Neural network applications

Plan for today:

- Introduce 3 deep learning research projects
- Investigate issues of bias, ethics, and fairness in AI.

Algorithms for Robust Lifelong Learning

Shixian Wen, Amanda Rios, Yunhao Ge, Laurent Itti

This work was supported in part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA

Lifelong Learning example: human brain

Day 1

Day 2

…

Day N

- Human brain learns many tasks from one or more domains over its lifetime.
- Each day the human brain is exposed to new tasks, it should learn about them
- The difficult part:
  - Should not forget tasks it has previously learned about
  - Selectively transfer the learned knowledge when learning a new task
  - Build a system that accumulatively learn about new things that we encounter everyday

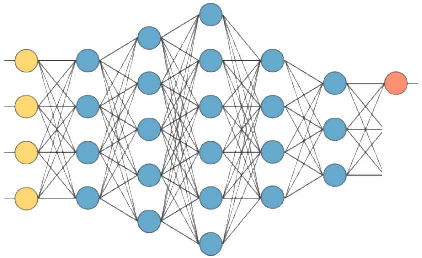
Incremental learning

Feed in:

Task 1
Cats + Dogs

Cats

Dogs



T1 classes

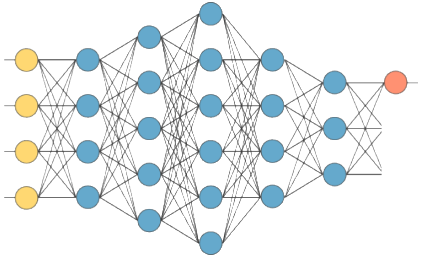
Incremental learning

Feed in:

Task 1
Cats + Dogs

Cats

Dogs



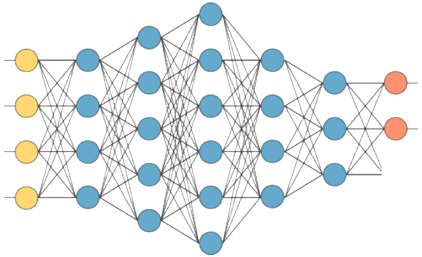
T1 classes

Task 2
Cars + Birds

Replay

Cats Dogs

Birds Cars



T1+T2 classes

Incremental learning

Very memory & compute expensive

Feed in:

Task 1
Cats + Dogs

Cats

Dogs

Need all data from all tasks

Task 2
Cars + Birds

Replay

Cats Dogs

Birds Cars

Cannot learn tasks sequentially and benefit from the training of old tasks

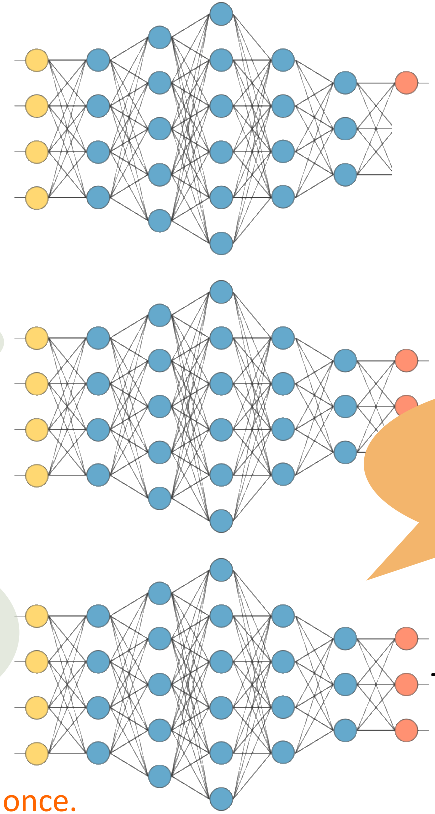
Task 3
Books + Trees

Replay

Cats Dogs

Cars Birds

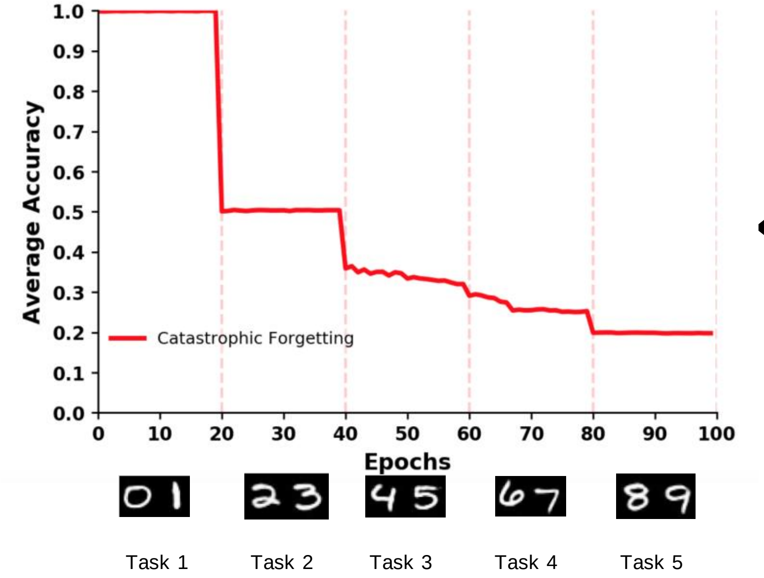
Books Trees



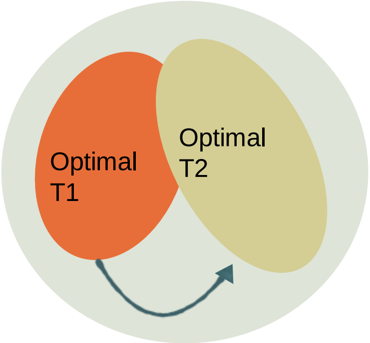
T1 + T2 + T3 classes

e.g., ImageNet: train on all 1,000 categories at once.

Catastrophic forgetting (Mccloskey & cohen, 1989)



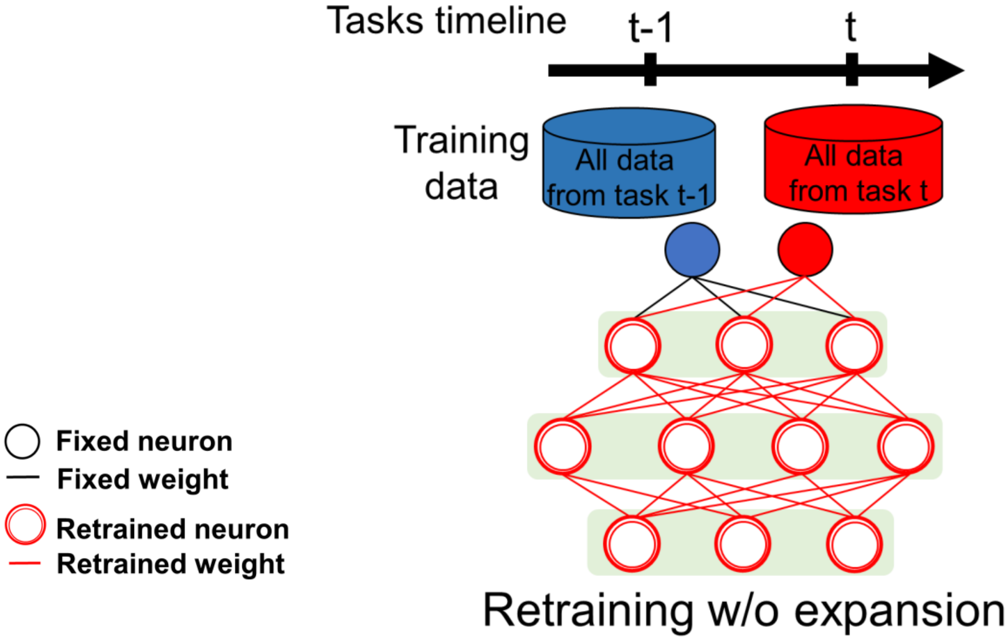
Optimal Weight Space:



Loss of supervision from T1 no longer guarantees W will be in the optimal space of T1

Four classes of approaches: type 1

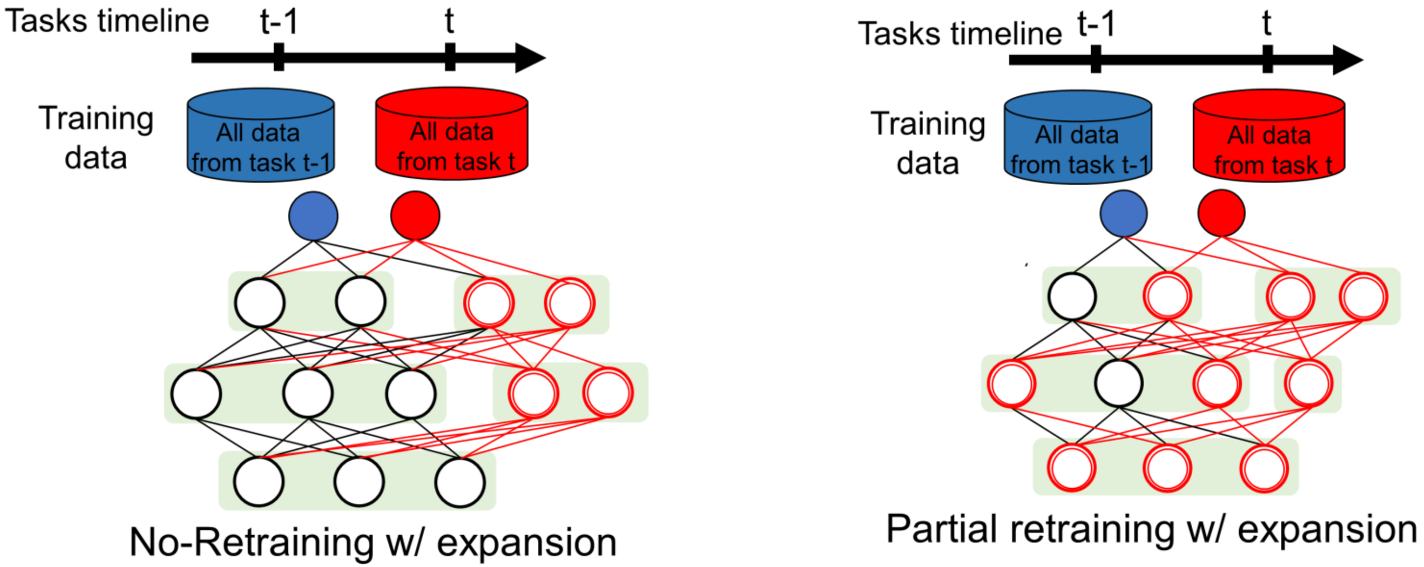
Type 1: constrain the network weights for previous tasks

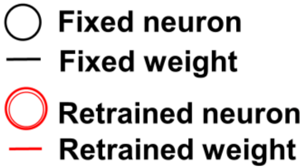


e.g., elastic weight consolidation (EWC), Kirkpatrick et al., 2017

Four classes of approaches: type 2

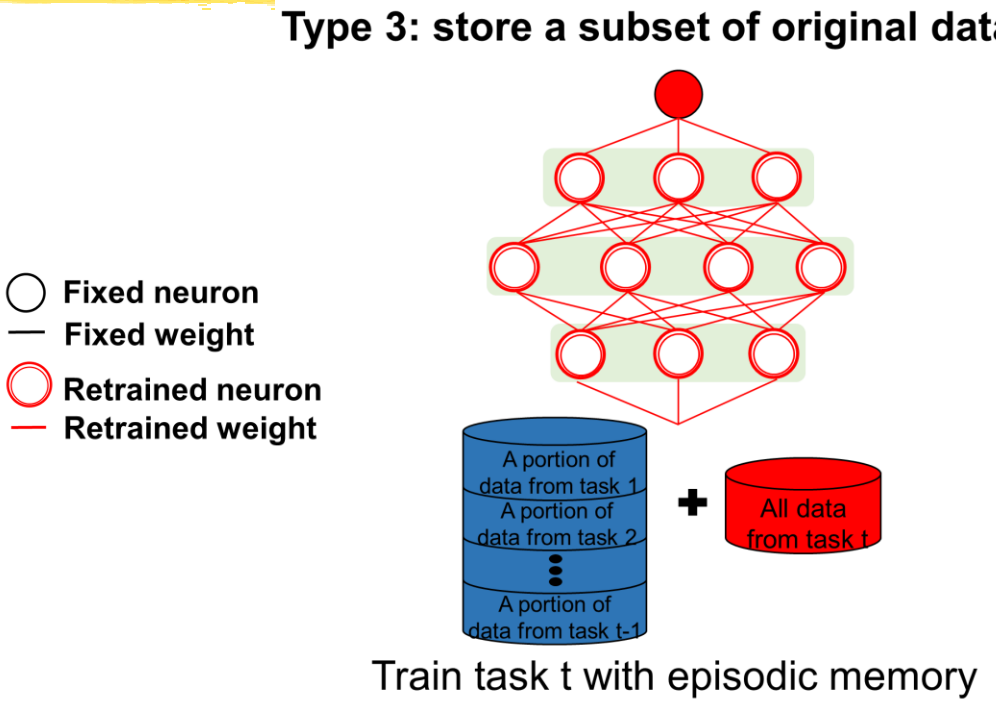
Type 2: expand the network capacity for new task and constrain the network weights for previous tasks





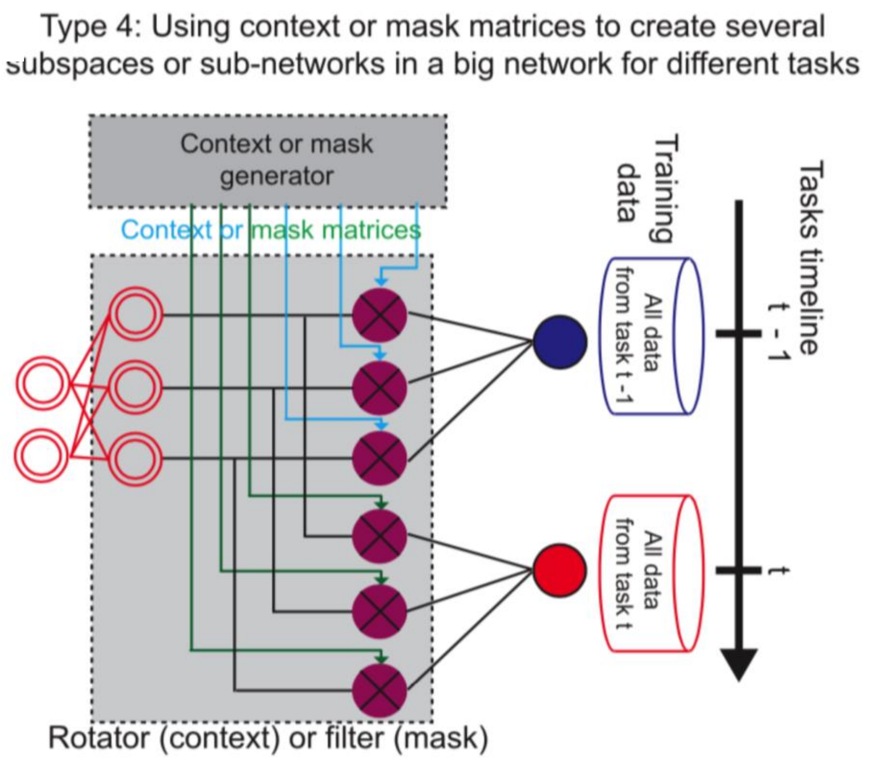
e.g., progressive neural nets, dynamically expandable nets, Rusu et al., 2017; Yoon et al., 2017

Four classes of approaches: type 3



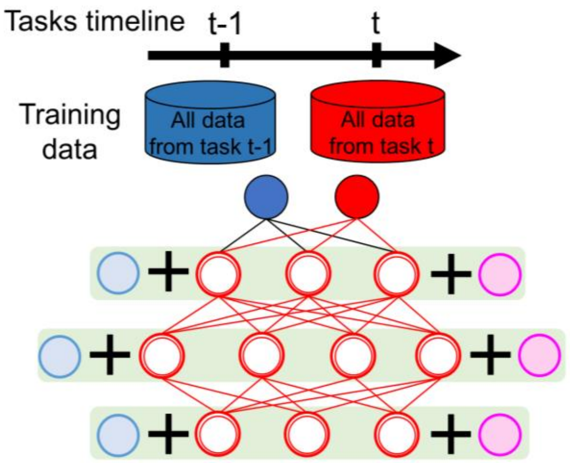
e.g., Incremental classifier and representation learning (ICARL), Gradient episodic memory (GEM), Rebuffi et al., 2017, Lopez-Paz et al., 2017

Four classes of approaches: type 4

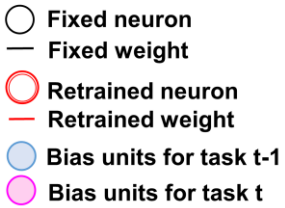


e.g., parameter super position (PSP), Cheung et al., 2019

Four classes of approaches: type 5

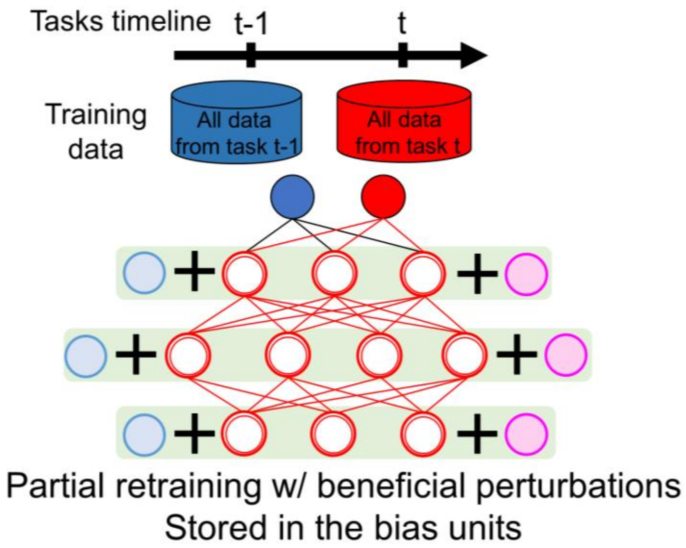


Partial retraining w/ beneficial perturbations
Stored in the bias units



New proposed approach,
Wen et al., 2021

Four classes of approaches: type 5

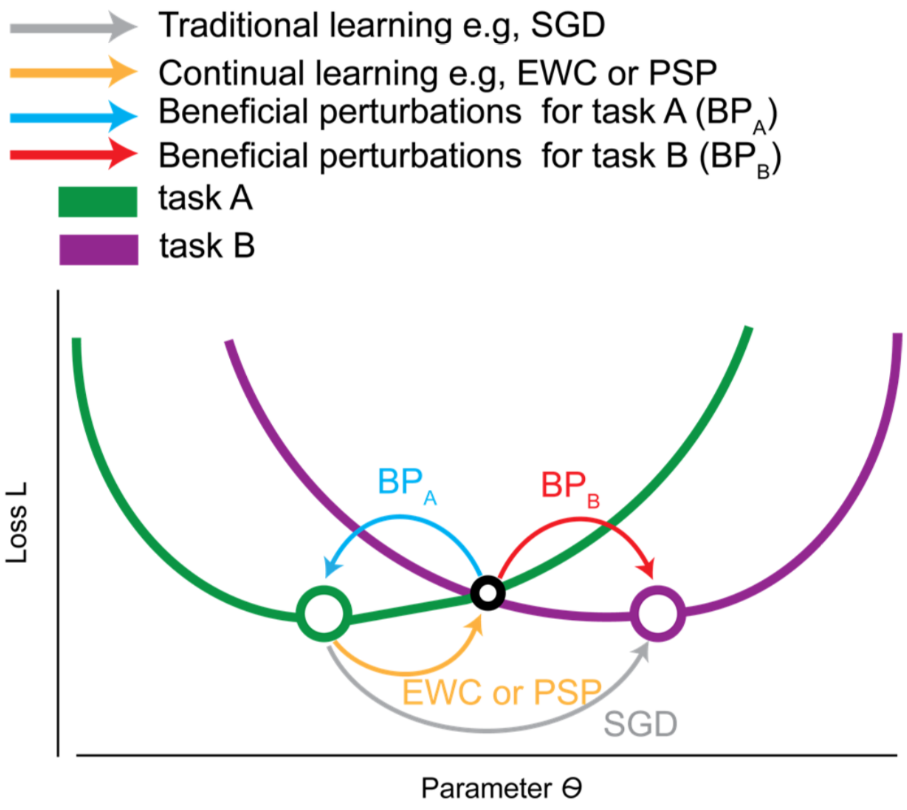


	Type 1	Type 2	Type 3	Type 4	Type 5 (ours)
Parameter efficient	✓	✗	✓	✗ / ✓	✓
Memory efficient	✓	✓	✗	✓	✓
Does not requires pervious task data	✓	✓	✗	✓	✓
Multiple input to output mappings	✗	✗	✗	✓	✓
unlimited capacity to accomondate new tasks	✓	✓	✓	✗	✓

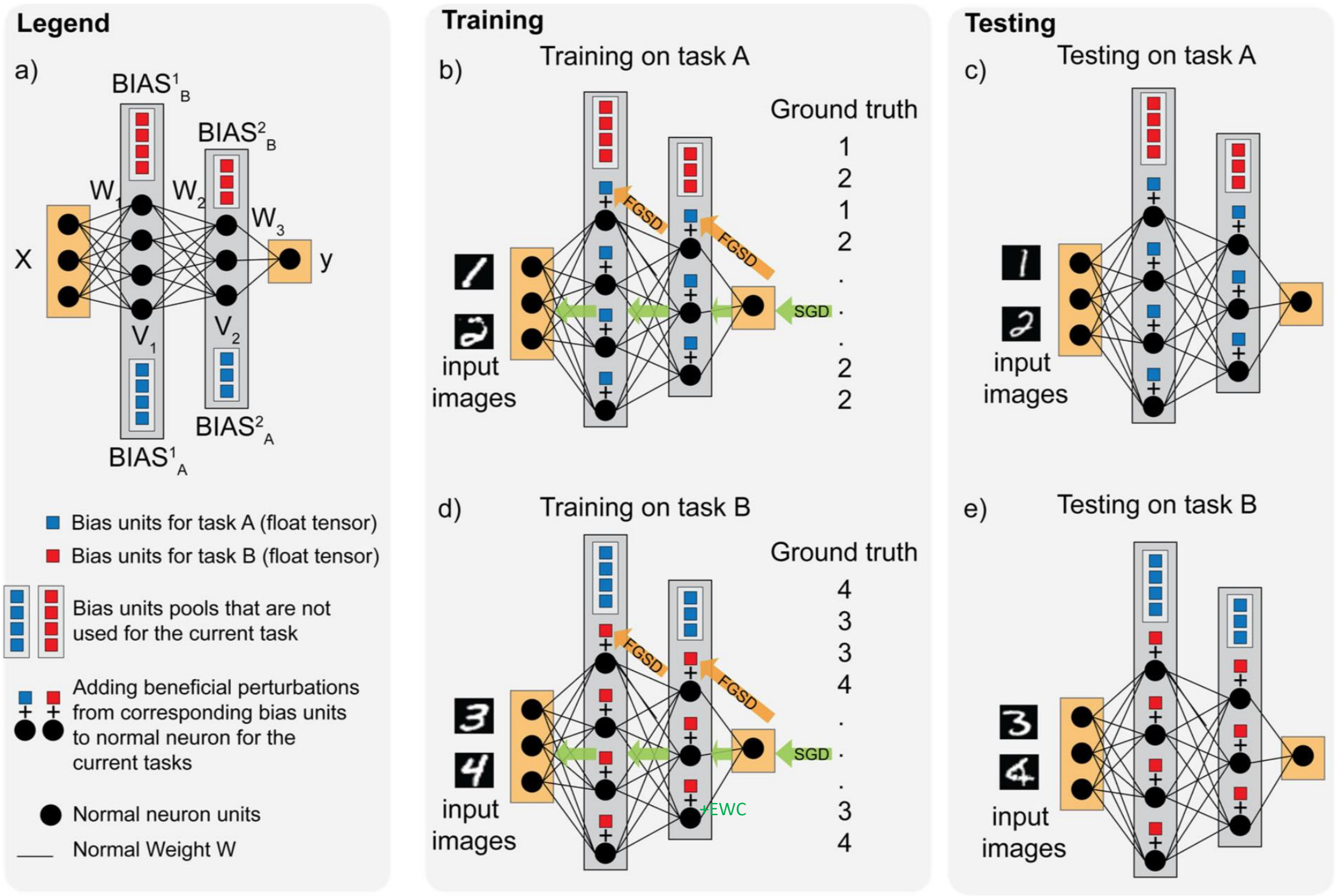
○ **Fixed neuron**
— **Fixed weight**
○ **Retrained neuron**
— **Retrained weight**
○ **Bias units for task t-1**
○ **Bias units for task t**

New proposed approach, Wen, Rios, Ge & Itti, 2021

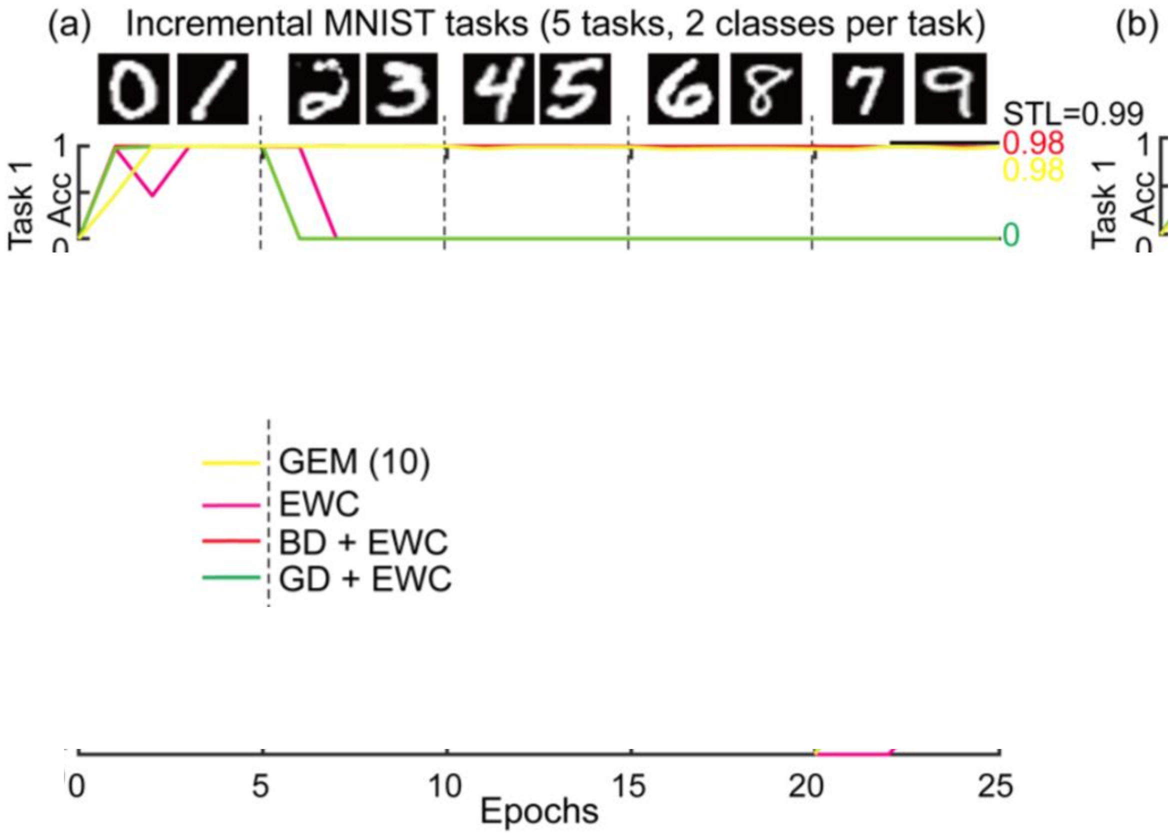
Overview of the approach



Overview of the approach

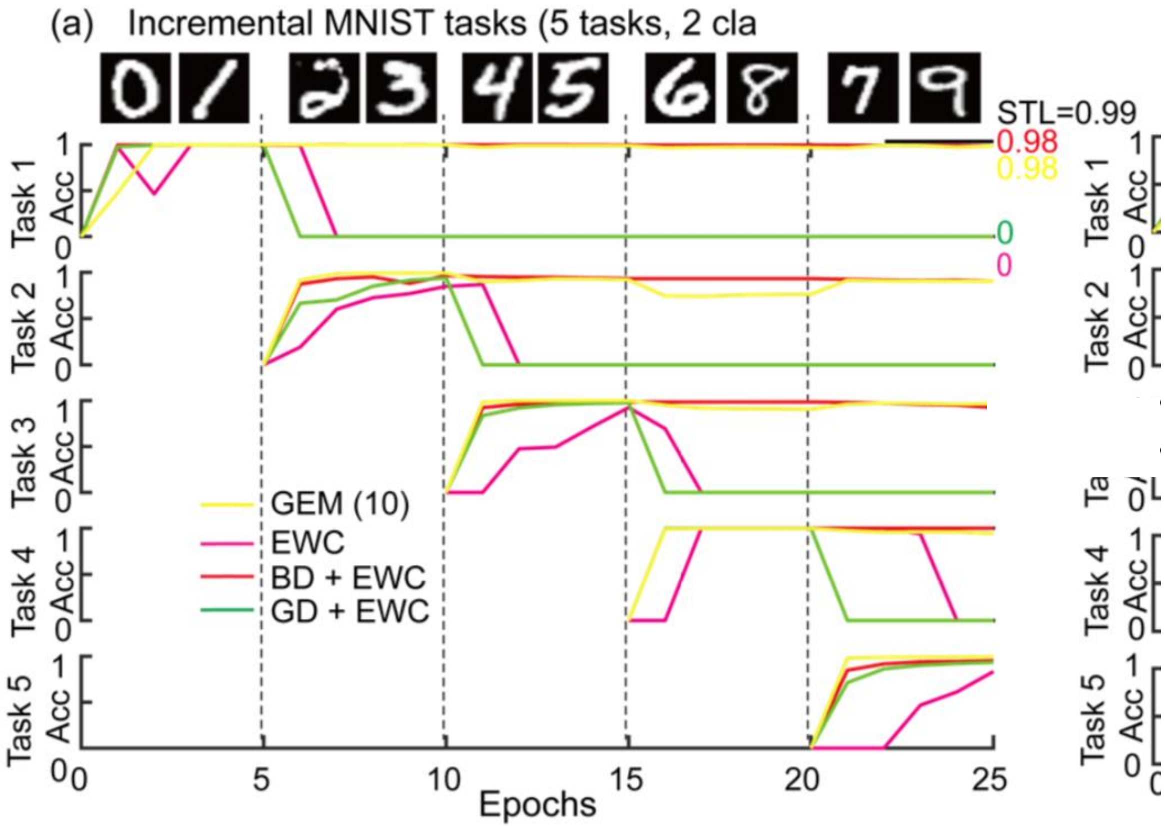


Results



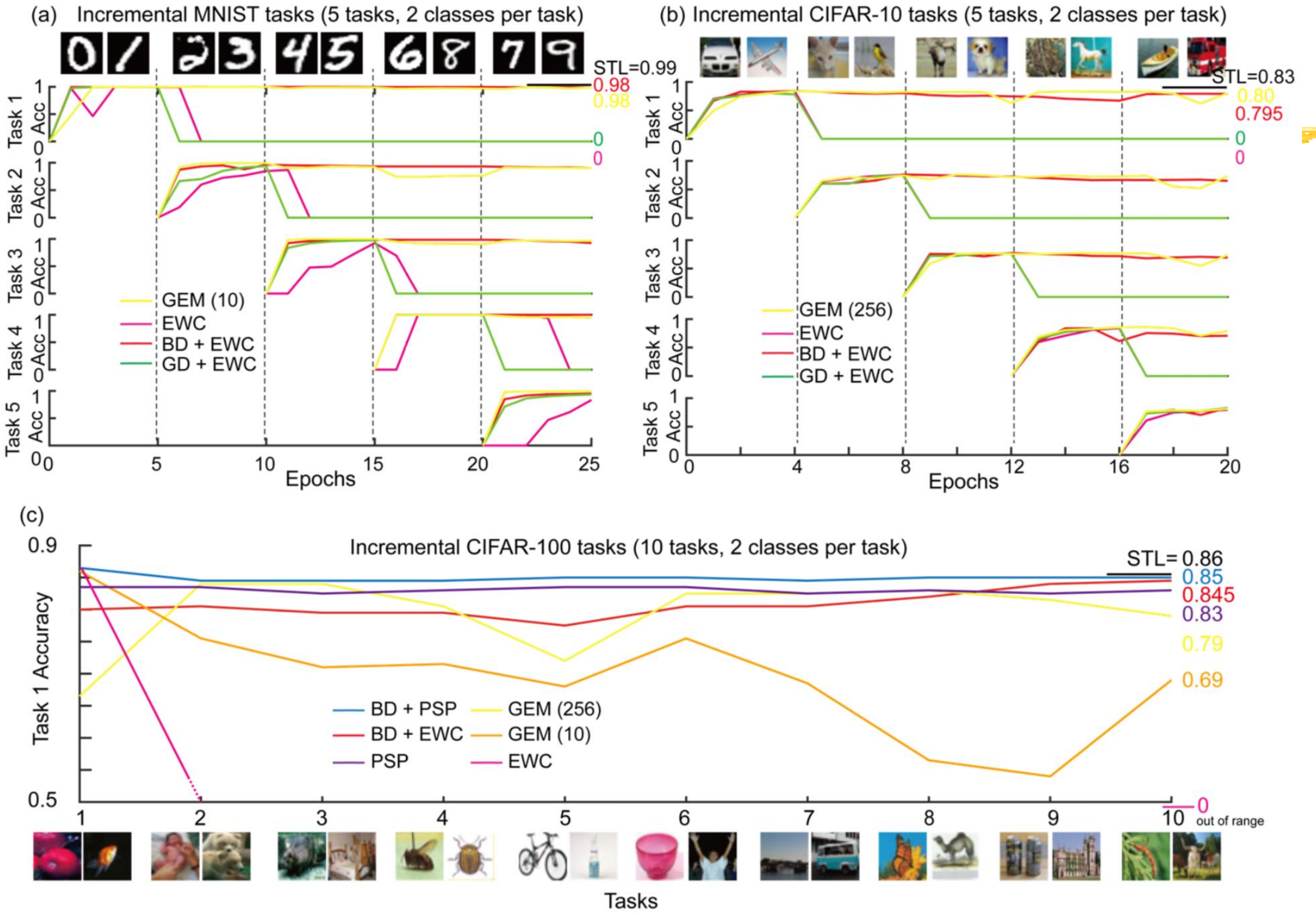
(a) Incremental MNIST tasks (5 tasks, 2 classes per task)
(b)
STL=0.99
0.98
0.98
0
Task 1 Acc
1
0
GEM (10)
EWC
BD + EWC
GD + EWC
0
5
10
15
20
25
Epochs

Results



(a) Incremental MNIST tasks (5 tasks, 2 cla
STL=0.99
0.98
0.98
0
0
Task 1 Acc
Task 2 Acc
Task 3 Acc
Task 4 Acc
Task 5 Acc
GEM (10)
EWC
BD + EWC
GD + EWC
0
5
10
15
20
25
Epochs

Results

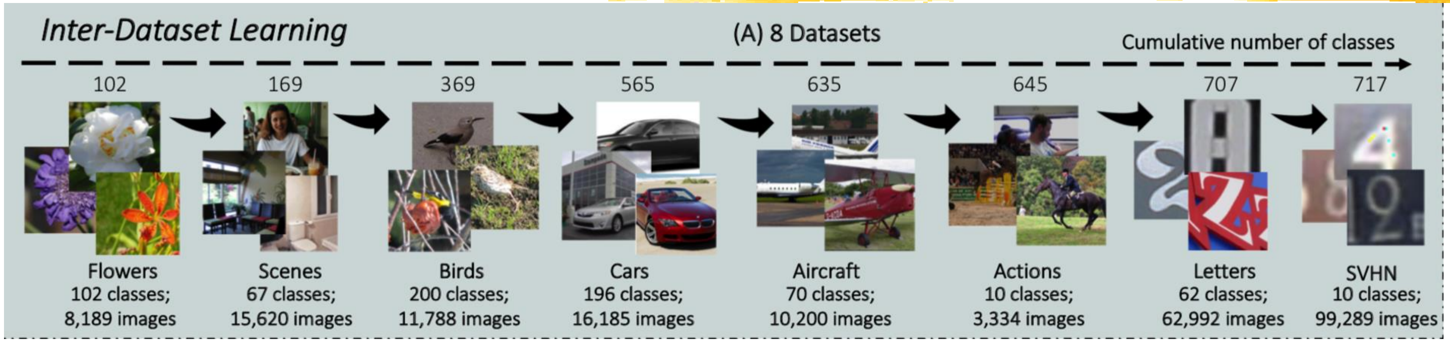


(a) Incremental MNIST tasks (5 tasks, 2 classes per task)
STL=0.99
0.98
0.98
0
0
Task 1 Acc
Task 2 Acc
Task 3 Acc
Task 4 Acc
Task 5 Acc
GEM (10)
EWC
BD + EWC
GD + EWC
Epochs
(b) Incremental CIFAR-10 tasks (5 tasks, 2 classes per task)
STL=0.83
0.80
0.795
0
0
GEM (256)
EWC
BD + EWC
GD + EWC
Epochs
(c)
Incremental CIFAR-100 tasks (10 tasks, 2 classes per task)
STL= 0.86
0.85
0.845
0.83
0.79
0.69
0
out of range
Task 1 Accuracy
BD + PSP
BD + EWC
PSP
GEM (256)
GEM (10)
EWC
Tasks

Results

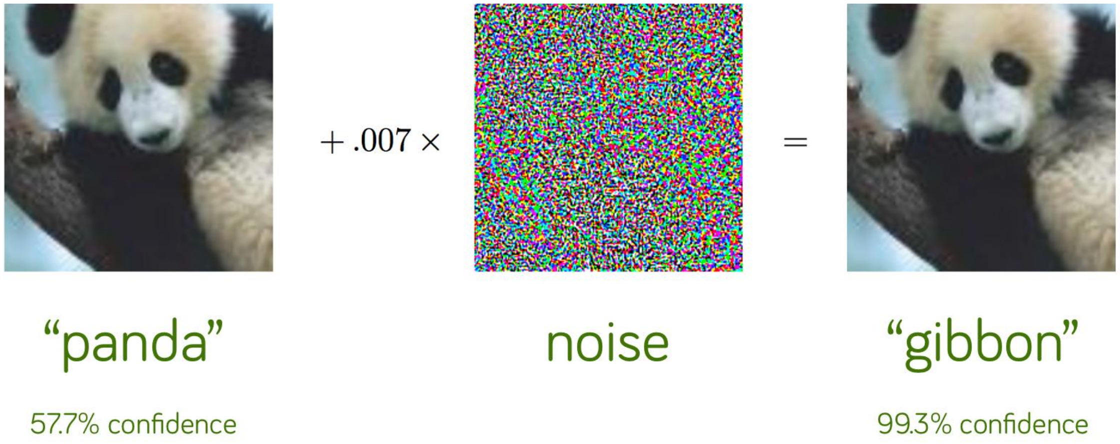
Dataset	Method	Task 1 performance after training all sequential tasks	Additional memory storage costs per task (Bytes)
Incremental MNIST (5 tasks, 2 classes per task)	GEM(10)	0.980	47,040
	BD+EWC	**0.980**	**4,808**
Incremental CIFAR-10 (5 tasks, 2 classes per task)	GEM(256)	**0.800**	4,718,592
	GEM(150)	0.698	2,764,800
	BD+EWC	0.795	**4,808**
Incremental CIFAR-100 (10 tasks, 2 classes per task)	GEM(256)	0.790	4,718,592
	GEM(209)	0.775	3,852,288
	BD+PSP	**0.850**	20,776
	PSP	0.830	15,968
	BD+EWC	0.845	**4,808**

Results



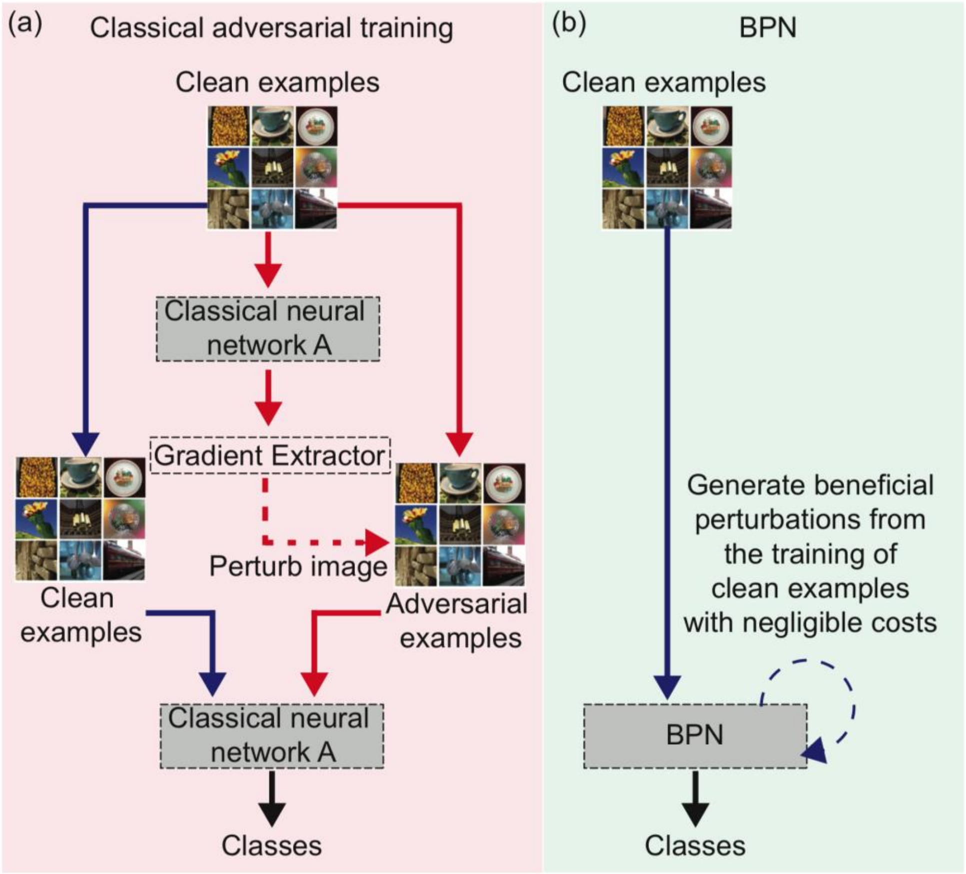
Method \ Dataset	Flower	Scenes	Birds	Cars	Aircraft	Actions	Letters	SVHN	Average
IMM [25]	67.44*	47.08*	42.28*	39.19*	18.93*	32.88*	46.35*	53.15*	43.41*
LwF [27]	65.68*	46.04*	42.47*	37.24*	30.75*	34.55*	50.31*	88.89*	49.49*
EWC [20]	–	–	–	–	–	–	–	–	50.00*
EBLL [45]	–	–	–	–	–	–	–	–	50.29*
SI [50]	74.58*	52.53*	48.65*	43.96*	32.19*	38.22*	47.7*	66.09*	50.49*
MAS [1]	76.34*	55.30*	49.03*	44.59*	33.48*	41.76*	51.93*	69.13*	52.69*
BD + EWC (ours)	65.67	47.99	42.70	27.96	40.17	47.61	66.97	83.24	52.79
SLNID [2]	–	–	–	–	–	–	–	–	54.50*
PSP [5]	75.36	52.39	40.85	28.08	42.45	48.64	66.52	83.74	54.75
BD + PSP (ours)	78.47	56.04	47.96	34.40	47.58	51.61	71.19	83.73	**58.87**

Application: Adversarial robustness

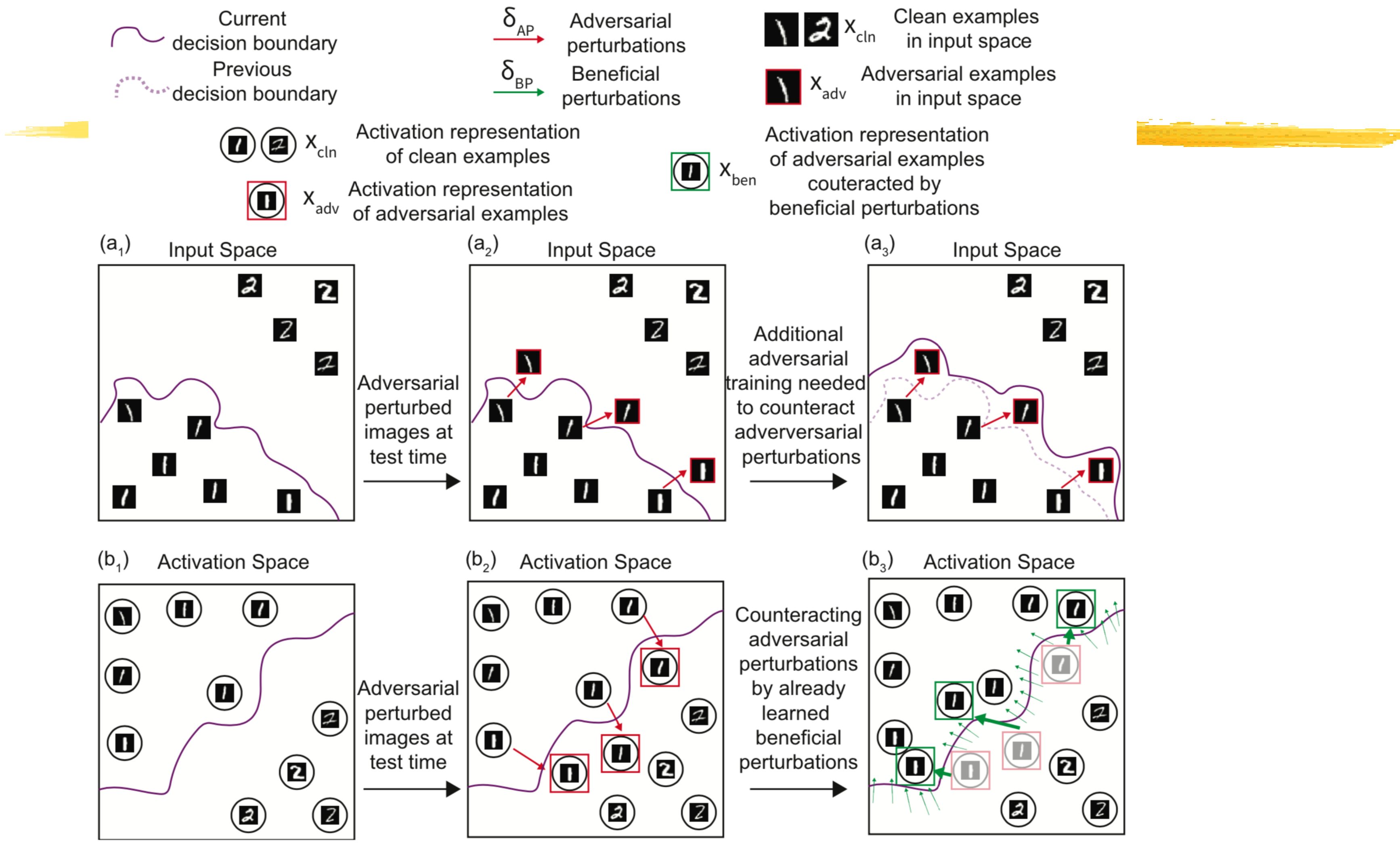


from https://towardsdatascience.com/breaking-neural-networks-with-adversarial-attacks-f4290a9a45aa

Overview of the approach



Wen & Itti, submitted



Current decision boundary
Previous decision boundary
δ AP
Adversarial perturbations
δ BP
Beneficial perturbations
X cln
Clean examples in input space
X adv
Adversarial examples in input space
X cln
Activation representation of clean examples
X adv
Activation representation of adversarial examples
X ben
Activation representation of adversarial examples couteracted by beneficial perturbations
(a1) Input Space
Adversarial perturbed images at test time
(a2) Input Space
Additional adversarial training needed to counteract adverversarial perturbations
(a3) Input Space
(b1) Activation Space
Adversarial perturbed images at test time
(b2) Activation Space
Counteracting adversarial perturbations by already learned beneficial perturbations
(b3) Activation Space

Results

Computation cost / network	Forward (FLOPS)	Backward (FLOPS)
Classical Network (ResNet-50)	51,112,224	51,112,225
BPN (ResNet-50)	51,112,224	51,115,321

Results

Table 2. Training on clean examples for BPN and classical network (CN). Testing on clean examples (Cln Ex) and adversarial examples (Adv Ex) (generated by FGSM, $\epsilon = 0.3$ for MNIST, FashionMNIST and TinyImageNet). CN does poorly on adversarial examples. While, BPN can successfully defend adversarial examples.

Datasets & network structure		MNIST LeNet	FasMNIST LeNet	TinyImageNet ResNet-18
Cln Ex	BPN	**99.17**	**89.53**	57.55
	CN	99.01	89.17	**64.30**
Adv Ex	BPN	**98.88**	**54.07**	**53.29**
	CN	18.08	11.87	1.45

Results

Table 5. Training on clean examples of MNIST and TinyImageNet for BPN and classical network (CN). Testing on adversarial examples generated by a variety of adversarial attack methods. BPN can successfully defend those adversarial examples.

Dataset & network (rows) / Attacks (columns)		PGD Linf	FGSM	Basic Iterative Attack L2	Aka Basic Iterative Attack
MNIST	BPN	**95.41**	**98.35**	**98.52**	**11.35**
	CN	2.18	17.53	97.24	9.74
TinyImageNet	BPN	**44.37**	**52.39**	**16.23**	**5.9**
	CN	0.00	1.29	15.11	0.5

CLR: Channel-wise Lightweight Reprogramming for Continual Learning

Yunhao Ge   Yuecheng Li*   Shuo Ni*   Jiaping Zhao   Ming-Hsuan Yang   Laurent Itti

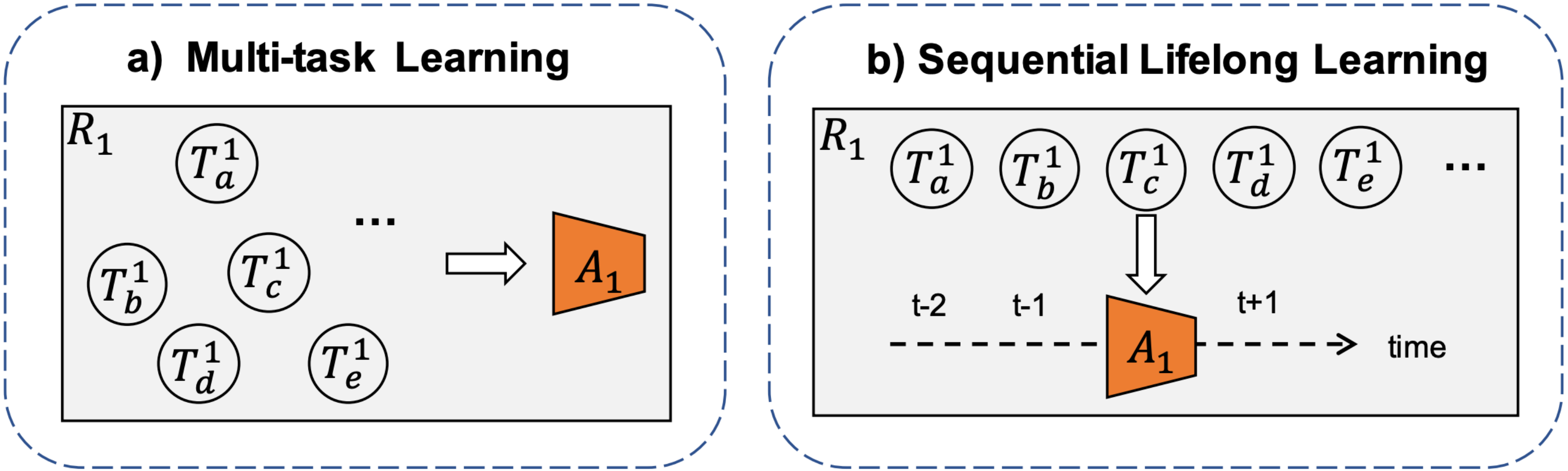
* co-second author

USC
Google Research

Paper   Code   SKILL-102 dataset

ICCV23
PARIS

Background: Continual Learning and Catastrophic Forgetting



Ge, Yunhao, Yuecheng Li, Di Wu, Ao Xu, Adam M. Jones, Amanda Sofie Rios, Iordanis Fostiropoulos et al. "**Lightweight Learner for Shared Knowledge Lifelong Learning**." TMLR (2023)

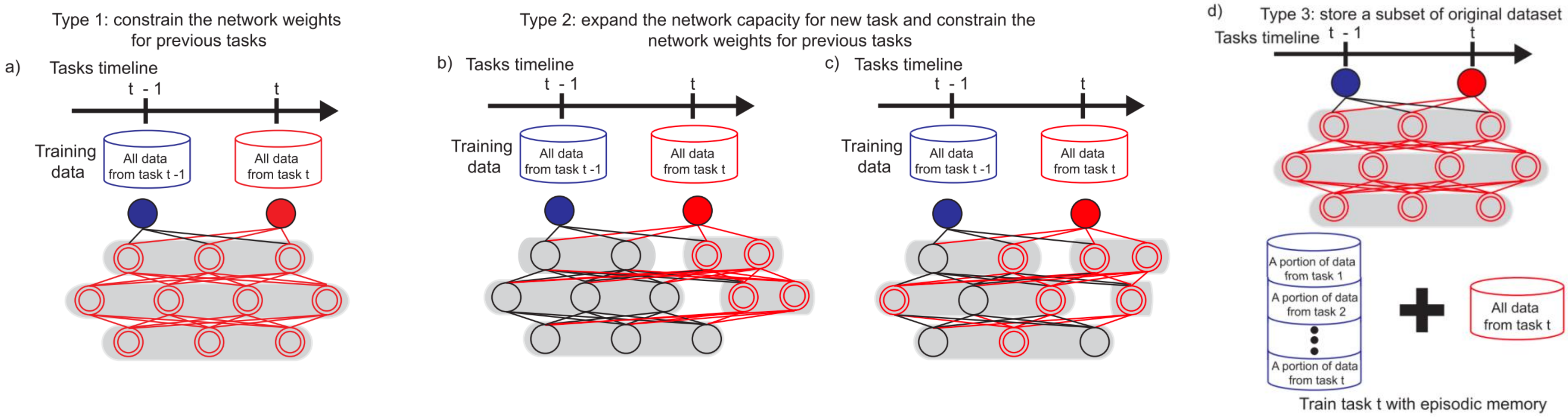
Continual (Lifelong) learning aims to emulate the human ability to continually accumulate knowledge over sequential tasks.

Background: 3 main types of Continual Learning methods

Regularization methods

Dynamic network methods

Replay methods



EWC, IMM, SI, MAS, LwF, LFL, DMC

SUPSUP, PSP, CCLL,EFTs, BPN

iCaRL, ER GSS, AGEM, AGEM-R, DER, DERPP

Wen, Shixian, Amanda Rios, Yunhao Ge, and Laurent Itti. "**Beneficial perturbation network for designing general adaptive artificial intelligence systems**." IEEE Transactions on Neural Networks and Learning Systems 33, no. 8 (2021): 3778-3791

Background: 3 main types of Continual Learning methods

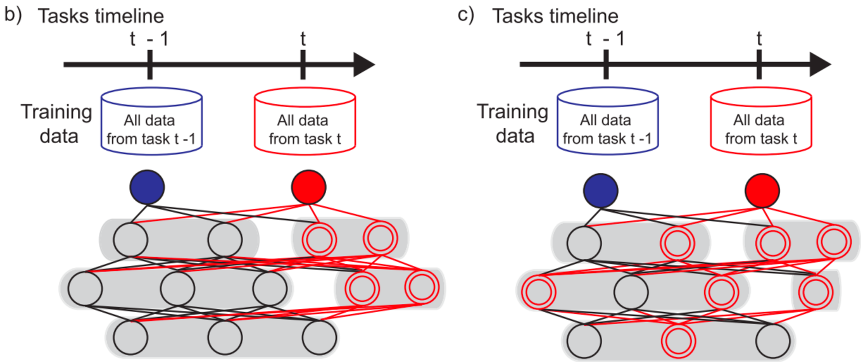
Regularization methods

Type 1: constrain the network weights for previous tasks

EWC, IMM, SI, MAS, LwF, LFL, DMC

Dynamic network methods

Type 2: expand the network capacity for new task and constrain the network weights for previous tasks



SUPSUP, PSP, CCLL,EFTs, BPN

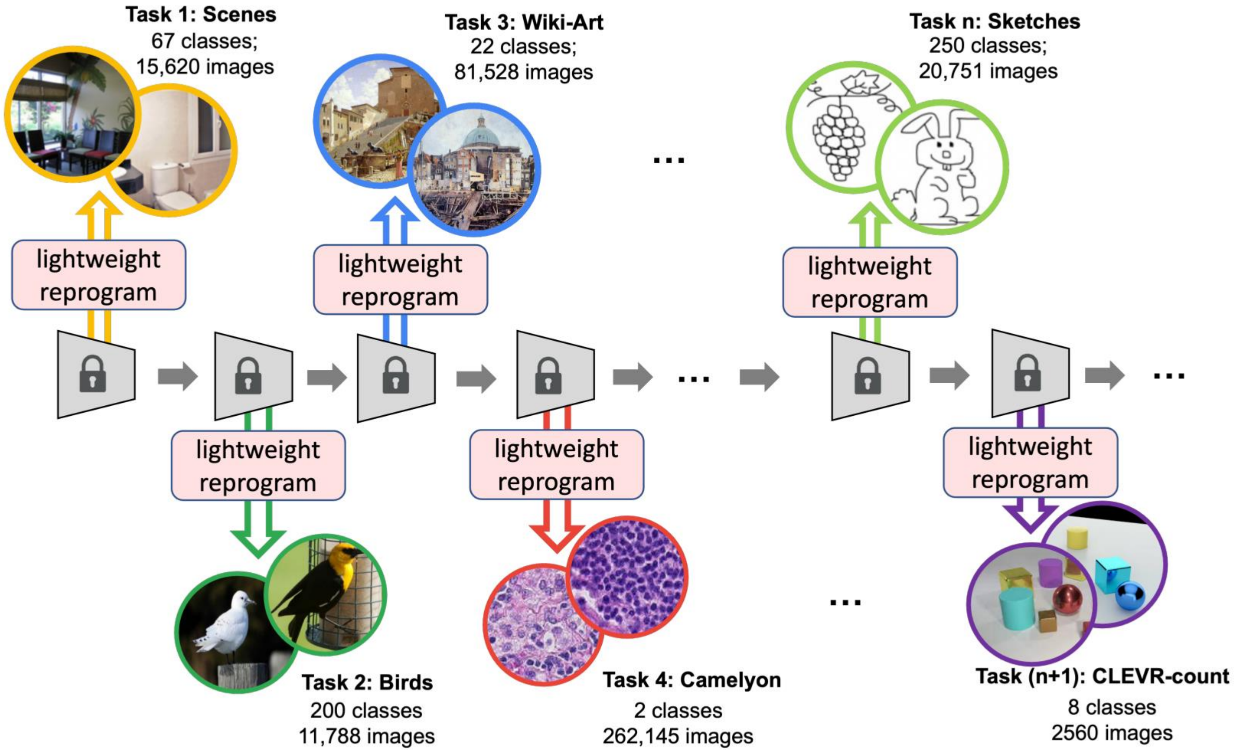
Replay methods

Type 3: store a subset of original dataset

iCaRL, ER GSS, AGEM, AGEM-R, DER, DERPP

Wen, Shixian, Amanda Rios, Yunhao Ge, and Laurent Itti. "**Beneficial perturbation network for designing general adaptive artificial intelligence systems**." IEEE Transactions on Neural Networks and Learning Systems 33, no. 8 (2021): 3778-3791

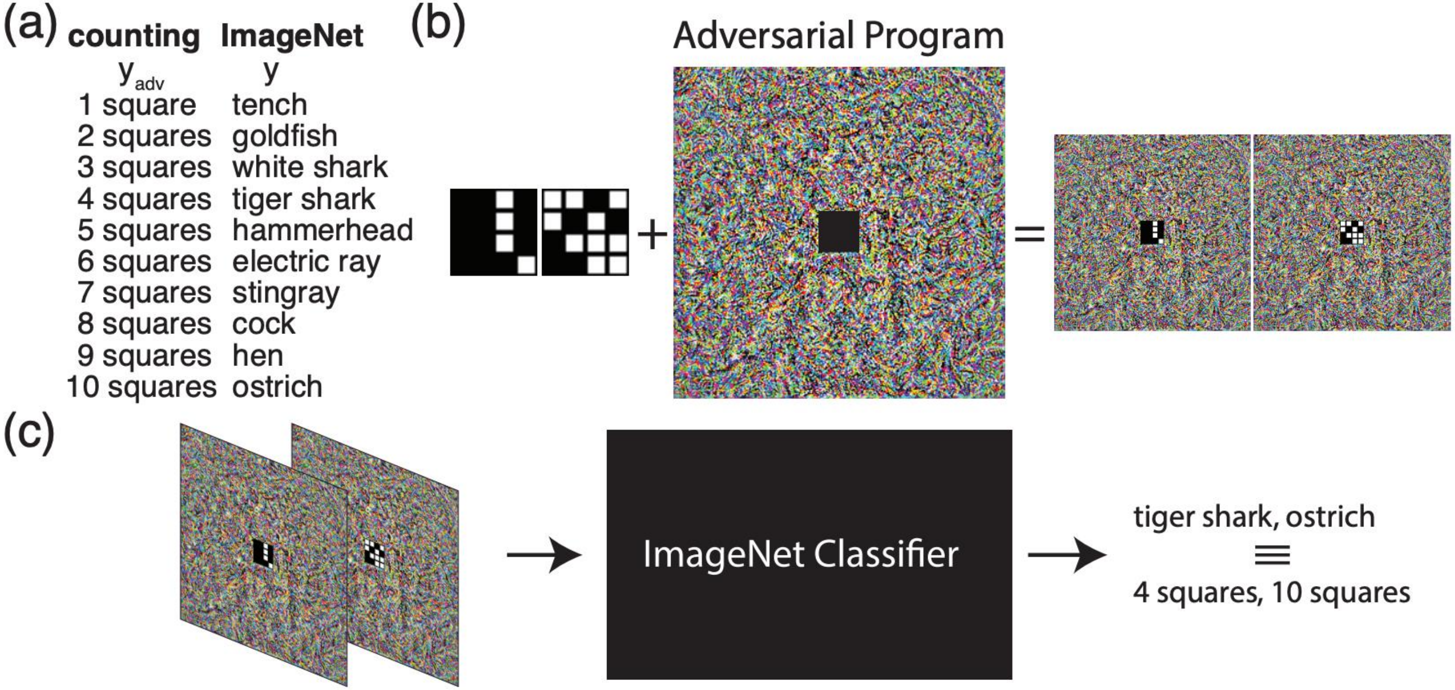
CLR: Channel-wise Lightweight Reprogramming



Ge, Yunhao, Yuecheng Li, Shuo Ni, Jiaping Zhao, Ming-Hsuan Yang, and Laurent Itti. "**CLR: Channel-wise Lightweight Reprogramming for Continual Learning**." ICCV (2023).

Motivation 1: Adversarial Reprogramming

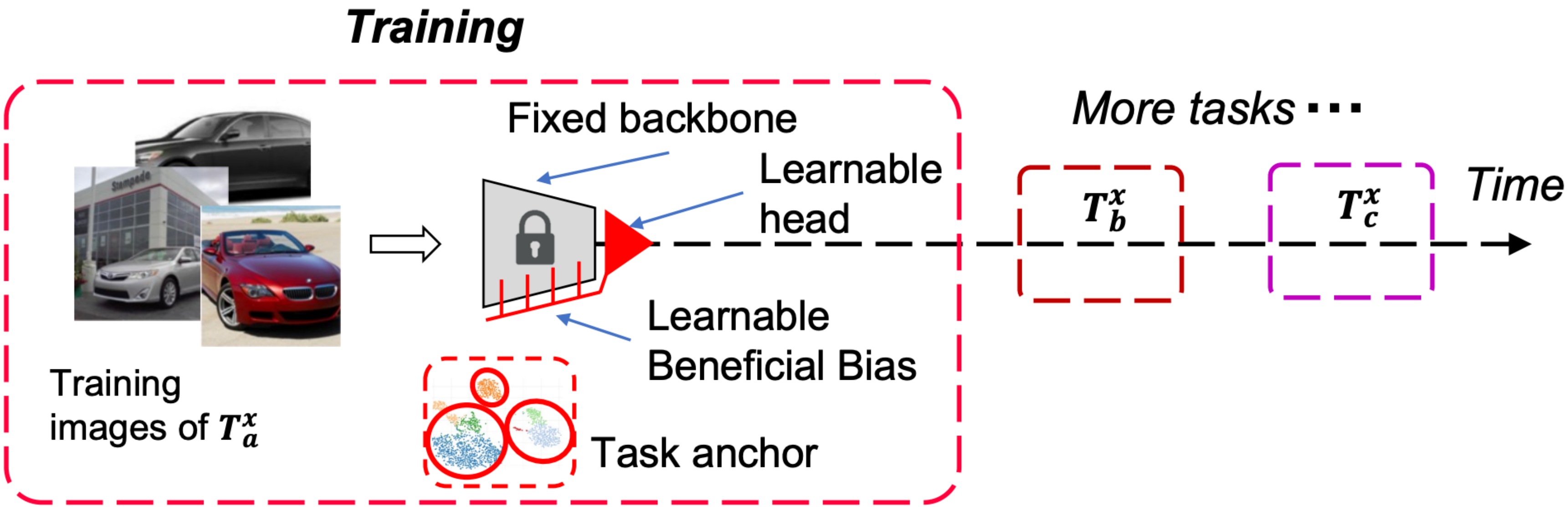
Reuse instead of re-learn



Elsayed, Gamaleldin F., Ian Goodfellow, and Jascha Sohl-Dickstein. "**Adversarial reprogramming of neural networks**." arXiv preprint arXiv:1806.11146 (2018).

Motivation 2: Light-weight Lifelong Learner

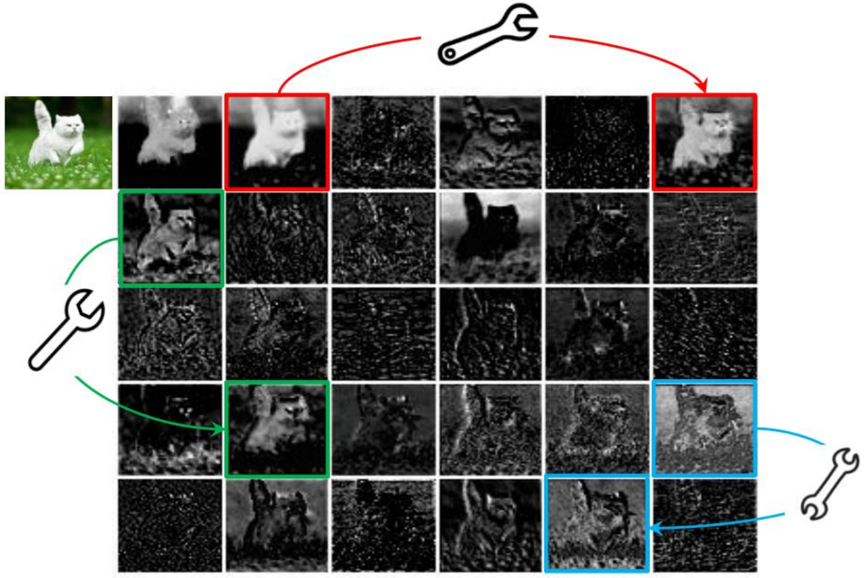
Lightweight parameters could shift model distribution

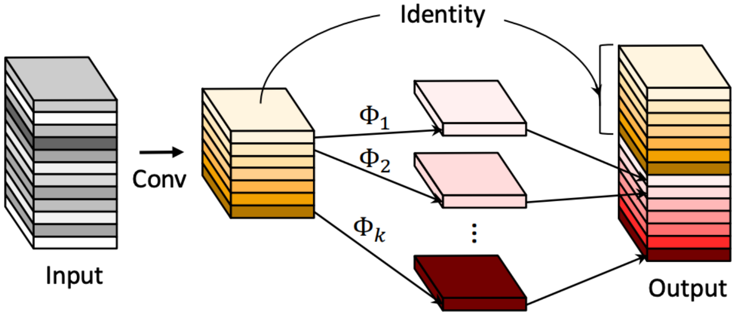


Ge, Yunhao, Yuecheng Li, Di Wu, Ao Xu, Adam M. Jones, Amanda Sofie Rios, Iordanis Fostiropoulos et al. "**Lightweight Learner for Shared Knowledge Lifelong Learning**." TMLR (2023)

Motivation 3: GhostNet

Channel-wise transformations may link two different kernels

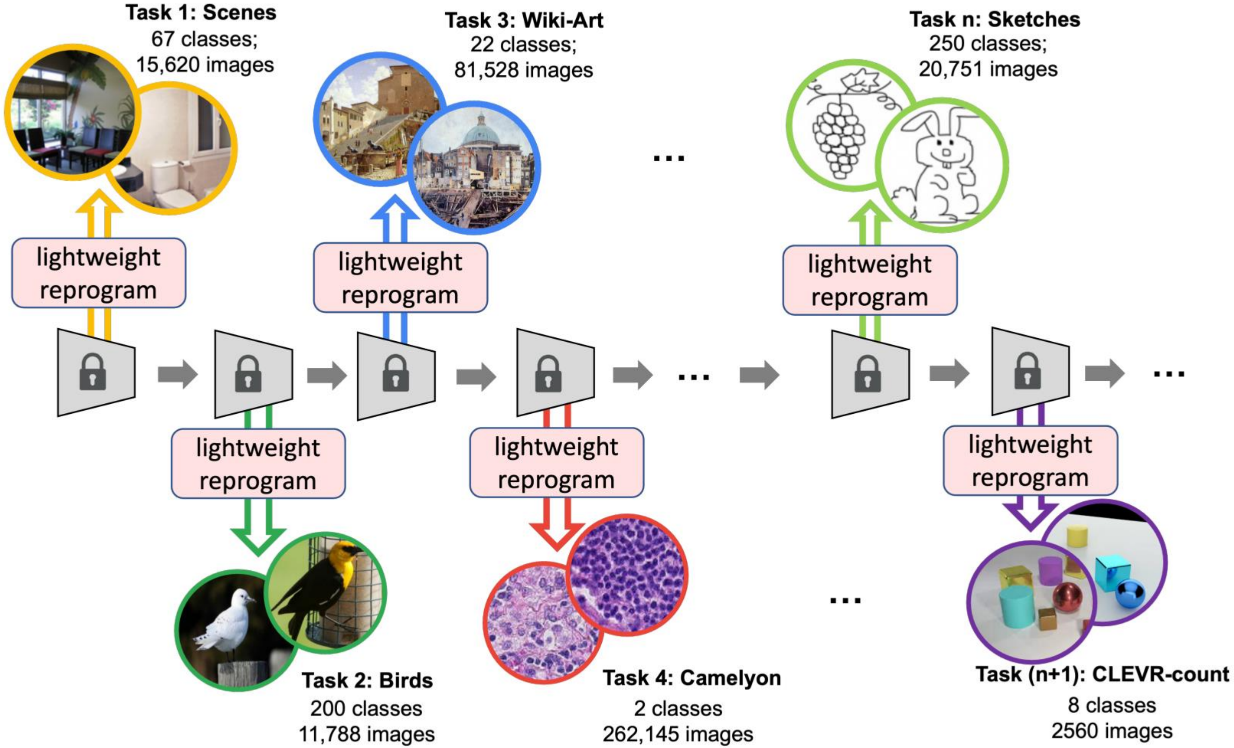




(b) The Ghost module.

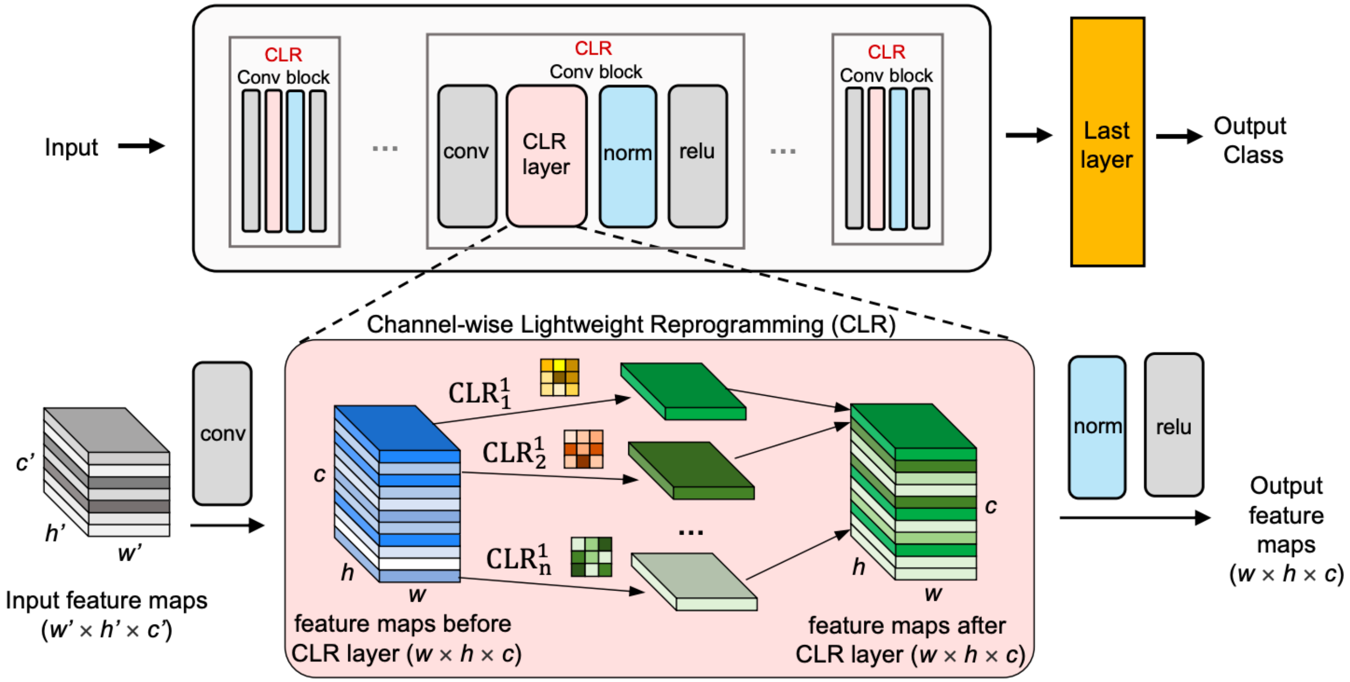
Han, Kai, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. "**Ghostnet: More features from cheap operations**." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1580-1589. 2020.

CLR: Channel-wise Lightweight Reprogramming



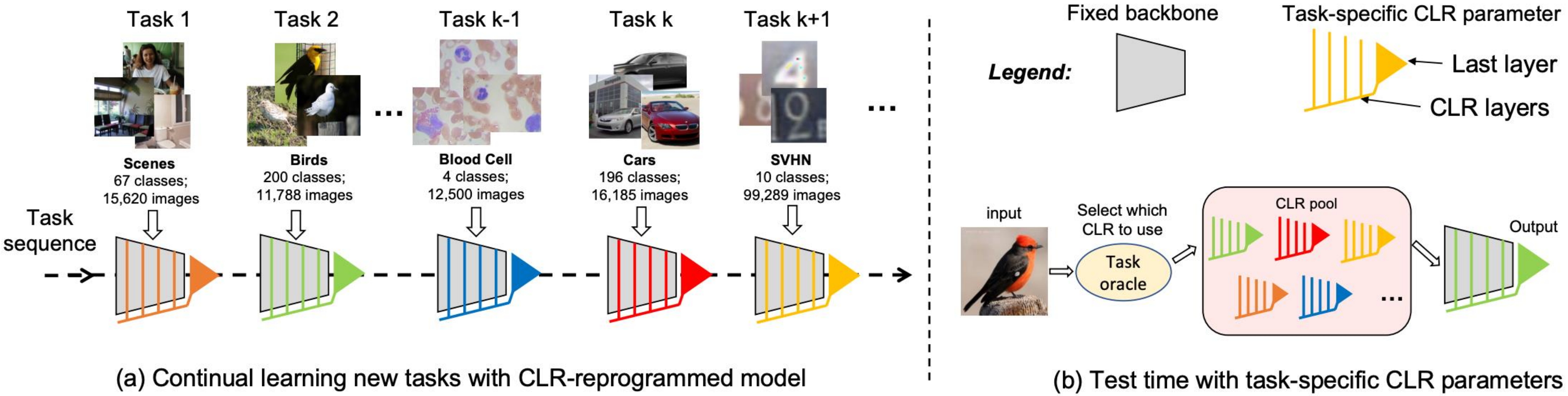
Ge, Yunhao, Yuecheng Li, Shuo Ni, Jiaping Zhao, Ming-Hsuan Yang, and Laurent Itti. "**CLR: Channel-wise Lightweight Reprogramming for Continual Learning**." ICCV (2023).

CLR: Channel-wise Lightweight Reprogramming



Ge, Yunhao, Yuecheng Li, Shuo Ni, Jiaping Zhao, Ming-Hsuan Yang, and Laurent Itti. "**CLR: Channel-wise Lightweight Reprogramming for Continual Learning**." ICCV (2023).

CLR: Channel-wise Lightweight Reprogramming



Ge, Yunhao, Yuecheng Li, Shuo Ni, Jiaping Zhao, Ming-Hsuan Yang, and Laurent Itti. "**CLR: Channel-wise Lightweight Reprogramming for Continual Learning**." ICCV (2023).

Experiments: SKILL-102 continual Learning dataset

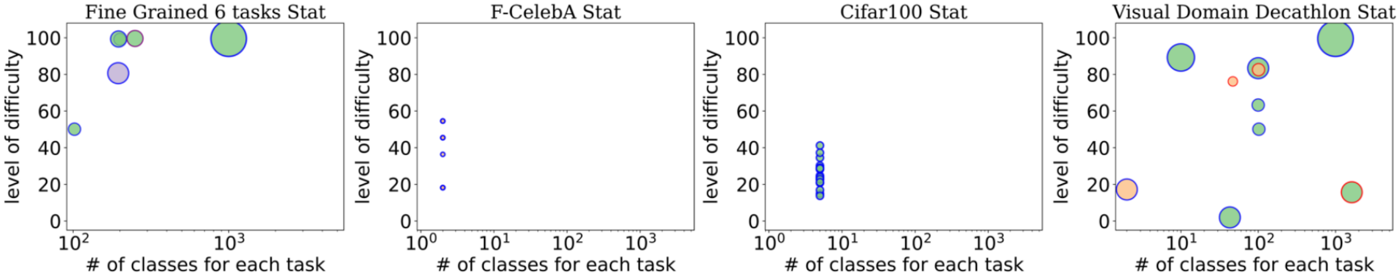
SKILL-102 dataset

a)

b)

Comparison	SKILL-102	Visual domain decathlon	Cifar100	F-CelebA	Fine grained 6 tasks
# Tasks	**102**	10	20	10	6
# Classes	**7,059**	3,128	100	20	1,943
# Images	**5,162,149**	1,659,142	60,000	1,180	1,440,086
Max(**Δ** accuracy/difficulty)	**99.60%**	97.56%	27.45%	36.36%	49.50%
# Different classification target	**5**	2	1	1	2
Mix image style (nature/artifact)	✓	✓	✓	✗	✓
Mix super/fine-class classification	✓	✓	✗	✗	✓

c)



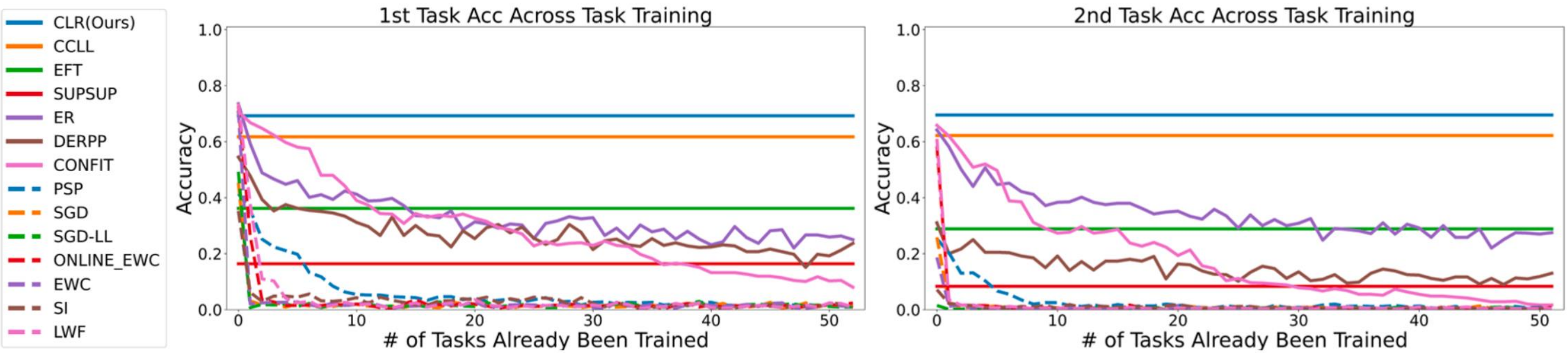
Ge, Yunhao, Yuecheng Li, Di Wu, Ao Xu, Adam M. Jones, Amanda Sofie Rios, Iordanis Fostiropoulos et al. "**Lightweight Learner for Shared Knowledge Lifelong Learning**." TMLR (2023)

Experiments: SKILL-102 continual Learning dataset

Comparison	53-dataset (ours)	8-dataset [1, 2]	ImageNet [47]	Fine grained 6 tasks [47] [35]	Cifar100 [34]	F-CelebA [43]
# tasks	**53**	8	20	6	20	10
# Classes	1,584	738	1000	**1943**	100	20
# Images	**1,811,028**	166,360	1,300,000	1,440,086	60,000	1,189
# different classification target	**5**	1	1	2	1	1
Mix image style (nature/artifact)	✓	✓	✗	✓	✗	✗
Mix super/fine-class classification	✓	✓	✓	✗	✗	✗

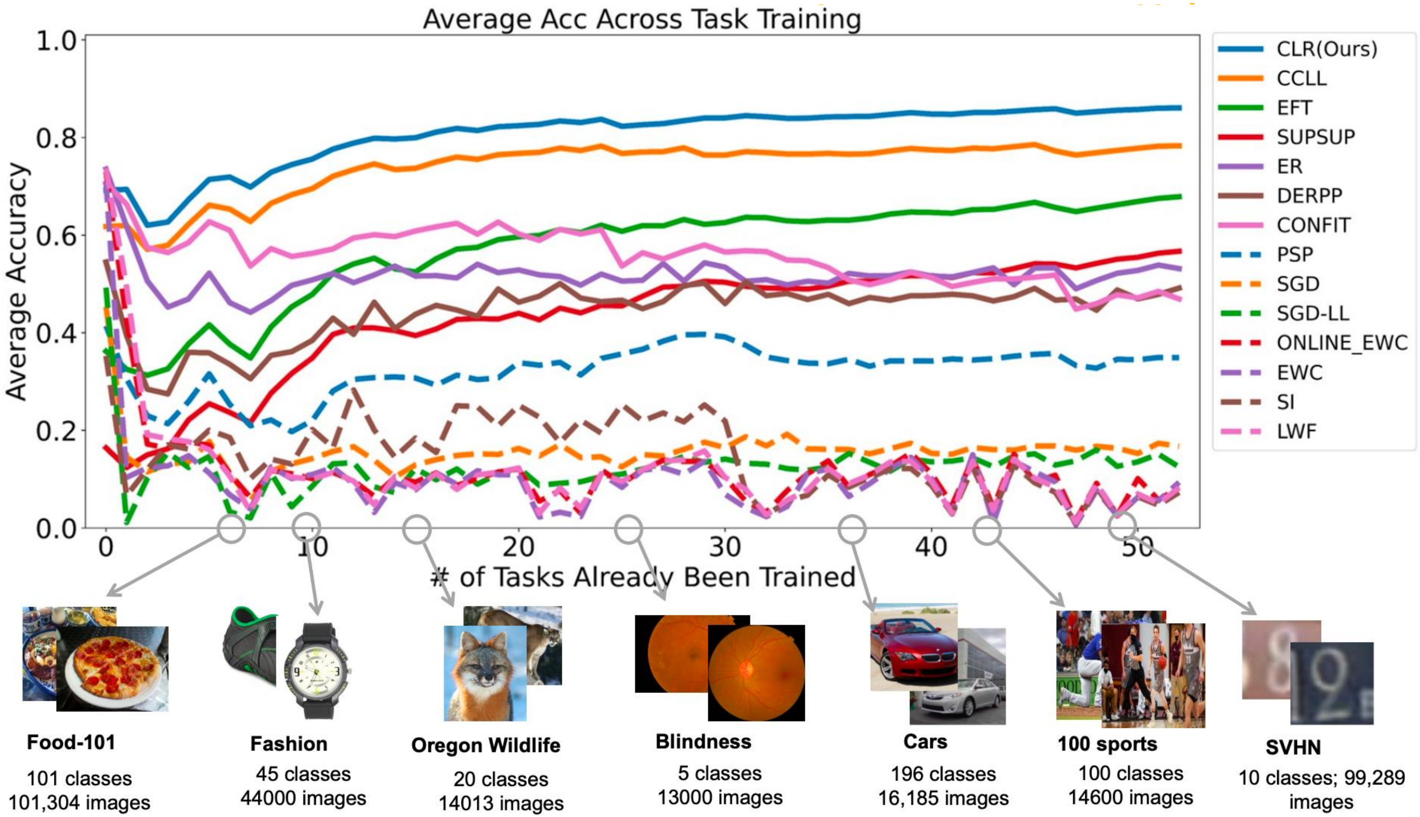
Ge, Yunhao, Yuecheng Li, Shuo Ni, Jiaping Zhao, Ming-Hsuan Yang, and Laurent Itti. "**CLR: Channel-wise Lightweight Reprogramming for Continual Learning**." ICCV (2023).

Experiments: First task performance during continual learning



Ge, Yunhao, Yuecheng Li, Shuo Ni, Jiaping Zhao, Ming-Hsuan Yang, and Laurent Itti. "**CLR: Channel-wise Lightweight Reprogramming for Continual Learning**." ICCV (2023).

Experiments: Average accuracy on all tasks learned so far



Ge, Yunhao, Yuecheng Li, Shuo Ni, Jiaping Zhao, Ming-Hsuan Yang, and Laurent Itti. "**CLR: Channel-wise Lightweight Reprogramming for Continual Learning**." ICCV (2023).

Experiments: Parameter and computation cost

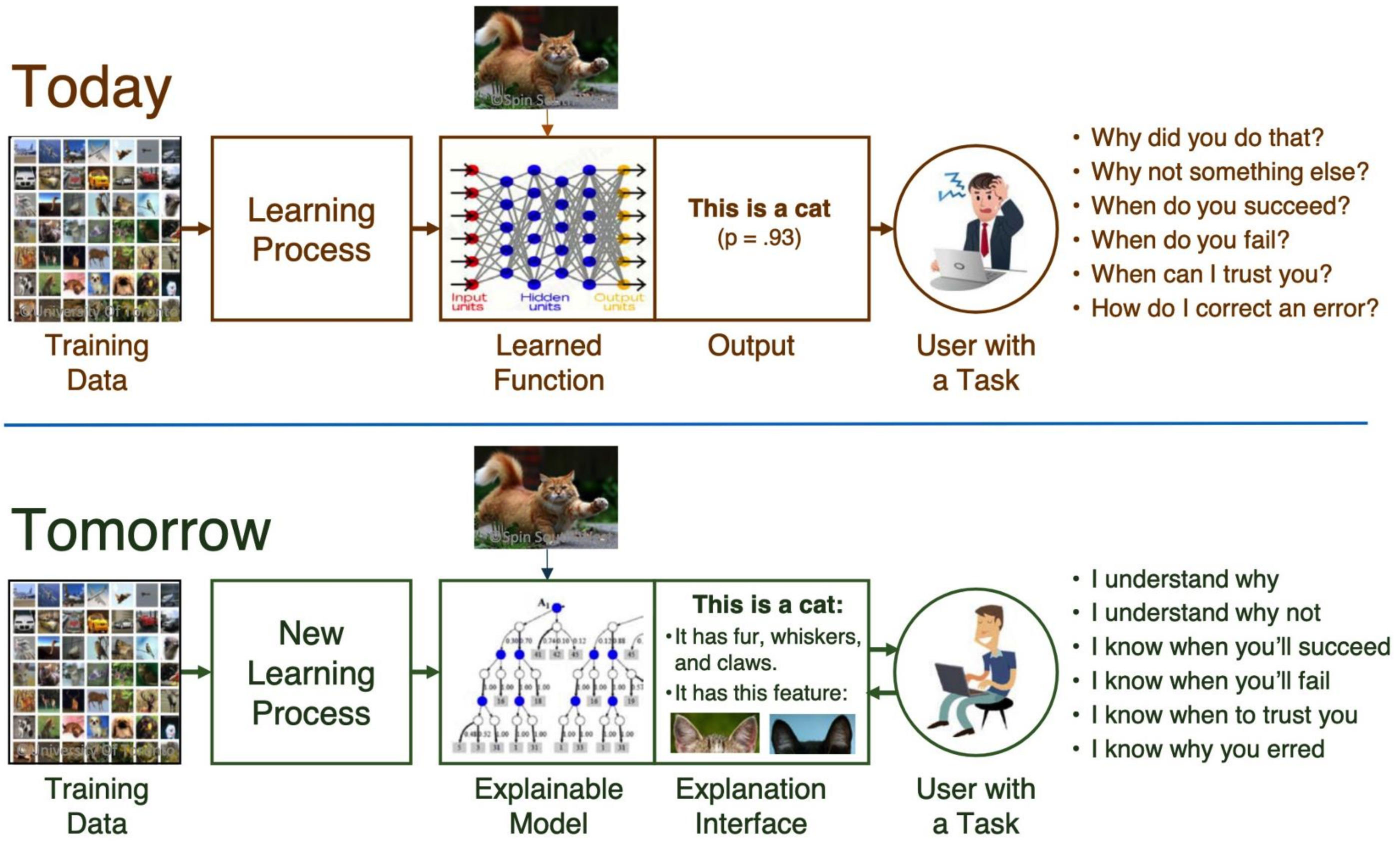
Method	Extra parameter to add 1 new task	Computation cost	Average Acc (53-datasets)
SGD	**0%**	1	16.71 %
PSP [12]	5.02%	0.828	34.91 %
EWC [32]	0%	1.160	9.36 %
ONLINE EWC [50]	0%	1.011	9.47* %
SGD-LL	**0%**	**0.333**	12.49 %
ER [45]	189.14%	3.99	53.13 %
SI [60]	0%	1.680	7.28%
LwF [36]	0%	1.333	8.23%
SUPSUP [58]	3.06%	1.334	56.69 %
EFT [55]	3.17%	1.078	67.8 %
CCLL [53]	0.62%	1.006	78.3 %
CLR (Ours)	0.59%	1.003	**86.05 %**

Ge, Yunhao, Yuecheng Li, Shuo Ni, Jiaping Zhao, Ming-Hsuan Yang, and Laurent Itti. "**CLR: Channel-wise Lightweight Reprogramming for Continual Learning**." ICCV (2023).

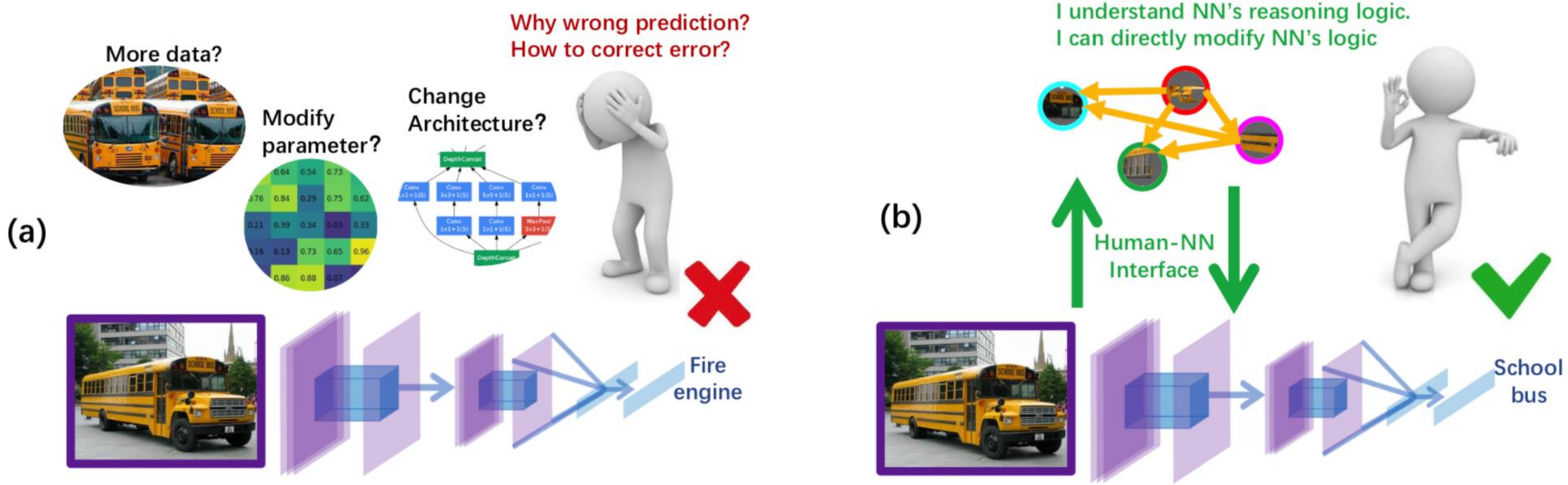
Towards explainable AI

Explainable AI – What Are We Trying To Do?

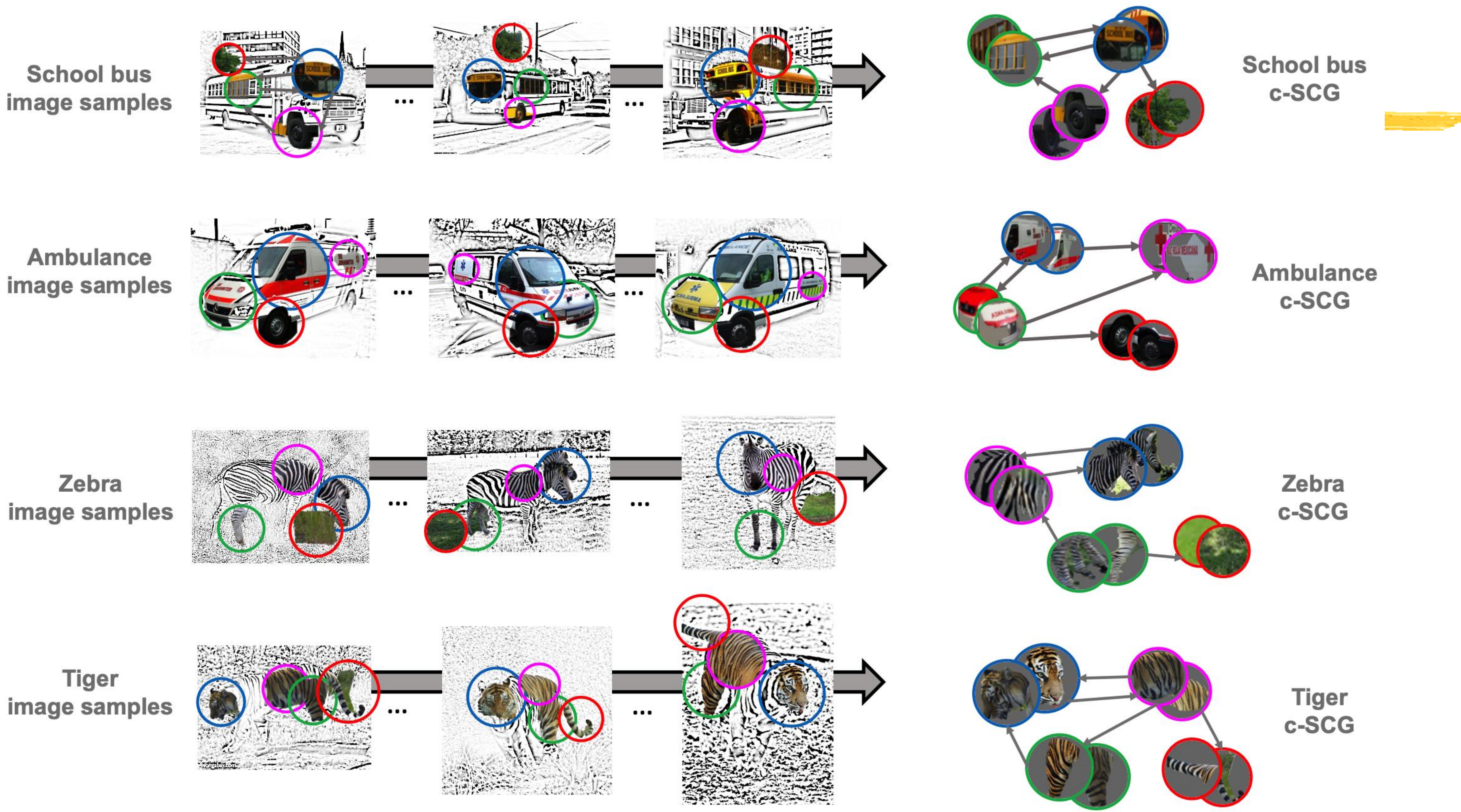
David Gunning, “Explainable Artifcial Intelliigence (XAI),” 2017



Robustness through explainability

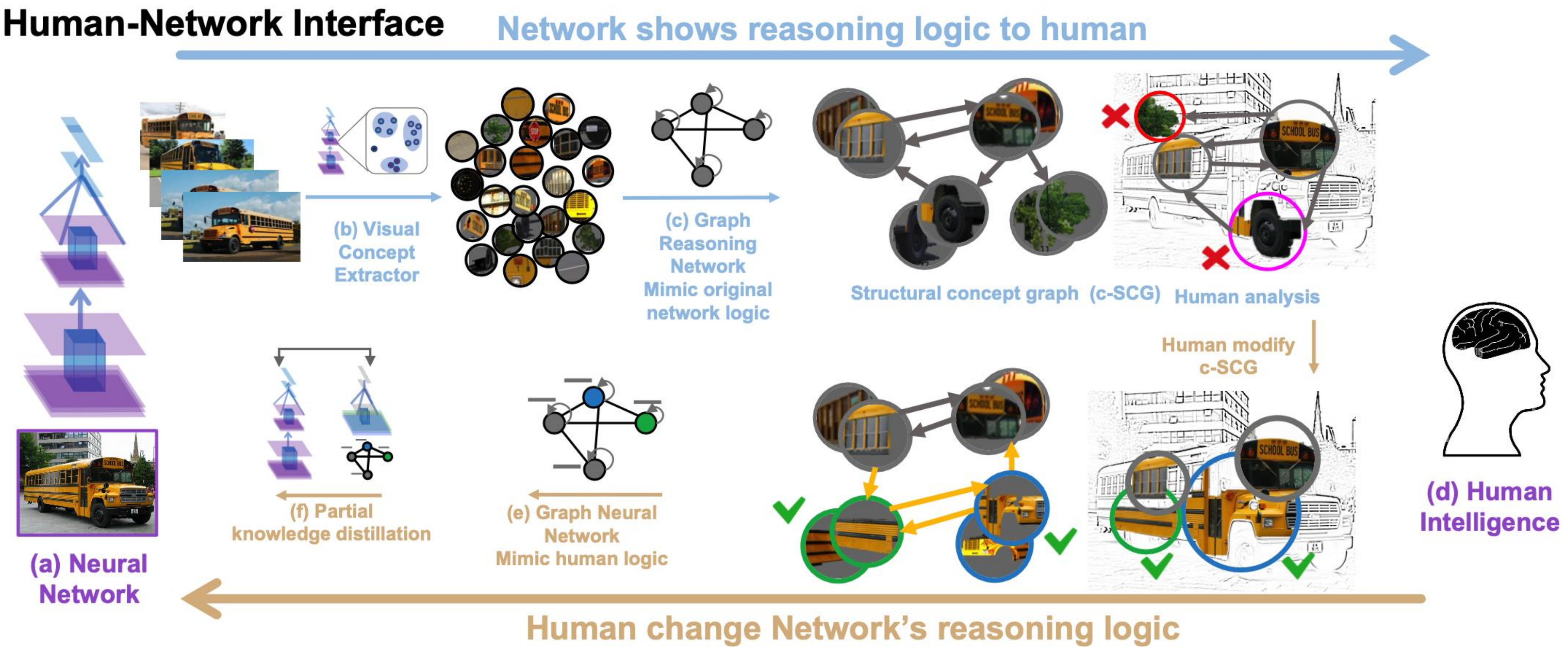


Ge et al., in prep

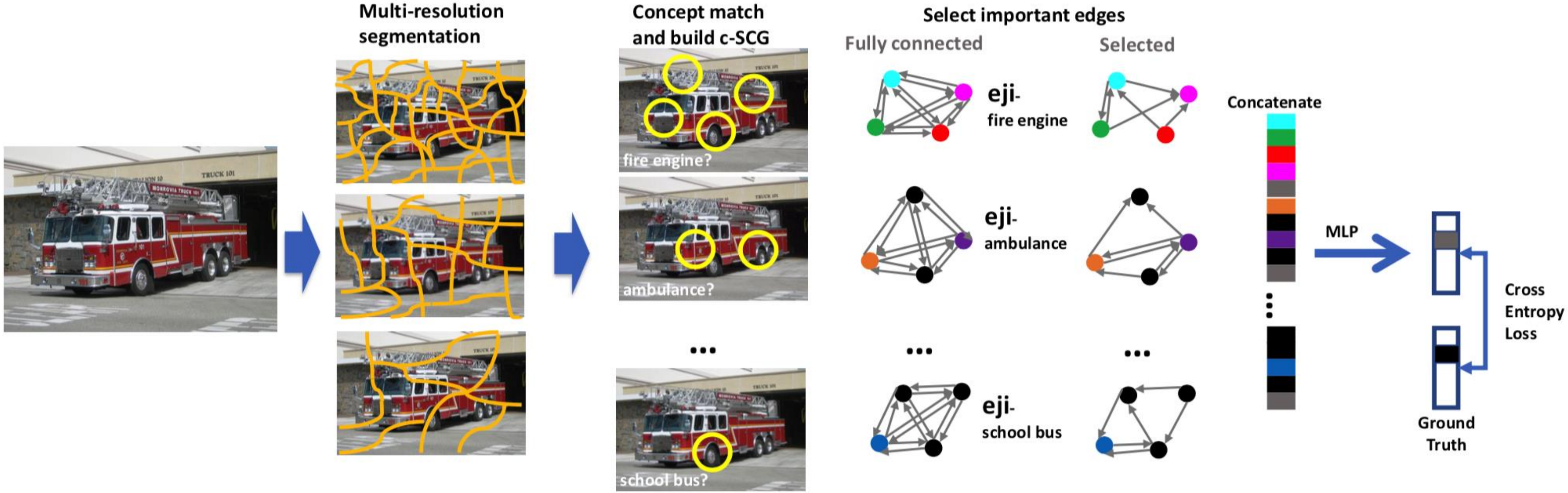


School bus
image samples
School bus
c-SCG
Ambulance
image samples
Ambulance
c-SCG
Zebra
image samples
Zebra
c-SCG
Tiger
image samples
Tiger
c-SCG

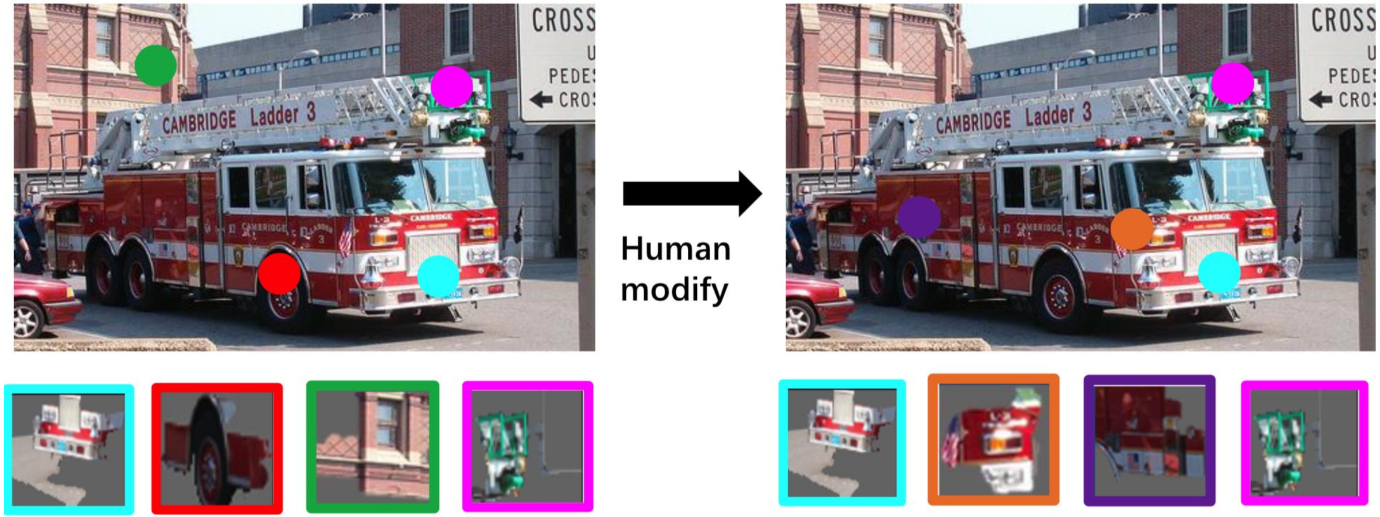
Robustness through explainability



Extracting visual concept graphs



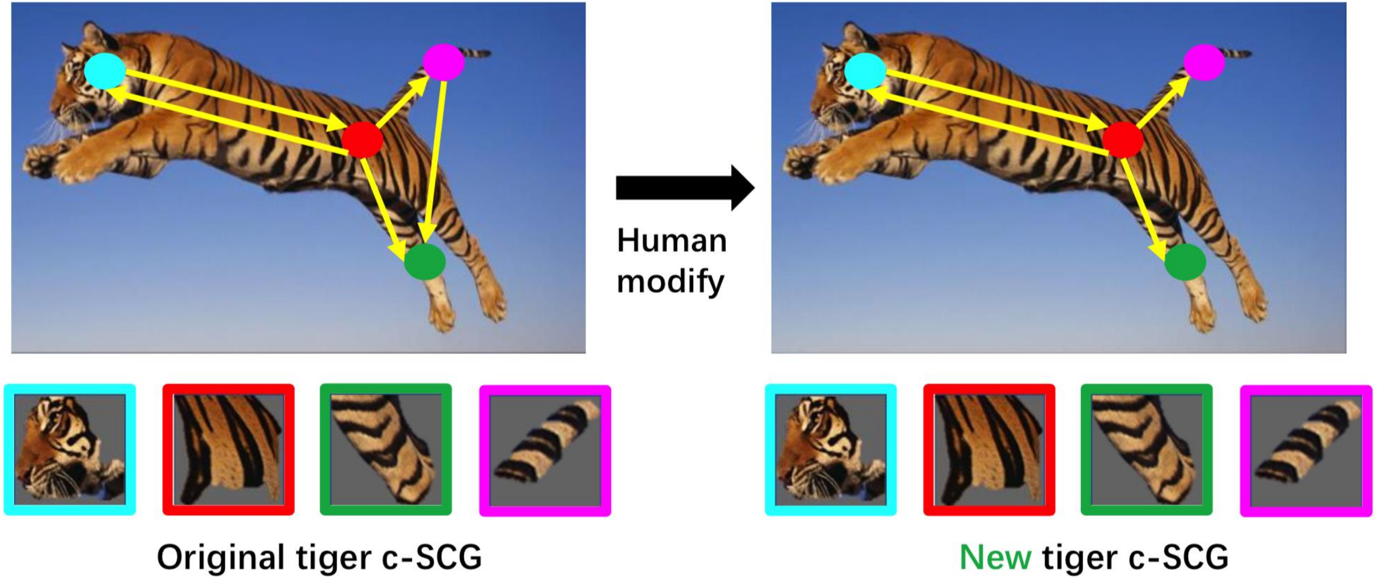
Example: human modifies nodes



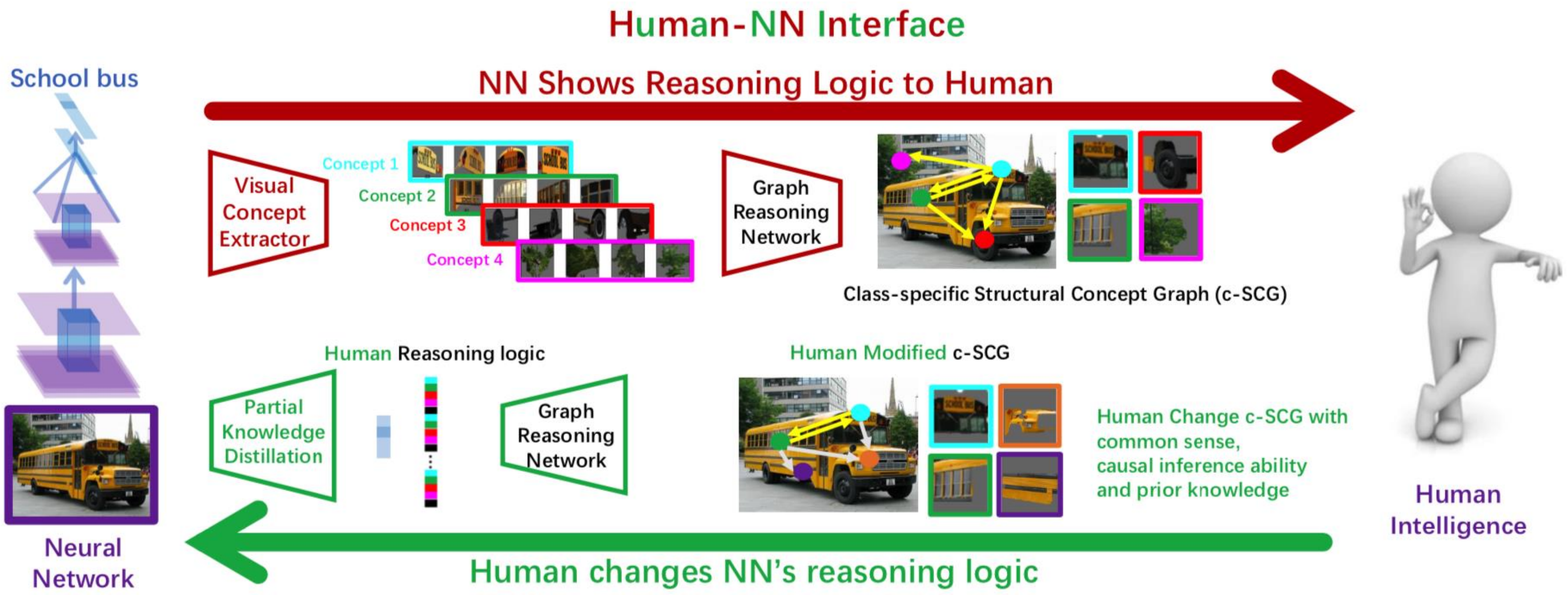
Original fire engine c-SCG

New Fire engine c-SCG

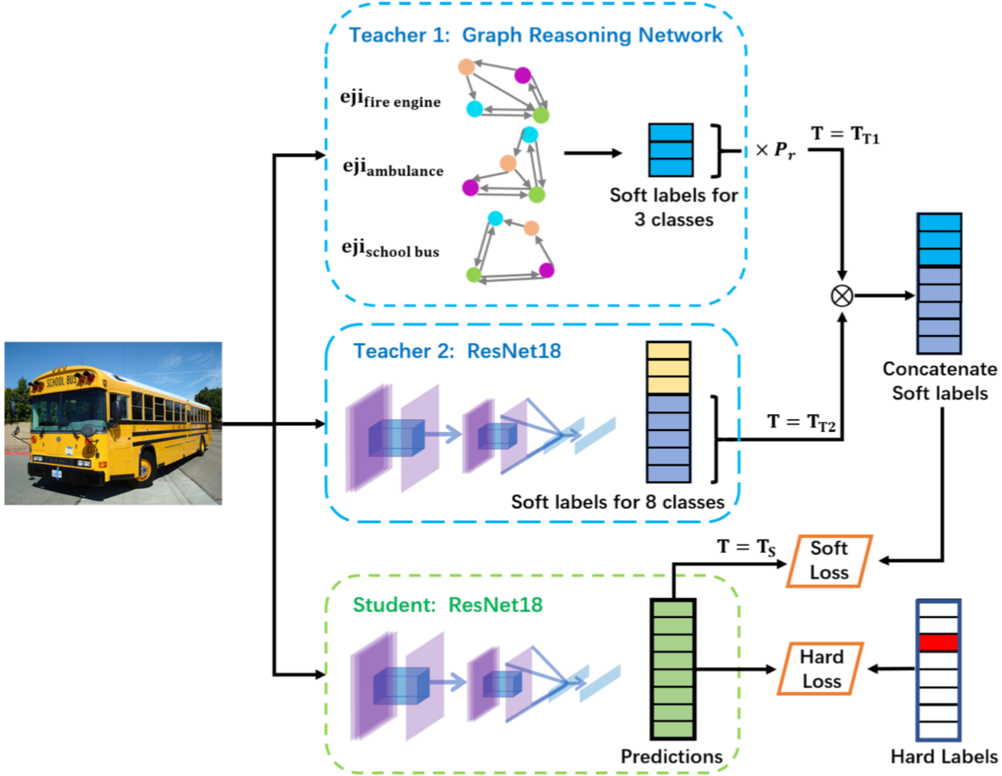
Example: human modifies edges



Robustness through explainability



Partial knowledge distillation



Results

Table 1: **Human improves a network's performance with HNI: experiments on six different image classification tasks (performance is tabulated as percent correct classification).**

Datasets	# images	Classes	# classes	original performance	performance with HNI
Cats	2382	modified classes	2	88.33	91.64 (+3.31)
		all classes	12	93.06	93.61 (+0.55)
Cars	8144	modified classes	2	83.33	91.67 (+8.34)
		all classes	10	86.33	88.33 (+2.00)
Monkeys	1642	modified classes	3	78.75	85.00 (+6.25)
		all classes	10	90.00	93.61 (+3.61)
Flowers	22267	modified classes	3	78.94	85.76 (+6.82)
		all classes	10	82.37	85.57 (+3.2)
Fashion	16186	modified classes	3	61.71	67.14 (+5.43)
		all classes	6	73.79	74.37 (+0.58)
Buildings	5063	modified classes	3	47.78	57.78 (+10.00)
		all classes	17	53.73	54.51 (+0.78)

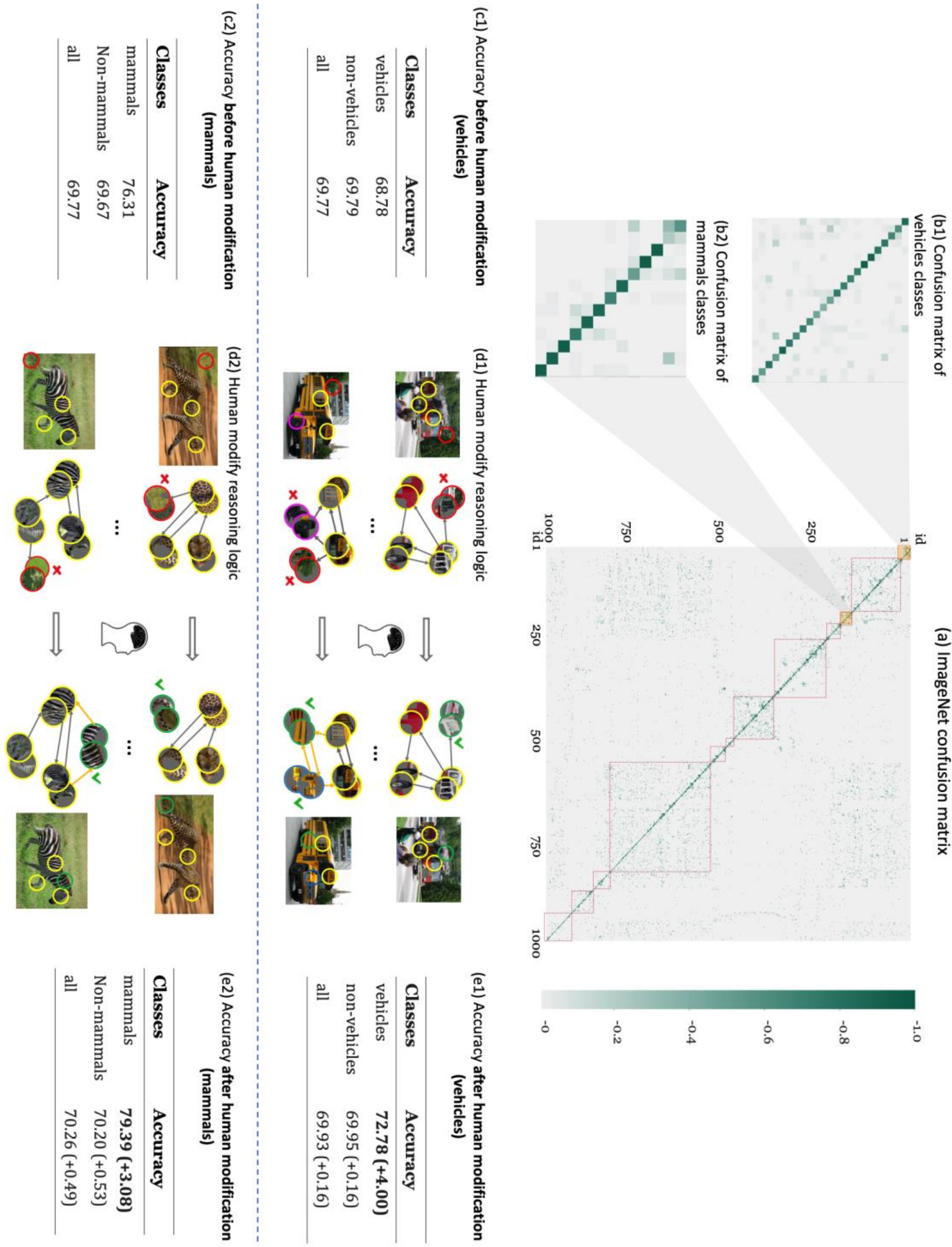


Figure 3: **Humans can improve a network's performance with HNI. We conduct large-scale experiments on the ImageNet dataset, which contains 1,000 real-world classes.** (a) Confusion matrix of a 1,000-class original GoogleNet image classification network trained on ImageNet. Most of the errors are within each of 12 super-classes that correspond to groups of related classes (e.g., mammals, vehicles, birds, etc). There are two main challenges: (1) How to correct the errors and improve accuracy within a super-class (local logic) with the help of human involvement? (2) How to maintain the performance of all other classes in the 1,000 classes? We show results of two large-scale experiments, for the superclasses of vehicles (23 classes, total 23,000 training images) and mammals (13 classes, 13,000 images), to show how one can use HNI to improve performance within a superclass without degrading performance of other classes. We first consider the super-class of vehicles (b1), with original accuracy over these 68.78% (c1). For each class, the network-to-human pass was used to show the reasoning logic of the original network as a c-SCG to a human operator (d1). The operator spotted and corrected any reasoning errors of the network. The human-to-network pass then distilled the human-modified logic back to the original network with the help of Graph neural network and partial knowledge distillation. Performance was improved on the vehicle classes, without degradation of non-vehicle classes (e1), demonstrating how humans could use their own knowledge to correct reasoning errors of the network and improve network accuracy. The same process is also shown for the superclass of mammals (b2, c2, d2, e2).

Conclusions

- New Beneficial Biases help not only with lifelong learning but also with robustness to adversarial attacks

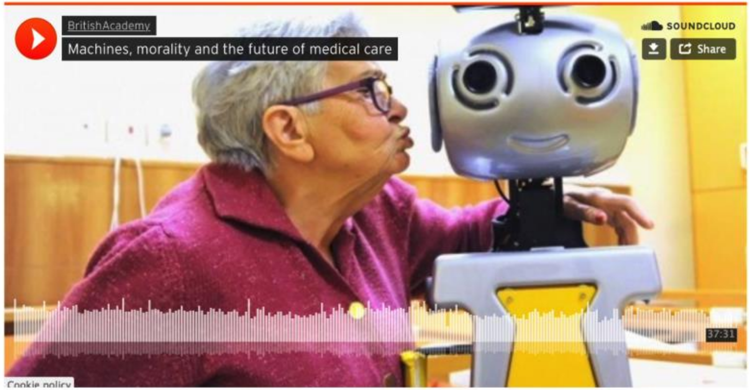
- Human/NN interaction through graph representations allows one to improve robustness and generalization by injecting human knowledge about natural vs spurious correlations in training data.

This work was supported by C-BRIC (one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA), DARPA (HR00112190134) and the Army Research Office (W911NF2020053). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

Some slides in this lecture are from

https://courses.grainger.illinois.edu/ece448/sp2021/slides/lec25.pdf

CS440/ECE448 Lecture 25:
AI Ethics



Modified by Mark Hasegawa-Johnson, 4/2021
Including slides by Svetlana Lazebnik, 10/2017
CC-BY-4.0: Copy at will, but cite the source

Image source: https://www.britac.ac.uk/audio/machines-morality-and-future-medical-care

Problem statement: Example 1

- Bickel, Hammel, and O'Connell, "Sex bias in graduate admissions: Data from Berkeley," Science 187(4175):398–404, 1975

- Women were being admitted to Berkeley at a far lower rate than men.

Problem statement: Example 1

- Bickel, Hammel, and O'Connell, "Sex bias in graduate admissions: Data from Berkeley," Science 187(4175):398–404, 1975

- Women were being admitted to Berkeley at a far lower rate than men.

- Women were applying to departments with lower acceptance rates. Within each department, admission rates for men and women were the same.

- Is this fair?

Problem statement: Example 2

A given job might discriminate based on:

- Upper body strength
- Habitual clothing
- Undergraduate major

All of these correlate with gender.

Is it fair to use them as a basis for employment?

CNBC

OpenAI faces complaint to FTC that seeks investigation and suspension of ChatGPT releases

The Center for AI and Digital Policy accuses OpenAI of violating a part of the FTC Act that prohibits unfair and deceptive business...

Reuters

Australian mayor readies world's first defamation lawsuit over ChatGPT content

A regional Australian mayor said he may sue OpenAI if it does not correct ChatGPT's false claims that he had served time in prison for...

5 days ago

Times of India

China's Baidu has sued Apple over fake versions of its ChatGPT rival Ernie bot

Chinese search engine giant Baidu has filed lawsuits against "relevant" app developers and Apple over fake copies of its ChatGPT rival Ernie...

2 days ago

AI ethics, fairness, and bias

Fairness
– Representation
– Algorithmic solutions
– Transparency as a solution to the fairness problem

Accountability
– Physical safety – Data safety
– Economic safety

Transparency
– Explainable AI
– Understanding the inner workings of a superintelligence

AI ethics, fairness, and bias

Fairness
– Representation
– Algorithmic solutions
– Transparency as a solution to the fairness problem

Accountability
– Physical safety – Data safety
– Economic safety

Transparency
– Explainable AI
– Understanding the inner workings of a superintelligence

AI ethics, fairness, and bias

Let's define the following random variables:

- A = protected attribute: An observable fact that should not be predictive of outcomes, e.g., gender, race, age, disability.

- X = observable data that we can use for our decision

- Y = the unknown correct label for this person (e.g., Y = 1 might mean "this person should receive a loan" or "should be admitted to USC")

- $\hat{Y}$ = a function of X, designed using probabilistic or neural methods to approximate Y as closely as possible

AI ethics, fairness, and bias

Demographic Parity:

The probability of a positive outcome is the same, regardless of protected attribute.

$$P\left(\hat{Y} = 1 | A = a\right) = P\left(\hat{Y} = 1 | A = a'\right) \quad \forall a, a'$$

Predictive Parity:

Precision is the same, regardless of protected attribute.

$$P\left(Y = 1 | \hat{Y} = 1, A = a\right) = P\left(Y = 1 | \hat{Y} = 1, A = a'\right) \quad \forall a, a'$$

Error Rate Balance:

Recall is the same, regardless of protected attribute.

$$P\left(\hat{Y} = 1 | Y = 1, A = a\right) = P\left(\hat{Y} = 1 | Y = 1, A = a'\right) \quad \forall \hat{y}, a, a'$$

AI ethics, fairness, and bias

You can't have all three

From Bayes' theorem
P(M|D)=P(D|M)P(M)/P(D)

Predictive parity

$$P(\hat{Y}=1|Y=1,A=a)=\frac{P(Y=1|\hat{Y}=1,A=a)P(\hat{Y}=1|A=a)}{P(Y=1|A=a)}$$

Demographic parity

Balanced error

Ground Truth

$$P(\hat{Y}=1|Y=1,A=a')=\frac{P(Y=1|\hat{Y}=1,A=a')P(\hat{Y}=1|A=a')}{P(Y=1|A=a')}$$

The balanced error, predictive parity, and demographic parity terms cannot all be independent of A unless Y is also independent of A.

In other words, if the current state of society is unfair (distribution of positive outcomes currently depends on protected attribute), then algorithmic solutions cannot make it fair (at least not in all three ways, all at once).

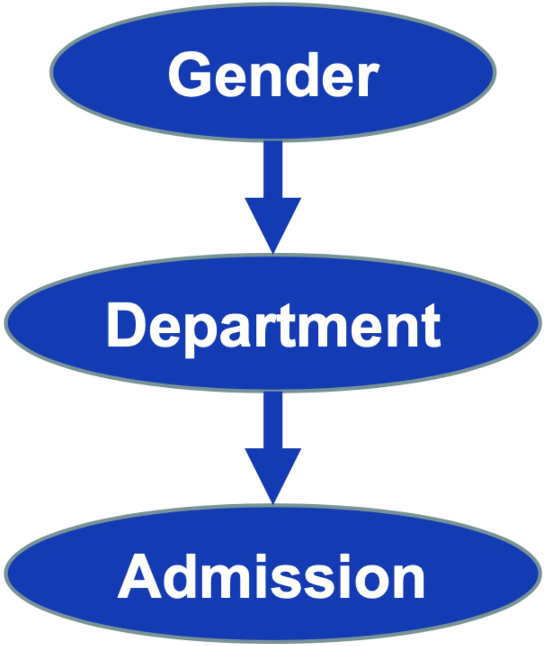
Other Useful Definitions of Fairness in AI

Individual Fairness:
The dissimilarity between two outcomes should be less than the dissimilarity between the people.

Counterfactual Fairness:
If a person's protected attribute were changed (and all their other attributes were possibly changed, according to their dependence on the protected attribute), then the outcome should not change.

State-of-the-art Solution: Explain Your Assumptions to Your Users

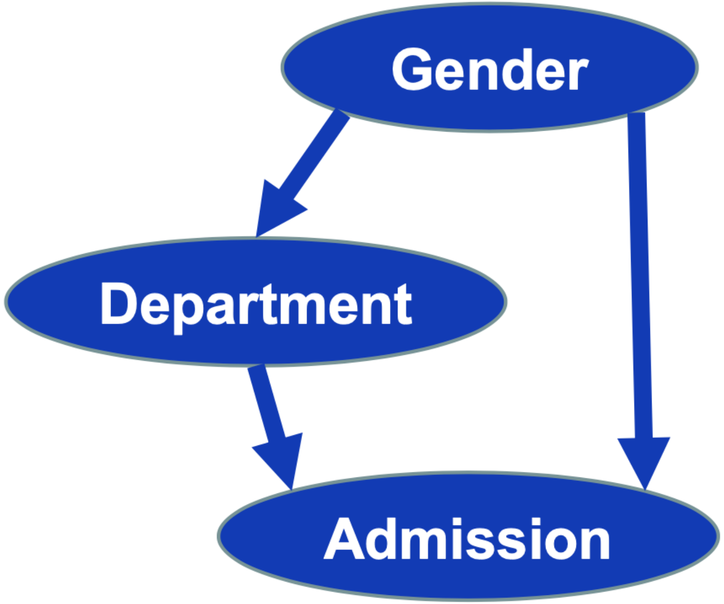
- The Transparency Dogma says that an algorithm's assumptions should be published in a way that users can understand, so that users can participate in a public debate about the fairness of the assumptions.
- A common way to do this is by drawing a Bayesian network.
- Example shown: the Bickel-Hammel-O'Connell data. Admission was conditionally independent of Gender, given Department.



State-of-the-art Solution: Explain Your Assumptions to Your Users

A possible solution: in order to better approximate demographic parity,

- Make admission explicitly dependent on gender.
- Admit women at a slightly higher percentage, in every department to which they apply, so that…
- …the total percentage of admitted women equals the total percentage of admitted men.



Public debate: is that more fair, or less fair?