Continual learning without forgetting what we already know

CENTER FOR Brains Minds+ Machines

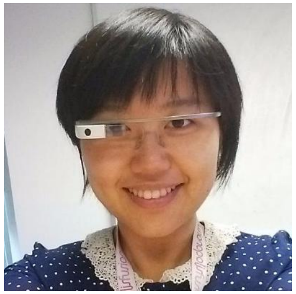
Aprendiendo cosas nuevas sin olvidar lo que ya sabemos

Gabriel Kreiman

klab.tch.harvard.edu

Trabajo en conjunto con:

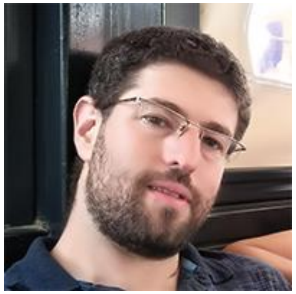
Mengmi Zhang



Shashi Gupta

Morgan Talbot

Nimrod Shaham



Haim Sompolinsky

Scan to download papers+data+code

Progreso en inteligencia artificial (AI)

Beating world
champions at chess

IBM's Watson

X47B

AlphaGo

Self-driving

Definiciones y desiderata

H

Perro

W

$$\mathbf{x_i} \in \mathbb{R}^{W \times H \times 3}$$

Arbol

Mesa

Perro

Perro

$$i = \{1, ..., N\}$$

$$(\mathbf{x_i}, y_i) \qquad y_i \in 1, ..., C$$

Tumores: C=2

Imagenet: C=1000

Definiciones y desiderata

Perro

$$\hat{y}_i = f(\mathbf{x}_i)$$

Desiderata

$$\hat{y}_i = f(\mathbf{x}_i)$$

1. Queremos aprender *f* a partir de ejemplos

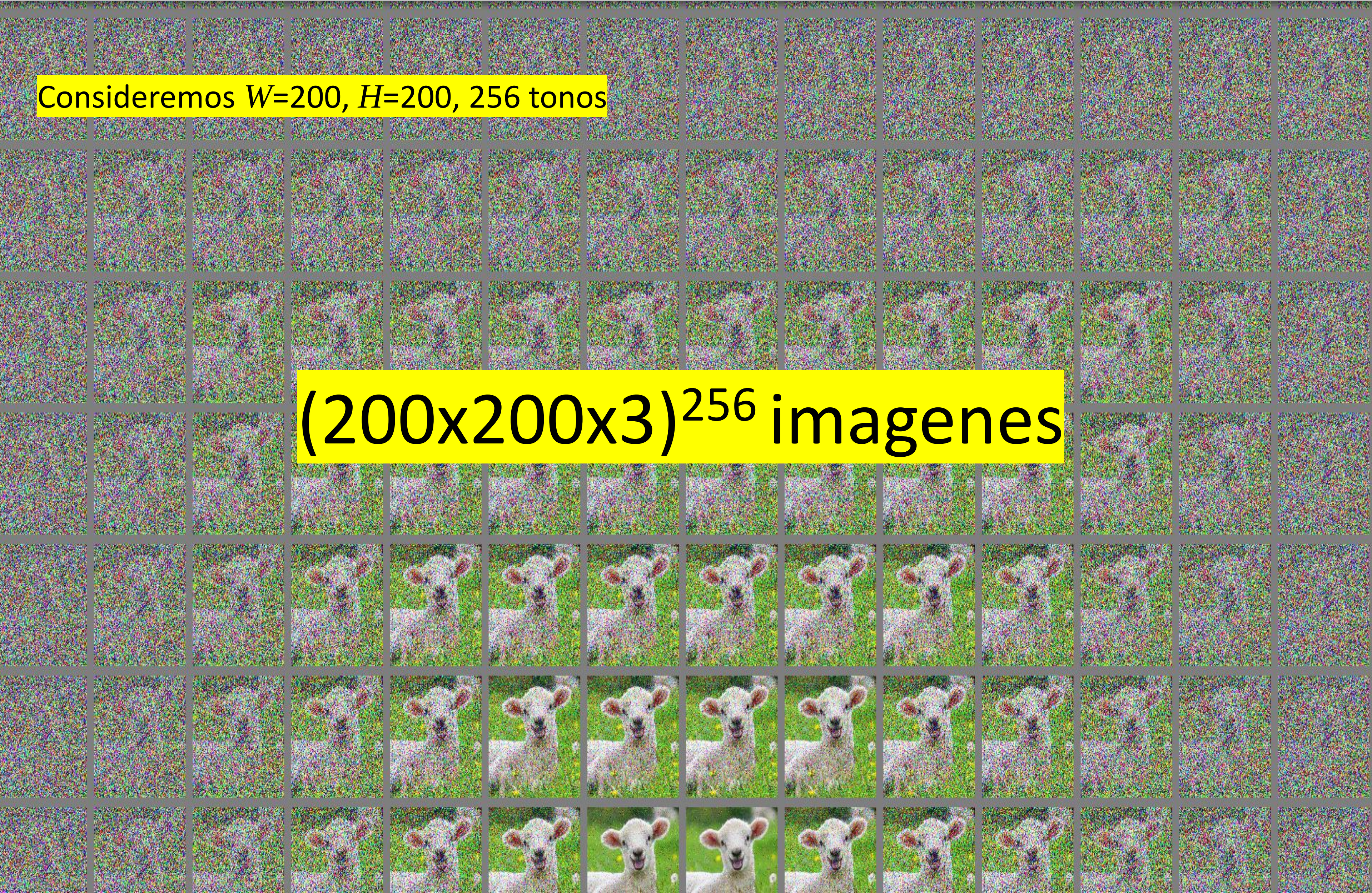
Desiderata

$$\hat{y}_i = f(\mathbf{x}_i)$$

1. Queremos aprender f a partir de ejemplos

2. N pequenio, N << numero de parametros

El numero de imagenes es astronomico



Consideremos W=200, H=200, 256 tonos

$(200x200x3)^{256}$ imagenes

Desiderata

$$\hat{y}_i = f(\mathbf{x}_i)$$

1. Queremos aprender f a partir de ejemplos

2. N pequenio, N << numero de parametros

Desiderata

$$\hat{y}_i = f(\mathbf{x}_i)$$

1. Queremos aprender f a partir de ejemplos

2. N pequenio, N << numero de parametros

3. Robustez a perturbaciones

$$f(\mathbf{x} + \epsilon) = f(\mathbf{x})$$

Desiderata

$$\hat{y}_i = f(\mathbf{x}_i)$$

1. Queremos aprender f a partir de ejemplos

2. N pequenio, N << numero de parametros

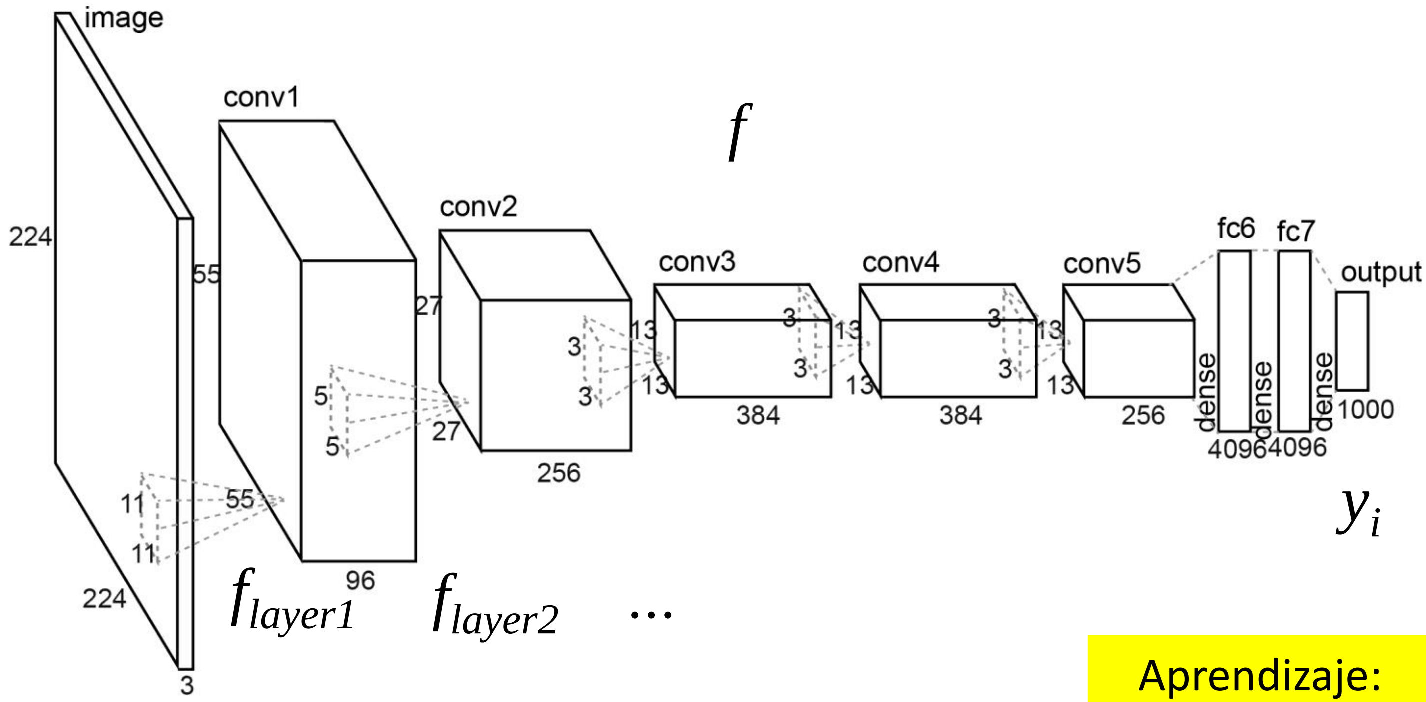
3. Robustez a perturbaciones

4. Aprendizaje continuo

$$f_C(\mathbf{x}) \rightarrow f_{C+1}(\mathbf{x})$$

Bisbita Llanera

f: Redes neuronales

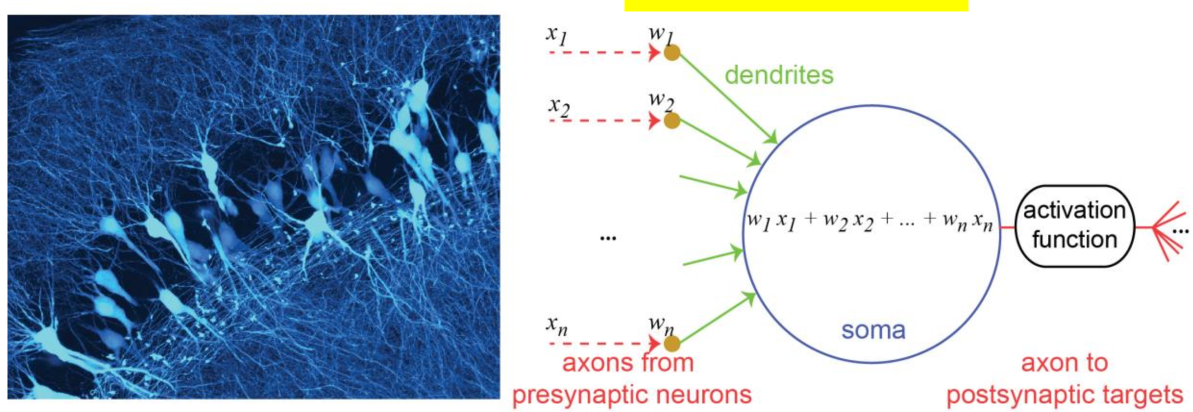


$$\mathbf{x_i} \in \mathbb{R}^{W \times H \times 3}$$

Aprendizaje:
Ajustar ***w***

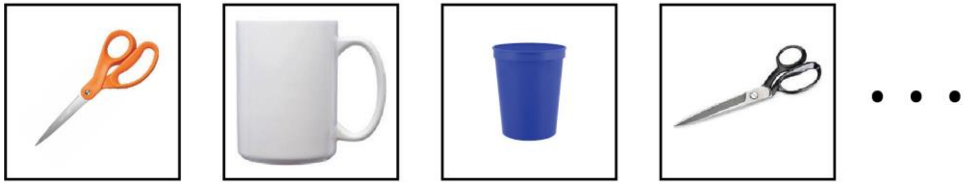
A

B



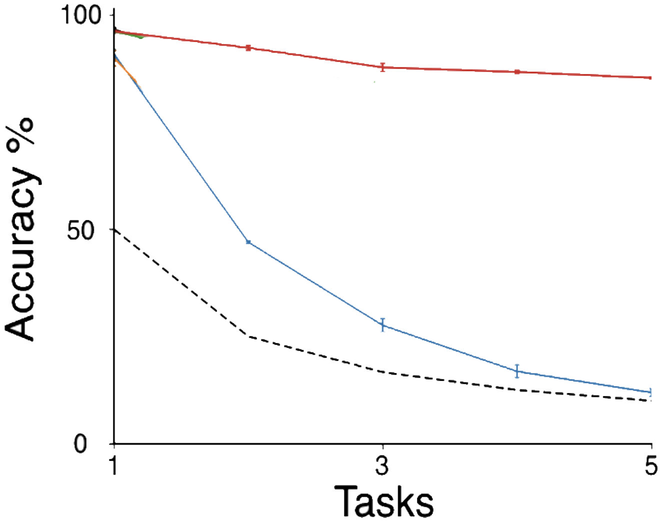
Aprendizaje continuo

Task 1



f_C

f_{C+2}



Zhang, M., Badkundri, R., Talbot, M. & Kreiman, G. Hypothesis-driven Stream Learning with Augmented Memory. *arXiv* **2104.02206** (2021).

Dos estrategias para aprendizaje continuo

1. Proteccion sinaptica

2. Repeticion [Replay]

Proteccion sinaptica

Elastic weight consolidation

RESEARCH ARTICLE | APPLIED MATHEMATICS

Overcoming catastrophic forgetting in neural networks

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, +10, and Raia Hadsell Authors Info & Affiliations

Edited by James L. McClelland, Stanford University, Stanford, CA, and approved February 13, 2017 (received for review July 19, 2016)

March 14, 2017 | 114 (13) 3521-3526 | https://doi-org.ezp-prod1.hul.harvard.edu/10.1073/pnas.1611835114

PNAS
Vol. 114 | No. 13

Significance
Abstract
Results
Fisher Overlap
Discussion
Materials and Methods
Random Patterns
MNIST Experiments

Significance

Deep neural networks are currently the most successful machine-learning technique for solving a variety of tasks, including language translation, image classification, and image generation. One weakness of such models is that, unlike humans, they are unable to learn multiple tasks sequentially. In this work we propose a practical solution to train such models sequentially by protecting the weights important for previous tasks. This approach, inspired by synaptic consolidation in neuroscience, enables state of the art results on multiple reinforcement learning problems experienced sequentially.

Sparse Distributed Memory + Multi-Layer Perceptron

Published as a conference paper at ICLR 2023

SPARSE DISTRIBUTED MEMORY IS A CONTINUAL LEARNER

Trenton Bricken
Systems, Synthetic, and Quantitative Biology
Harvard University
trentonbricken@g.harvard.edu

Xander Davies & Deepak Singh
Computer Science
Harvard College
{alexander_davies, tejasvisingh}@college.harvard.edu

Dmitry Krotov
MIT-IBM Watson AI Lab
IBM Research
krotov@ibm.com

Gabriel Kreiman
Harvard Medical School
Programs in Biophysics and Neuroscience
Gabriel.Kreiman@childrens.harvard.edu

ABSTRACT

Continual learning is a problem for artificial neural networks that their biological counterparts are adept at solving. Building on work using Sparse Distributed Memory (SDM) to connect a core neural circuit with the powerful Transformer model, we create a modified Multi-Layered Perceptron (MLP) that is a strong continual learner. We find that every component of our MLP variant translated from biology is necessary for continual learning. Our solution is also free from any memory replay or task information, and introduces novel methods to train sparse networks that may be broadly applicable.

Dos estrategias para aprendizaje continuo

1. Proteccion sinaptica

2. Repeticion [Replay]

Repeticion [Replay]

1. Que almacenar? Almacenamiento = $$$

2. Cuan a menudo repetir informacion?

3. Que informacion repetir?

Estudio analitico de aprendizaje continuo

1. Recurrent neural network with attractor dynamics
2. Forgetting
3. Replay

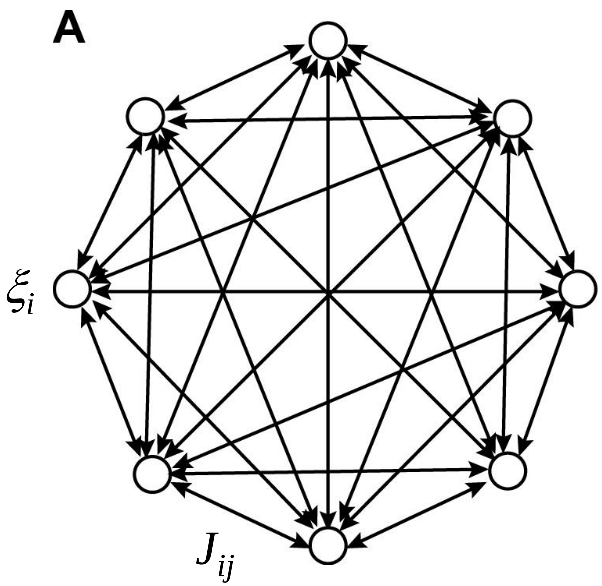
OPEN

Stochastic consolidation of lifelong memory

Nimrod Shaham[1], Jay Chandra[1], Gabriel Kreiman[2] & Haim Sompolinsky[1,3]✉

Humans have the remarkable ability to continually store new memories, while maintaining old memories for a lifetime. How the brain avoids catastrophic forgetting of memories due to interference between encoded memories is an open problem in computational neuroscience. Here we present a model for continual learning in a recurrent neural network combining Hebbian learning, synaptic decay and a novel memory consolidation mechanism: memories undergo stochastic rehearsals with rates proportional to the memory's basin of attraction, causing self-amplified consolidation. This mechanism gives rise to memory lifetimes that extend much longer than the synaptic decay time, and retrieval probability of memories that gracefully decays with their age. The number of retrievable memories is proportional to a power of the number of neurons. Perturbations to the circuit model cause temporally-graded retrograde and anterograde deficits, mimicking observed memory impairments following neurological trauma.

Redes neuronal recurrentes: el modelo de Hopfield



ξ_i (binary) activity of unit i

J_{ij} synaptic strength between i and j

Hopfield, PNAS 1982
Kreiman, Cambridge University Press 2021, Chapter 7

Introduciendo olvido

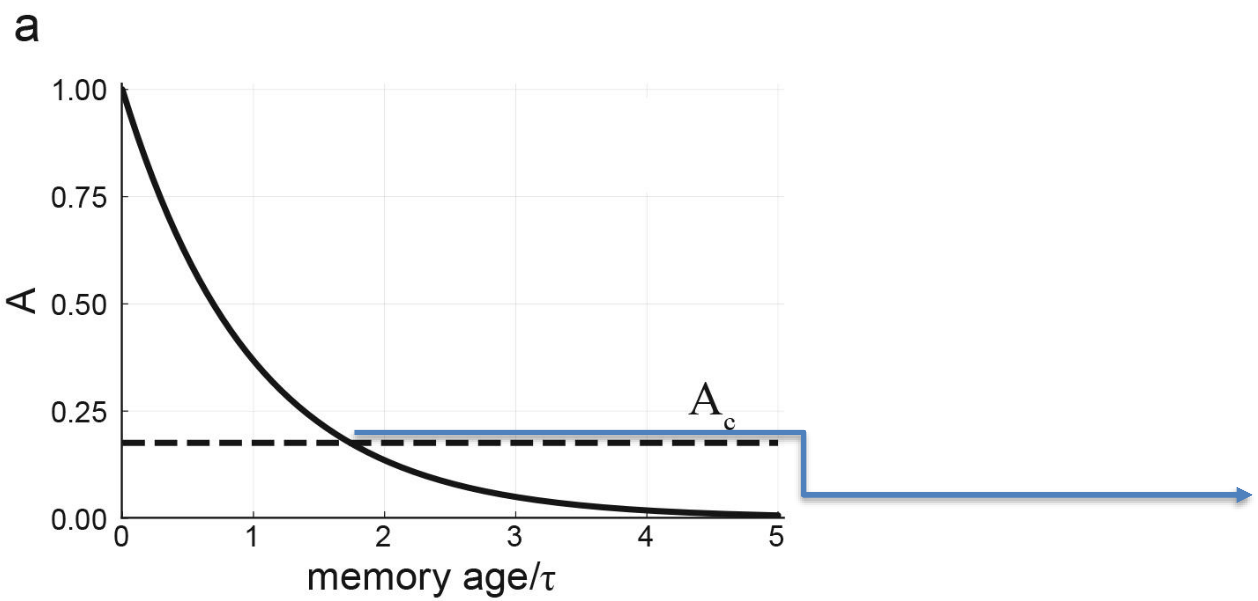
$$J_{ij}(t + \Delta t) = \sum_l \xi_i^l \xi_j^l \delta_{t,l}$$

$$J_{ij}(t) = \sum_l A_l(t) \xi_i^l \xi_j^l$$

$$\frac{dA_l}{dt} = -\frac{1}{\tau} A_l$$

A_l: efficacy of memory l at time t

Regimen de olvido completo



Unrealistic: recent memories are retrieved perfectly, old memories are all lost

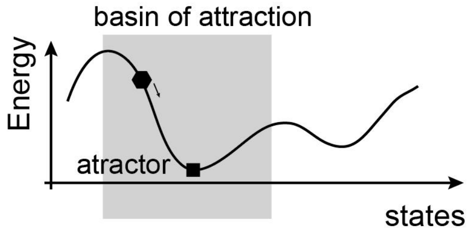
Practicar, practicar, practicar

$$J_{ij}(t+\Delta t) = (1-\Delta t/\tau)J_{ij}(t) + \sum_l \xi_i^l \xi_j^l \delta_{t,l}$$

$$J_{ij}(t) = \sum_l A_l(t)\xi_i^l \xi_j^l$$

$$\frac{dA_l}{dt} = -\frac{1}{\tau}A_l$$

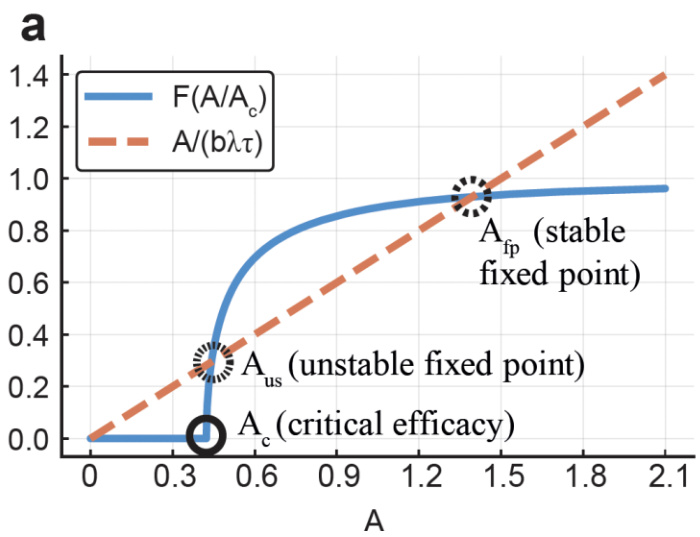
B



$$R_l(t) = \sum_{\{t_k\}} \delta(t-t_k)$$

$$r_l(t) = \lambda F(A_l(t)/A_c(t))$$

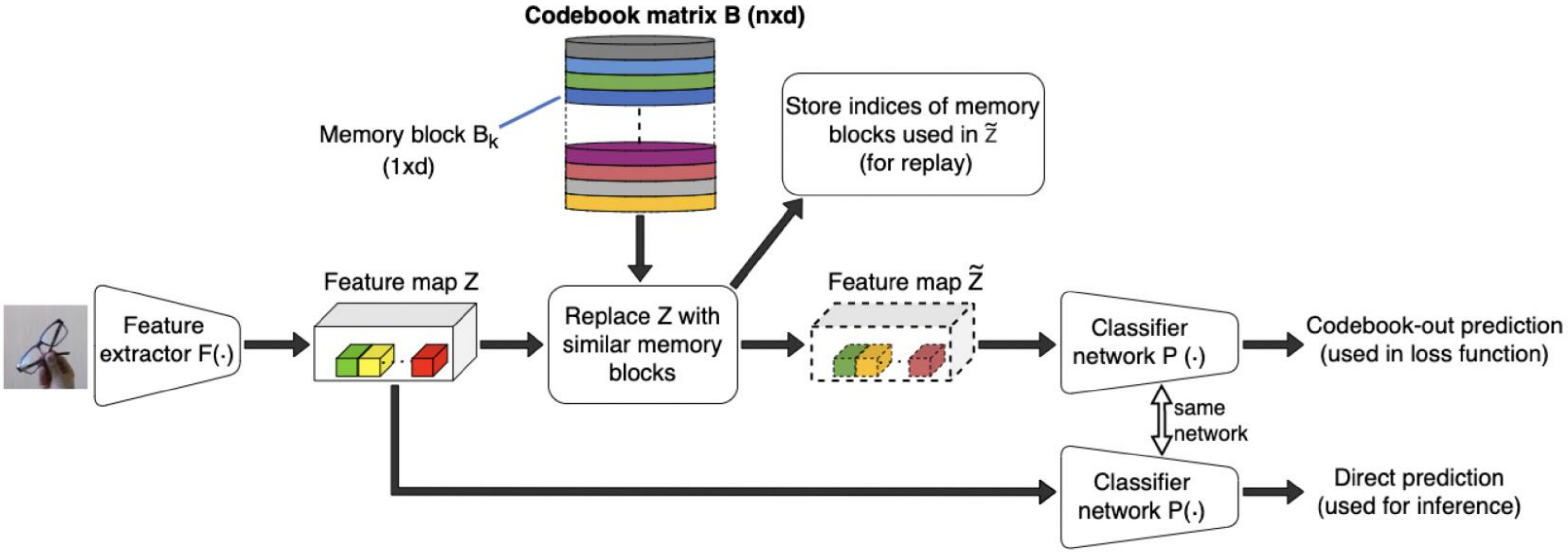
Dinamica de practica y olvido



1. Calculo de capacidad
2. Prediccion de vida media de memorias
3. Deficiencias de memoria

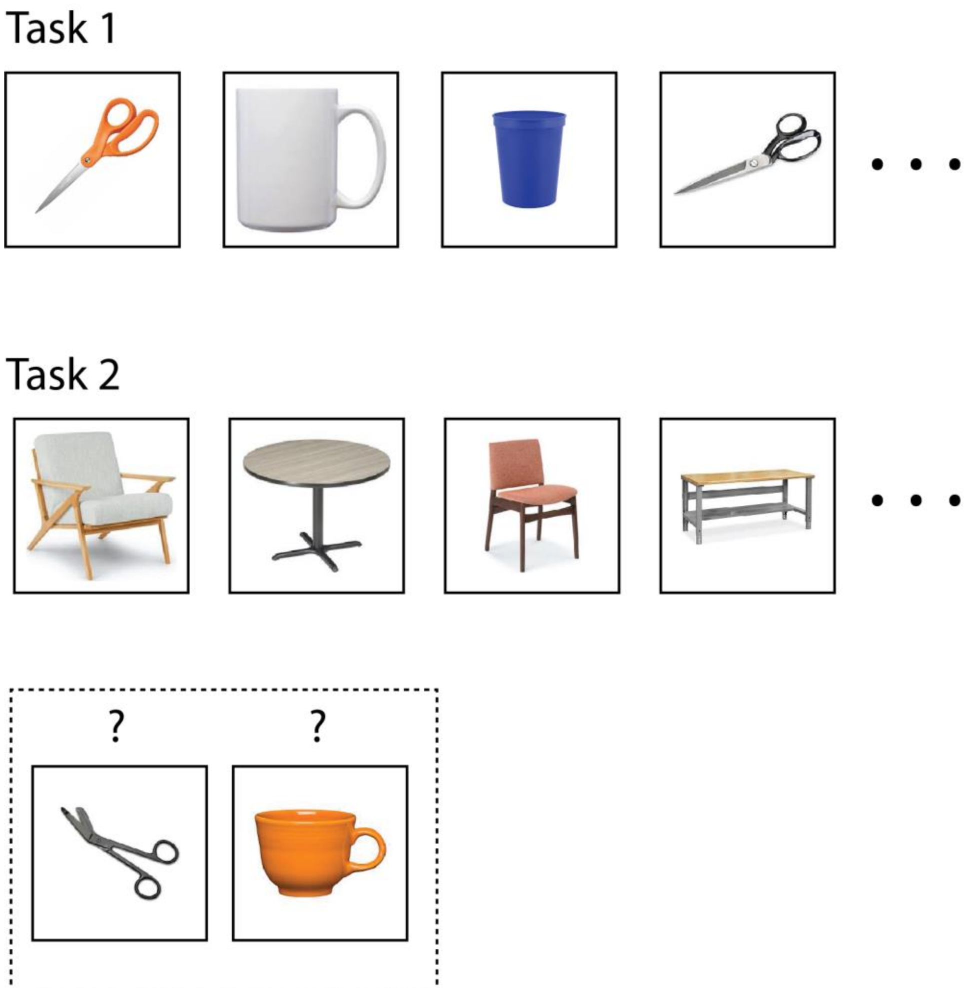
CRUMB: Algoritmo de aprendizaje continuo en neural networks

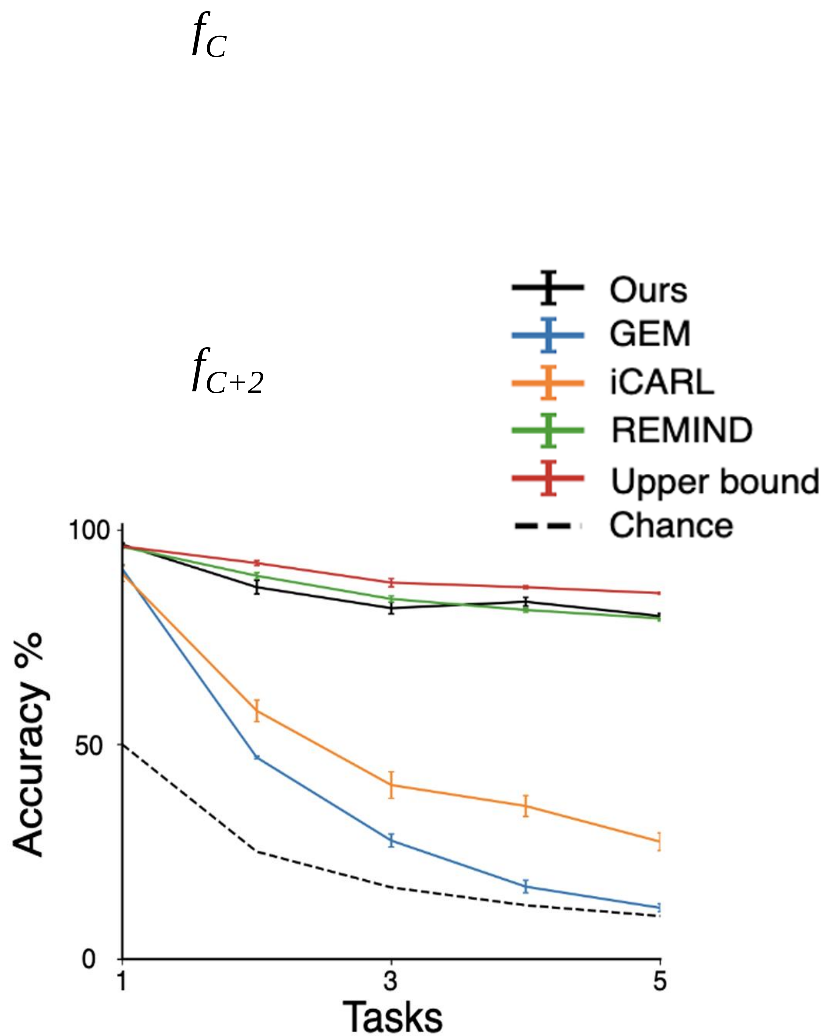
CRUMB: **C**ompositional **R**eplay **U**sing **M**emory **B**locks



Talbot et al, arXiv 2022

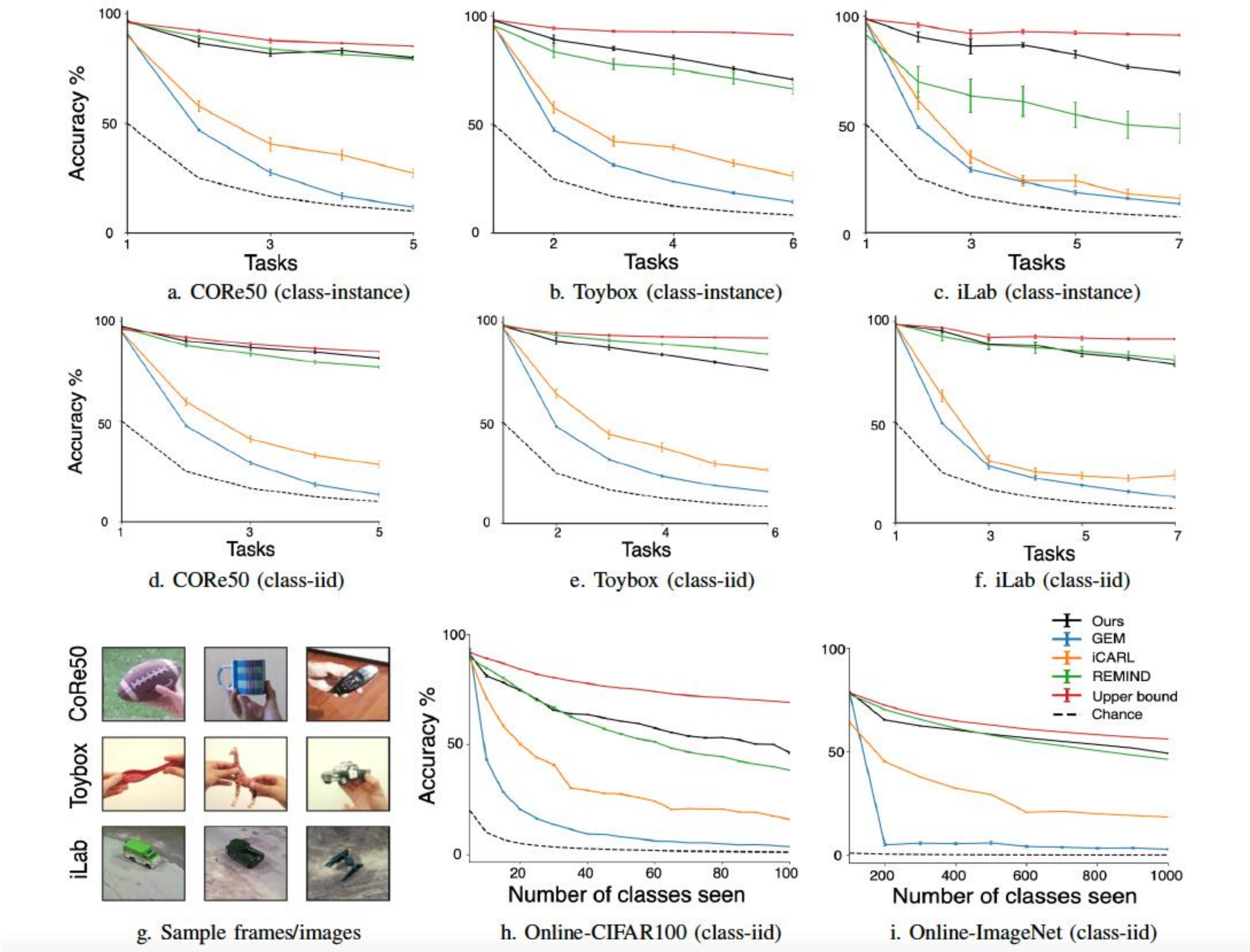
CRUMB aprende en forma continua





Zhang, M., Badkundri, R., Talbot, M. & Kreiman, G. Hypothesis-driven Stream Learning with Augmented Memory. *arXiv* **2104.02206** (2021).

CRUMB generaliza a otros datos y condiciones



a. CORe50 (class-instance) b. Toybox (class-instance) c. iLab (class-instance)

d. CORe50 (class-iid) e. Toybox (class-iid) f. iLab (class-iid)

g. Sample frames/images h. Online-CIFAR100 (class-iid) i. Online-ImageNet (class-iid)

CRUMB es superior a otros algorithmos

	CORe50 [39] class-instance	CORe50 [39] class-iid	Toybox [10] class-instance	Toybox [10] class-iid	iLab [11] class-instance	iLab [11] class-iid	CIFAR100 [41] class-iid	ImageNet [40] class-iid
Ours	**79.9**	**81.4**	**70.6**	76.0	**73.8**	78.6	**46.2**	**49.2**
GEM [29]	11.9	13.5	14.3	15.7	13.0	12.8	3.5	2.9
iCARL [26]	27.0	28.5	27.3	26.5	15.6	23.6	15.9	18.5
REMIND [4]	77.0	76.0	66.2	**84.1**	48.1	**81.0**	38.2	46.2
EWC [8]	12.2	12.4	14.3	15.7	13.5	13.0	3.9	0.1
MAS [47]	14.4	17.4	18.9	19.2	20.5	22.1	5.5	0.1
SI [15]	12.0	12.9	14.3	15.5	12.8	13.0	3.6	8.8
Stable SGD [45]	13.7	13.2	13.5	13.8	9.8	6.9	7.3	-
GSS [27]	15.0	15.6	14.7	15.0	13.0	12.8	3.2	-
BiC [25]	10.2	11.8	11.0	10.2	11.2	10.9	4.0	-
CoPE [44]	16.6	16.3	21.7	22.4	17.6	18.6	8.8	-
LwF [12]	12.5	12.4	21.9	20.9	10.5	11.9	4.2	-
RM [30]	12.0	12.4	9.8	20.8	18.2	9.3	4.2	-
AAN [43]	14.0	15.6	13.2	17.6	10.6	15.0	6.6	-
Lower bound	12.1	12.8	15.5	16.9	12.8	16.4	3.5	3.0
Upper bound	85.3	84.6	91.0	92.0	91.3	91.4	69.0	56.1

TABLE II
CRUMB USES ONLY 3-4% OF REMIND'S PEAK RAM USAGE, AND ITS RUNTIME IS ONLY 22-25% OF REMIND'S.

Dataset	Peak RAM (GB) Ours	Peak RAM (GB) REMIND	Runtime (hours) Ours	Runtime (hours) REMIND
CIFAR100	**0.036**	0.87	**0.29**	1.91
Imagenet	**1.66**	44.34	**7.86**	35.64

Discussion

1. La posibilidad de aprender cosas nuevas sin olvidar lo que ya sabemos es critica en Biologia y en AI

2. El aprendizaje continuo es complicado para redes neuronales que reusan los mismos circuitos

3. Dos estrategias: proteccion de sinapsis y repeticion

4. Teoria dinamica de aprendizaje continuo en redes neuronales con atractores

5. CRUMB: una red neuronal con memoria usa repeticion eficientemente para aprender en forma continua

Continual learning without forgetting what we already know

CENTER FOR Brains Minds+ Machines

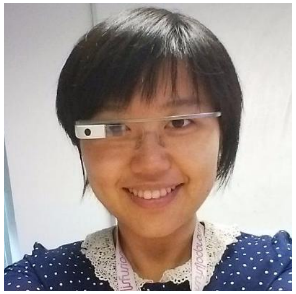
Aprendiendo cosas nuevas sin olvidar lo que ya sabemos

Gabriel Kreiman

klab.tch.harvard.edu

Trabajo en conjunto con:

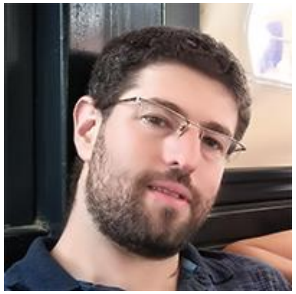
Mengmi Zhang



Shashi Gupta

Morgan Talbot

Nimrod Shaham



Haim Sompolinsky

Scan to download papers+data+code