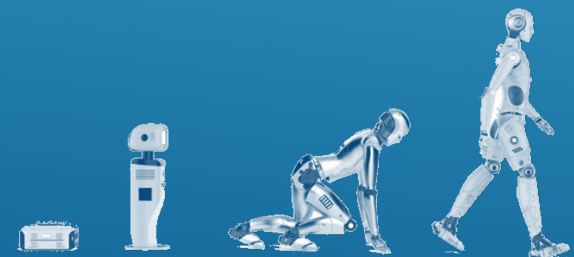
CONTINUAL LEARNING

Constant Marks

PhD Oral Exam

February 5th, 2021



GOALS OF CONTINUAL LEARNING

Efficient

Models need not be trained from scratch, re-learning things we have already learned.

Adaptive

Fast learning leads to adaptable and customizable models.

Scalable

Bounded computational and memory overhead

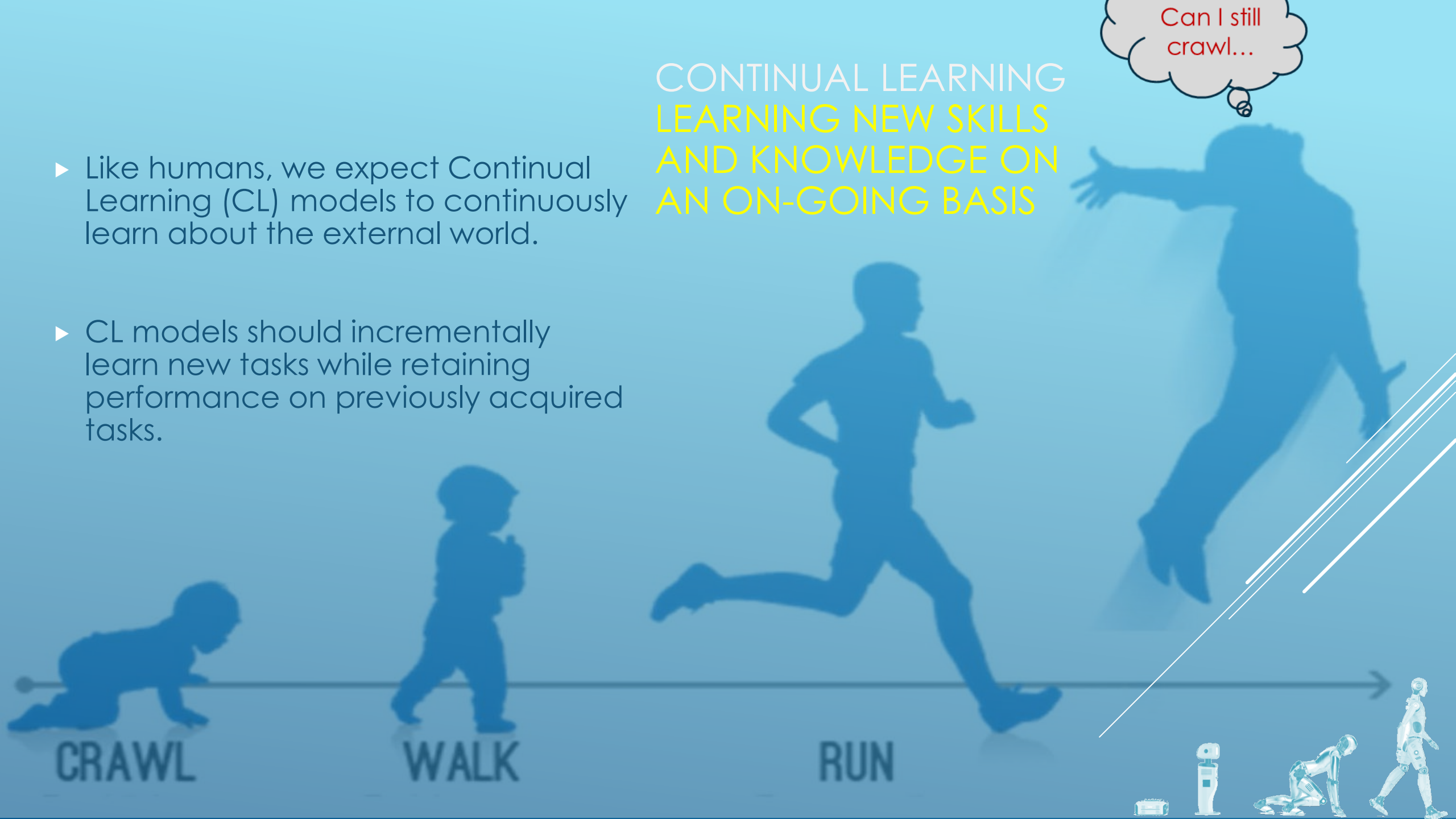
" … the resultant of the process of acquiring, storing in memory, retrieving, combining, comparing, and using in new contexts information and conceptual skills."

--Lloyd Humphreys
The construct of general intelligence

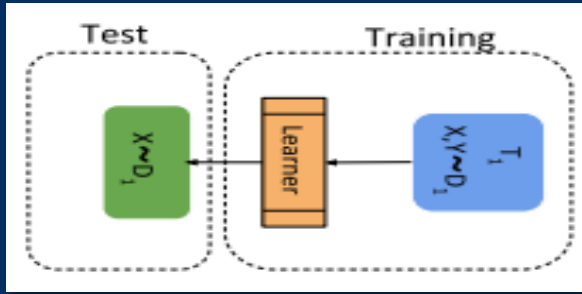
Lamonaco, 2019 https://medium.com/continual-ai/continual-learning-for-production-systems-304cc9f60603

CONTINUAL LEARNING
LEARNING NEW SKILLS AND KNOWLEDGE ON AN ON-GOING BASIS

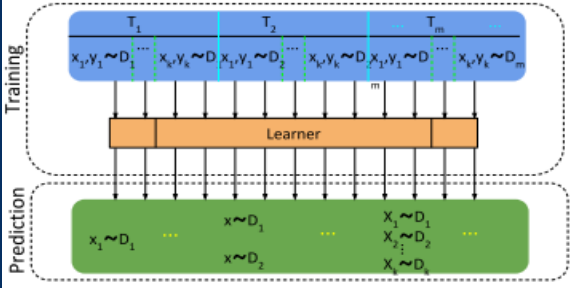
- Like humans, we expect Continual Learning (CL) models to continuously learn about the external world.
- CL models should incrementally learn new tasks while retaining performance on previously acquired tasks.



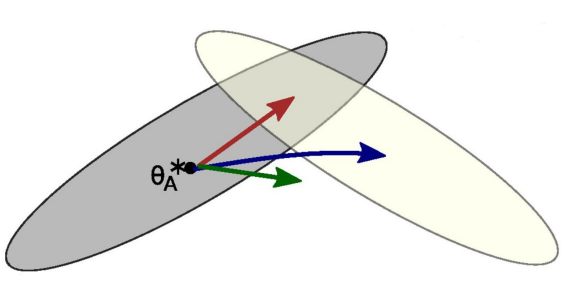
TRADITIONAL ARTIFICIAL NEURAL NETWORK (ANN) TRAINING



Trained with data sampled from a **representative** and **static** distribution, ANNs have surpassed human abilities in many increasingly complex tasks.



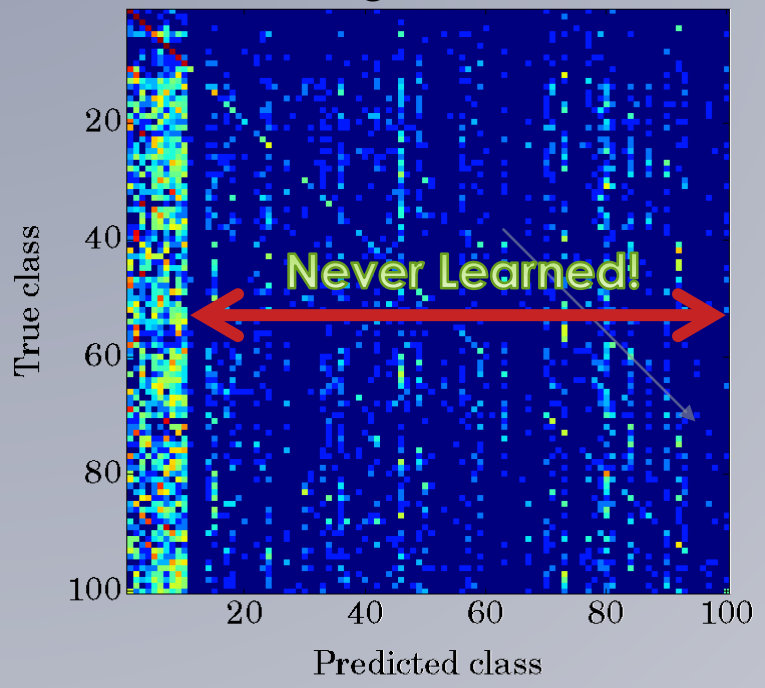
Unlike most curated datasets, real-world data is rarely static but rather **shifting** as new tasks are encountered over time.



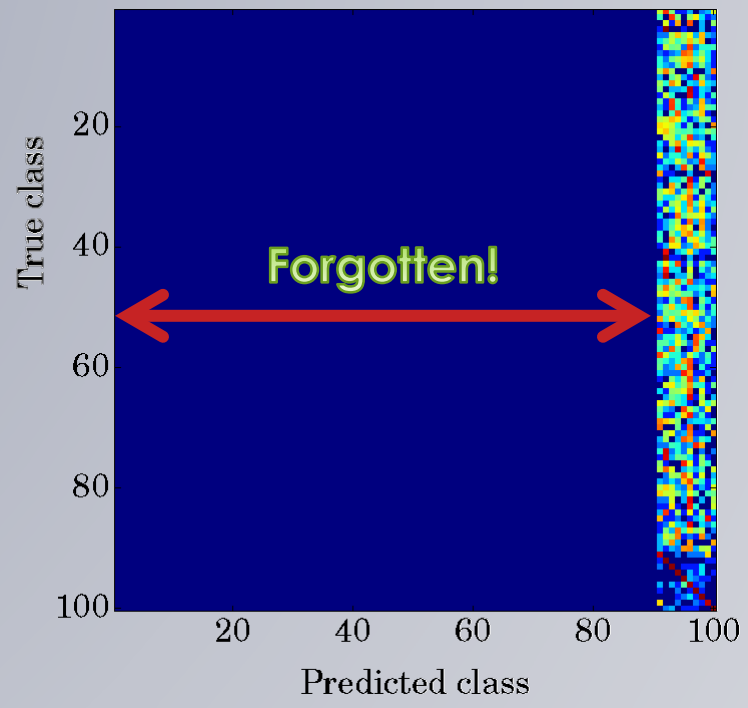
Trained sequentially from shifting distributions, ANNs are are prone to **catastrophic forgetting** of previous tasks.

CATASTROPHIC FORGETTING

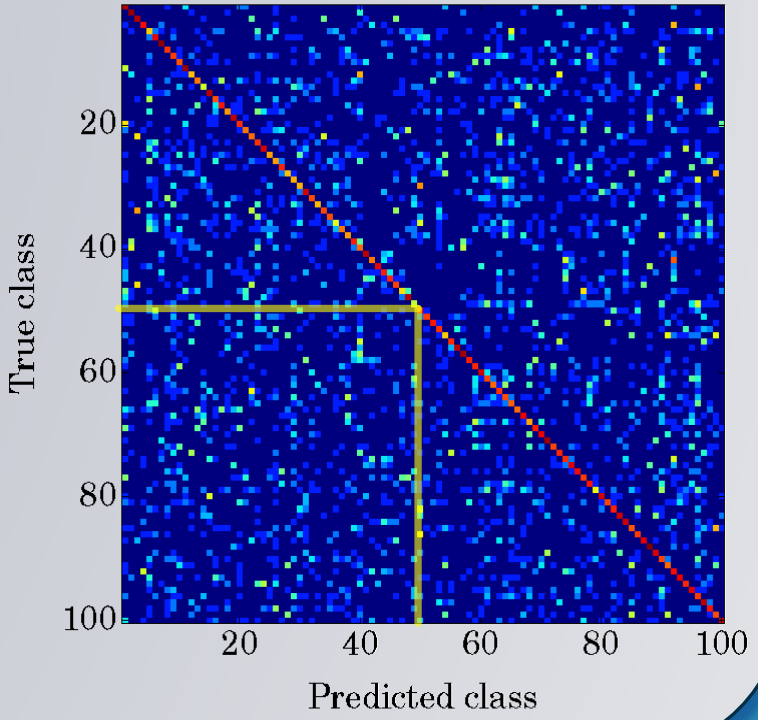
Fixed weights after learning first 10 classes



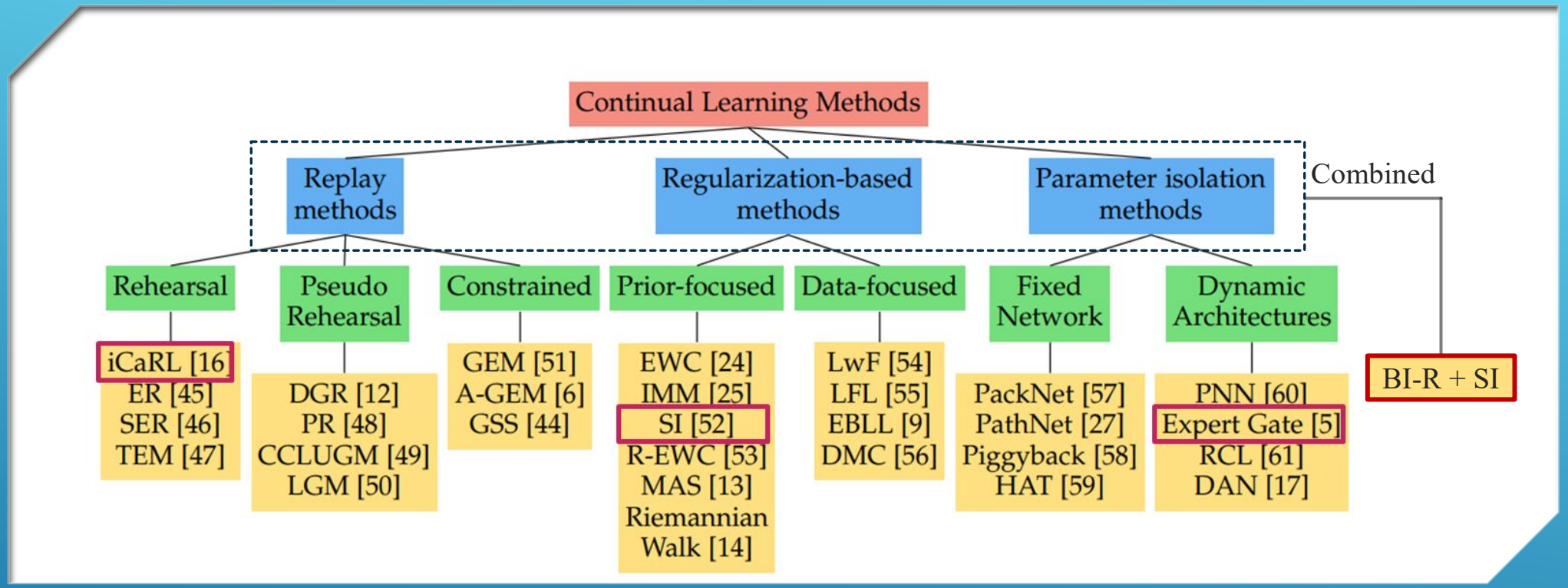
No fixing of weights (finetuning on last 10 classes)



Continual Learning (iCaRL)

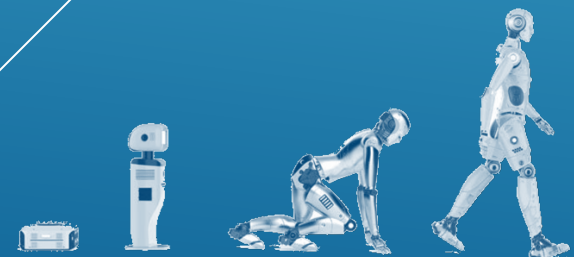


Rebuffi, 2017

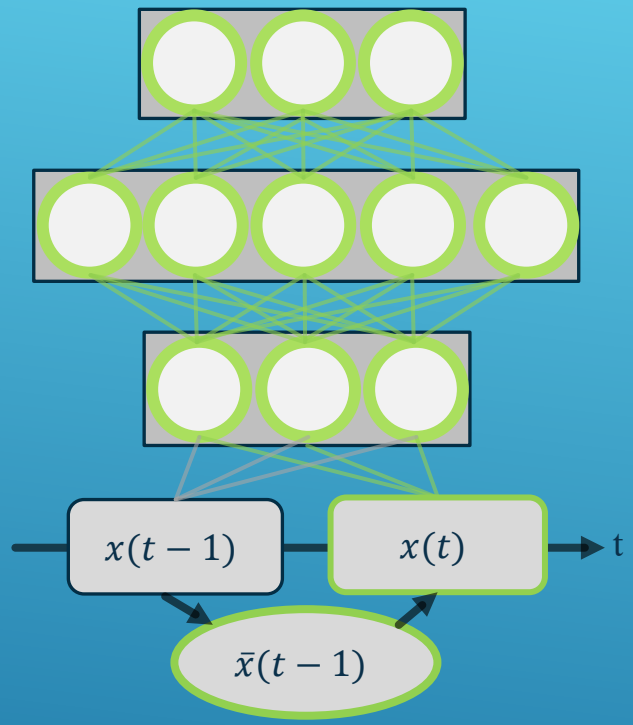


CURRENT APPROACHES

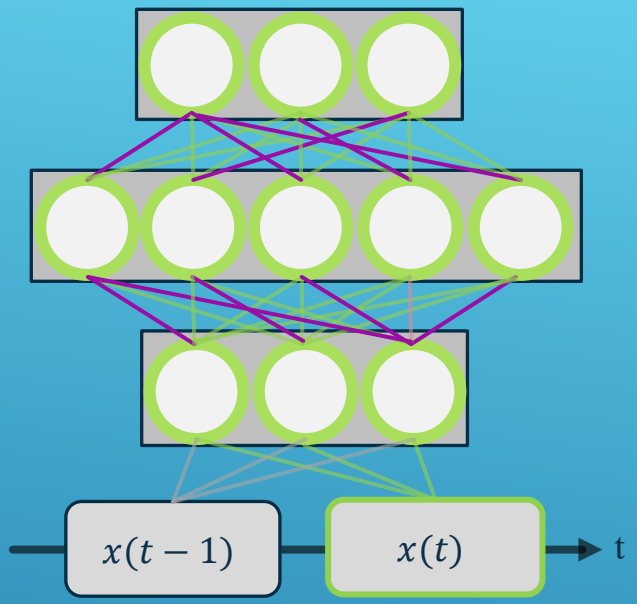
De Lange, 2020 https://arxiv.org/abs/1909.08383



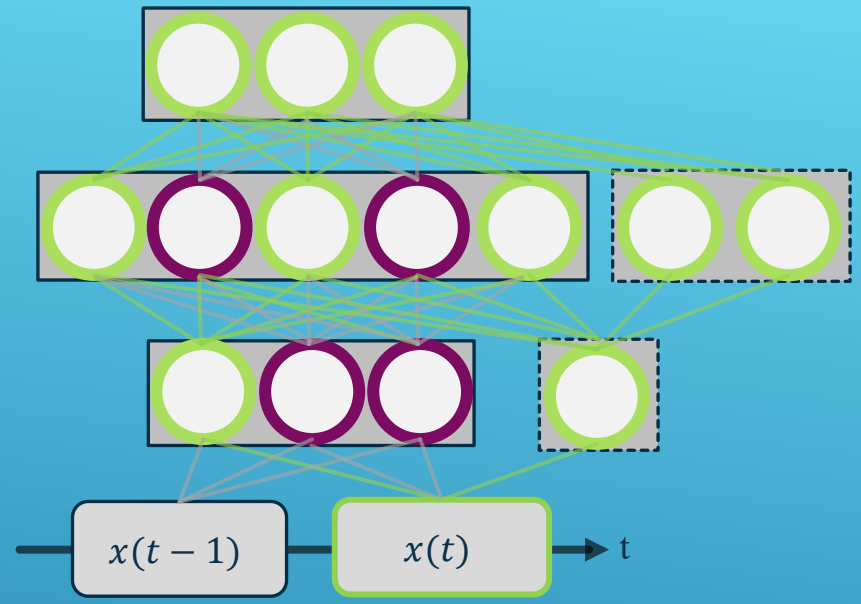
OVERVIEW OF CURRENT APPROACHES



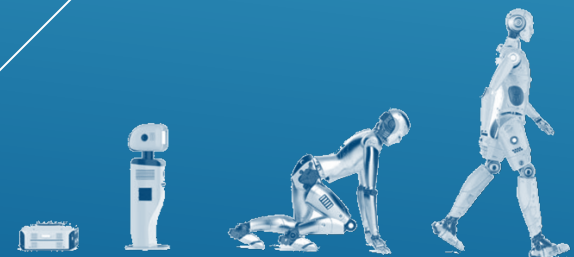
Retraining with Replay



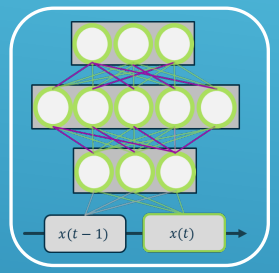
Retraining with Regularization



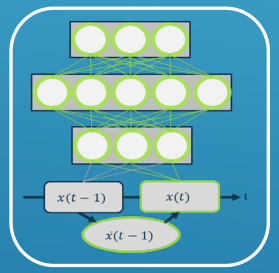
Retraining with Parameter Isolation/Expansion



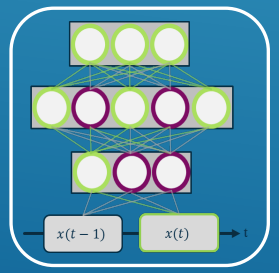
Continuous Learning History in Three Models



Regularization:
Synaptic Intelligence(SI)
Zenke et al., PMLR, 2018



Replay
Incremental Classifier and Representation Learning (iCaRL)
Rebuffi et al., CVPR, 2017



Parameter Isolation
Context-dependent gating (XdG)
Masse et al., PNAS, 2018

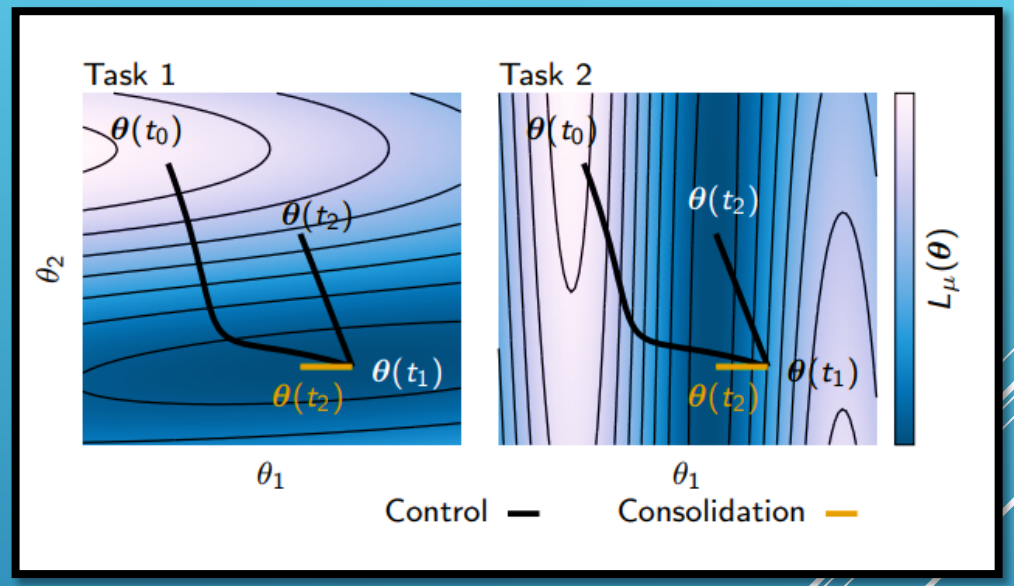
REGULARIZATION – SI

Goal

Find a solution to a new task **in the neighborhood** of an older one.

Observation

- For an over parameterized model many configurations of θ will result in in the same performance.
- It is likely then that there is a solution for task B, θ_B^* that is close to that of task A , θ_A^*.



Problem

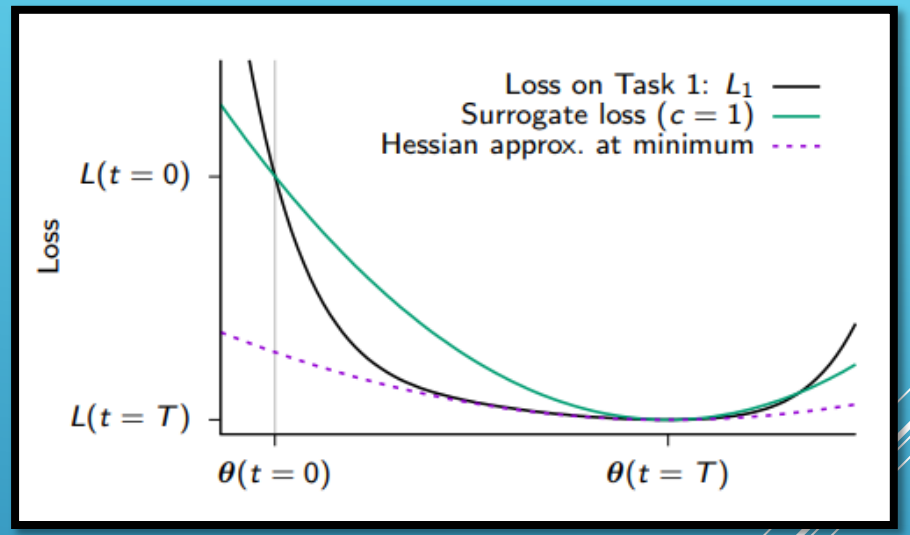
Protects the performance on task A by constraining the parameters to stay in a region of low error for task A centered around θ_A^*

Zenke, F., Poole, B. and Ganguli, S., 2017, July. Continual learning through synaptic intelligence. In *International Conference on Machine Learning* (pp. 3987-3995). PMLR.

REGULARIZATION – SI

Method

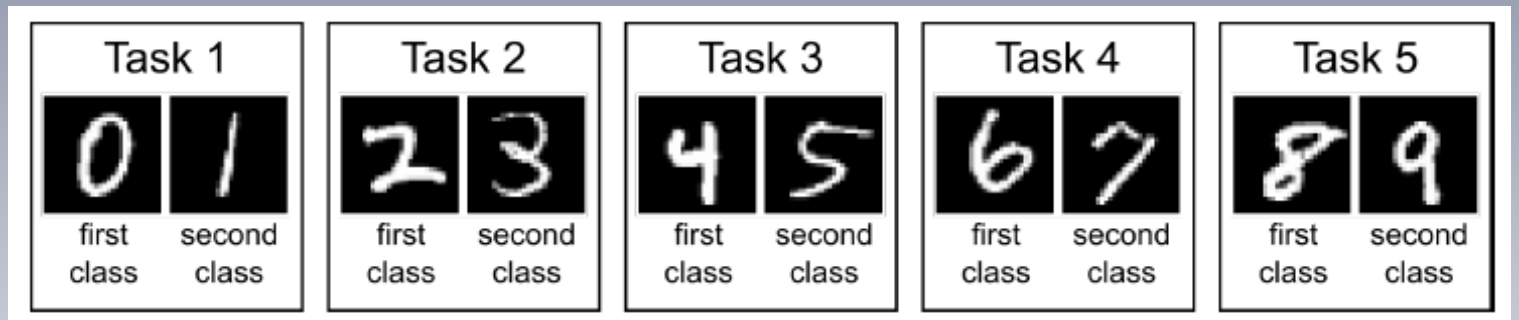
- Importance of a parameter θ_k for a single task is determined by two quantities:
  1. How much an individual parameter contributed to a drop in the loss ω_k^ν over the entire trajectory of training
     - Approximated by running sum of the product of the gradient with the parameter update
  2. How far it moved $\Delta_k^\nu \equiv \theta_k(t^\nu) - \theta_k(t^{\nu-1})$



$$\tilde{L}_\mu = L_\mu + \underbrace{c \sum_k \Omega_k^\mu \left(\tilde{\theta}_k - \theta_k\right)^2}_{\text{surrogate loss}} \qquad \Omega_k^\mu = \sum_{\nu<\mu} \frac{\omega_k^\nu}{(\Delta_k^\nu)^2 + \xi}$$

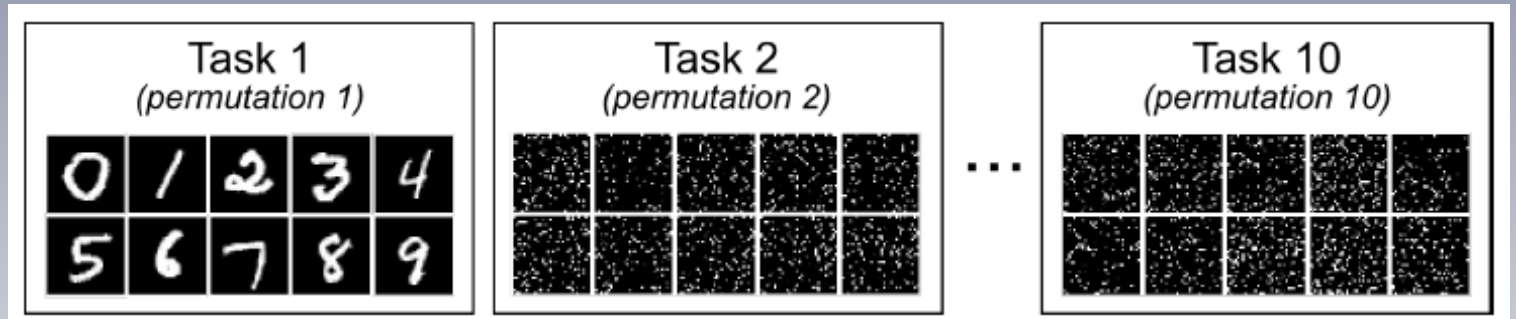
Zenke, F., Poole, B. and Ganguli, S., 2017, July. Continual learning through synaptic intelligence. In *International Conference on Machine Learning* (pp. 3987-3995). PMLR.

DATA SETS AND THREE CONTINUOUS LEARNING SCENARIOS



Split MNIST task protocol

Task-IL	Given task, 1st or 2nd class?
Domain-IL	With task unknown, 1st or 2nd class?
Class-IL	With task unknown, which digit is it?



Permuted MNIST task protocol

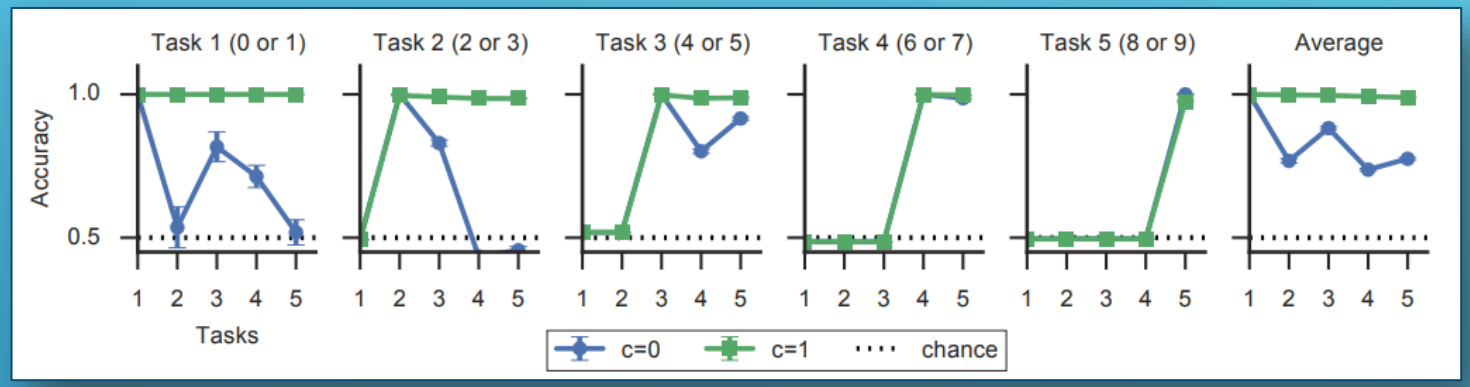
Task-IL	Given permutation, which digit?
Domain-IL	With permutation unknown, which digit?
Class-IL	Which digit and which permutation?

van de Ven, 2019 https://arxiv.org/abs/1904.07734

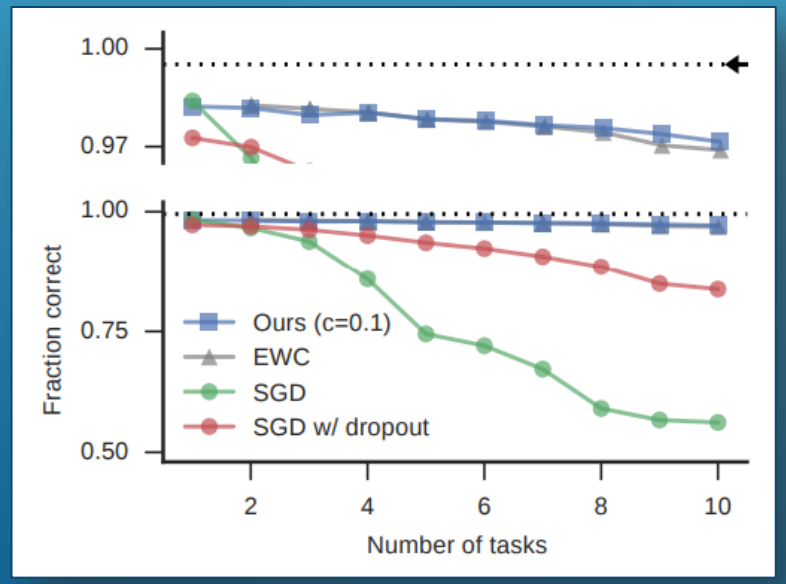
REGULARIZATION – SI

Results

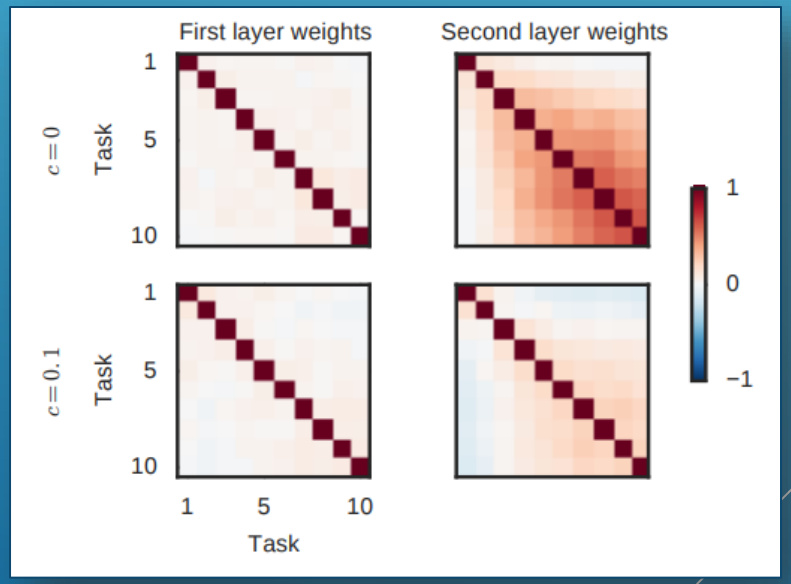
Task-IL - Split MNIST accuracy as a function of task trained so far



Task-IL – Permuted MNIST accuracy



Weight Importance correlation



Zenke, F., Poole, B. and Ganguli, S., 2017, July. Continual learning through synaptic intelligence. In *International Conference on Machine Learning* (pp. 3987-3995). PMLR.

REPLAY - iCaRL

Goal

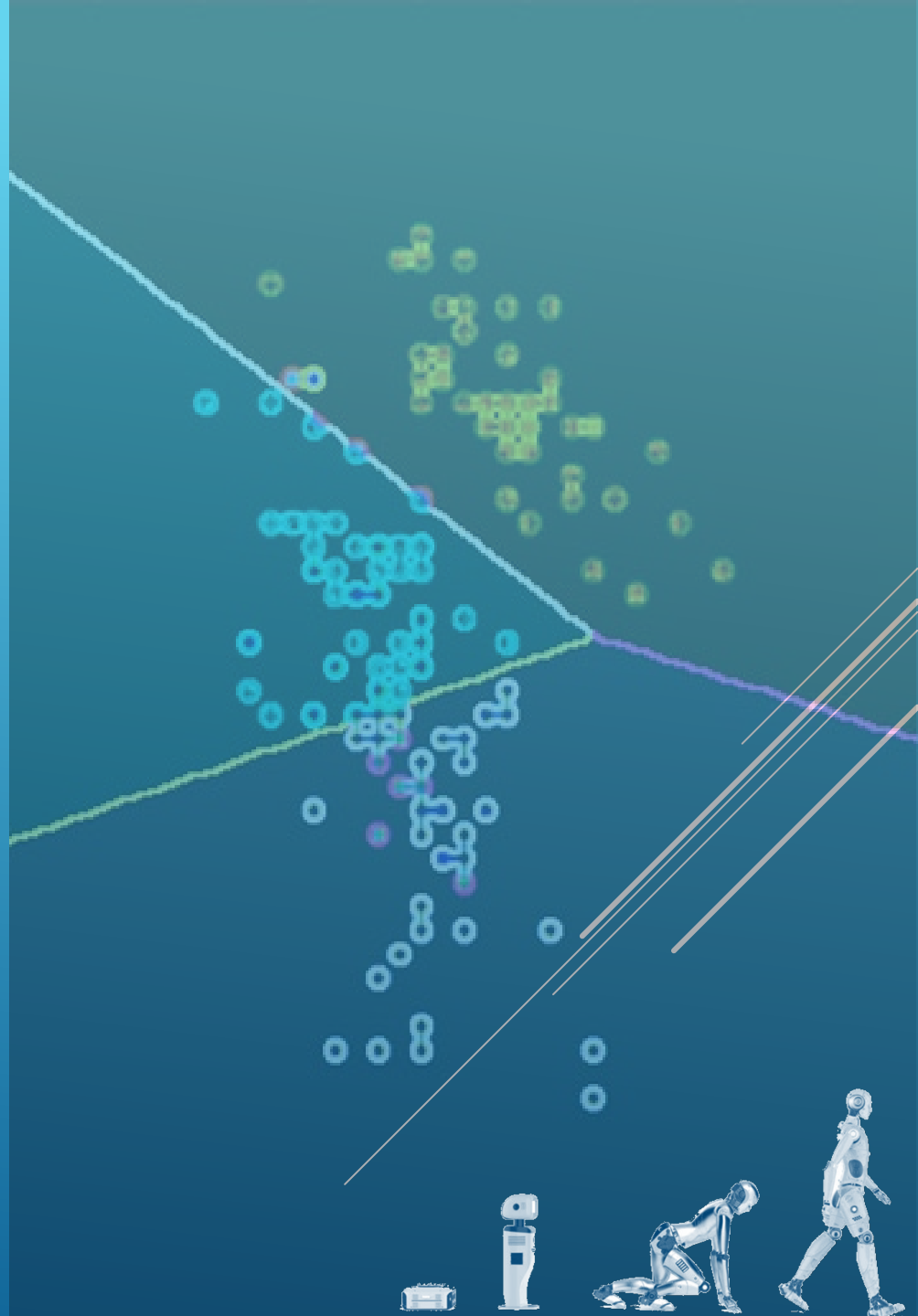
Identify most relevant samples to keep for replay and replace traditional softmax with nearest-mean classifier.

Observation

- The typical classification rule is equivalent to the use of a linear classifier with non-linear feature map φ and weight vectors.
- If φ changes, all weights must be updated as well, or the network outputs will change uncontrollably - observable as catastrophic forgetting.

Problem

- Develop nearest-mean-of-exemplars classifier that remains useful as class-prototypes automatically change whenever the feature representation changes.



Rebuffi, S.A., Kolesnikov, A., Sperl, G. and Lampert, C.H., 2017. iCaRL: Incremental classifier and representation learning. CVPR.

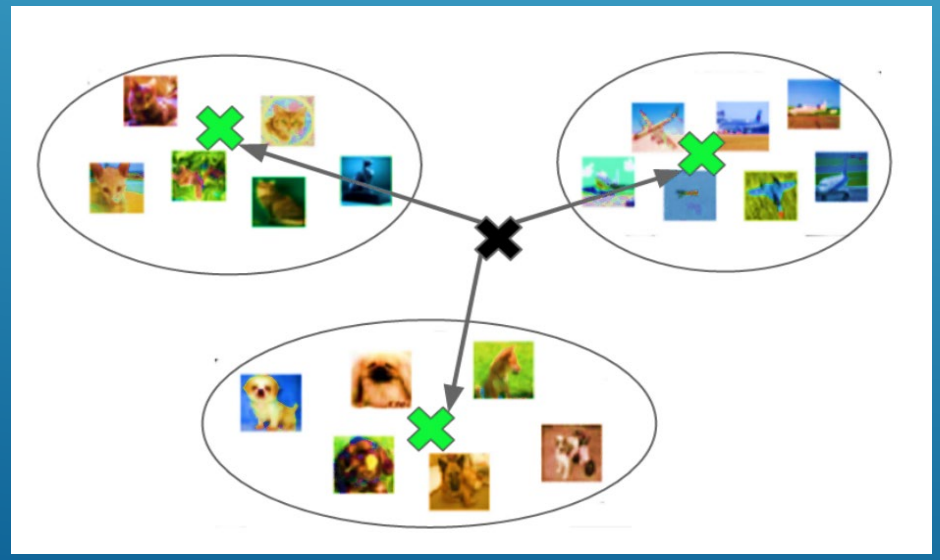
REPLAY - iCaRL

Method

Classifying a new sample image:

1. Nearest-Mean-of-Exemplars Classification

1. compute a prototype vector for each class observed so far
2. compute the feature vector of the image that should be classified
3. assigns the class label with most similar prototype



For each new task:

1. Representation Learning

1. Augment current training samples with exemplar set
2. Test and save network outputs for old classes
3. Updated network parameters by minimizing combined **classification** and **distillation** loss function

2. Exemplar Selection

1. Select exemplars that cause the average feature vector to best approximate the average over all training examples
2. Prune exemplars to satisfy buffer constraint

Rebuffi, S.A., Kolesnikov, A., Sperl, G. and Lampert, C.H., 2017. iCaRL: Incremental classifier and representation learning. CVPR.

Split CIFAR-100

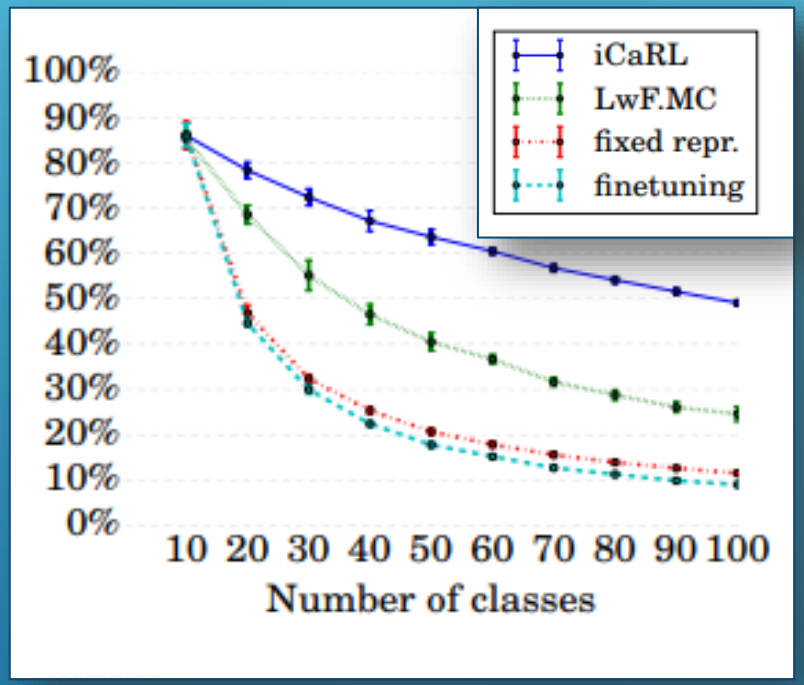
Task-IL	Given task, class 1-10?
Domain-IL	With task unknown, class 1-10?
Class-IL	With task unknown, which class is it (1-100)?

CIFAR-100 DATA SET AND THREE CONTINUOUS LEARNING SCENARIOS

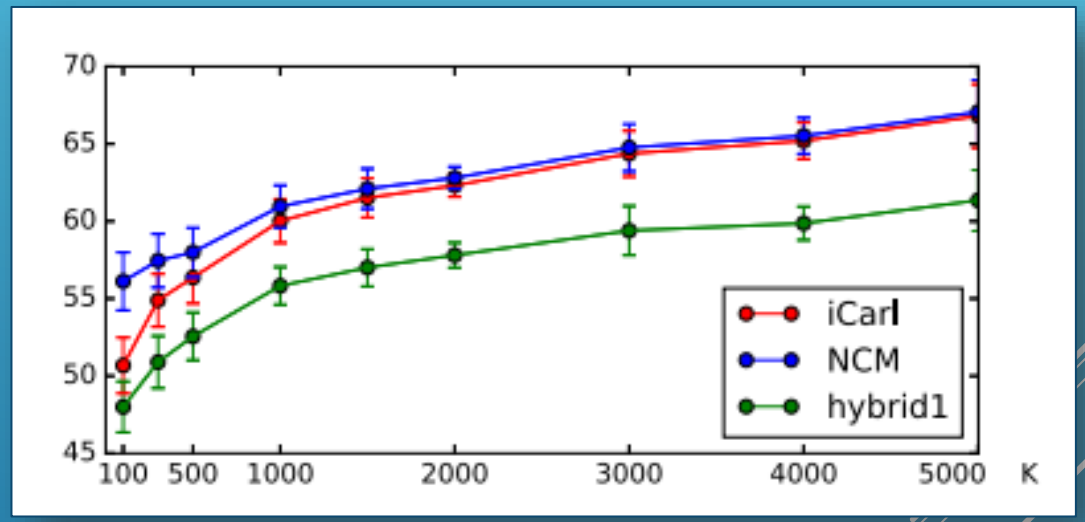
van de Ven, 2019 https://arxiv.org/abs/1904.07734

Results

Class-IL – Split CIFAR-100



Effect of Memory Budget



Rebuffi, S.A., Kolesnikov, A., Sperl, G. and Lampert, C.H., 2017. iCaRL: Incremental classifier and representation learning. CVPR.

DYNAMIC ARCH – XdG

Goal

Adding a context-dependent gating signal, such that only sparse, mostly nonoverlapping patterns of units are active for any one task.

Observation

- It is uncertain whether synaptic stabilization (e.g., regularization) alone can support continual learning across large numbers of tasks.

- Neuroscience studies have proposed that diverse mechanisms act to stabilize learned information, raising the question as to whether several complementary algorithms are required to support continual learning in ANNs.

Problem

- Develop a complementary system for guiding task incremental learning through the main ANN.

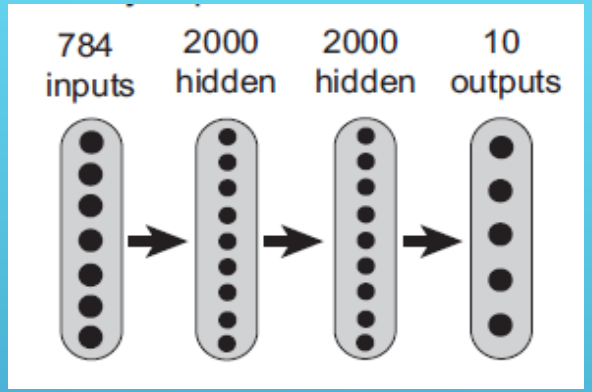
Masse, N.Y., Grant, G.D. and Freedman, D.J., 2018. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *PNAS*.

DYNAMIC ARCH – XdG

Method

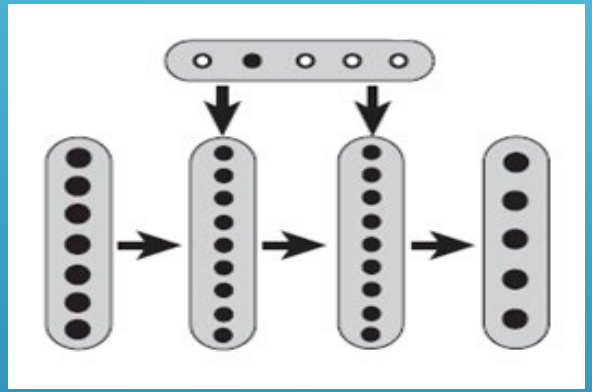
1. Network with Stabilization

- Implemented with EWC or SI



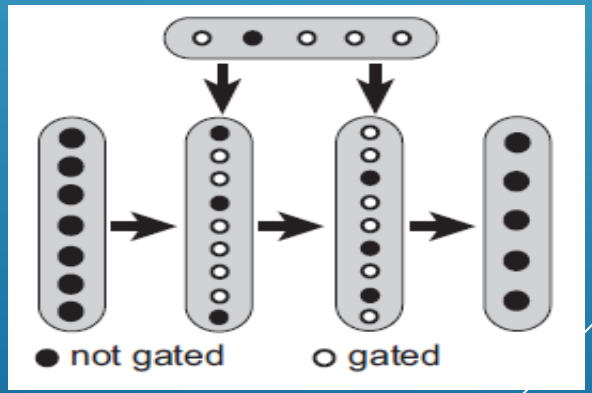
2. Learning context signals

- Task information is provided as secondary input (One-hot vector) to hidden layers with
- Context input weights are trainable



3. Context-dependent gating

- Random selected gating of hidden units - 80%/20% split for each task
- Identity of gated units fixed for each task for training and testing

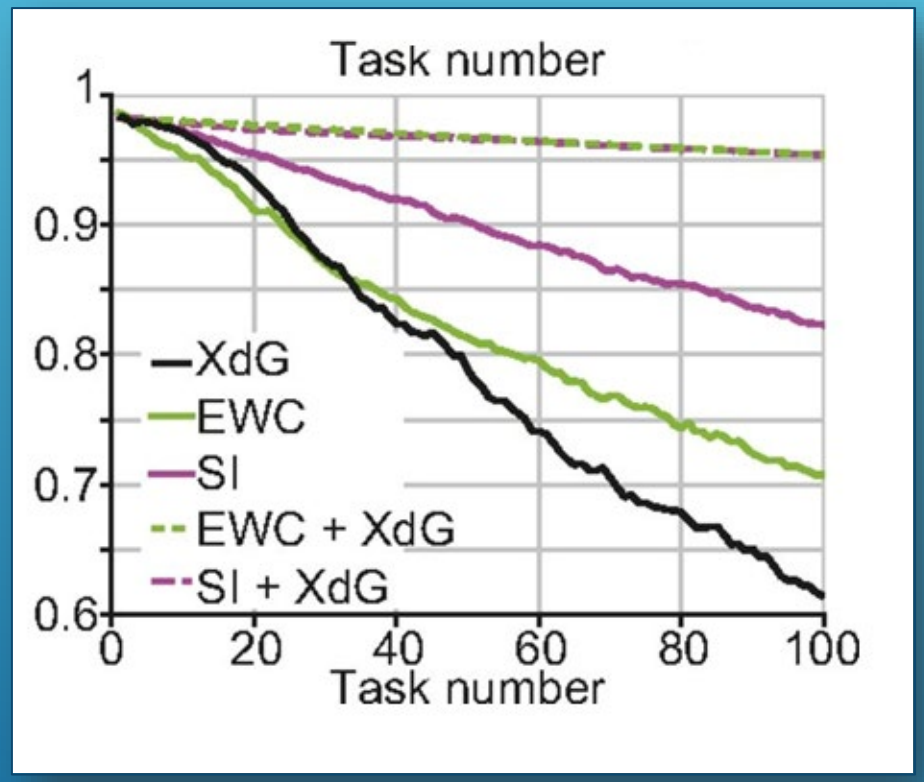


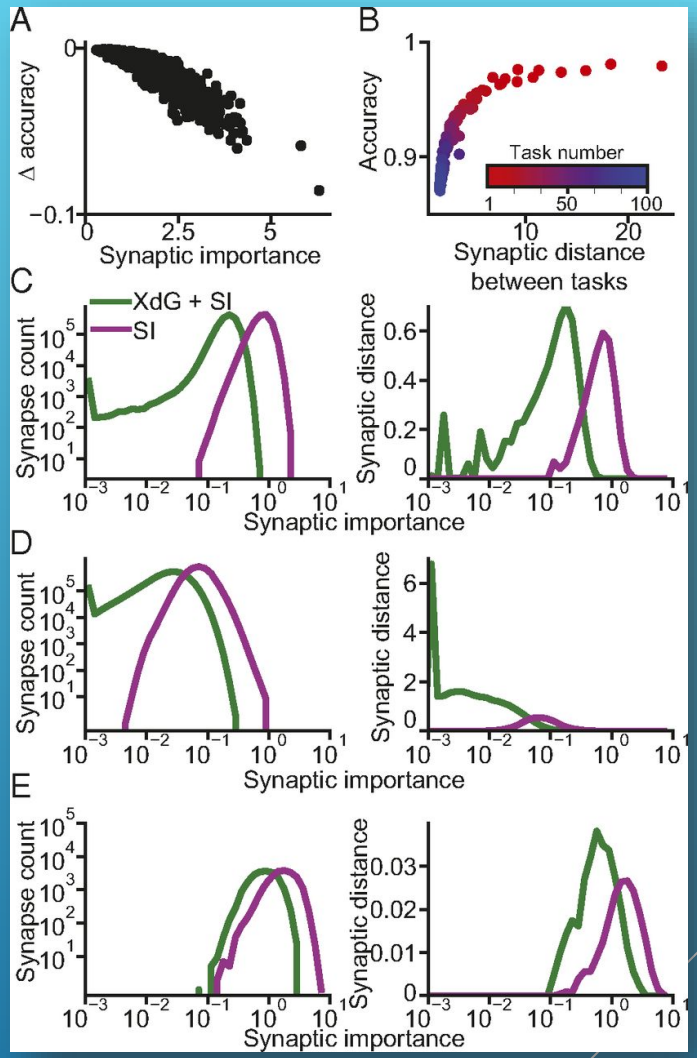
Masse, N.Y., Grant, G.D. and Freedman, D.J., 2018. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *PNAS.*

DYNAMIC ARCH – XdG

Results

Task-IL – Permuted MNIST accuracy





Masse, N.Y., Grant, G.D. and Freedman, D.J., 2018. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *PNAS.*

MODEL COMPARISON

SPLIT MNIST

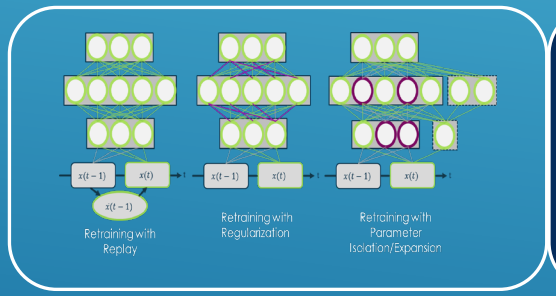
Approach	Method	Task-IL	Domain-IL	Class-IL
Baselines	*None – lower bound*	*87.19 (± 0.94)*	*59.21 (± 2.04)*	*19.90 (± 0.02)*
	Offline – upper bound	*99.66 (± 0.02)*	*98.42 (± 0.06)*	*97.94 (± 0.03)*
Task-specific	XdG	99.10 (± 0.08)	-	-
Regularization	EWC	98.64 (± 0.22)	63.95 (± 1.90)	20.01 (± 0.06)
	Online EWC	99.12 (± 0.11)	64.32 (± 1.90)	19.96 (± 0.07)
	SI	99.09 (± 0.15)	65.36 (± 1.57)	19.99 (± 0.06)
Replay	LwF	99.57 (± 0.02)	71.50 (± 1.63)	23.85 (± 0.44)
	DGR	99.50 (± 0.03)	95.72 (± 0.25)	90.79 (± 0.41)
	DGR+distill	99.61 (± 0.02)	96.83 (± 0.20)	91.79 (± 0.32)
Replay + Exemplars	iCaRL (budget = 2000)	-	-	94.57 (± 0.11)

PERMUTED MNIST

Approach	Method	Task-IL	Domain-IL	Class-IL
Baselines	*None – lower bound*	*81.79 (± 0.48)*	*78.51 (± 0.24)*	*17.26 (± 0.19)*
	Offline – upper bound	97.68 (± 0.01)	97.59 (± 0.01)	97.59 (± 0.02)
Task-specific	XdG	91.40 (± 0.23)	-	-
Regularization	EWC	94.74 (± 0.05)	94.31 (± 0.11)	25.04 (± 0.50)
	Online EWC	95.96 (± 0.06)	94.42 (± 0.13)	33.88 (± 0.49)
	SI	94.75 (± 0.14)	95.33 (± 0.11)	29.31 (± 0.62)
Replay	LwF	69.84 (± 0.46)	72.64 (± 0.52)	22.64 (± 0.23)
	DGR	92.52 (± 0.08)	95.09 (± 0.04)	92.19 (± 0.09)
	DGR+distill	97.51 (± 0.01)	97.35 (± 0.02)	96.38 (± 0.03)
Replay + Exemplars	iCaRL (budget = 2000)	-	-	94.85 (± 0.03)

van de Ven, G.M. and Tolias, A.S., 2019. Three scenarios for continual learning. arXiv preprint arXiv:1904.07734.

Current State of the Art Continuous Learning



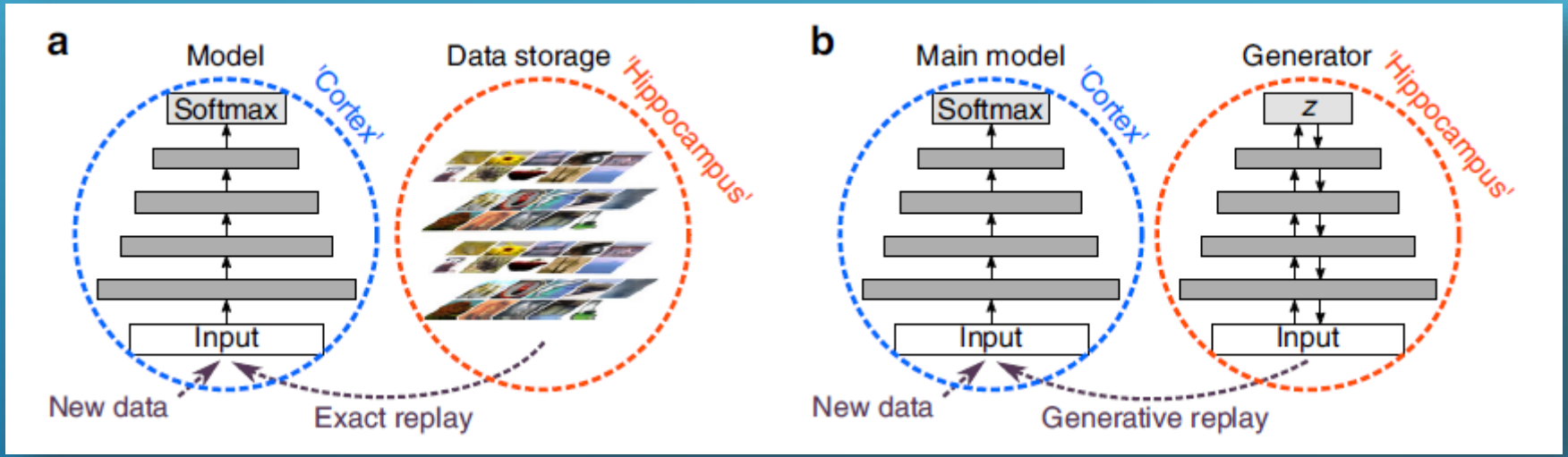
Complementary Systems

Brain Inspired Replay(BI-R)

van de Ven et al., Nature Communications, 2020

Brain-inspired replay - Generative Replay Model

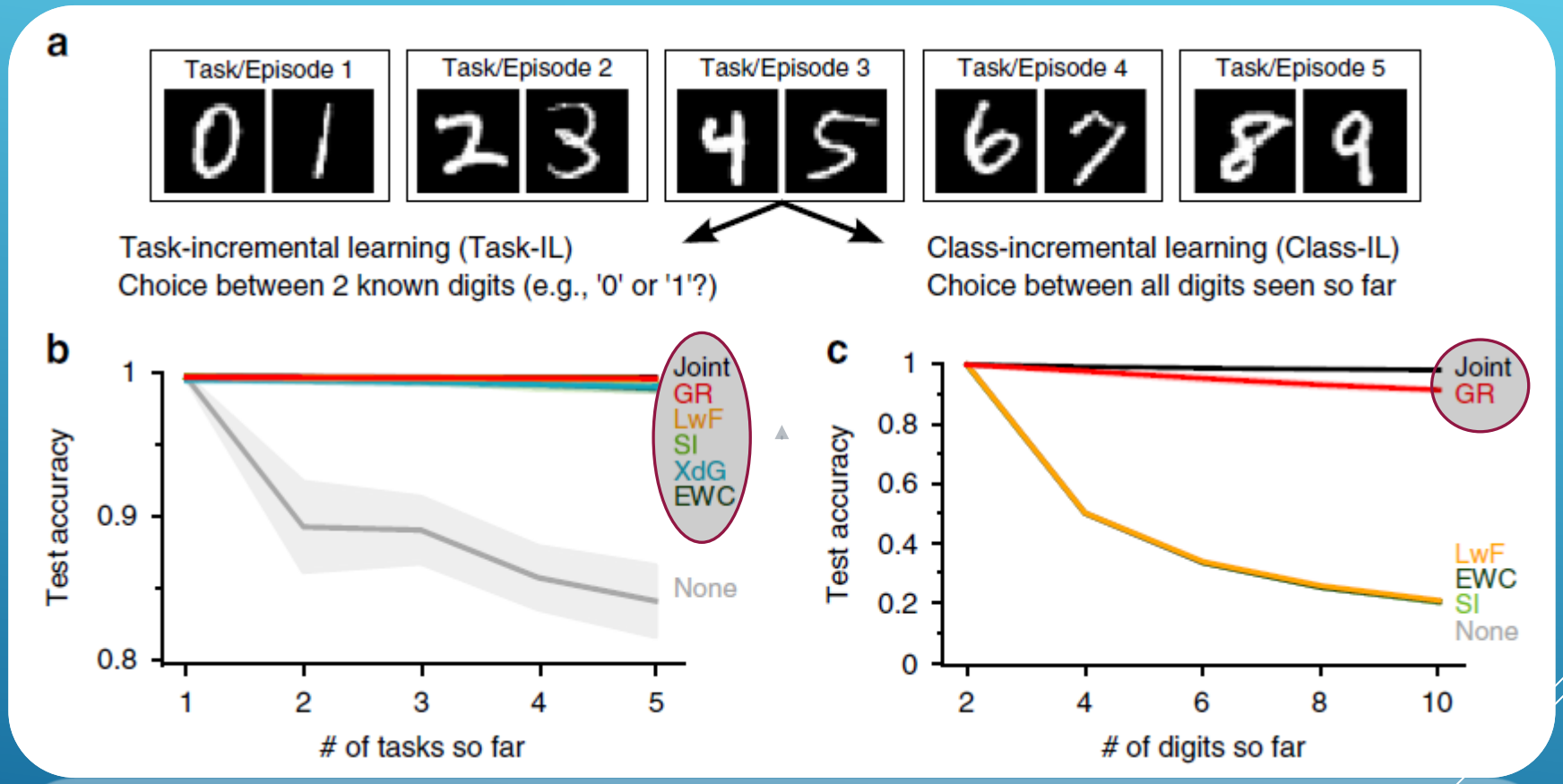
- Rehearsal is key… but storage is problematic.
- Build a separate generative model to generate replay data



a Exact or experience replay, which views the hippocampus as a memory buffer in which experiences can simply be stored
b Generative replay with a separate generative model, which views the hippocampus as a generative neural network

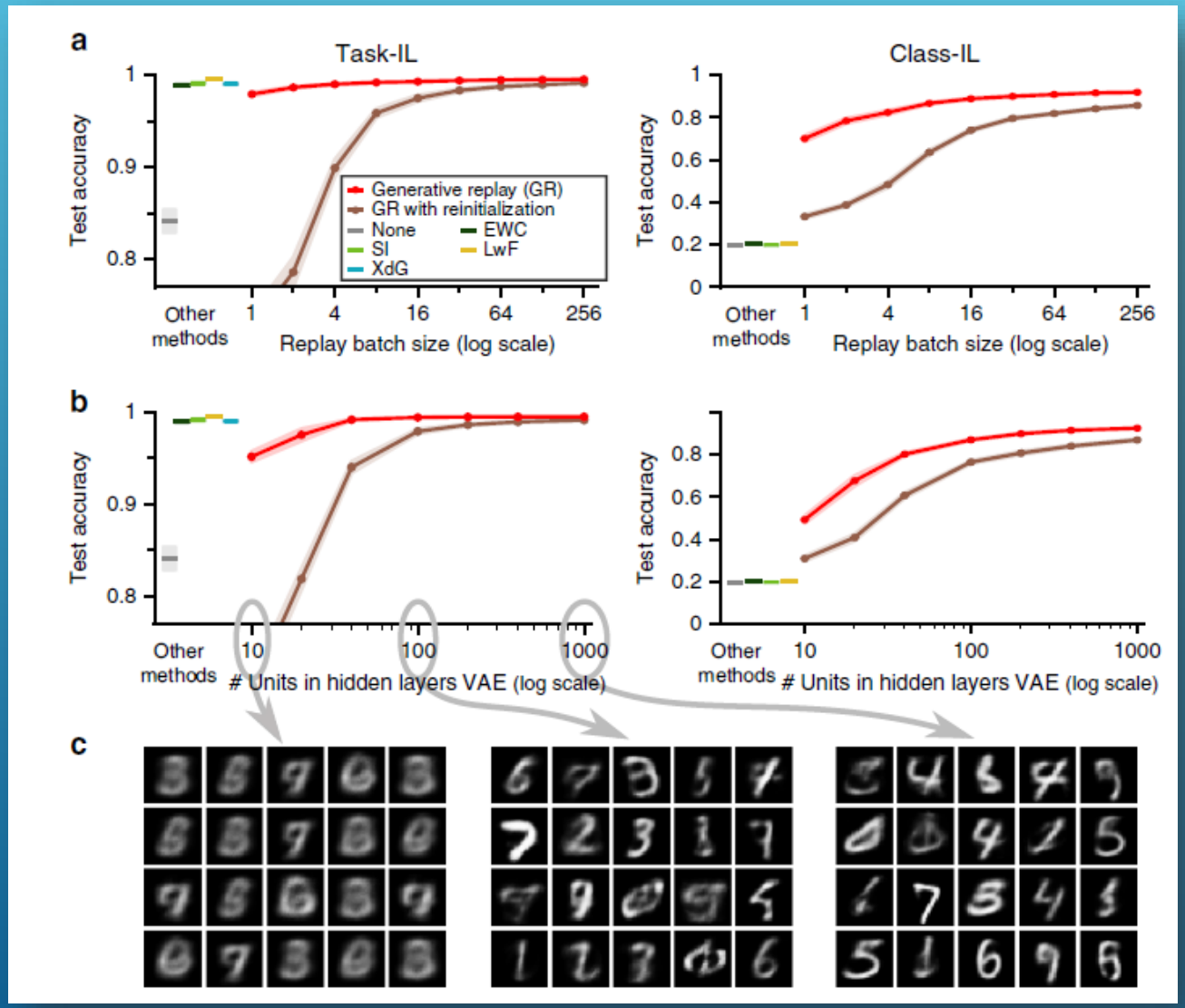
van de Ven, G.M., Siegelmann, H.T. and Tolias, A.S., 2020. Brain-inspired replay for continual learning with artificial neural networks. Nature communications.

Results on Split MNIST



van de Ven, G.M., Siegelmann, H.T. and Tolias, A.S., 2020. Brain-inspired replay for continual learning with artificial neural networks. Nature communications.

Performance – Generative Replay is robust and Efficient

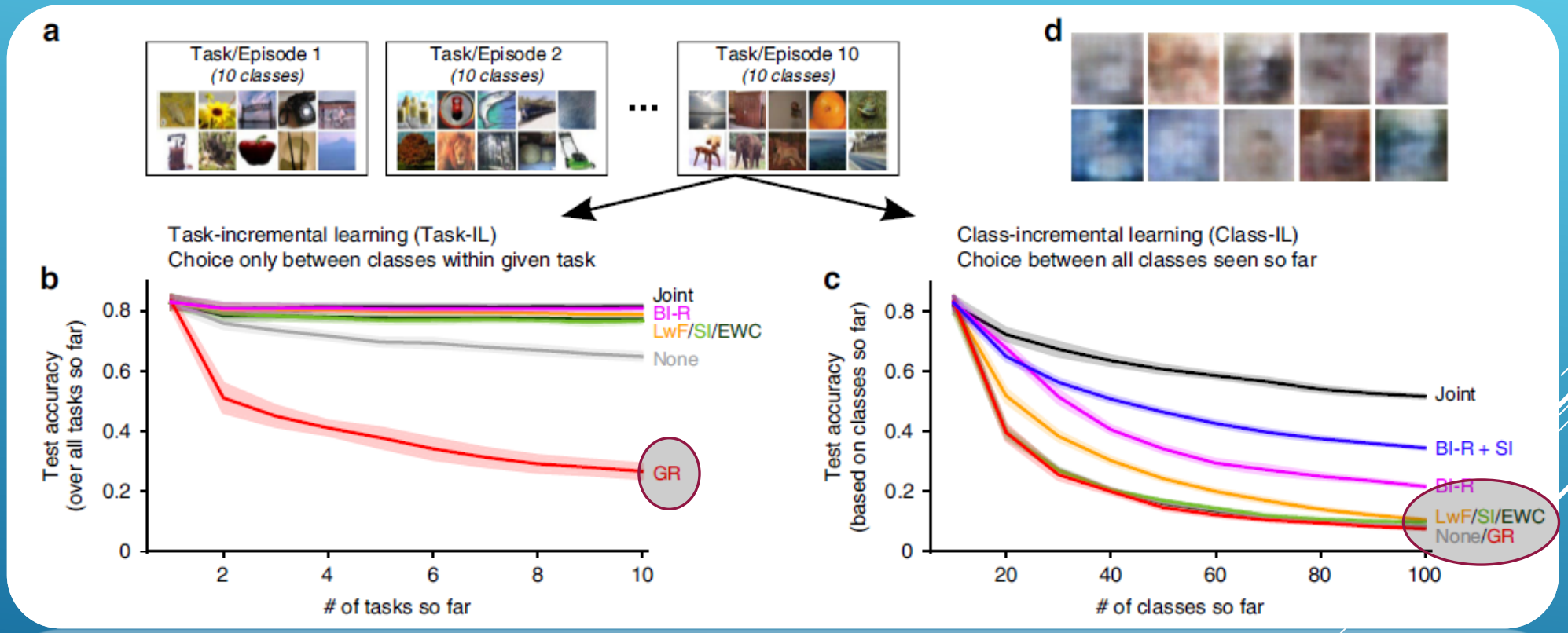


- Substantially reducing the quantity or the quality of replay does not severely affect performance
- Panel c shows random samples from the generative model after finishing training on the fourth task (i.e., examples of what is replayed during training on the final task).

van de Ven, 2020 https://www.nature.com/articles/s41467-020-17866-2

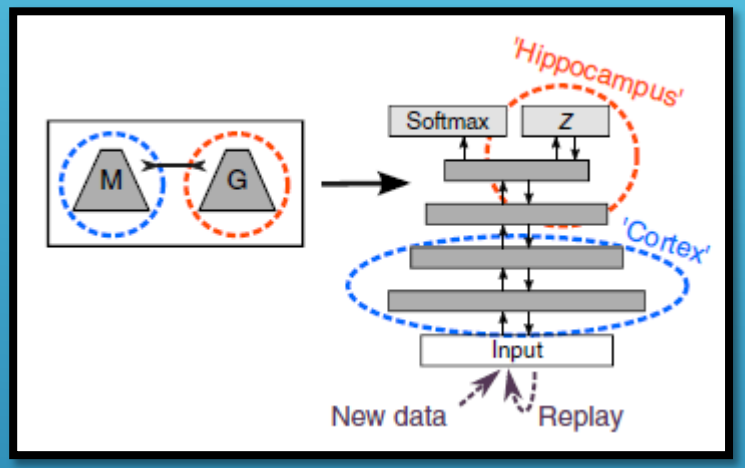
CURRENT STATE — BI-R

MNIST is not that hard…. CIFAR-100?



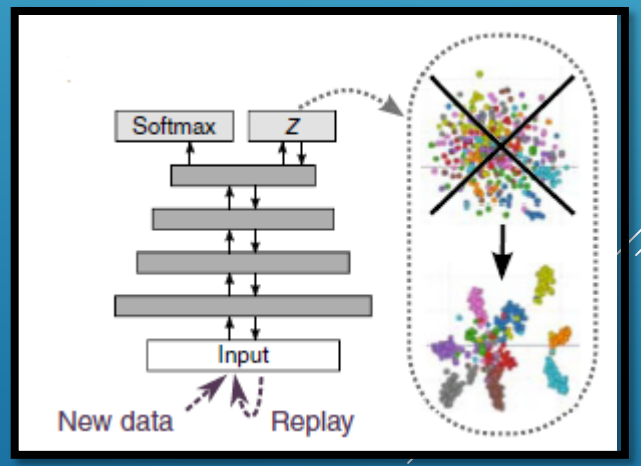
van de Ven, G.M., Siegelmann, H.T. and Tolias, A.S., 2020. Brain-inspired replay for continual learning with artificial neural networks. Nature communications.

Brain-inspired Modifications to the GR model

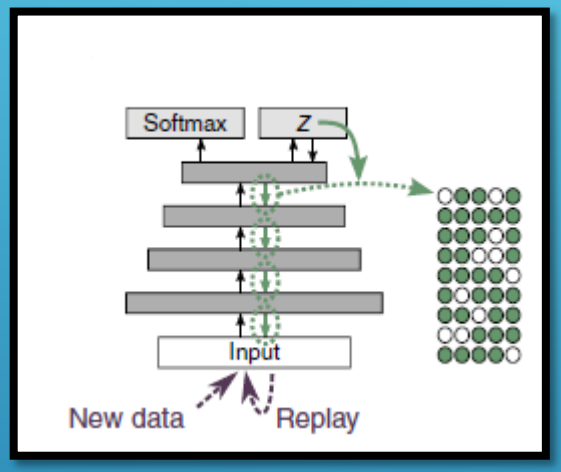


1. Replay-through-feedback
   Merge the main and generator models by adding backward connections

2. Conditional Replay
   - The standard normal prior is replaced by a Gaussian mixture with a separate mode for each class.
   - Restricting the sampling of the latent variables to their corresponding modes.



Brain-inspired Modifications to the GR model

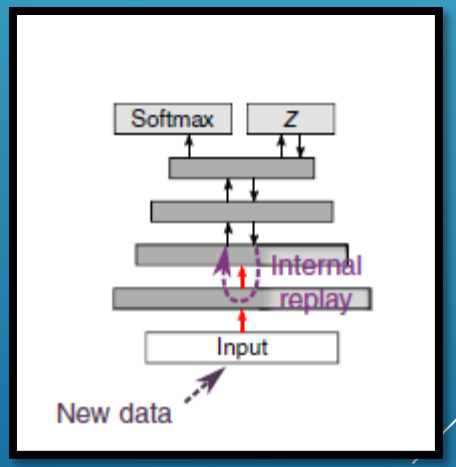


3. Gating based on internal context

- Random task-specific gating (XdG) is not possible when the task is known, as in Class-IL.
- Only the decoder part uses context gates by conditioning on an internal context.
- The internal context is the given by the class to be generated.

4. Internal Replay

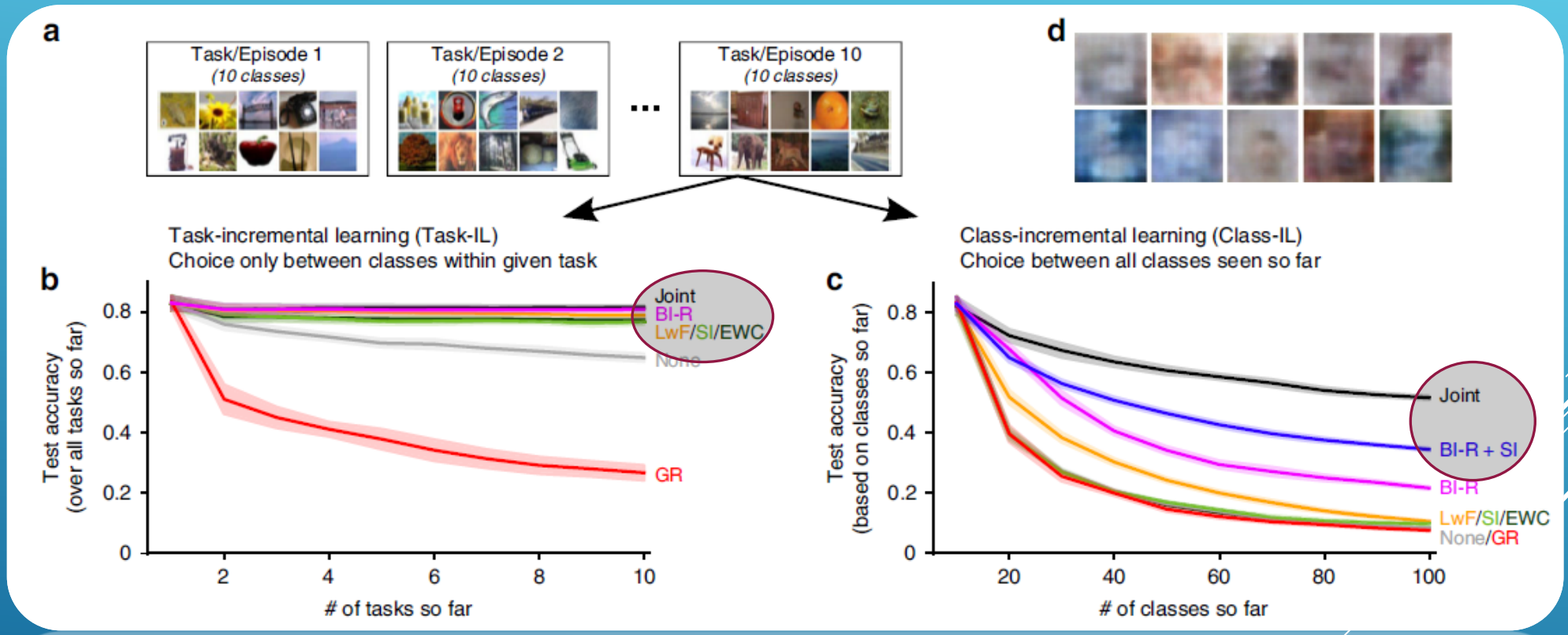
- Internal representations are replayed instead of pixel-level representations.
- Assumes the first few layers of the network do not change.



van de Ven, 2020 https://www.nature.com/articles/s41467-020-17866-2

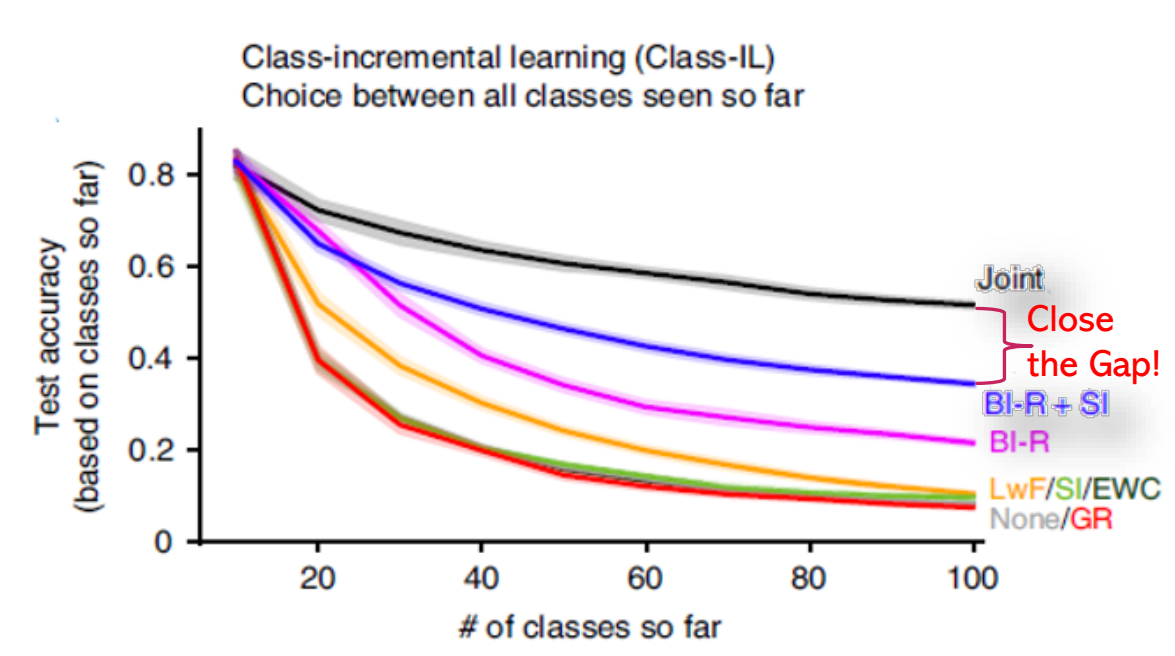
CURRENT STATE — BI-R

Brain-Inspired Replay + Synaptic Intelligence



van de Ven, G.M., Siegelmann, H.T. and Tolias, A.S., 2020. Brain-inspired replay for continual learning with artificial neural networks. Nature communications.

RESEARCH DIRECTION



Ideas

Improved gating

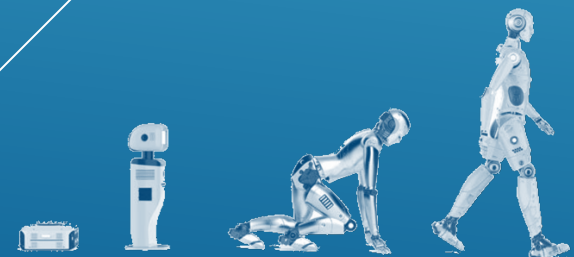
- Introduce more intelligent control over random selection, such as through node importance metrics

Network architecture search

- Incorporate ideas from NAS such as co-trained controller network to find task specific subgraphs

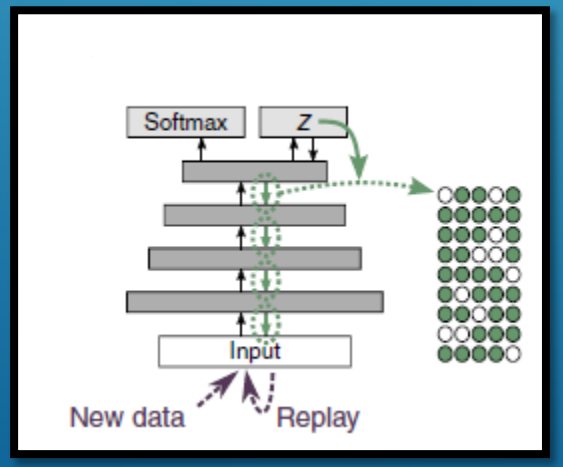
Explore more drastic IL problems

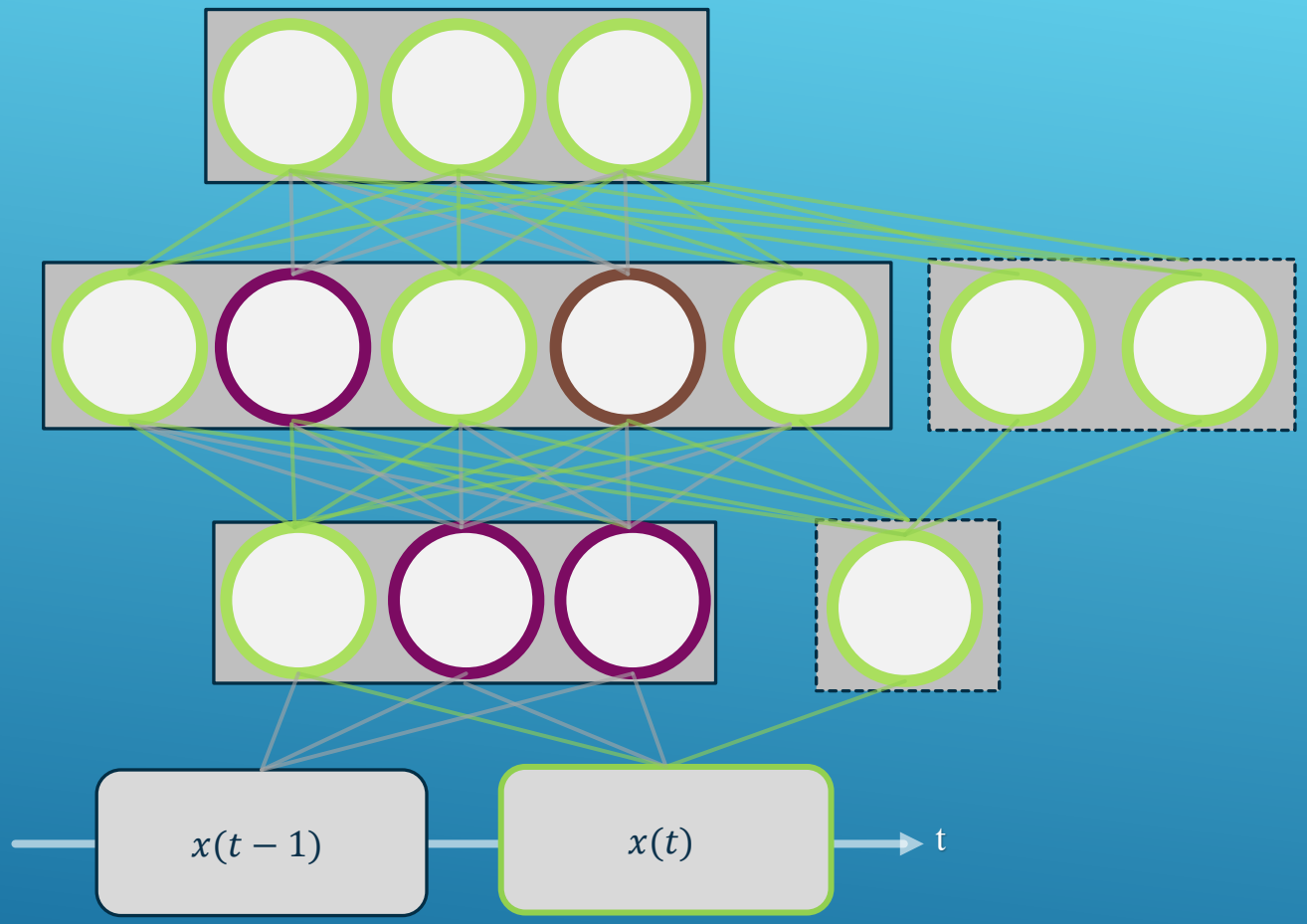
- e.g. MNIST-> CIFAR



- In the continual learning framework, the goal is to find a model that preserves performance on previously learned tasks while learning to perform well on new tasks.
- Intuitively we can make use of either overparameterized networks and keep a fixed network size or we can grow the network to accommodate more knowledge.
- Current state-of-the-art CL models (see van de Ven) use an instance aware controller to gate certain neurons, but this gating is determined randomly.
- Propose 1) develop strategy to select optimal subgraphs instead of random subgraphs

IMPROVED GATING

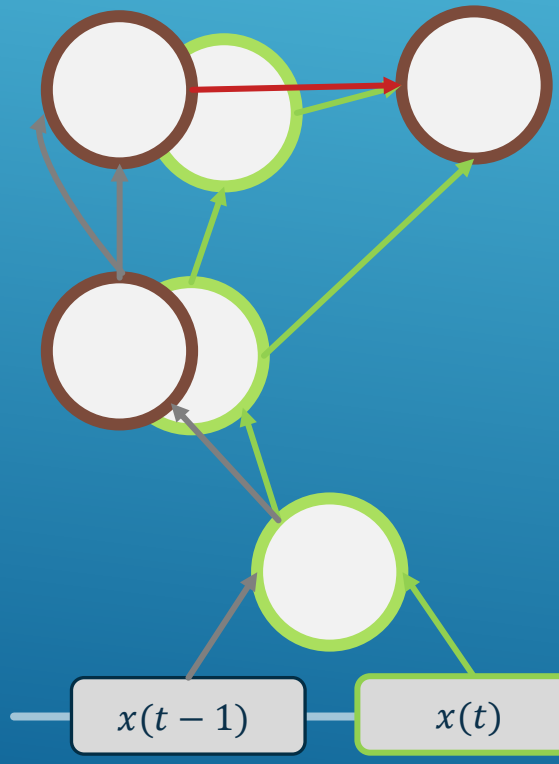




IMPROVED GATING

Micro-gating:
Select most important paths for previous task to protect from current task, adding new nodes to hidden layers when necessary.

Macro-gating:
Select most important layers for previous task for pruning, adding new layers when necessary



REFERENCES

- Rebuffi, S.A., Kolesnikov, A., Sperl, G. and Lampert, C.H., 2017. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 2001-2010).
- Lomonaco, V. 2019. *Continual Learning for Production Systems*, Medium, viewed 2 January 2021, <https://medium.com/continual-ai/continual-learning-for-production-systems-304cc9f60603>.
- De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G. and Tuytelaars, T., 2019. A continual learning survey: Defying forgetting in classification tasks. arXiv preprint arXiv:1909.08383.
- Zenke, F., Poole, B. and Ganguli, S., 2017, July. Continual learning through synaptic intelligence. In *International Conference on Machine Learning* (pp. 3987-3995). PMLR.
- Masse, N.Y., Grant, G.D. and Freedman, D.J., 2018. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences*, *115*(44), pp.E10467-E10475.
- van de Ven, G.M. and Tolias, A.S., 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.
- van de Ven, G.M., Siegelmann, H.T. and Tolias, A.S., 2020. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, *11*(1), pp.1-14.