just about finding a minimum

It is about finding a minimum that generalizes well

Let us review what makes a minimum good

Shallow

Deep

Potential problems

a) No gradients

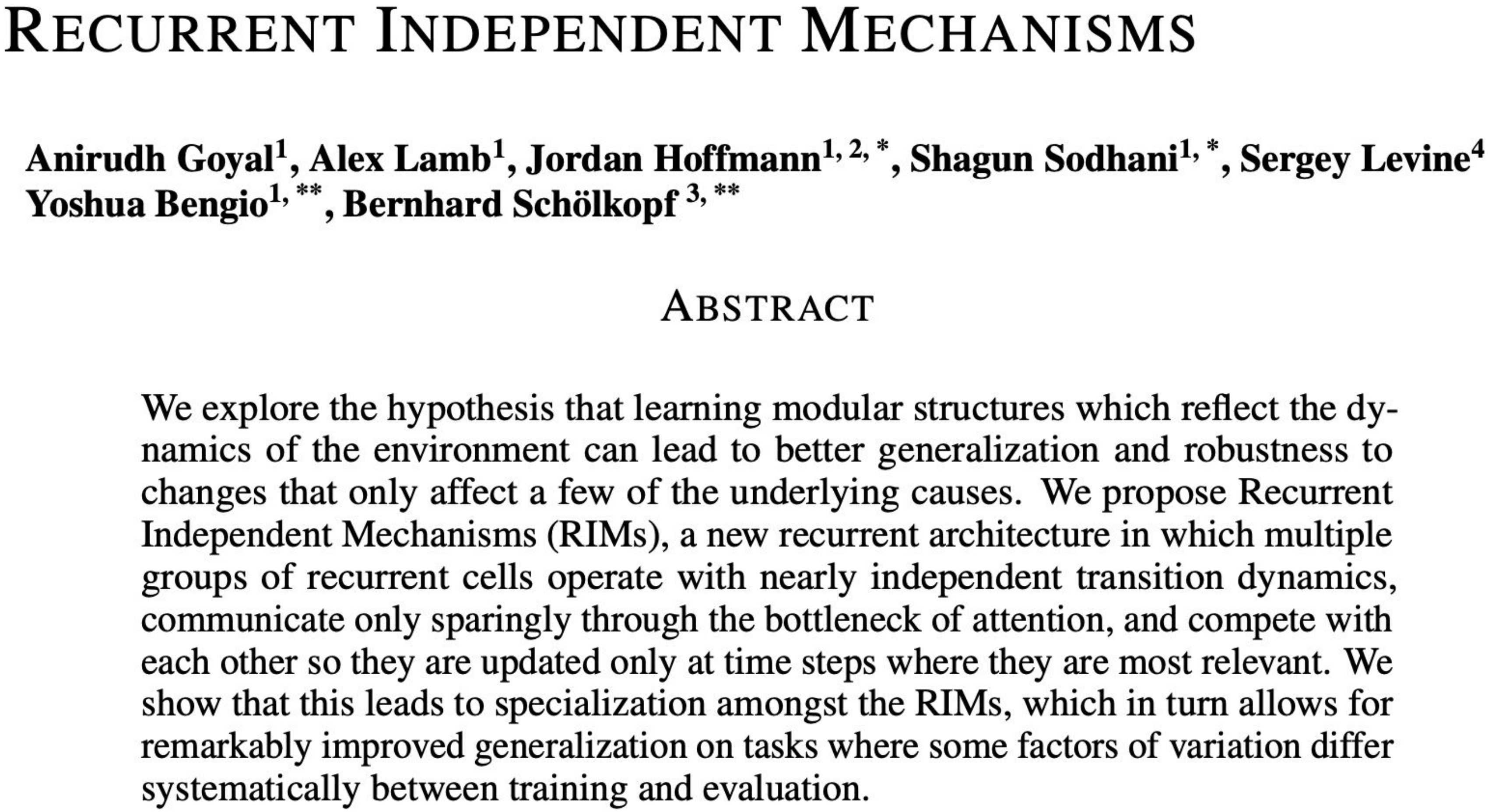
b) Infinite gradients

c) All gradients in low dimensional space

Lets understand the field

https://60years.vizhub.ai/

Causality in atari games

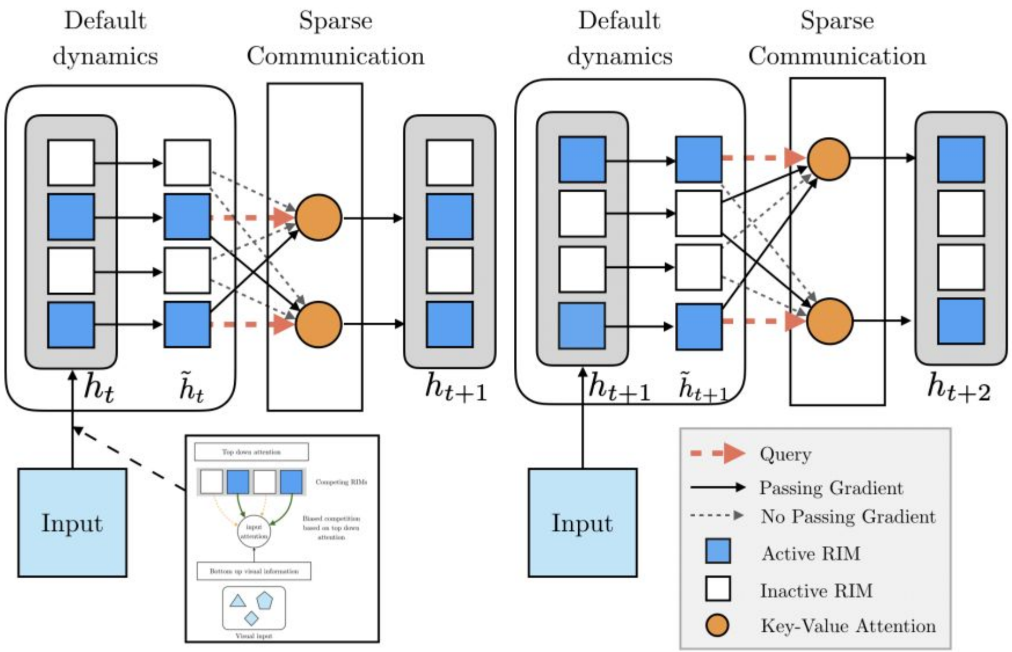


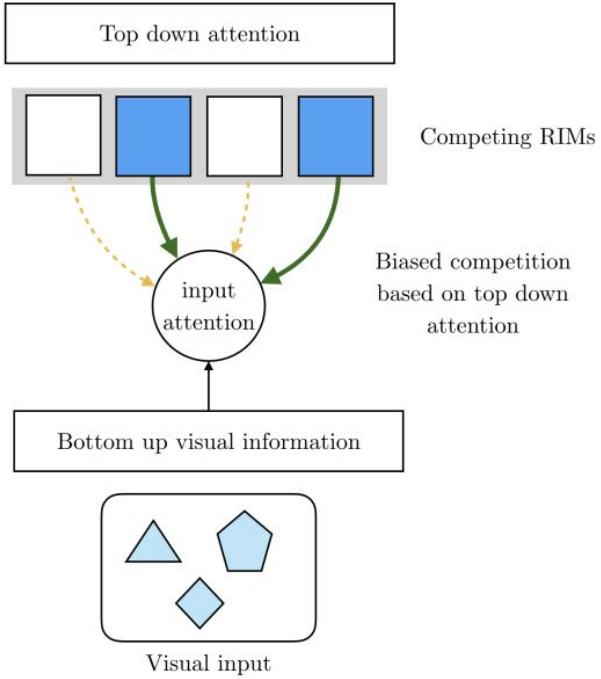
Recurrent Independent Mechanisms

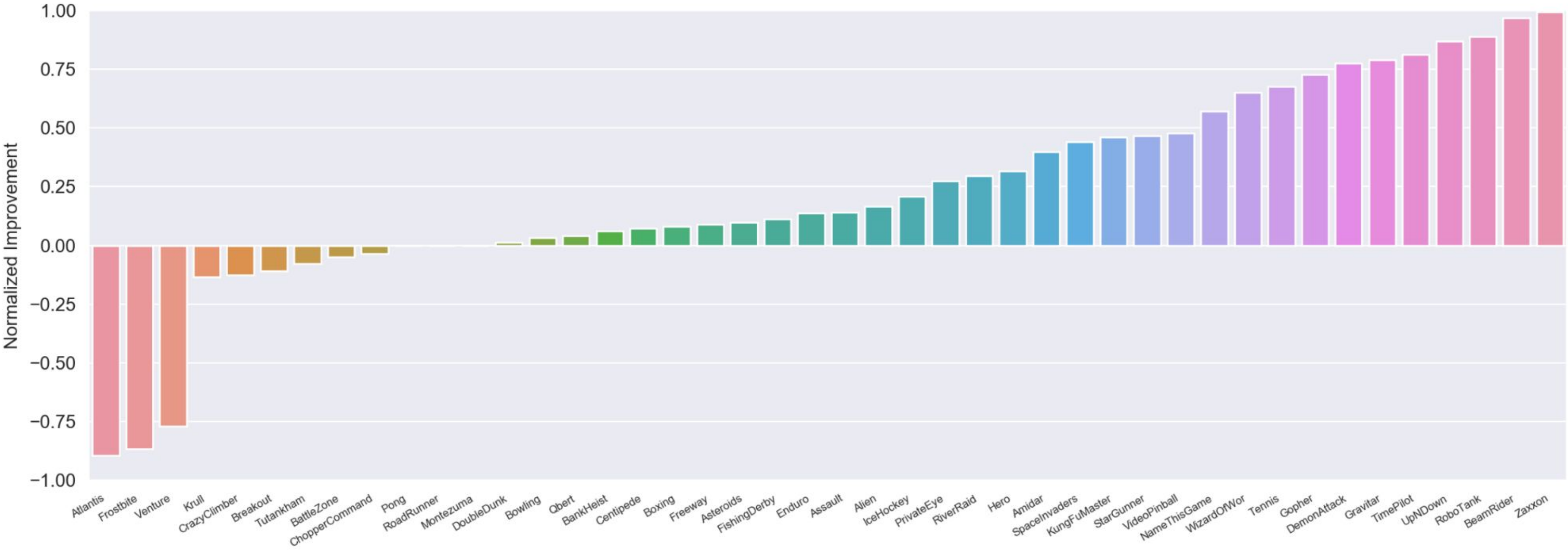
Anirudh Goyal[1], Alex Lamb[1], Jordan Hoffmann[1, 2, *], Shagun Sodhani[1, *], Sergey Levine[4]
Yoshua Bengio[1, **], Bernhard Schölkopf [3, **]

Abstract

We explore the hypothesis that learning modular structures which reflect the dynamics of the environment can lead to better generalization and robustness to changes that only affect a few of the underlying causes. We propose Recurrent Independent Mechanisms (RIMs), a new recurrent architecture in which multiple groups of recurrent cells operate with nearly independent transition dynamics, communicate only sparingly through the bottleneck of attention, and compete with each other so they are updated only at time steps where they are most relevant. We show that this leads to specialization amongst the RIMs, which in turn allows for remarkably improved generalization on tasks where some factors of variation differ systematically between training and evaluation.


Default dynamics
Sparse Communication
Default dynamics
Sparse Communication
h_t
$\tilde{h}_t$
h_{t+1}
h_{t+1}
$\tilde{h}_{t+1}$
h_{t+2}
Input
Input
Top down attention
Competing RIMs
input attention
Biased competition based on top down attention
Bottom up visual information
Visual input
Query
Passing Gradient
No Passing Gradient
Active RIM
Inactive RIM
Key-Value Attention


Top down attention
Competing RIMs
Biased competition based on top down attention
input attention
Bottom up visual information
Visual input


Normalized Improvement
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Atlantis
Frostbite
Venture
Krull
CrazyClimber
Breakout
Tutankham
BattleZone
ChopperCommand
Pong
RoadRunner
Montezuma
DoubleDunk
Bowling
Qbert
BankHeist
Centipede
Boxing
Freeway
Asteroids
FishingDerby
Enduro
Assault
Alien
IceHockey
PrivateEye
RiverRaid
Hero
Amidar
SpaceInvaders
KungFuMaster
StarGunner
VideoPinball
NameThisGame
WizardOfWor
Tennis
Gopher
DemonAttack
Gravitar
TimePilot
UpNDown
RoboTank
BeamRider
Zaxxon