Department of Computer Science
香港城市大學
City University of Hong Kong

IJCAI/2023 MACAO

IJCAI 2023 Tutorial on

Continual Learning and Its Extension to Pre-trained Models

Ying Wei
City University of Hong Kong

Da-Wei Zhou
Nanjing University

Han-Jia Ye
Nanjing University

Outline

- Introduction (20 minutes)
  - Problem definition & history of continual learning
  - Evaluation metrics & benchmarks in continual learning
- Conventional Continual Learning (125 minutes)
  - Approaches for catastrophic forgetting (60 minutes)
  - Approaches for forward transfer (65 minutes)
- Continual Learning of Pre-trained Models (45 minutes)
  - Challenges and opportunities
  - Approaches
- Conclusion and Future Discussion (10 minutes)
- Q&A Session (10 minutes)

Da-Wei Zhou

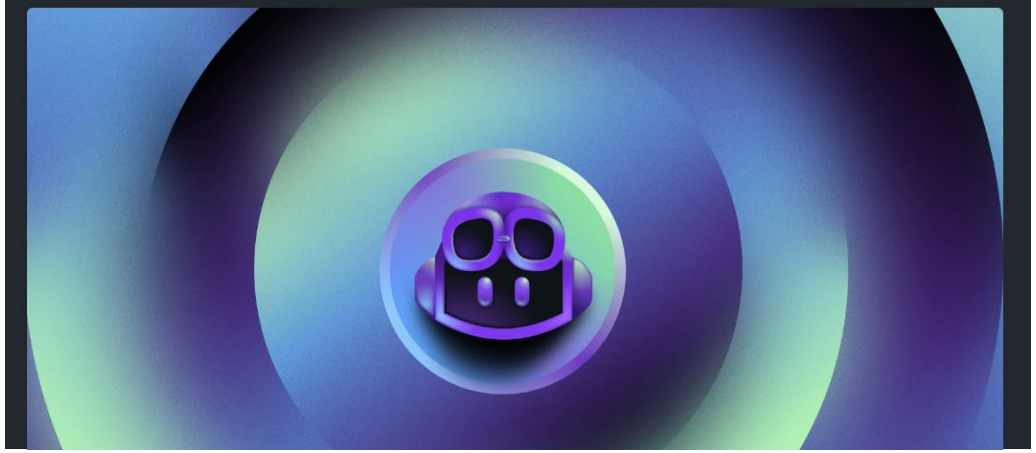
Han-Jia Ye

Ying Wei

Recent Advances in AI

GitHub Copilot X: The AI-powered developer experience

GitHub Copilot is evolving to bring chat and voice interfaces, support pull requests, answer questions on docs, and adopt OpenAI's GPT-4 for a more personalized developer experience.

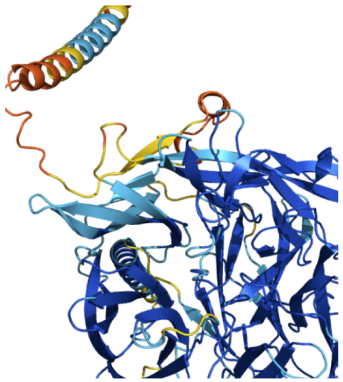


ChatGPT: New AI chatbot has everyone talking to it

7 December 2022

AlphaFold is an AI system developed by DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.

DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) have partnered to create AlphaFold DB to make these predictions freely available to the scientific community. The latest database release contains over 200 million entries, providing broad coverage of UniProt (the standard repository of protein sequences and annotations). We provide individual downloads for the human proteome and for the proteomes of 47 other key organisms important in research and global health. We also provide a download for the manually curated subset of UniProt (Swiss-Prot).



Q8I3H7: May protect the malaria parasite against attack by the immune system. Mean pLDDT 85.57.

Recent Advances in AI

User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

OpenAI ✓
@OpenAI

"A photo of an astronaut riding a horse" #dalle

Stable Diffusion ✓
@ai_diffusion

On stage 🎤

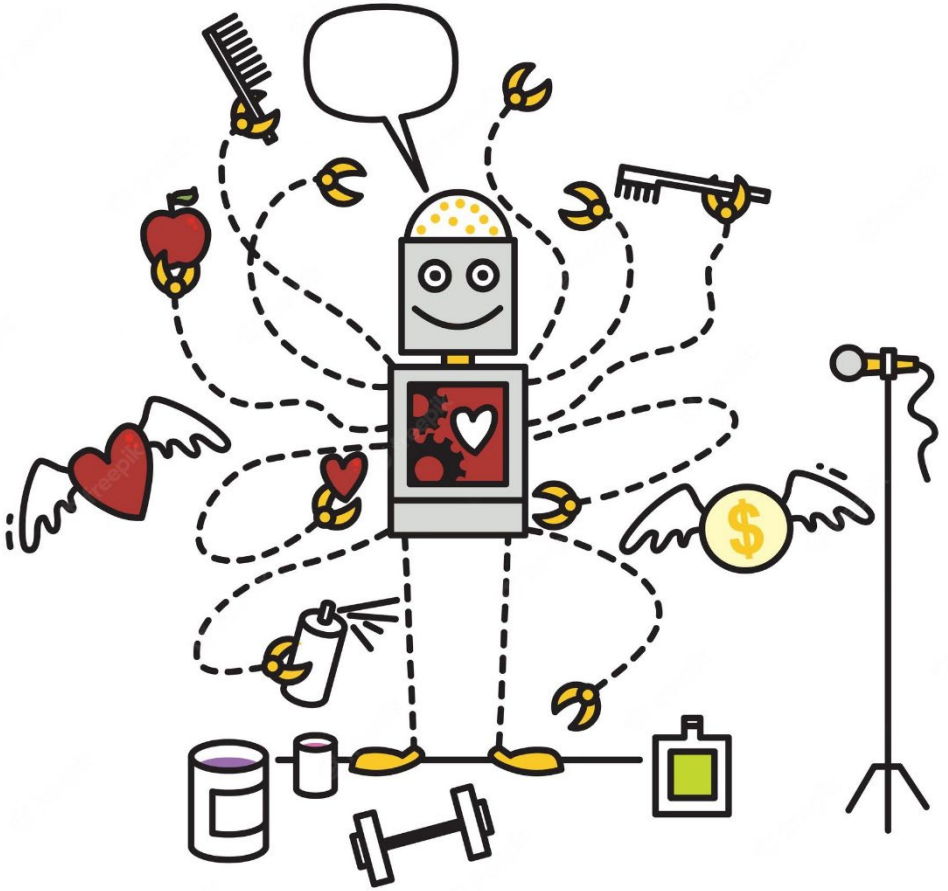
#CoolPope #PopeFrancis @Pontifex
#stabledifussion #detailedprompt #prompt #stablediffusionart
#digitalartwork #aigenerated

detailed prompt in image description

Recent Advances in AI

Uriel Singer et al. *Make-A-Video: Text-to-Video Generation without Text-Video Data*. ICLR 2023.
Jay Zhangjie Wu et al. *Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation*. ICCV 2023.
Fu-Yun Wang et al. *Gen-L-Video: Multi-Text to Long Video Generation via Temporal Co-Denoising*. CoRR 2023.

Can AI Manage Multiple Tasks Like Human?

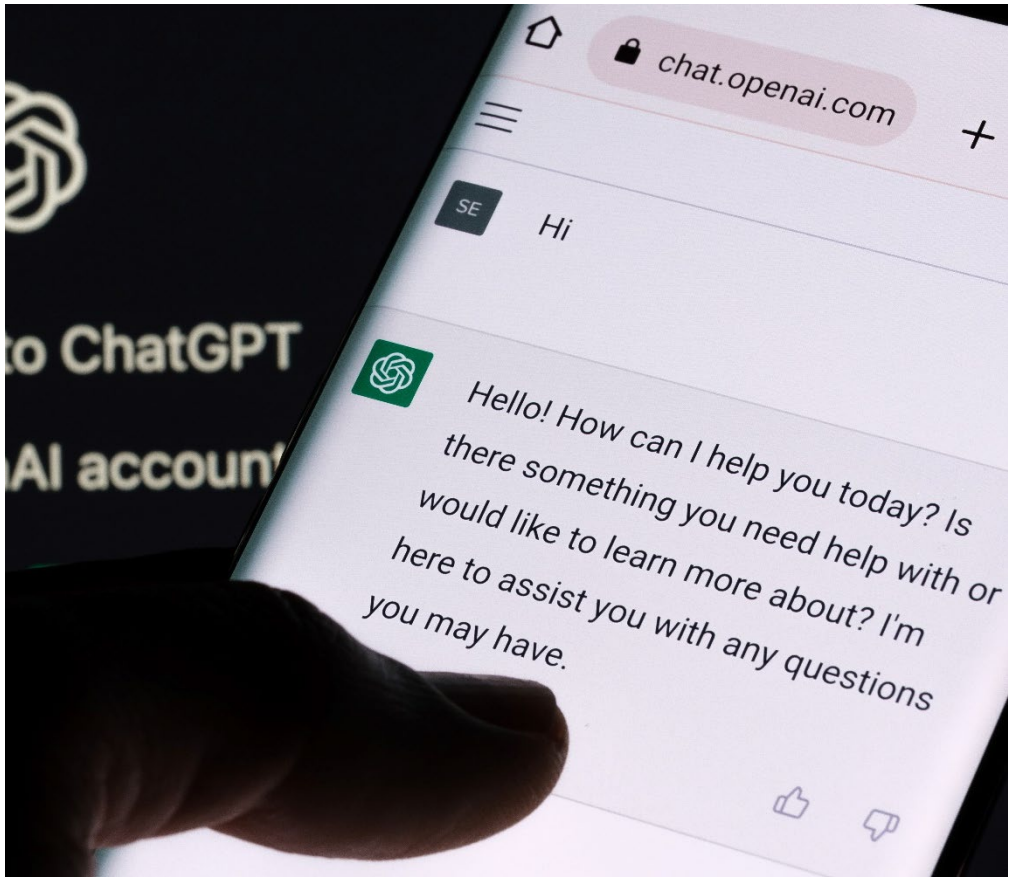


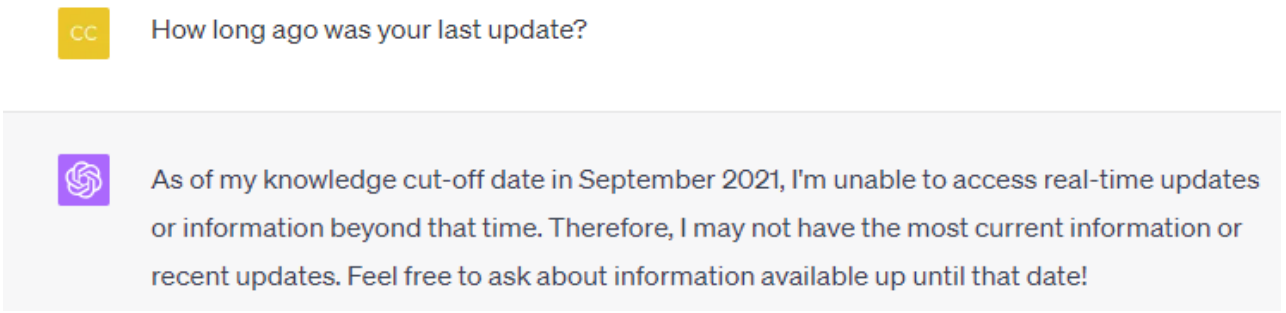
Human do Continual Learning

Continual learning is the constant development of ***increasingly complex behaviors***; the process of ***building more complicated skills*** on top of those already developed.

MARK B. RING. *CHILD: A First Step Towards Continual Learning.* Machine Learning 1997.

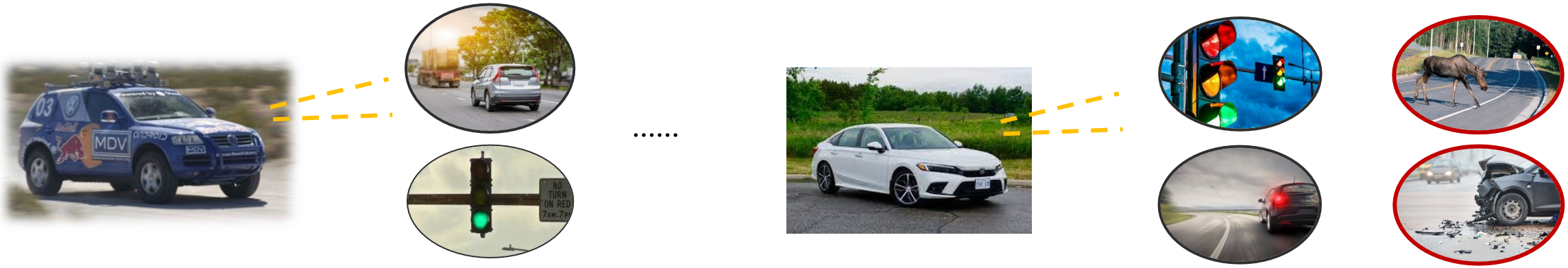
Real World APPs Require Continual Learning

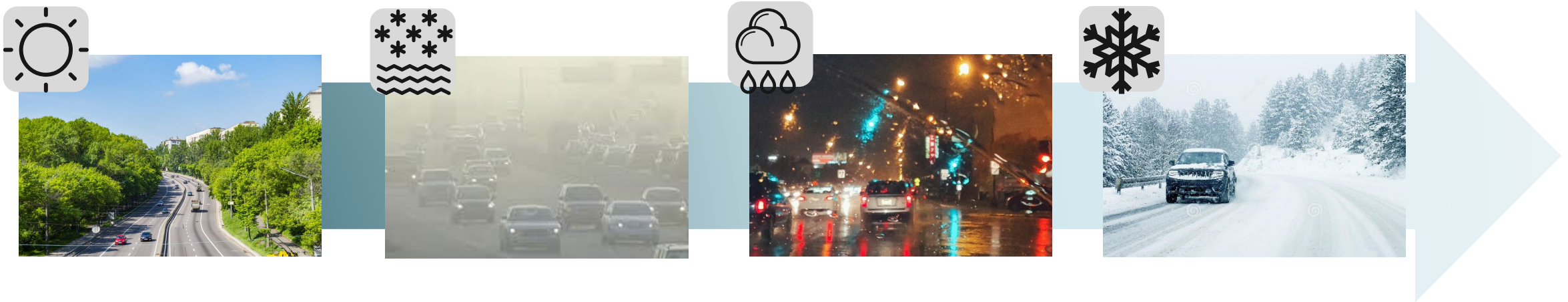




News will occur, and chatbot needs updating

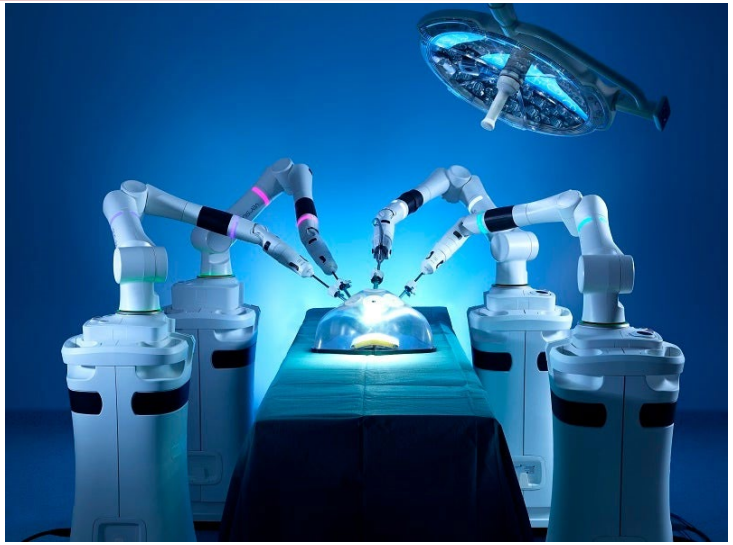
Real World APPs Require Continual Learning

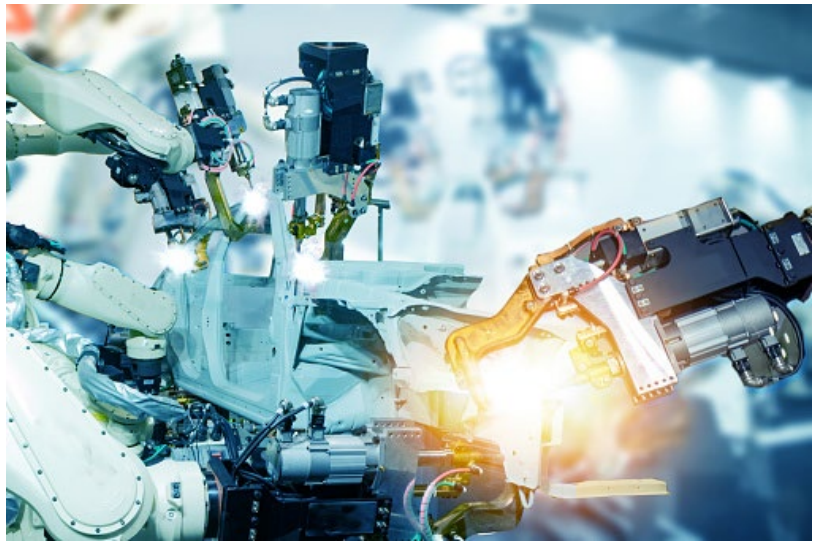


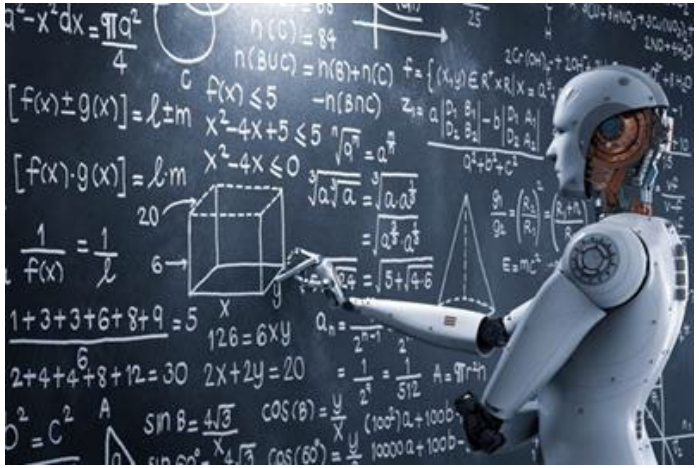


New scenes will emerge, and self-driving vehicles need updating.

Real World APPs Require Continual Learning





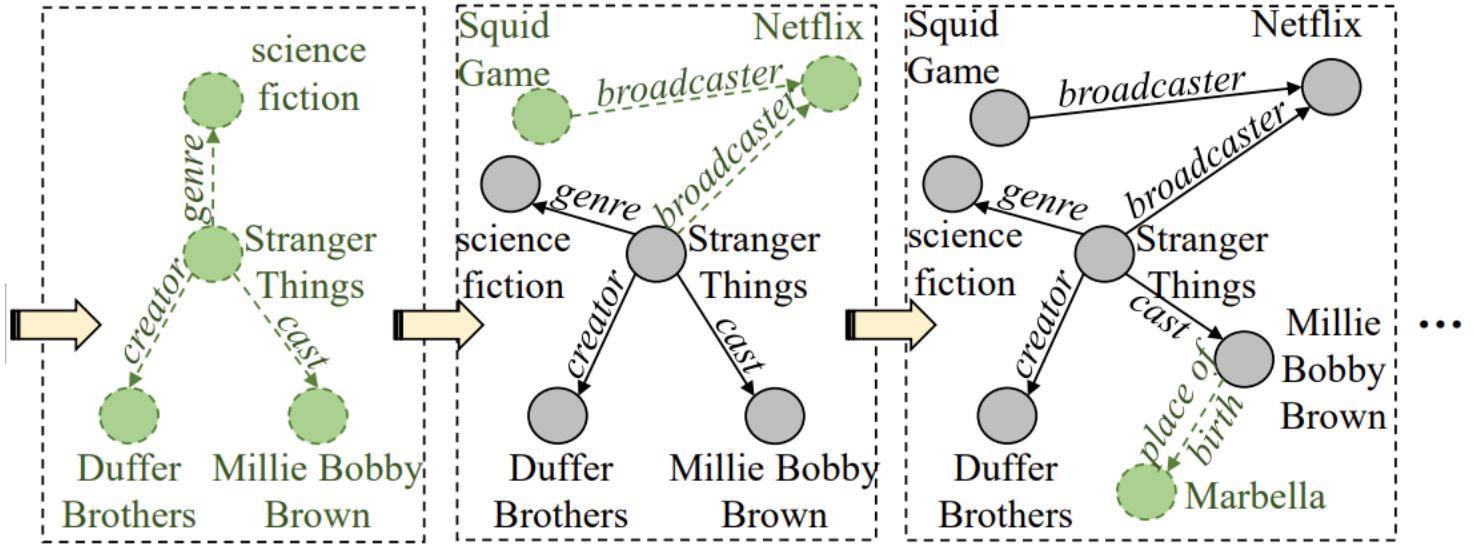


Robot needs to learn new skills

Real World APPs Require Continual Learning

Face recognition system needs to authenticate new users

Real World APPs Require Continual Learning

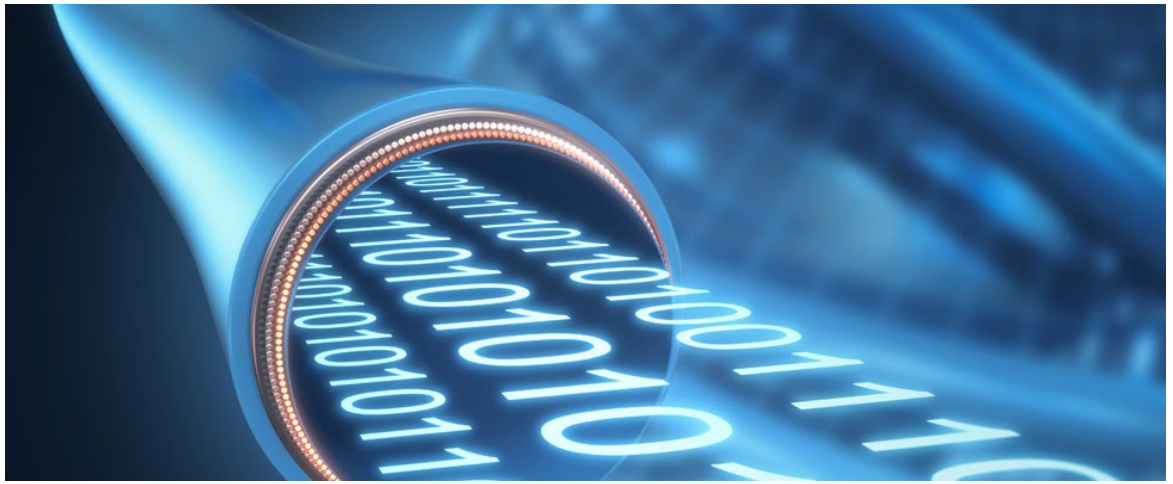


Social network faces new relationships

Yuanning Cui et al. *Lifelong Embedding Learning and Transfer for Growing Knowledge Graphs.* AAAI 2023.

Problem Definition

- *Continual Learning (a.k.a. incremental learning/lifelong learning)* refers to the scenario where the model needs to be continually updated with new data.
- Data comes in the stream format, which cannot be held for storage or privacy issues.



Zhiyuan Chen, Bing Liu. *Lifelong Machine Learning.* Morgan & Claypool 2018.

Problem Definition

- Given a sequence of tasks $(\mathcal{X}^t, \mathcal{Y}^t)$ drawn from distribution $\mathcal{D}^t$
- During the training process of task $\mathcal{T}$, we have ***limited or no access to former data*** $(\mathcal{X}^t, \mathcal{Y}^t), \mathrm{t} < \mathcal{T}$
- **Goal**: minimize the expected risk of all seen tasks

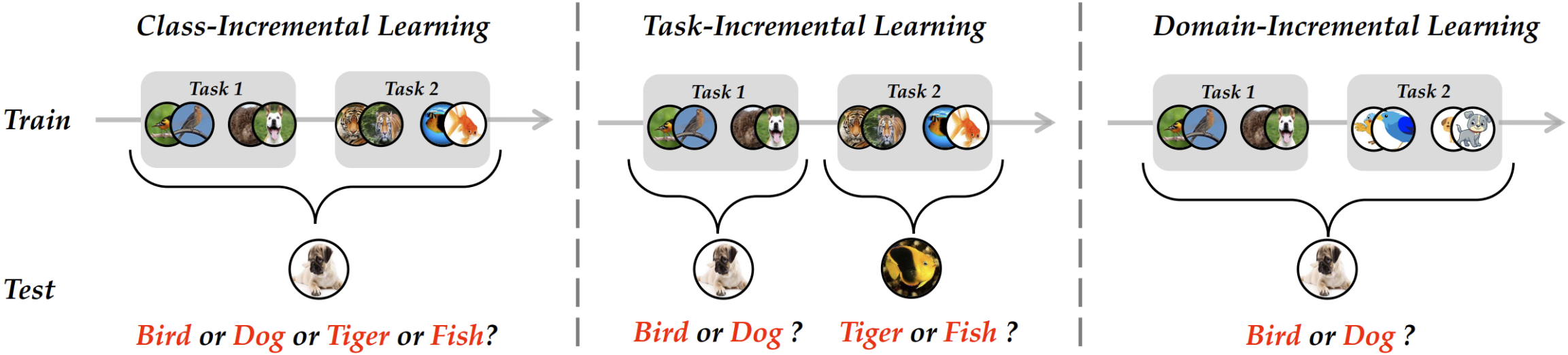
$$\min \sum_{t=1}^{\mathcal{T}} \mathbb{E}_{(\mathcal{X}^t, \mathcal{Y}^t)} [\ell(f(\mathcal{X}^t; \theta), \mathcal{Y}^t]$$

The specific setting counts on the ***definition of "task"*** $(\mathcal{X}^t, \mathcal{Y}^t)$

Matthias De Lange et al. *A continual learning survey: Defying forgetting in classification tasks.* TPAMI 2021.

Problem Definition

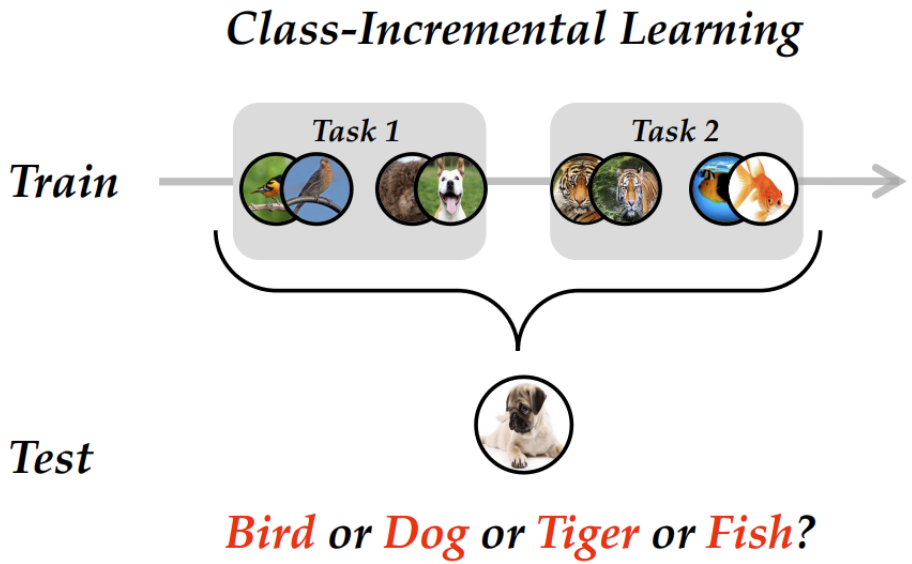
- Class-Incremental Learning (CIL): classify among all seen classes
- Task-Incremental Learning (TIL): classify among each (given) task
- Domain-Incremental Learning (DIL): classify among all distributions



van de Ven, Tuytelaars, Tolias. *Three types of incremental learning*. NMI 2022.

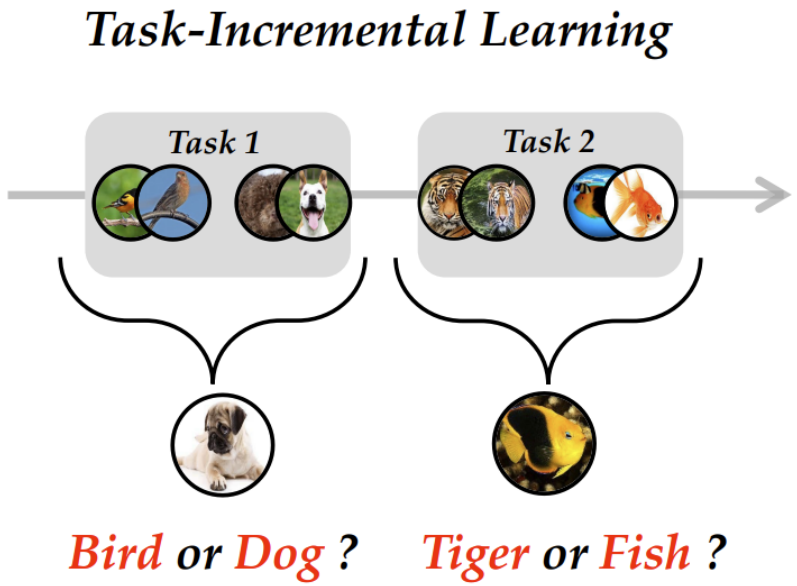
Class-Incremental Learning (CIL)

- Observed class labels are increasing $\{Y^t\} \subset \{Y^{t+1}\}$
- Class distributions are changing $P(Y^t) \neq P(Y^{t+1})$
- Input distribution is changing $P(X^t) \neq P(X^{t+1})$
- **Incrementally learn new classes to build a unified classifier**



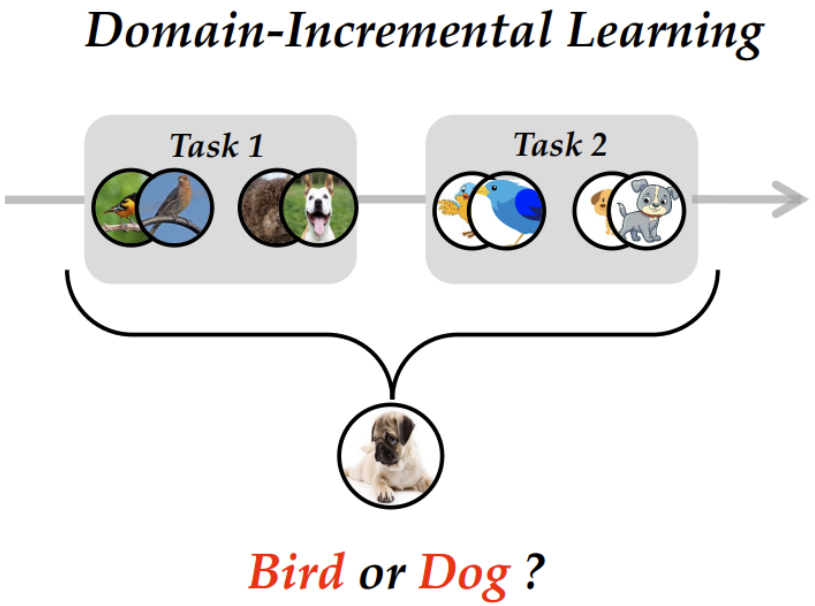
Task-Incremental Learning (TIL)

- Training target is changing from task to task $\{Y^t\} \neq \{Y^{t+1}\}$
- Input distribution is changing $P(X^t) \neq P(X^{t+1})$
- **The task id (t) is known during testing**
- An easier setting compared to CIL



Domain-Incremental Learning (DIL)

- Class labels are constant $\{Y^t\} = \{Y^{t+1}\}$
- Class distributions are constant $P(Y^t) = P(Y^{t+1})$
- Input distribution is changing $P(X^t) \neq P(X^{t+1})$



Data split in CIL/TIL

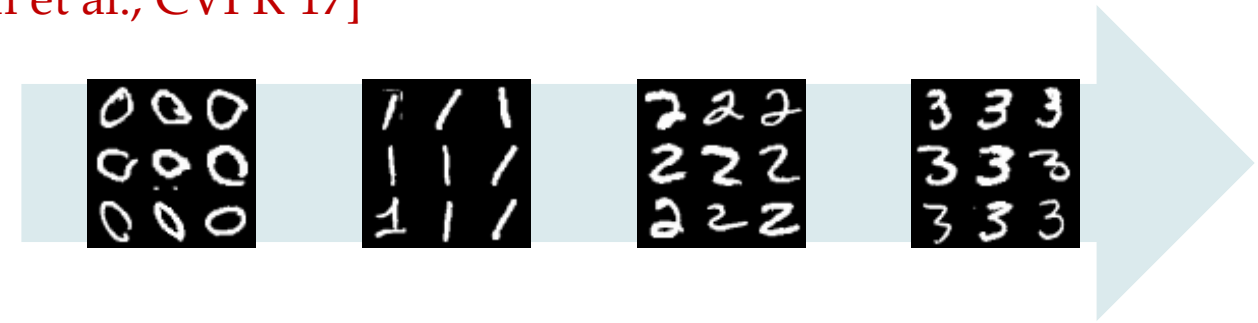
- Denote the total number of classes as C and the total number of tasks as N, there are two policies for class spilts:
  - Training from scratch (TFS) → Equally divide C classes into N tasks
  - Training from half (TFH) → Train C/2 classes in the first stage, and equally divide the other classes into the rest (N-1) tasks.

TFS, 5 tasks

TFH, 3 tasks

Benchmarks datasets in CIL/TIL

- Image classification
  - **CIFAR100** [Rebuffi et al., CVPR'17]
  - **ImageNet1000/ImageNet100** [Rebuffi et al., CVPR'17]
  - MNIST [Lopez-Paz et al., NIPS'17]
  - etc

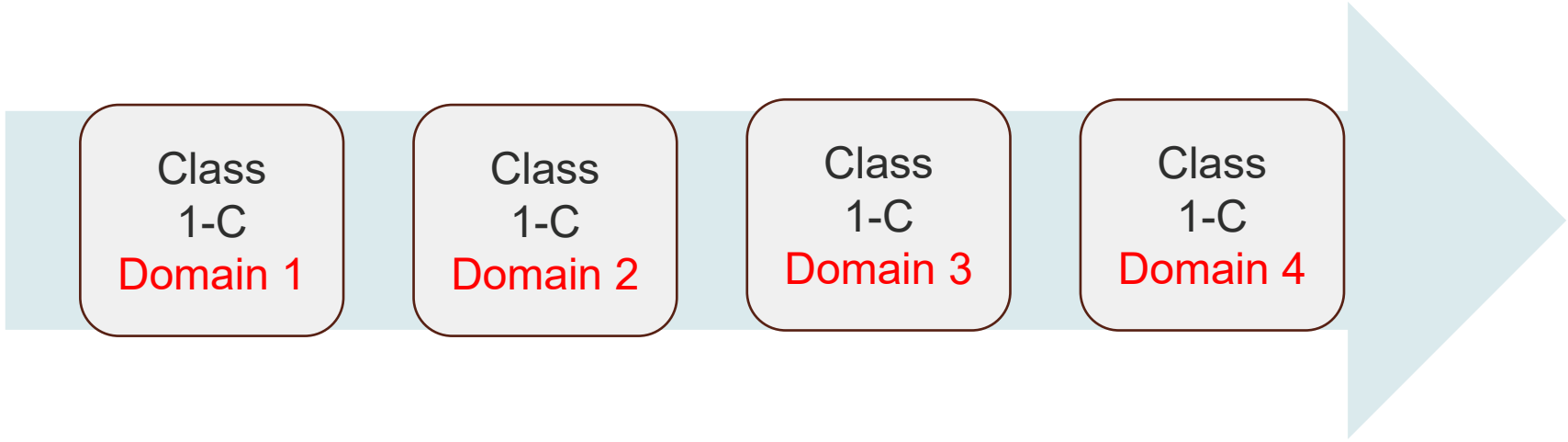


- NLP
  - Named-entity recognition: CoNLL-03, OntoNotes [Monaikul et al., AAAI'21]
  - Intent classification: CLINC150, HWU64, BANKING77 [Varshney et al., ACL'22]
  - Classification: Stackoverflow, FewRel [Varshney et al., ACL'22]
  - etc

Da-Wei Zhou et al. *Deep Class-Incremental Learning: A Survey*. CoRR 2023.
Zixuan Ke, Bing Liu. *Continual Learning of Natural Language Processing Tasks: A Survey*. CoRR 2022.

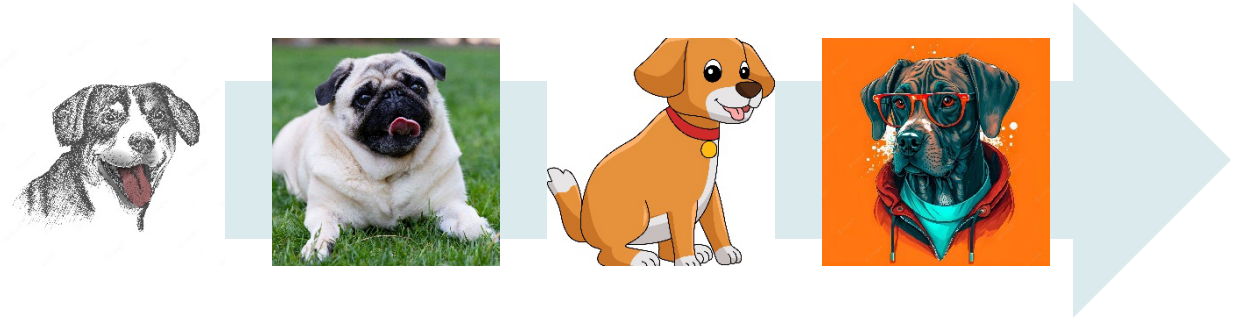
Data split in DIL

- Denote the total number of classes as C and the total number of tasks as N, each training task is combined of the C classes from a different domain:



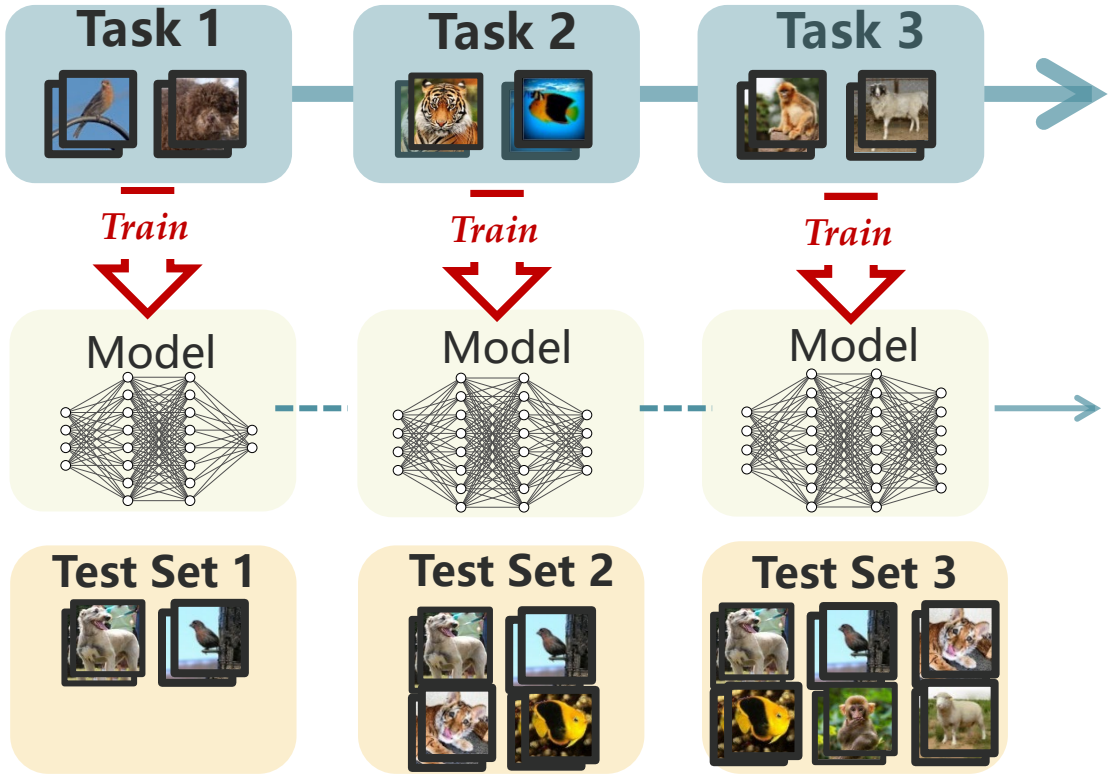
Benchmark datasets in DIL

- Image classification
  - MNIST-permutation [van de Ven et al., NMI'22]
  - CDDB [Wang et al., NeurIPS'22]
  - CORe50 [Wang et al., NeurIPS'22]
  - DomainNet [Wang et al., NeurIPS'22]
  - etc
- NLP
  - Paraphrase: Quora, Twitter and Wiki data [Li et al., NAACL'22]
  - Dialogue state tracking: TaskMaster, Schema Guided Dialogue, MultiWoZ [Madotto et al., EMNLP'21]
  - Question answering: AGNews, DBPedia, Yahoo [Wang et al., EMNLP'20]
  - etc



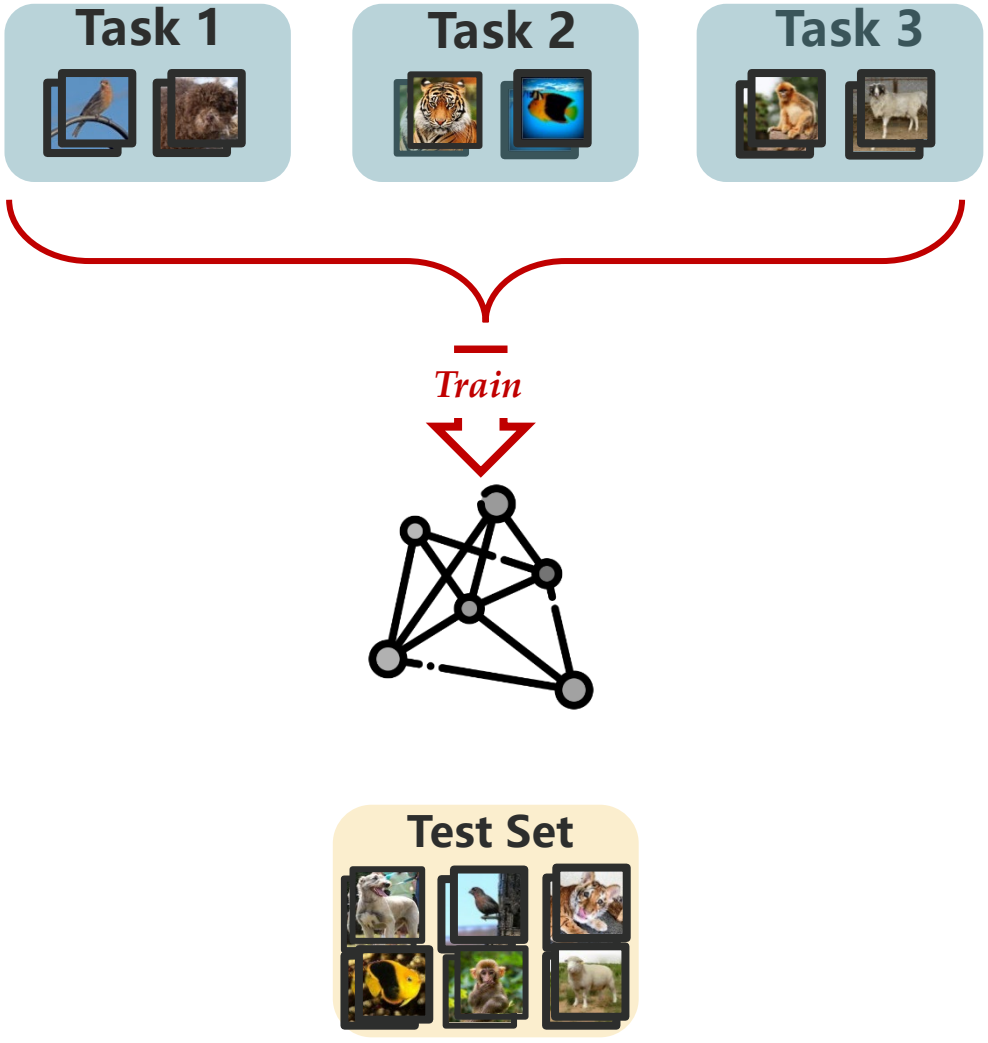
Da-Wei Zhou et al. *Deep Class-Incremental Learning: A Survey*. CoRR 2023.
Zixuan Ke, Bing Liu. *Continual Learning of Natural Language Processing Tasks: A Survey*. CoRR 2022.

Relationship with other topics



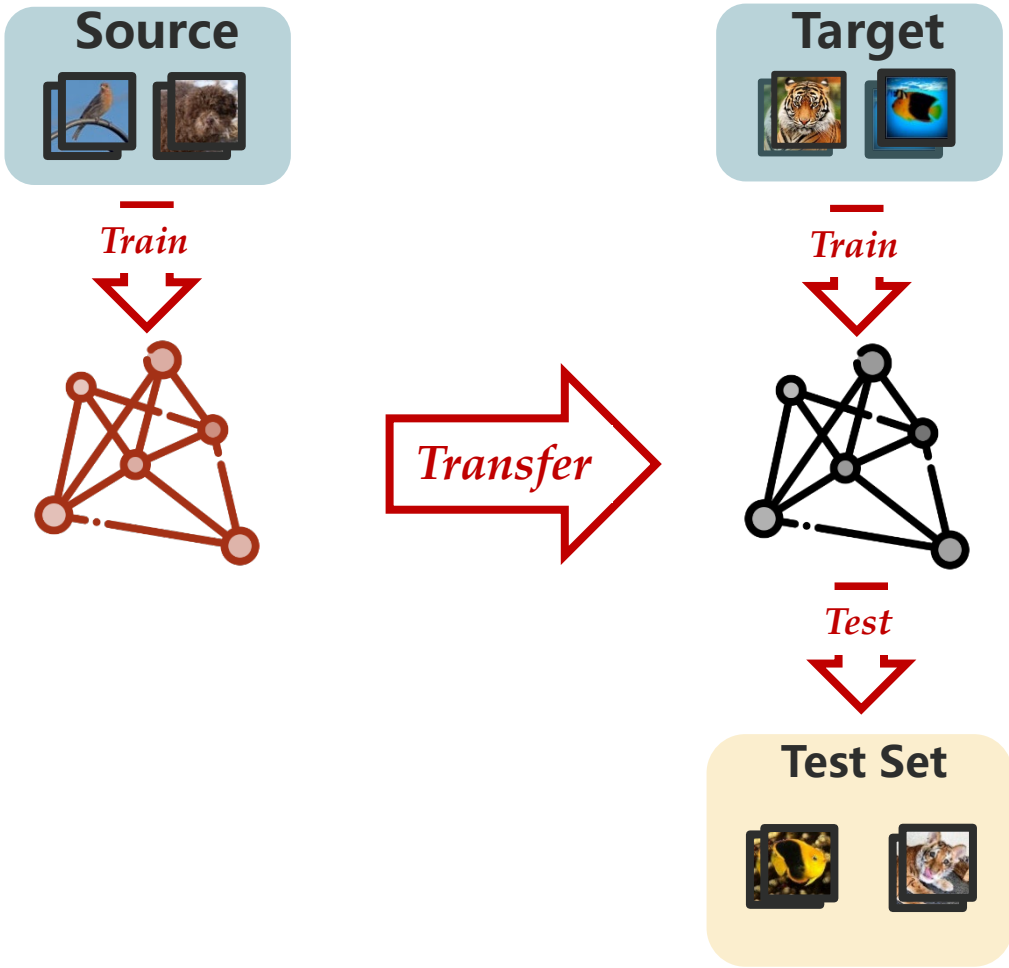
Multi-task Learning
Transfer Learning
Meta-Learning
…

Multi-task Learning



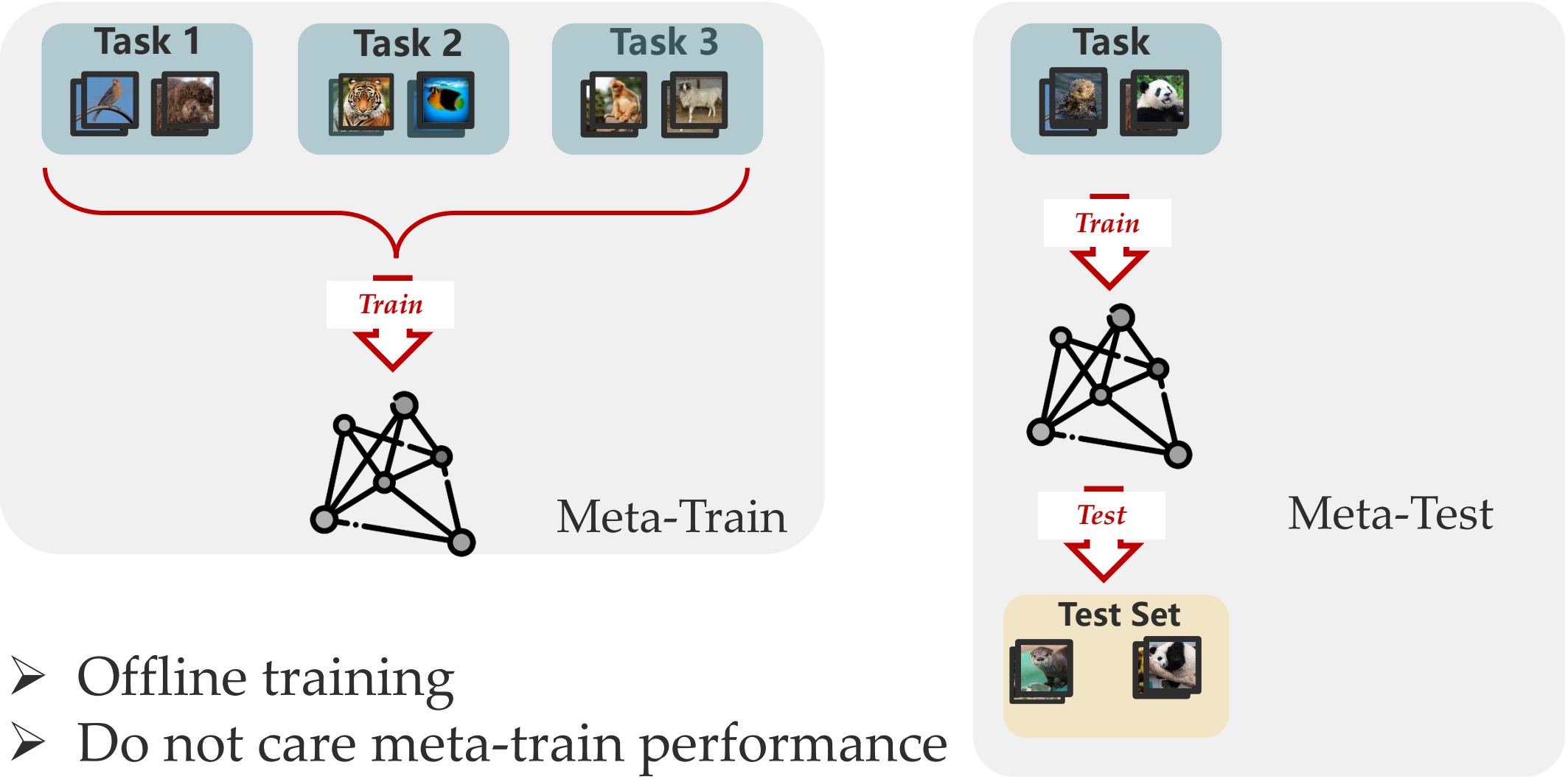
- Get all data at once
- Offline training

Transfer Learning



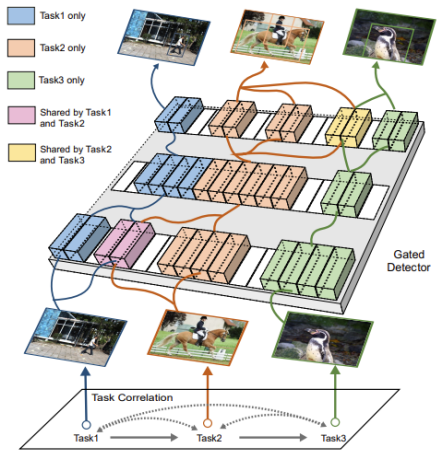
- Only two stages
- Do not care source performance

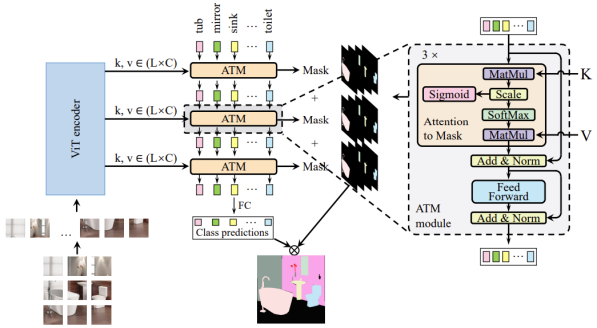
Meta-Learning

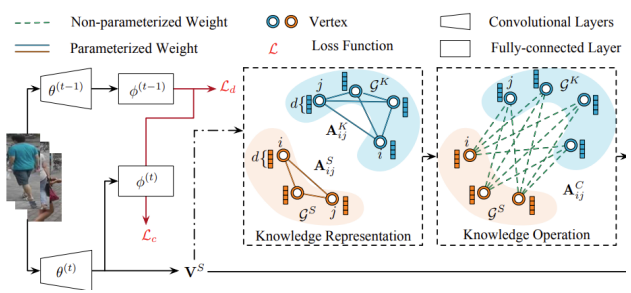


- Offline training
- Do not care meta-train performance

Applications of Continual Learning







Object Detection
[Yang et al. CVPR'22]

Semantic Segmentation
[Zhang et al. NeurIPS'22]

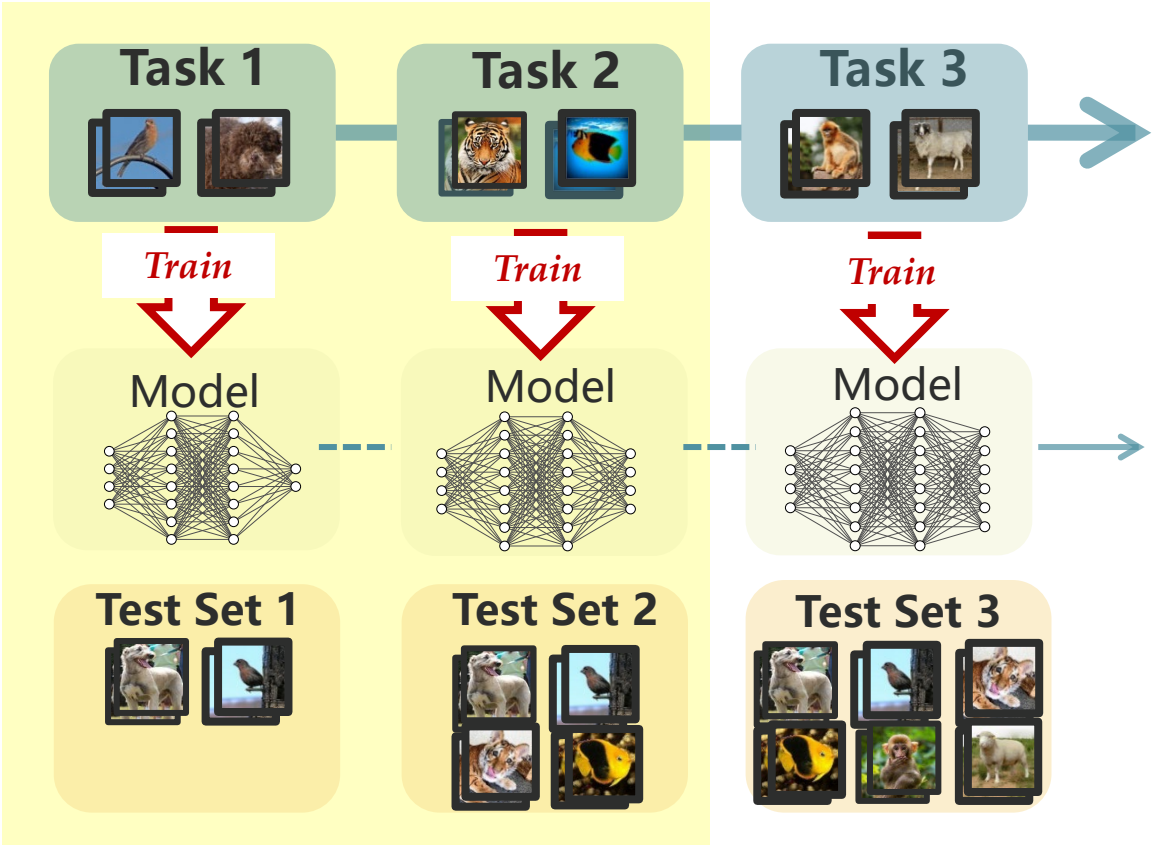
Re-ID
[Pu et al. CVPR'21]

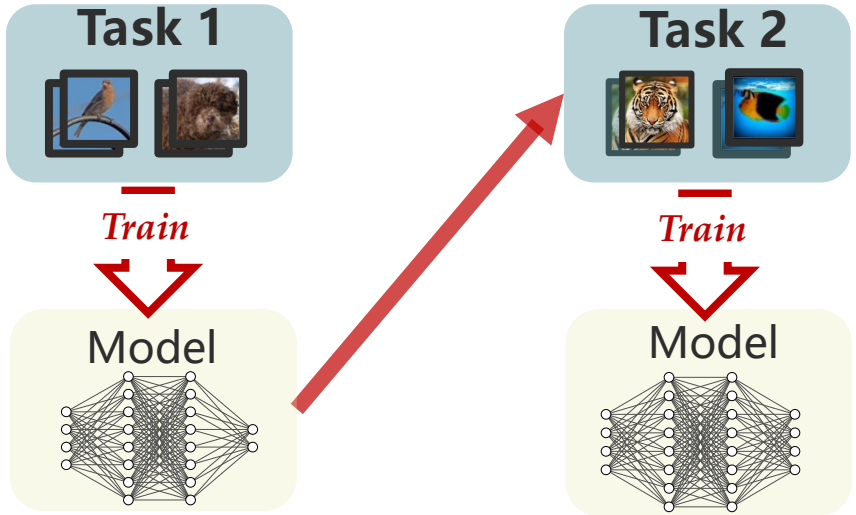
Diffusion
[Smith et al. CoRR'23]

Binbin Yang et al. *Continual Object Detection via Prototypical Task Correlation Guided Gating Mechanism*. CVPR 2022.
Bowen Zhang et al. *SegViT: Semantic Segmentation with Plain Vision Transformers*. NeurIPS 2022.
Nan Pu et al. *Lifelong Person Re-Identification via Adaptive Knowledge Accumulation*. CVPR 2021.
James Seale Smith et al. *Continual Diffusion: Continual Customization of Text-to-Image Diffusion with C-LoRA*. CoRR 2023.

Expected capabilities of continual learning

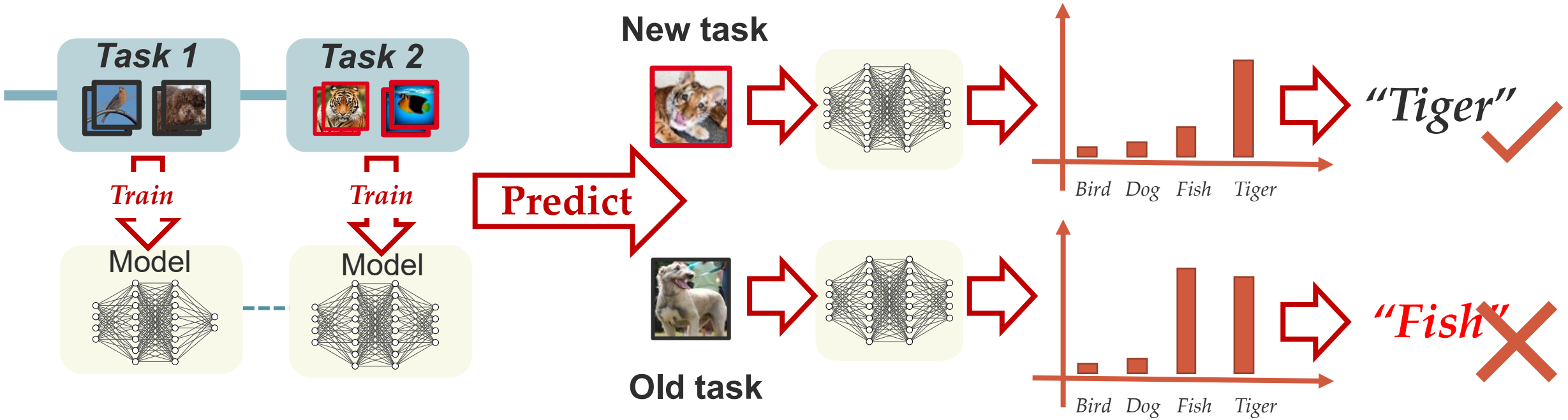
- Target: obtain the knowledge of all tasks seen so far





How about *finetuning* the current model with new data?

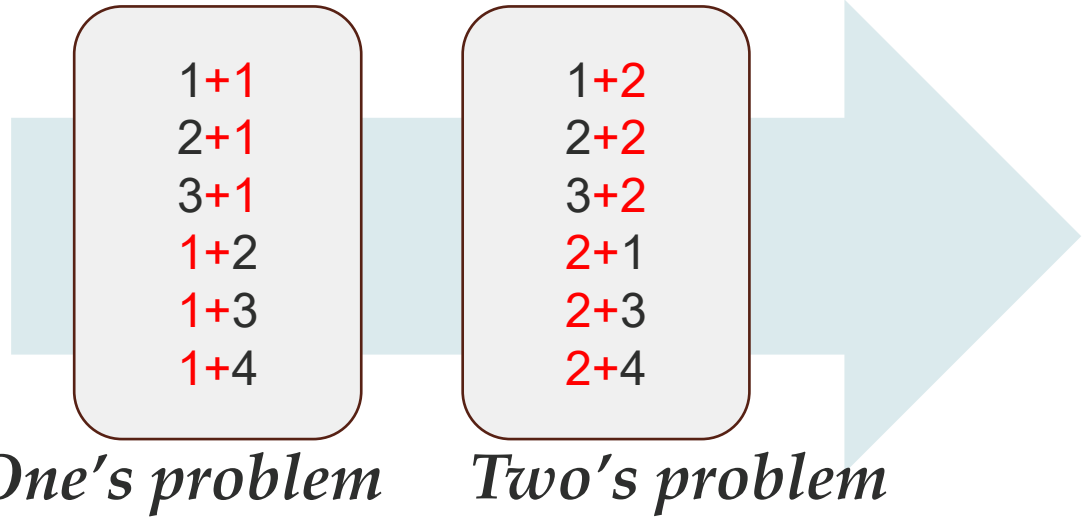
Catastrophic forgetting

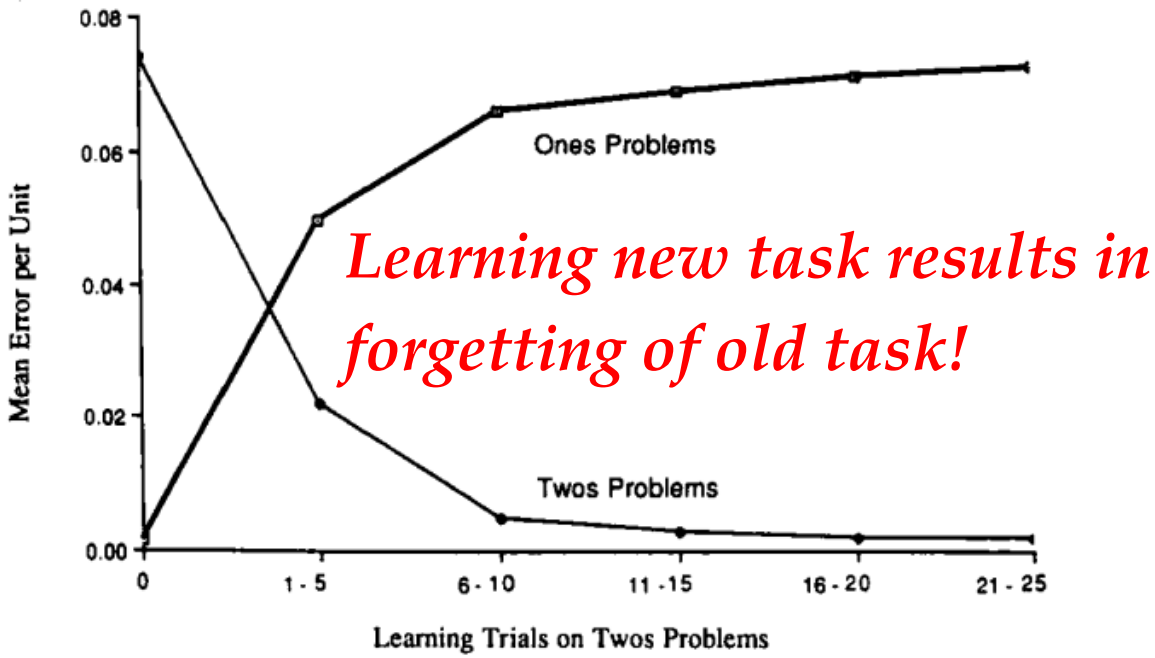


- Continual learning of new tasks will ***erase*** the semantic information of former tasks when learning new tasks.

Catastrophic forgetting

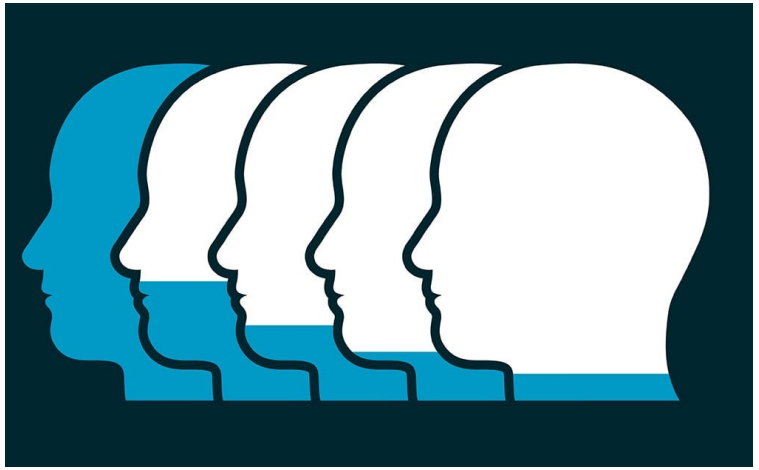
- "the process of learning a new set of patterns suddenly and completely *erased a network's knowledge of what it had already learned*" [French 1999]
- [McCloskey and Cohen, 1989] identifies catastrophic forgetting phenomena:



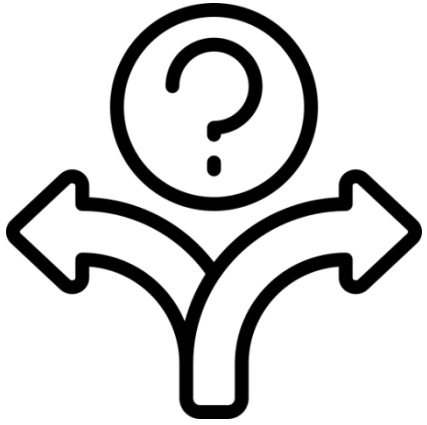


Robert M. French. *Catastrophic forgetting in connectionist networks*. Trends in Cognitive Sciences 1999.
Michael McCloskey, Neal J. Cohen. *Catastrophic interference in connectionist networks: The sequential learning problem.* Psychology of Learning and Motivation 1989.

Catastrophic forgetting



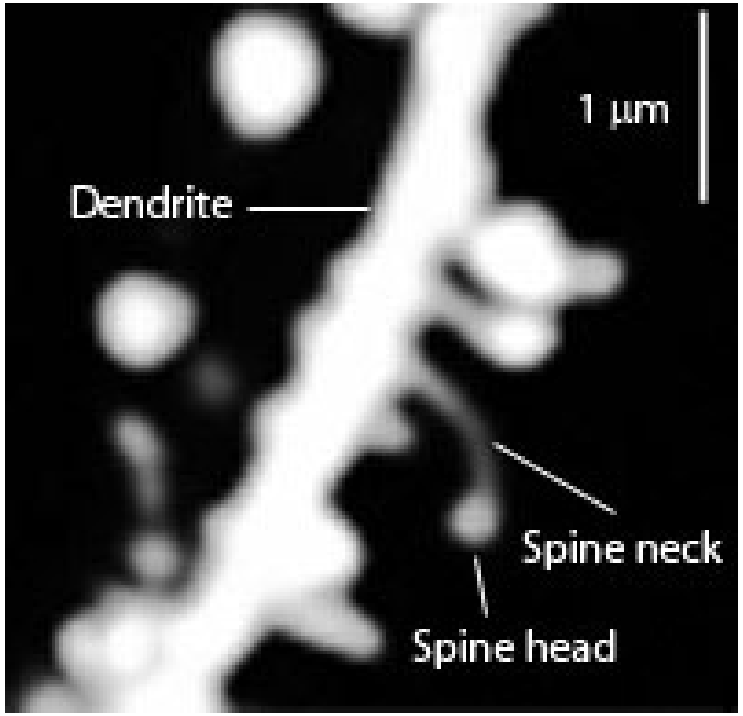
- In continual learning, the model is expected to *learn without forgetting*.
  - **Stability**: the ability to maintain old knowledge
  - **Plasticity**: the ability to learn new knowledge

Stability  **Plasticity**

Stephen Grossberg. *Studies of Mind and Brain Neural Principles of Learning, Perception, Development, Cognition, and Motor Control* 1982.
G A Carpenter, S Grossberg. *ART 2: self-organization of stable category recognition codes for analog input patterns.* Pattern Recognition by Self-Organizing Neural Networks 1987.

Catastrophic forgetting in Neuroscience

- Synaptic plasticity, the ability of ***neurons to modify their connections***, is involved in brain network remodeling following different types of brain damage.
- ***Dendritic spines*** are the major loci of synaptic plasticity and are considered as possible structural correlates of memory.



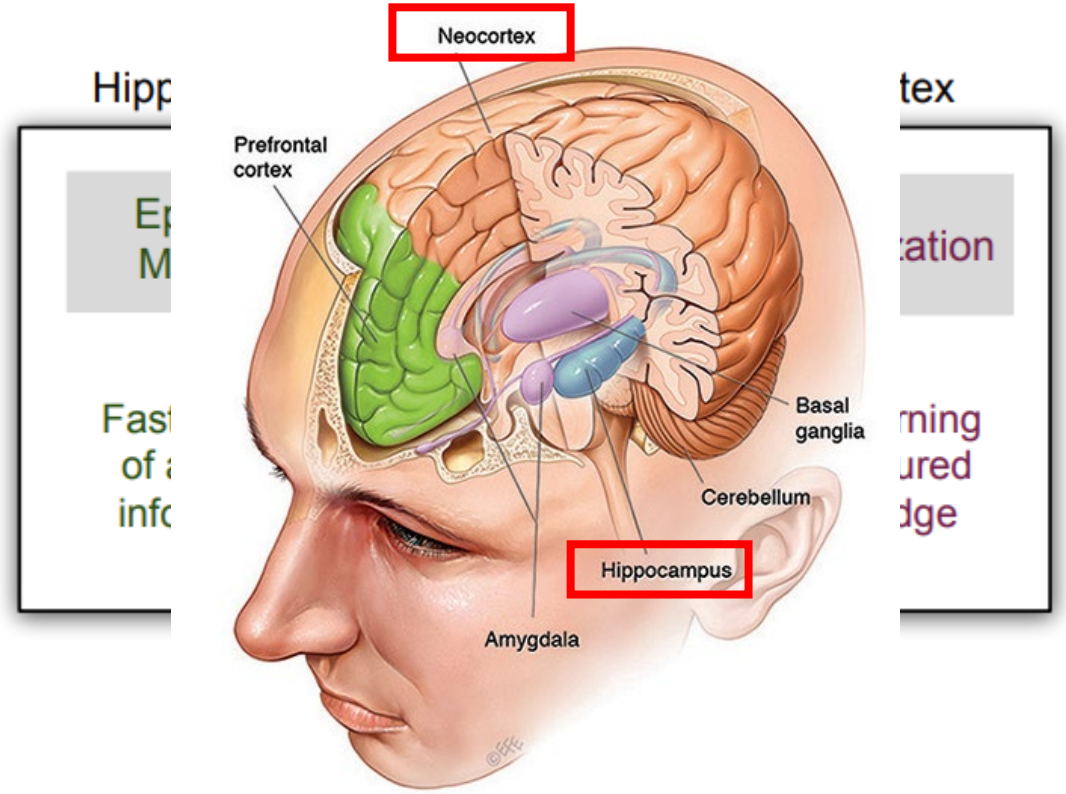
Observation: ***important parts in the network can be frozen or regularized to resist forgetting***

Mario Stampanoni Bassi et al. *Synaptic Plasticity Shapes Brain Connectivity: Implications for Network Topology.* International Journal of Molecular Sciences 2019.
Akiko Hayashi-Takagi et al. *Labelling and optical erasure of synaptic memory traces in the motor cortex.* Nature 2015.

Catastrophic forgetting in Neuroscience

- Complementary Learning Systems
  - ***Hippocampus*** for fast learning
  - ***Neocortex*** for slow learning

  - Hippocampus consolidates knowledge in neocortex by "replay"



Observation: *former knowledge can be recovered by experience replay*

German I. Parisi et al. *Continual Lifelong Learning with Neural Networks: A Review.* Neural Networks 2019.
James L McClelland et al. *Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory.* Psychological Review 1995.

Evaluation metrics in Continual Learning

- Splitting the data into T tasks, denote $a_{i,j}$ as the accuracy of task j after learning task i.

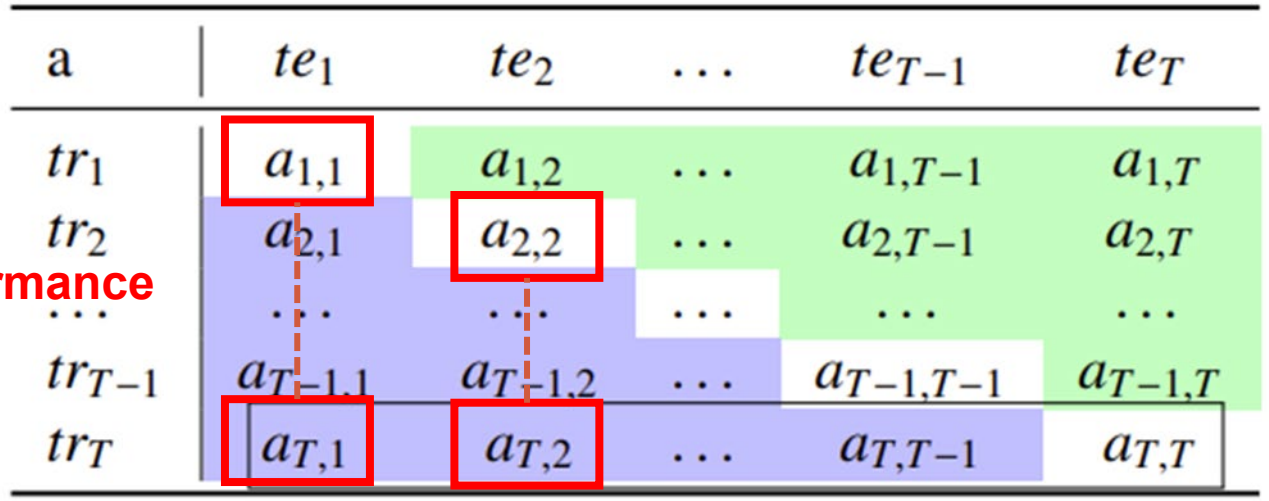
a	te_1	te_2	...	te_{T-1}	te_T
tr_1	$a_{1,1}$	$a_{1,2}$	...	$a_{1,T-1}$	$a_{1,T}$
tr_2	$a_{2,1}$	$a_{2,2}$	...	$a_{2,T-1}$	$a_{2,T}$
...	...	...	...	...	...
tr_{T-1}	$a_{T-1,1}$	$a_{T-1,2}$	...	$a_{T-1,T-1}$	$a_{T-1,T}$
tr_T	$a_{T,1}$	$a_{T,2}$	...	$a_{T,T-1}$	$a_{T,T}$

- Accuracy $\text{ACC} = \frac{1}{T}\sum_{i=1}^{T} a_{T,i}$ → average accuracy of all tasks after the last incremental stage

Arslan Chaudhry et al. *Riemannian walk for incremental learning: Understanding forgetting and intransigence*. ECCV 2018.

Evaluation metrics in Continual Learning

Gap between best performance and last performance

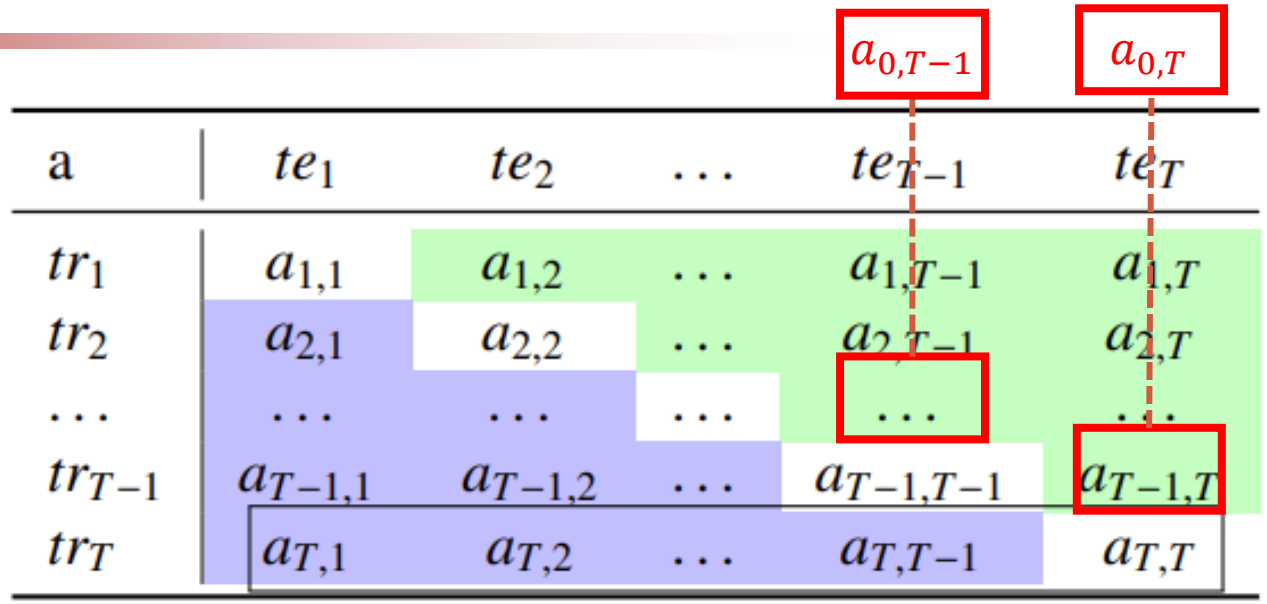


a	te_1	te_2	...	te_{T-1}	te_T
tr_1	$a_{1,1}$	$a_{1,2}$	...	$a_{1,T-1}$	$a_{1,T}$
tr_2	$a_{2,1}$	$a_{2,2}$	...	$a_{2,T-1}$	$a_{2,T}$
...	...	...	...	...	...
tr_{T-1}	$a_{T-1,1}$	$a_{T-1,2}$	...	$a_{T-1,T-1}$	$a_{T-1,T}$
tr_T	$a_{T,1}$	$a_{T,2}$	...	$a_{T,T-1}$	$a_{T,T}$

- Forgetting $= \frac{1}{T-1}\sum_{j=1}^{T-1} f_{T,j}, f_{k,j} = \max_{l\in\{1,\cdots,k-1\}}\left(a_{l,j}\right) - a_{k,j}, \forall j < k$
- Backward Transfer (BWT) $= \frac{1}{T-1}\sum_{j=1}^{T-1} a_{T,j} - a_{j,j}$
- BWT is (usually) the negative forgetting measure

Arslan Chaudhry et al. *Riemannian walk for incremental learning: Understanding forgetting and intransigence*. ECCV 2018.
David Lopez-Paz, Marc'Aurelio Ranzato. *Gradient episodic memory for continual learning*. NIPS 2017.

Evaluation metrics in Continual Learning



Randomly initialized model $a_{0,j}$

- Intransigence $\mathrm{IM_j} = a_j^* - a_{j,j}$ → Gap between offline training and CL
- Forward Transfer (FWT) $= \frac{1}{T-1}\sum_{j=2}^{T} a_{j-1,j} - a_{0,j}$ → The "zero-shot" ability gained from the continual learning process

Arslan Chaudhry et al. *Riemannian walk for incremental learning: Understanding forgetting and intransigence*. ECCV 2018.
David Lopez-Paz, Marc'Aurelio Ranzato. *Gradient episodic memory for continual learning*. NIPS 2017.

Evaluation metrics in CIL

a	te_1	te_2	...	te_{T-1}	te_T
tr_1	$a_{1,1}$	$a_{1,2}$	...	$a_{1,T-1}$	$a_{1,T}$
tr_2	$a_{2,1}$	$a_{2,2}$	...	$a_{2,T-1}$	$a_{2,T}$
...	...	...	...	...	...
tr_{T-1}	$a_{T-1,1}$	$a_{T-1,2}$	...	$a_{T-1,T-1}$	$a_{T-1,T}$
tr_T	$a_{T,1}$	$a_{T,2}$	...	$a_{T,T-1}$	$a_{T,T}$

- In Class-Incremental Learning, apart from the accuracy of the last stage, we also consider the <u>average accuracy along each stage</u>

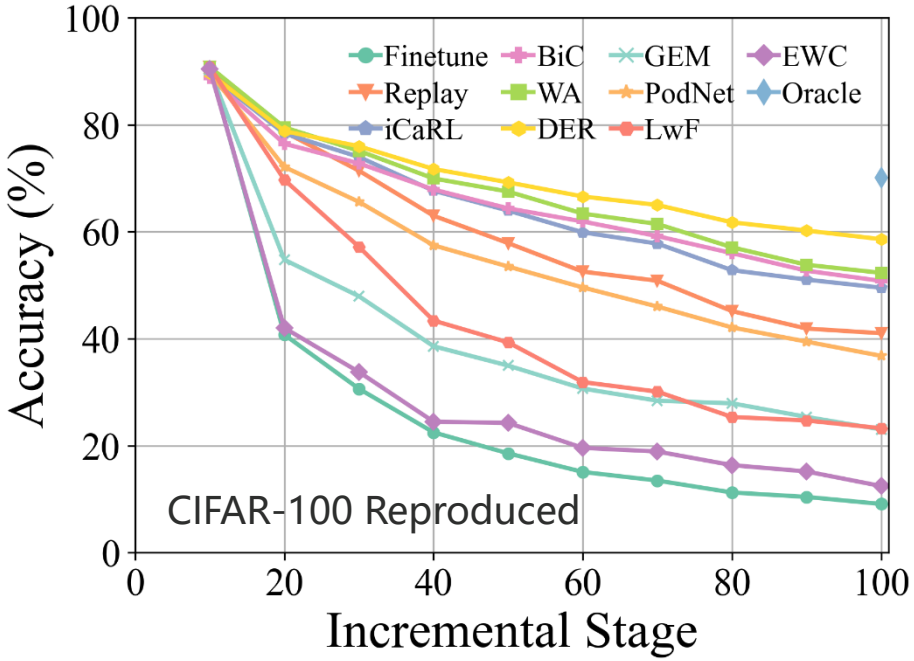
$$\bar{A} = \frac{1}{T}\sum_{i=1}^{T} ACC_i$$

Da-Wei Zhou et al. *Deep Class-Incremental Learning: A Survey*. CoRR 2023.

Resources for Continual Learning

PyCIL

https://github.com/G-U-N/PyCIL



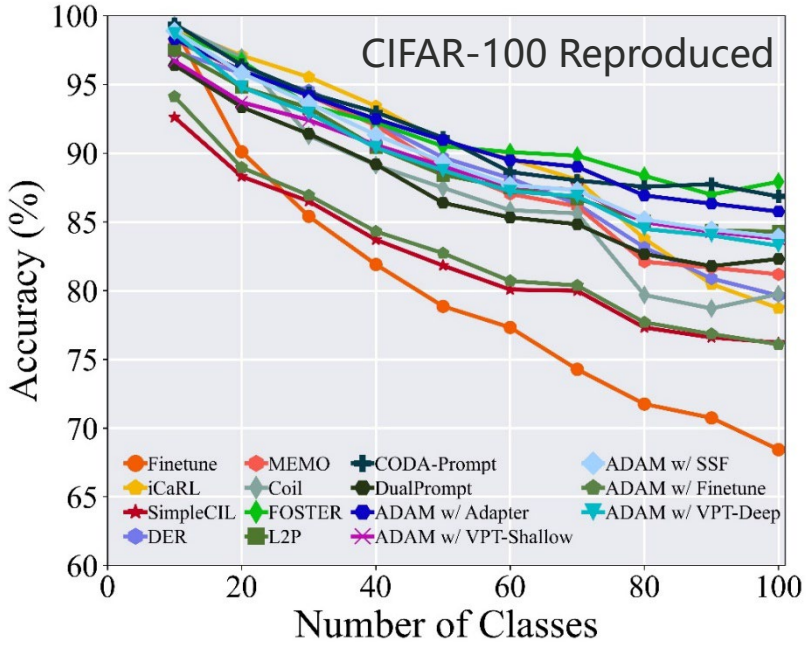
Comprehensive · Benchmark · Extendable · Maintained

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, De-Chuan Zhan. *PyCIL: A Python Toolbox for Class-Incremental Learning.* **SCIENCE CHINA Information Sciences** 2023.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, Ziwei Liu. *Deep class-incremental learning: A survey.* CoRR 2023.

Resources for Pre-trained Continual Learning

https://github.com/sun-hailong/LAMDA-PILOT



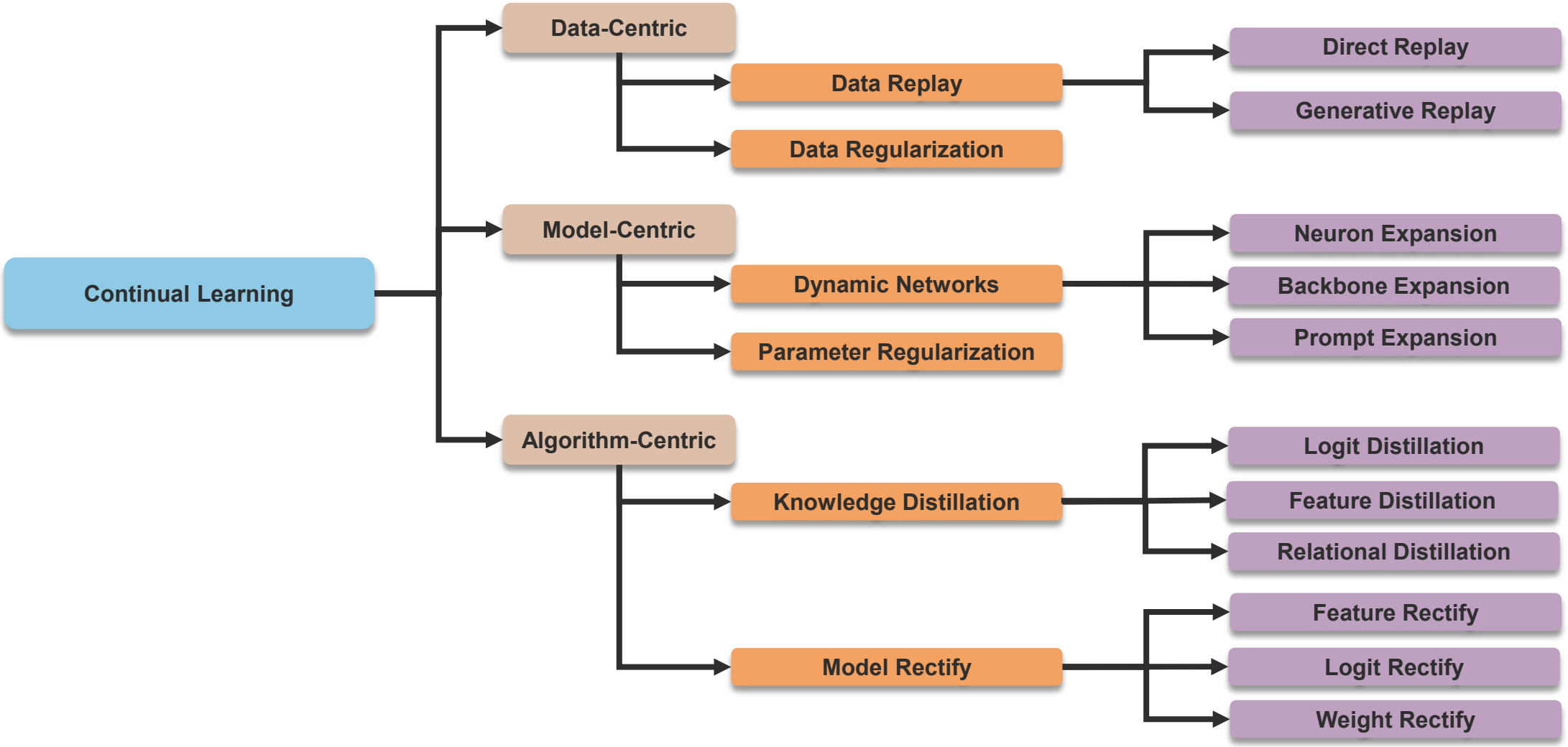
Comprehensive · Benchmark · Extendable · Maintained

Hai-Long Sun, Da-Wei Zhou, Han-Jia Ye, De-Chuan Zhan. *PILOT: A Pre-Trained Model-Based Continual Learning Toolbox*. CoRR 2023.

Outline

- Introduction (20 minutes)
  - Problem definition & history of continual learning
  - Evaluation metrics & benchmarks in continual learning
- Conventional Continual Learning (125 minutes)
  - Approaches for catastrophic forgetting (60 minutes)
  - Approaches for forward transfer (65 minutes)
- Continual Learning of Pre-trained Models (45 minutes)
  - Challenges and opportunities
  - Approaches
- Conclusion and Future Discussion (10 minutes)
- Q&A Session (10 minutes)

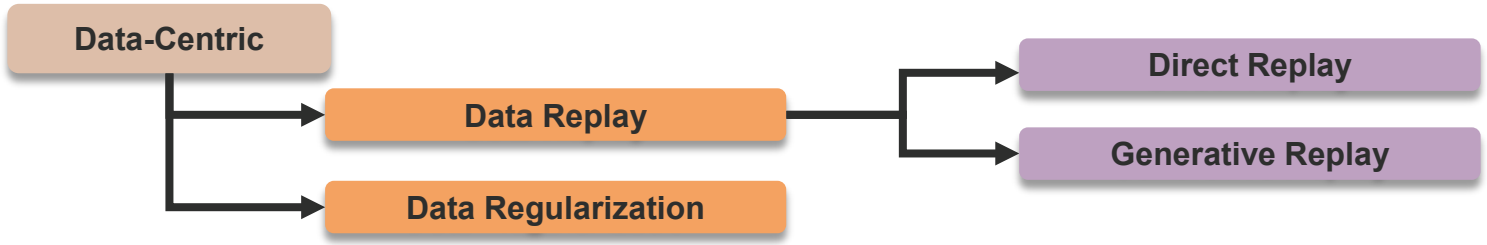
Taxonomy



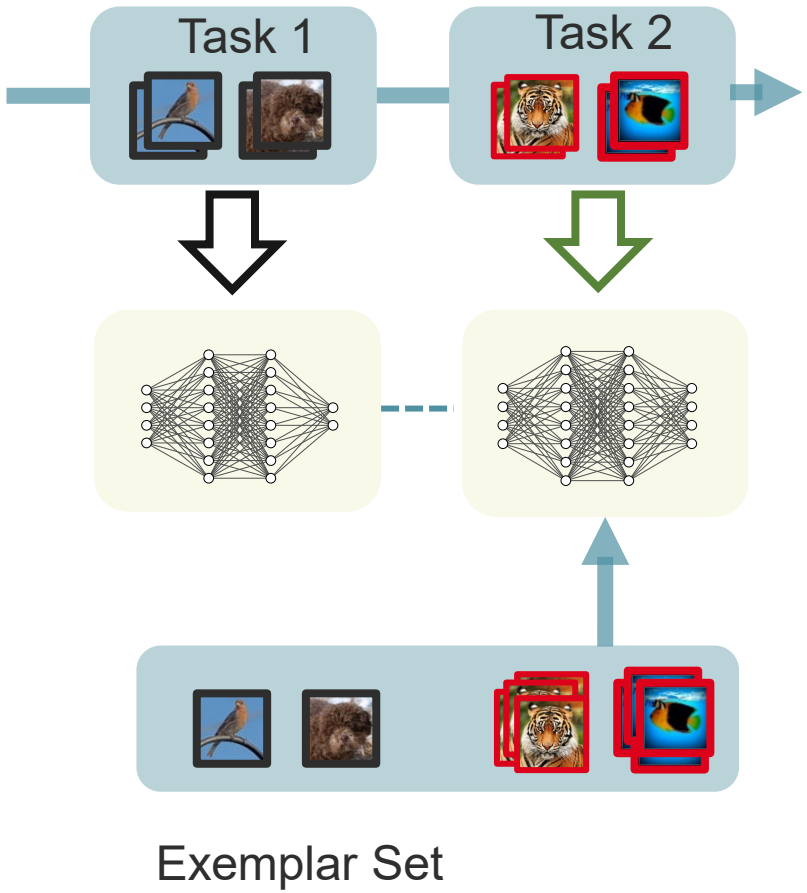
Da-Wei Zhou et al. *Deep Class-Incremental Learning: A Survey*. CoRR 2023.

Data-Centric Methods

- **Core Idea**: hosting the data to *replay* former knowledge when learning new, or *exert regularization terms* with former data
  - Data Replay: save a limited number of former instances for rehearsal when learning new
  - Data Regularization: regularize the optimization direction of the model with restriction of former instances



Direct replay



- Saving a ***fixed-size*** subset of "old data" to replay when learning new knowledge [Robins. Connection Science'95] [Rebuffi et al. CVPR'17]

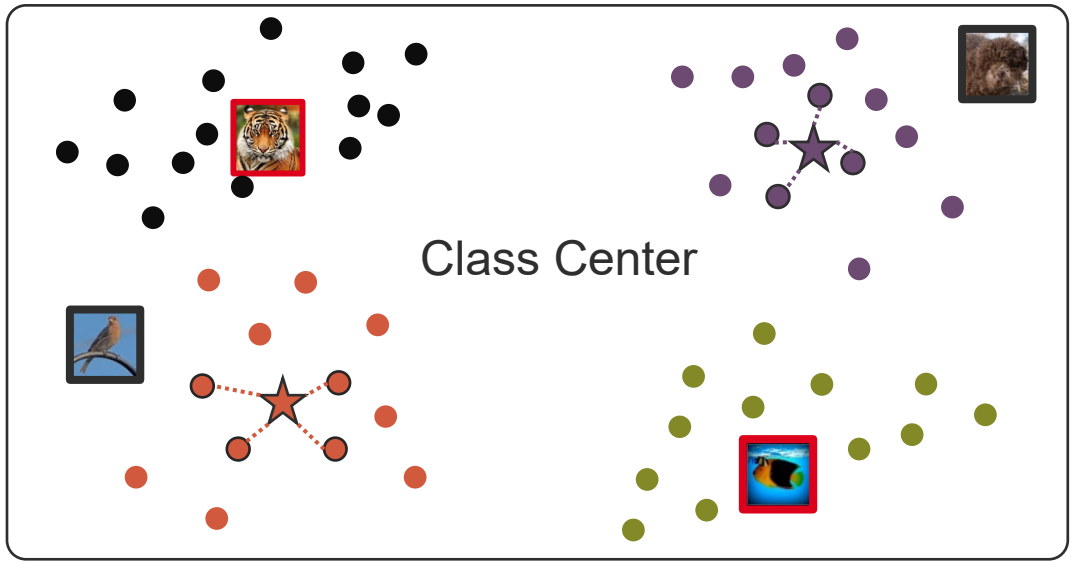
$$\min_{\theta} \sum_{x \in x_{new} \cup x_{old}} \ell(f(x), y)$$

- How to host the exemplar set?

Sylvestre-Alvise Rebuffi et al. *iCaRL: Incremental Classifier and Representation Learning*. CVPR 2017.

Hosting exemplar set

- How to choose the exemplars [Welling ICML'09]
- How to maintain a limited number of exemplars as data evolves



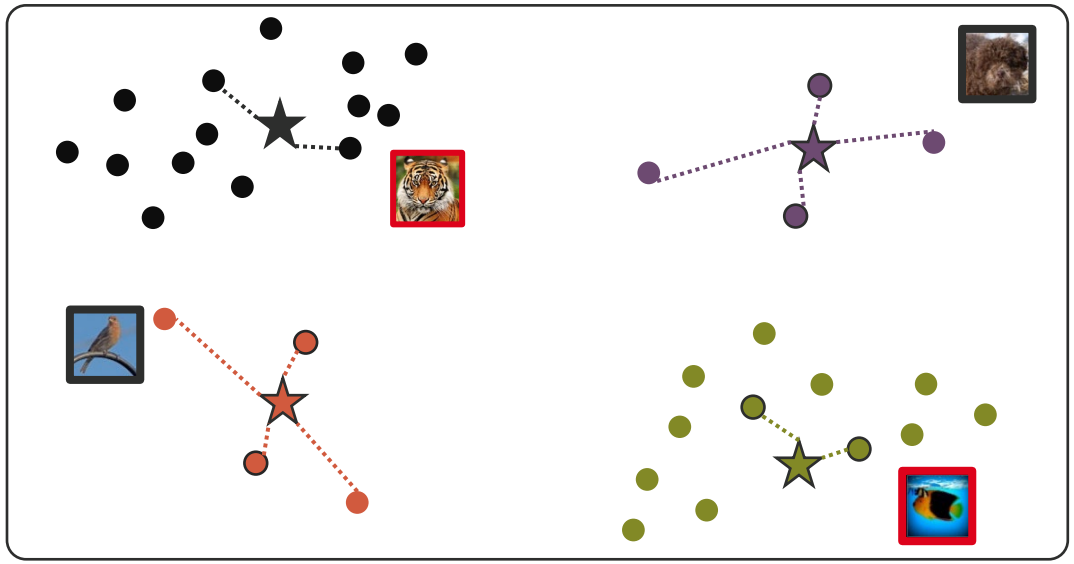
➢ Calculate the class center in the embedding space, and rank the instances via the distance to the center

$$\mu \leftarrow \frac{1}{n}\sum_{x \in X} \varphi(x)$$

Max Welling. *Herding dynamical weights to learn*. ICML 2009.
Sylvestre-Alvise Rebuffi et al. *iCaRL: Incremental Classifier and Representation Learning*. CVPR 2017.

Hosting exemplar set

- How to choose the exemplars [Welling ICML'09]
- How to maintain a ***limited number*** of exemplars as data evolves

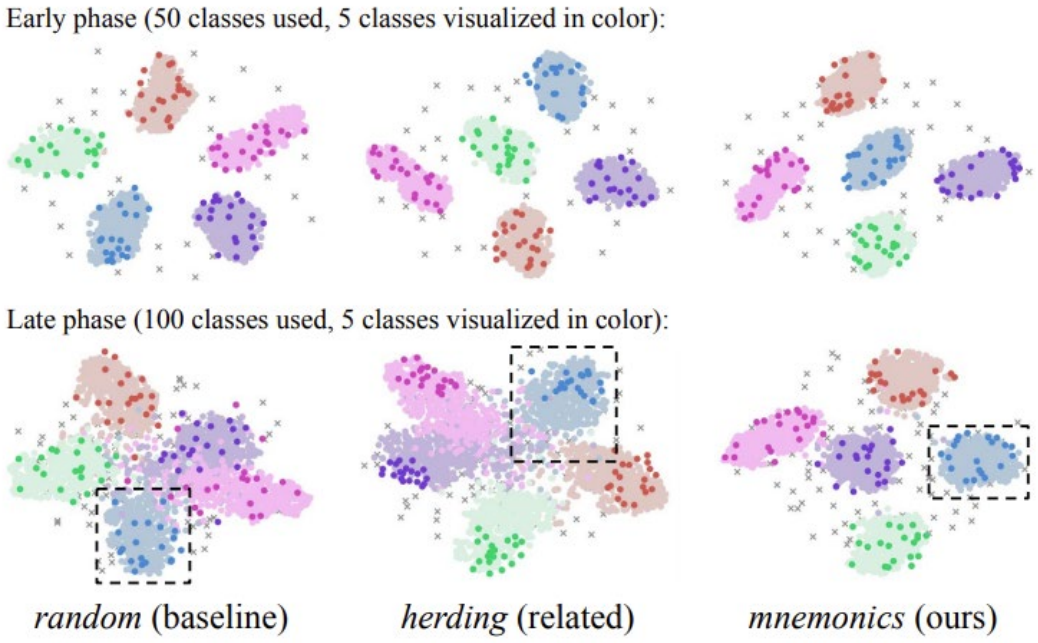


- ➢ Delete the old and include the new ones by the distance to the class center
- ➢ We keep 8 exemplars in total.
  In the first stage, we have 4 exemplars/class.
  In the second stage, we have 2 exemplars/class

Sylvestre-Alvise Rebuff et al. *iCaRL: Incremental Classifier and Representation Learning*. CVPR 2017.

Replay with optimizable exemplars

- Can we optimize the exemplar selection process?



- Formulating the selection process as bilevel optimization

$$\min_{\Theta_i} \mathcal{L}_c(\Theta_i; \mathcal{E}^*_{0:i-1} \cup D_i)$$
$$\text{s.t. } \mathcal{E}^*_{0:i-1} = \arg\min_{\mathcal{E}_{0:i-1}} \mathcal{L}_c\big(\Theta_{i-1}(\mathcal{E}_{0:i-1}); \mathcal{E}_{0:i-2} \cup D_{i-1}\big),$$

- Learned exemplars are situated near the decision boundary

Yaoyao Liu et al. *Mnemonics Training: Multi-Class Incremental Learning without Forgetting*. CVPR 2020.

Memory-Efficient Replay

- Exemplars are raw images; can we make the saving process memory-efficient?

A raw image costs 3*224*224 Bytes

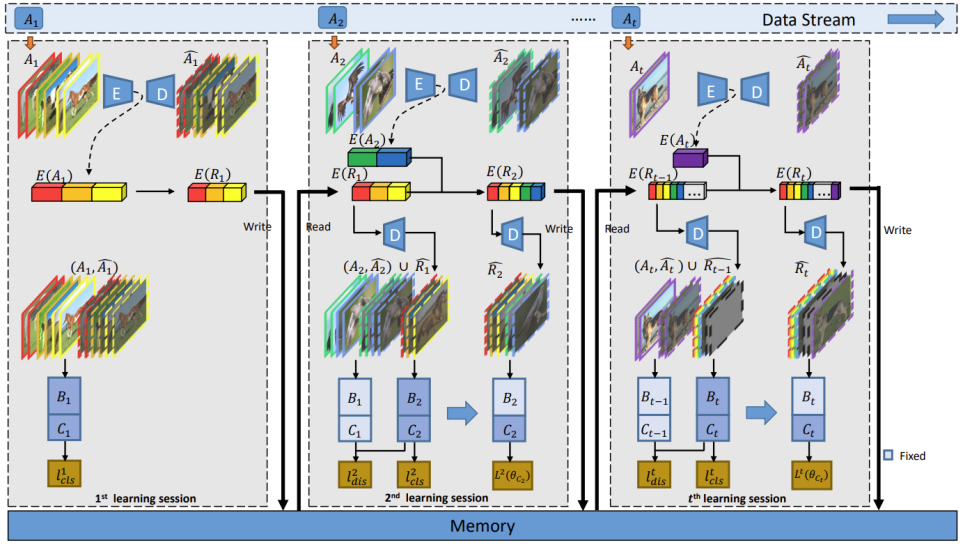
A low-resolution image costs 3*32*32 Bytes

An extracted feature costs 512 Bytes

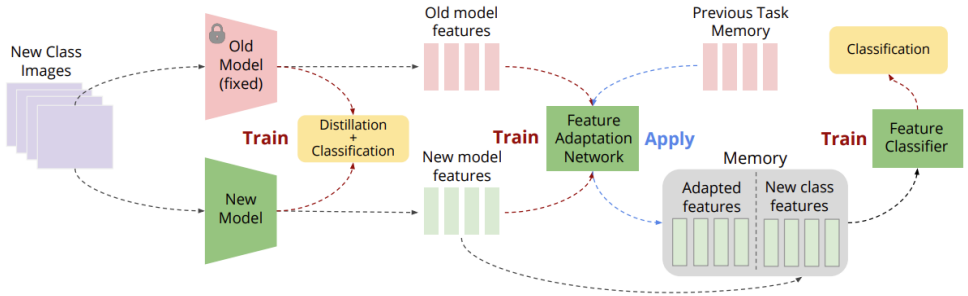
- However, corresponding learning algorithms need to be designed to overcome the ***distribution gap***

Memory-Efficient Replay

- Learn an extra encoder-decoder to map the low-resolution images into raw-images



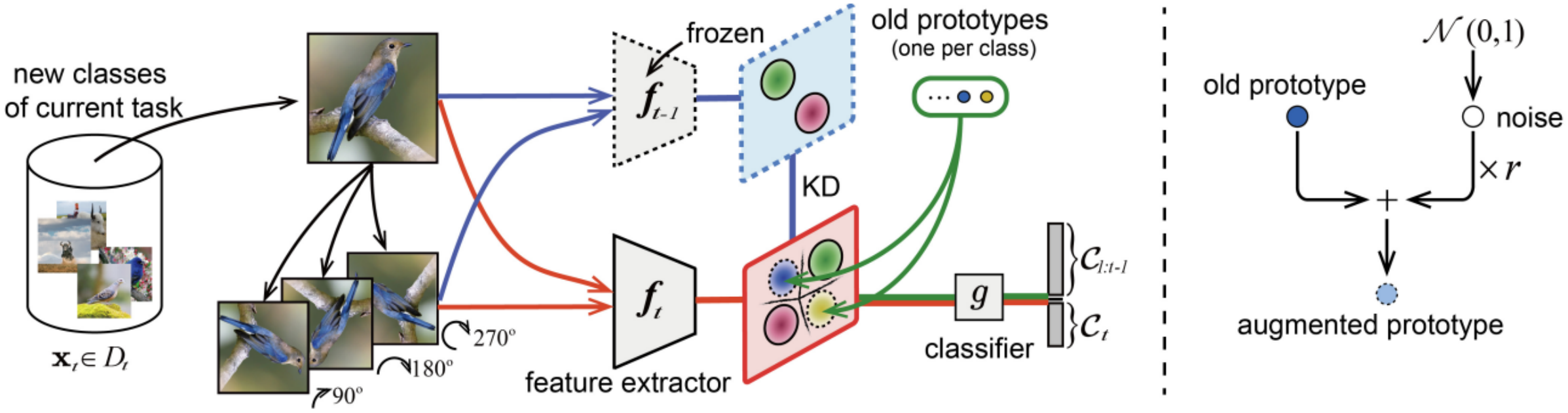
- Learn an extra adaptation lay to map old features into new features



Hanbin Zhao et al. *Memory Efficient Class-Incremental Learning for Image Classification*. TNNLS 2021.
Ahmet Iscen et al. *Memory-Efficient Incremental Learning Through Feature Adaptation*. ECCV 2020.

Feature Replay

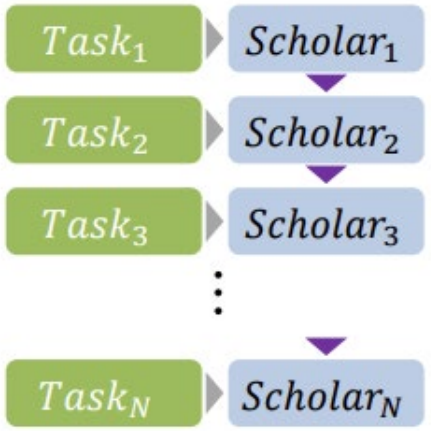
- Modelling the distribution of former classes as Gaussian distribution



Fei Zhu et al. *Prototype augmentation and self-supervision for incremental learning*. CVPR 2021.

Generative Replay

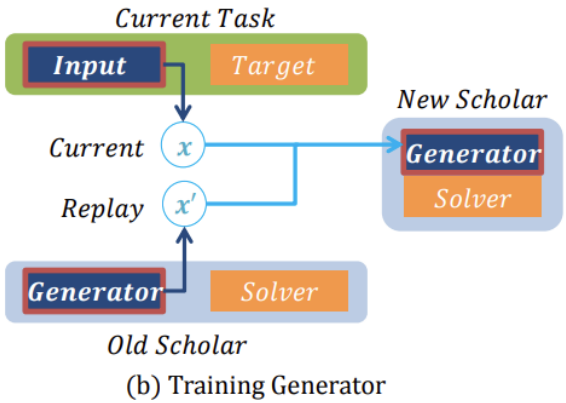
- Can we utilize generative models to learn the distributions of old classes, so that we have inexhaustible exemplars



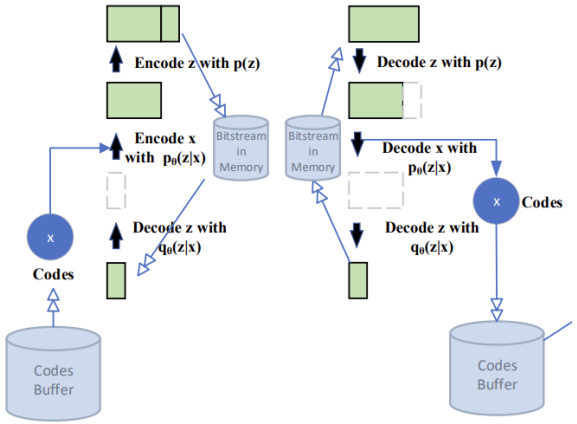
- The model (Scholar) is combined of a **generator G** and a **classification model C**
- In each new task, utilize the generated old instances and current task instances to train a new G and new C

Generative Replay

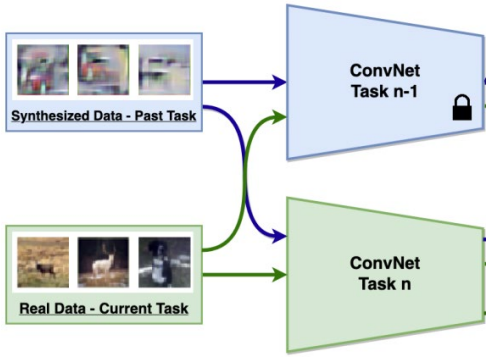
- We can combine various kinds of generative models…



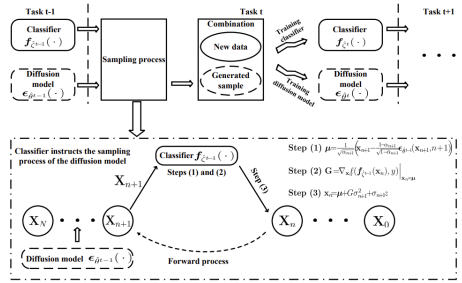
GAN
[Shin et al. NIPS'17]



VQ-VAE
[Jiang et al. CVPRW'21]



Deep Inversion
[Smith et al. ICCV'21]



Diffusion
[Gao et al. ICML'23]

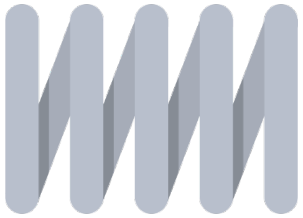
Hanul Shin et al. *Continual learning with deep generative replay*. NIPS 2017.
Jian Jiang et al. *IB-DRR - Incremental Learning with Information-Back Discrete Representation Replay.* CVPRW 2021.
James Smith et al. *Always Be Dreaming: A New Approach for Data-Free Class-Incremental Learning.* ICCV 2021.
Rui Gao et al. *DDGR: Continual Learning with Deep Diffusion-based Generative Replay*. ICML 2023.

Data Regularization

- Data replay directly rehearsals the instances during training, aiming to lower the loss of both old and new instances.
- Can we achieve this goal from another perspective?

$$\begin{aligned} &\text{minimize}_{\theta} \quad \ell(f_\theta(x), y) \\ &s.t. \quad \ell(f_\theta, \mathcal{M}_k) \leq \ell\left(f_\theta^{t-1}, \mathcal{M}_k\right) \ \ for\ all\ k < t \end{aligned}$$

Training loss of the current task
Obtaining the new knowledge



Maintaining the loss of former tasks
Remembering the old knowledge

David Lopez-Paz, Marc'Aurelio Ranzato. *Gradient episodic memory for continual learning*. NIPS 2017.

Data Regularization

- Ensuring the loss of former tasks not increase

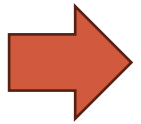
$$\text{minimize}_{\theta} \quad \ell(f_\theta(x), y)$$
$$s.t. \quad \ell(f_\theta, \mathcal{M}_k) \leq \ell\left(f_\theta^{t-1}, \mathcal{M}_k\right) \ for\ all\ k < t$$

- Making the angle between losses to be acute

$$\langle g, g_k \rangle := \left\langle \frac{\partial \ell(f_\theta(x), y)}{\partial \theta}, \frac{\partial \ell(f_\theta, \mathcal{M}_k)}{\partial \theta} \right\rangle \geq 0, for\ all\ k < t$$

- Projecting the gradient if violated

$$\text{minimize}_{\tilde{g}} \quad \frac{1}{2} \| g - \tilde{g} \|_2^2$$
$$s.t. \quad \langle \tilde{g}, g_k \rangle \geq 0, \quad for\ all\ k < t$$



QP problem

David Lopez-Paz, Marc'Aurelio Ranzato. *Gradient episodic memory for continual learning*. NIPS 2017.

Data Regularization

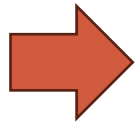
- Relaxing t-1 restrictions

$$\text{minimize}_{\theta} \quad \ell(f_\theta(x), y)$$
$$s.t. \quad \ell(f_\theta, \mathcal{M}_k) \le \ell\left(f_\theta^{t-1}, \mathcal{M}_k\right) \; for\ all\ k < t$$

$$\text{minimize}_{\theta} \quad \ell(f_\theta(x), y)$$
$$s.t. \quad \ell(f_\theta, \mathcal{M}) \le \ell\left(f_\theta^{t-1}, \mathcal{M}\right) \qquad \mathcal{M} = \cup_{k<t} \mathcal{M}_k$$

- A random batch in exemplar set as the regularization is enough

$$\text{minimize}_{\tilde{g}} \frac{1}{2} \| g - \tilde{g} \|_2^2, \qquad \tilde{g}^\top g_{\text{ref}} \ge 0$$

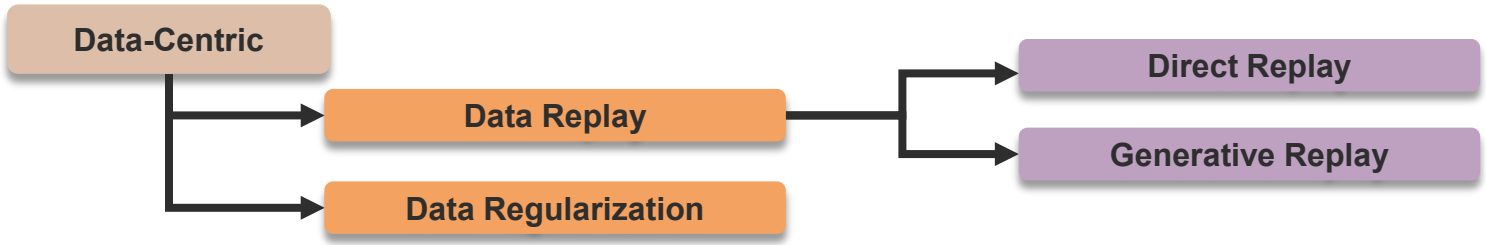


No need for QP

Arslan Chaudhry et al. *Efficient Lifelong Learning with A-GEM.* ICLR 2019.

Summary of Data-Centric Methods

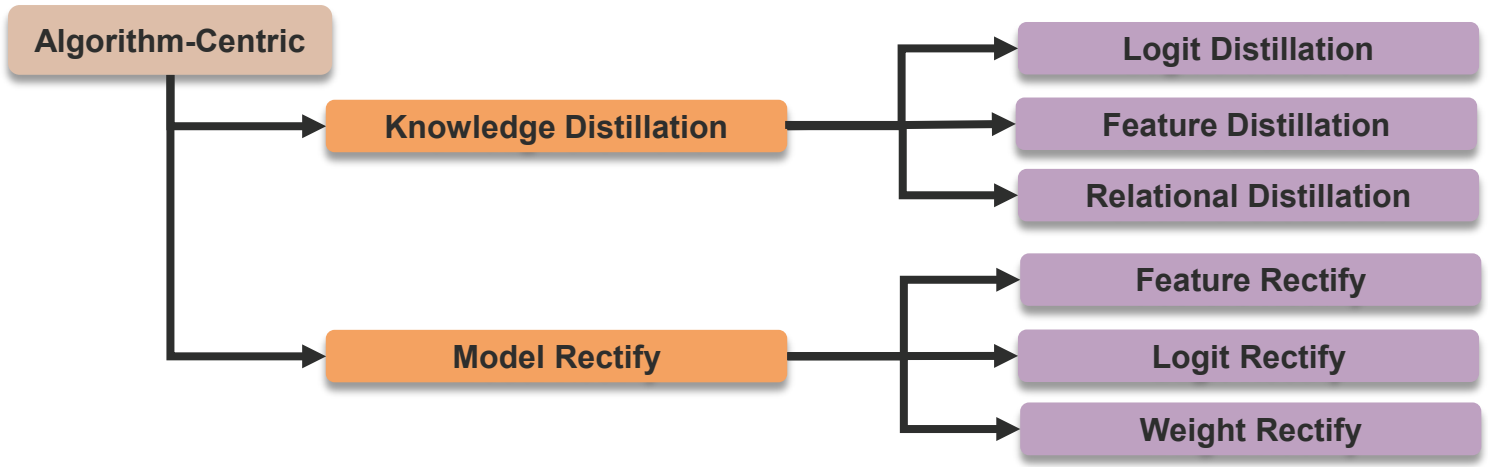
- Data replay is ***simple yet effective***!
- Data replay shall encounter ***overfitting*** [Verwimp et al. ICCV'21]
- The performance of generative replay is restricted by generative models
- Generative models also suffer from catastrophic forgetting [Wu et al. NeurIPS'18]
- Saving exemplars may encounter privacy or storage issues



Eli Verwimp et al. *Rehearsal revealed: The limits and merits of revisiting samples in continual learning*. ICCV 2021.
Chenshen Wu et al. *Memory replay gans: learning to generate images from new categories without forgetting*. NeurIPS 2018.

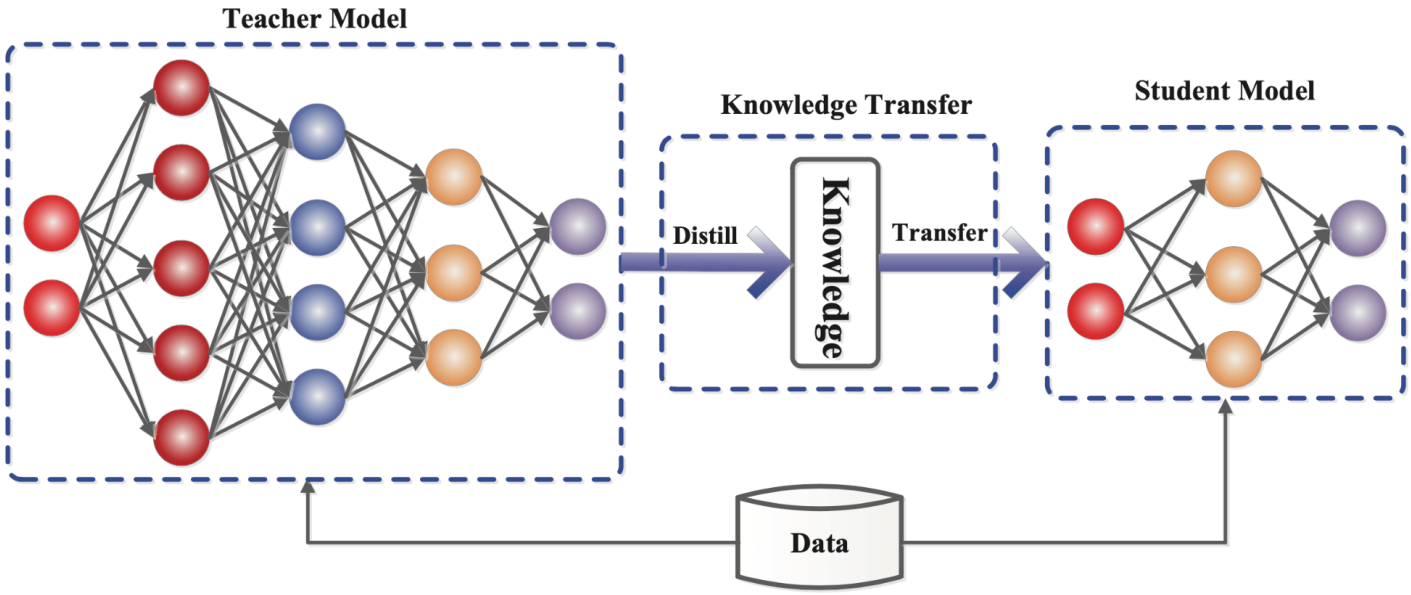
Algorithm-Centric Methods

- **Core Idea**: design training mechanisms to prevent the forgetting of old model
  - Knowledge Distillation: build the mapping between old and new model to reflect the semantic information of old classes in the updated model
  - Model Rectify: rectify the inductive bias of the incremental model to reflect the universal classifier



Knowledge Distillation in CL

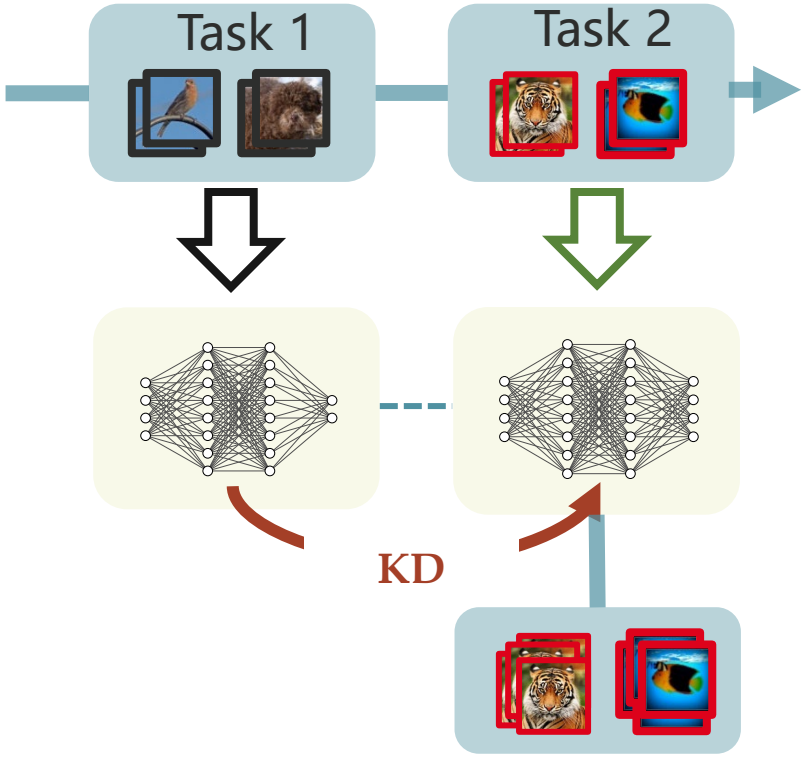
- Knowledge distillation [Hinton CoRR'15] is proposed to transfer the knowledge between the teacher model to the student
- Can we build the "teaching" process to resist forgetting?



Geoffrey Hinton et al. *Distilling the knowledge in a neural network.* CoRR 2015.

Logit Distillation

- To resist forgetting in the learning process, we can treat the ***model after the previous stage as the teacher*** to teach the current model.
- To reflect the former knowledge in the updated model



➢ Logit Distillation [Li et al. TPAMI'17]

$$\min_{\theta_{new}} \ell + \sum_{x_{new}} SIM\left(f_{\theta_{old}}(\boldsymbol{x}_{new}), f_{\theta_{new}}(\boldsymbol{x}_{new})\right)$$

"Learning new"

"Remembering old"

Zhizhong Li et al. *Learning without forgetting.* ECCV 2016.

Logit Distillation

- Exemplars can be utilized in the distillation process [Rebuffi et al. CVPR'17]

$$\min_{\theta_{new}} \ell + \sum_{x \in x_{new} \cup x_{old}} SIM\left(f_{\theta_{old}}(x), f_{\theta_{new}}(x)\right)$$

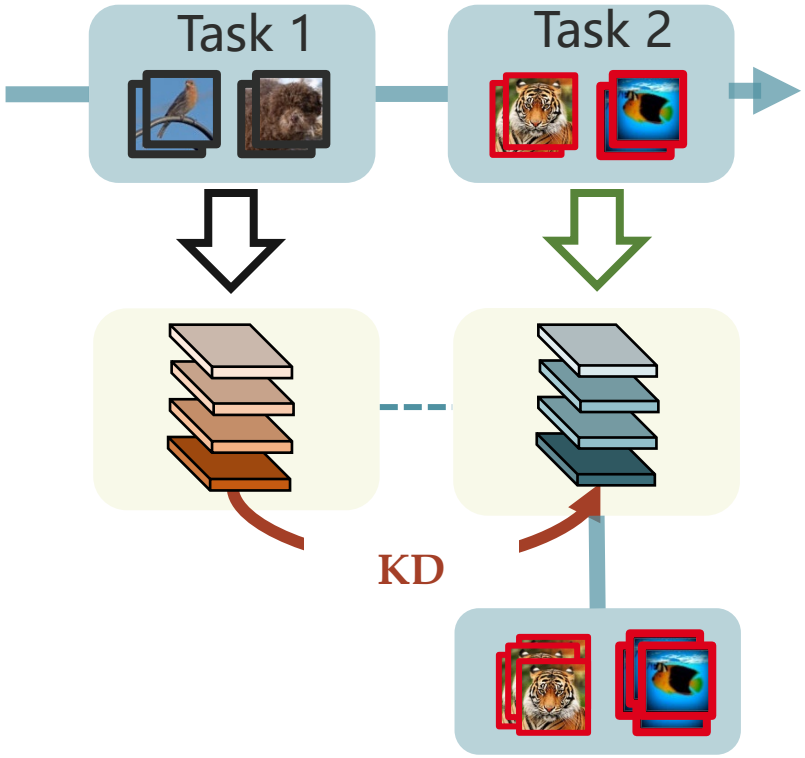
- The target of learning new knowledge can also be formulated by distilling from an expert model [Hou et al. ECCV'18]

$$\min_{\theta_{new}} \sum_{x \in x_{new}} SIM\left(f_{\theta_{expert}}(x), f_{\theta_{new}}(x)\right) + \sum_{x \in x_{new} \cup x_{old}} SIM\left(f_{\theta_{old}}(x), f_{\theta_{new}}(x)\right)$$

Sylvestre-Alvise Rebuffi et al. *icarl: Incremental classifier and representation learning.* CVPR 2017.
Saihui Hou et al. *Lifelong learning via progressive distillation and retrospection.* ECCV 2018.

Feature Distillation

- Apart from distilling the logits produced by the FC layer, we can also build the mapping after the embedding module



➢ Feature Distillation [Hou et al. CVPR'19]

$$\min_{\theta_{new}} \ell + \sum_{x_{new}} SIM\left(\phi_{\theta_{old}}(x_{new}), \phi_{\theta_{new}}(x_{new})\right)$$

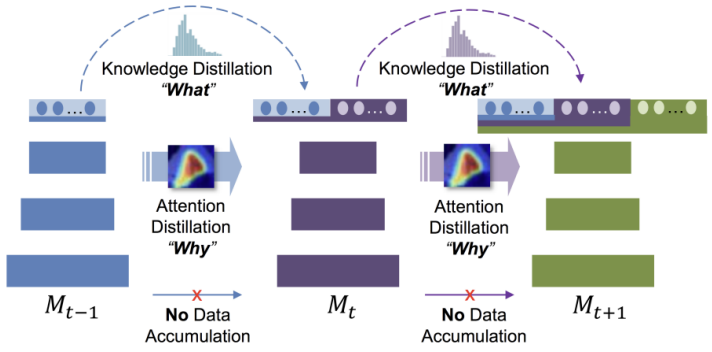
Forcing the updated model to produce the same features as the old one

Saihui Hou et al. *Learning a unified classifier incrementally via rebalancing.* CVPR 2019.

Feature Distillation

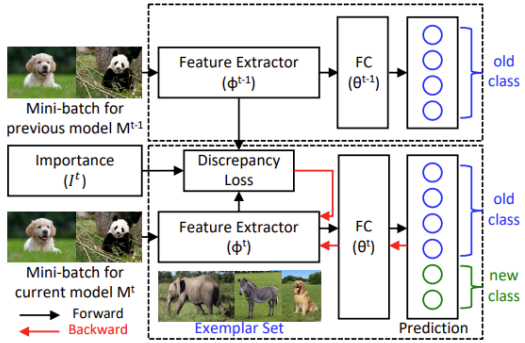
- Other feature products can also be distilled

$$\min_{\theta_{new}} \ell + \sum_{x_{new}} SIM\left(g(\phi_{\theta_{old}}(x_{new})), g(\phi_{\theta_{new}}(x_{new}))\right)$$



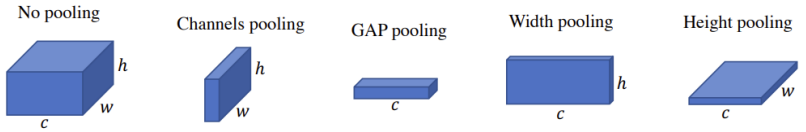
Attention Map

[Dhar et al. CVPR'19]



Weighted Feature Map

[Kang et al. CVPR'22]



Pooled features

[Douillard et al. ECCV'20]

Prithviraj Dhar et al. *Learning without memorizing.*. CVPR 2019.

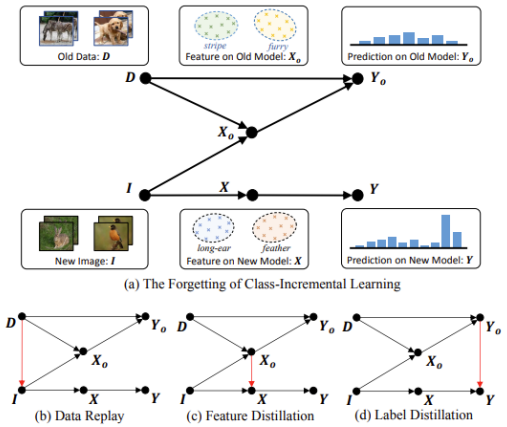
Minsoo Kang et al. *Class-incremental learning by knowledge distillation with adaptive feature consolidation.* CVPR 2022.

Arthur Douillard et al. *Podnet: Pooled outputs distillation for small-tasks incremental learning*. ECCV 2020.

Feature Distillation

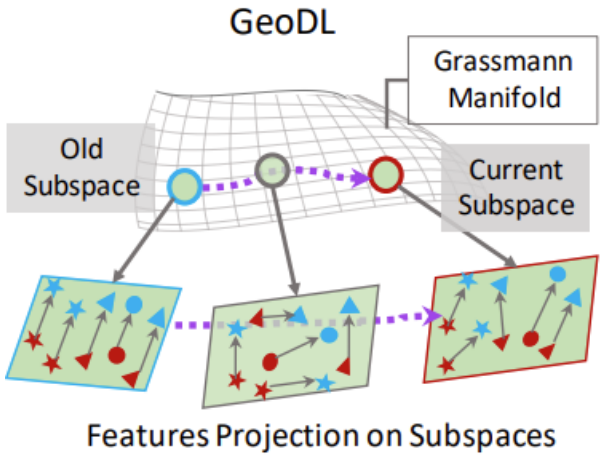
- Other feature products can also be distilled

$$\min_{\theta_{new}} \ell + \sum_{x_{new}} SIM\left(g(\phi_{\boldsymbol{\theta}_{old}}(\boldsymbol{x}_{new})), g(\phi_{\boldsymbol{\theta}_{new}}(\boldsymbol{x}_{new}))\right)$$



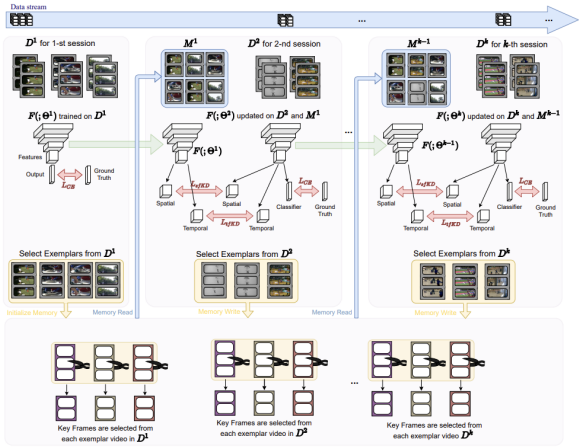
Causal effect

[Hu et al. CVPR'21]



Subspace feature

[Simon et al. CVPR'21]



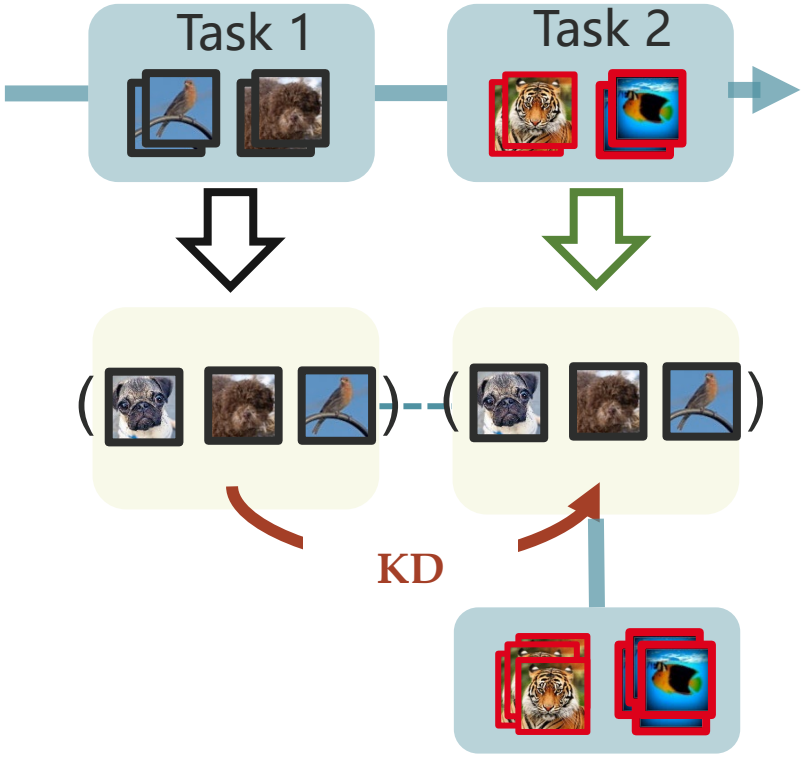
Spatial & Temporal feature

[Zhao et al. MM'21]

Xinting Hu et al. *Distilling causal effect of data in class-incremental learning.* CVPR 2021.
Christian Simon et al. *On learning the geodesic path for incremental learning.* CVPR 2021.
Hanbin Zhao et al. *When video classification meets incremental classes*. MM 2021.

Relational Distillation

- Apart from the instance-wise mapping, we can also build the mapping between **a group of instances (G)** among different models



➢ Relational Distillation [Gao et al. ECCV'22] [Dong et al. AAAI'21]

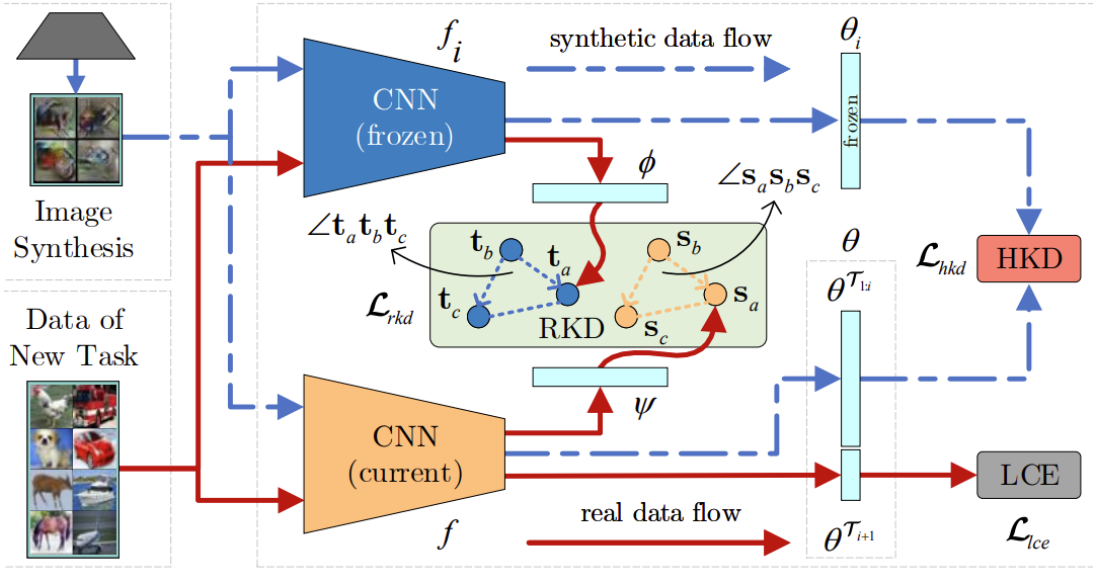
$$\min_{\theta_{new}} \ell + \sum_{\boldsymbol{G} \in \boldsymbol{x_{new}} \cup \boldsymbol{x_{old}}} SIM\left(f_{\boldsymbol{\theta}_{old}}(\boldsymbol{G}), f_{\boldsymbol{\theta}_{new}}(\boldsymbol{G})\right)$$

Forcing the updated model to **produce the same relationship among groups** as the old model

Songlin Dong et al. *Few-shot class-incremental learning via relation knowledge distillation.* AAAI 2021.
Qiankun Gao et al. *R-DFCIL: relation-guided representation learning for data-free class incremental learning.* ECCV 2022.

Relational Distillation

- Extract triplets as the instance group (x_i, x_j, x_k)
- How to reflect the relationship among the instance group?



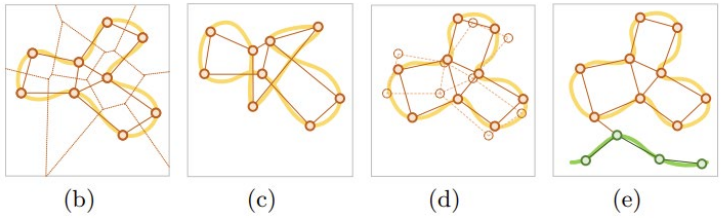
➢ Matching the angle [Gao et al. ECCV'22] [Dong et al. AAAI'21]

$$\min_{\theta_{new}} \ell + \sum_{G \in x_{new} \cup x_{old}} SIM\left(\angle_{old}(ijk), \angle_{new}(ijk)\right)$$

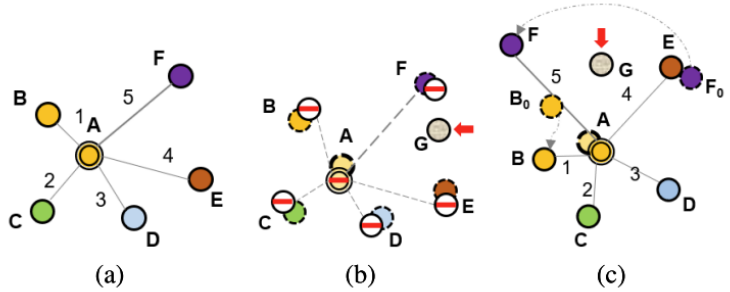
Songlin Dong et al. *Few-shot class-incremental learning via relation knowledge distillation.* AAAI 2021.
Qiankun Gao et al. *R-DFCIL: relation-guided representation learning for data-free class incremental learning.* ECCV 2022.

Relational Distillation

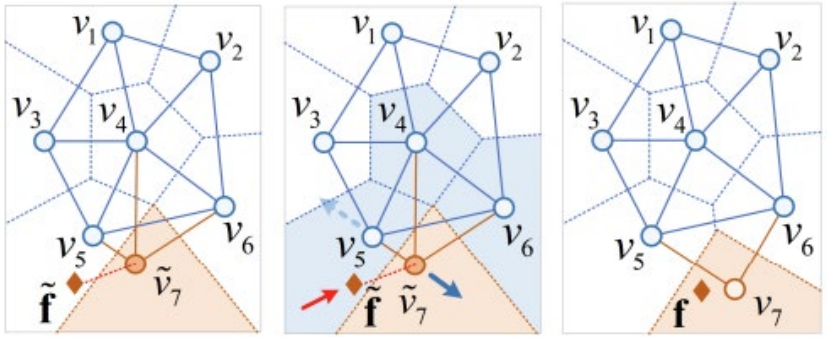
- Other ways to build the relationship?
- We can build the local graph as the instance group



Hebbian graph
[Tao et al. ECCV'20]



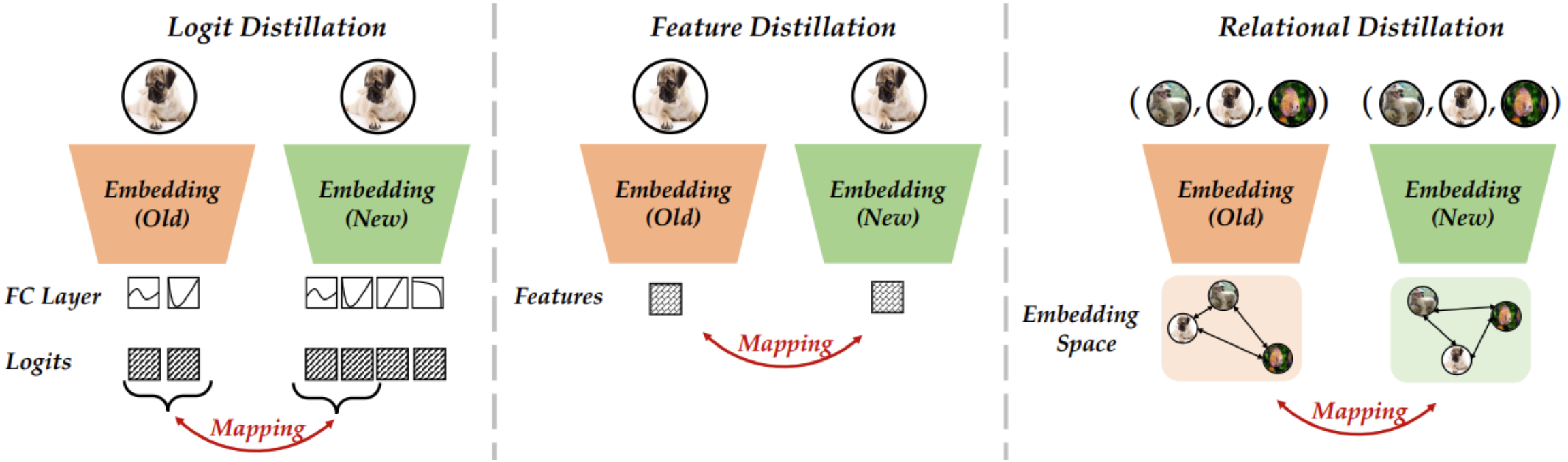
Local neighborhood
[Liu et al. TNNLS'22]



Neuron gas
[Tao et al. CVPR'20]

Xiaoyu Tao et al. *Topology-preserving class-incremental learning*. ECCV 2020.
Yu Liu et al. *Model behavior preserving for class-incremental learning*. TNNLS 2022.
Xiaoyu Tao et al. *Few-shot class-incremental learning*. CVPR 2020.

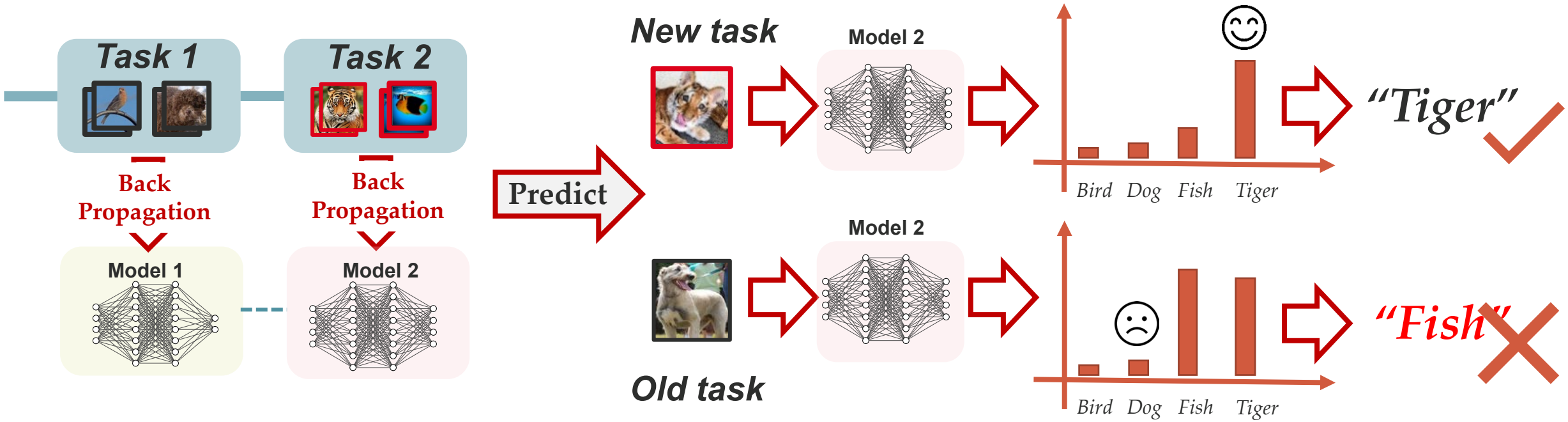
Knowledge distillation in CL



Forcing the updated model to produce the same *"behaviour"* as the old one

Model Rectification

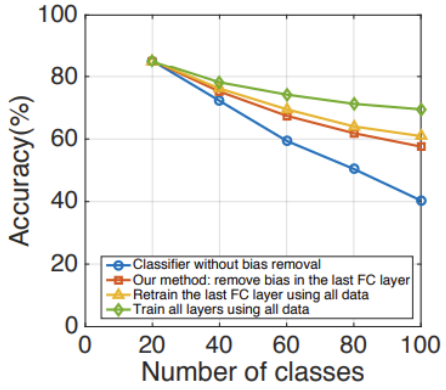
- Rectify the *"inductive bias"* of the incremental model

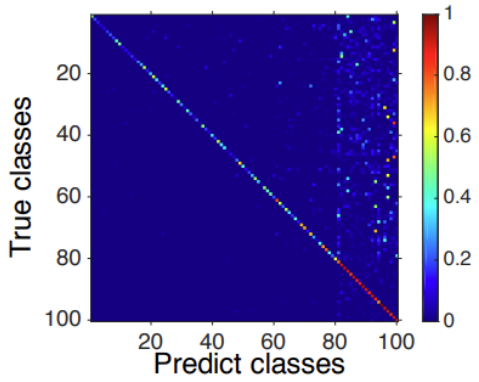


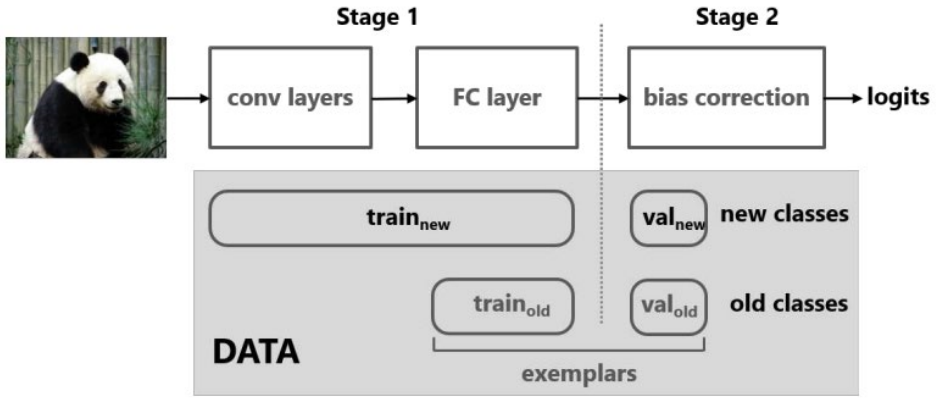
Learning new tasks will erase the discriminability of former tasks.
Model tends to predict all instances ***as the newly learned task.***

Logit Rectify

- Biased FC layers result in the imbalanced prediction
- We can ***learn to rescale*** the logits with a balanced validation set!







➢ Logit Rectify [Wu et al. CVPR'19] [Belouadah et al. ICCV'19] [Castro et al. ECCV'18]

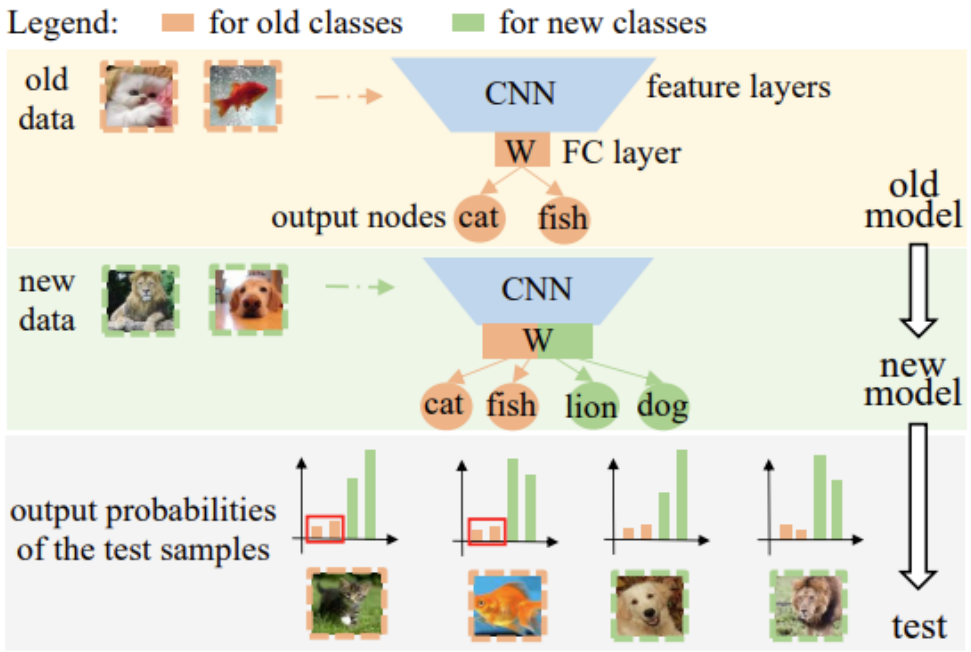
$$\hat{f}(x)_k = \begin{cases} \alpha f(x)_k + \beta, & k \in Y_b \\ f(x)_k, & otherwise \end{cases}$$

Decrease the logits of new classes for an ***unbiased prediction***

Yue Wu et al. *Large scale incremental learning.* CVPR 2019.
Eden Belouadah et al. *Il2m: Class incremental learning with dual memory.* ICCV 2019.
Francisco M Castro et al. *End-to-end incremental learning.* ECCV 2018.

Weight Rectify

- Biased FC layers result in the imbalanced prediction
- We can directly ***normalize the FC layers of new classes***!



➢ Weight Rectify [Zhao et al. CVPR'20]

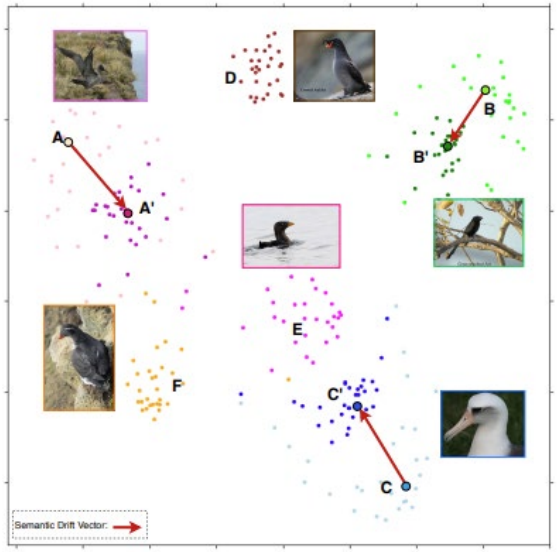
$$\widehat{W}_{new} = \frac{Mean(norm_{old})}{Mean(norm_{new})} W_{new}$$

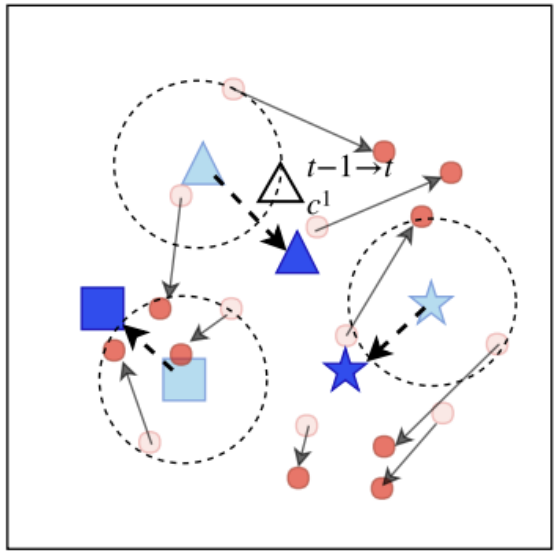
Decrease the weight norm of new classes for an ***unbiased prediction***

Bowen Zhao et al. *Maintaining discrimination and fairness in class incremental learning*. CVPR 2020.

Feature Rectify

- Updating the embeddings incrementally results in the feature drift of former classes





➢ Feature Rectify [Yu et al. CVPR'20]

$$\mu_k^t = \mu_k^{t-1} + \Delta\mu_k^t$$

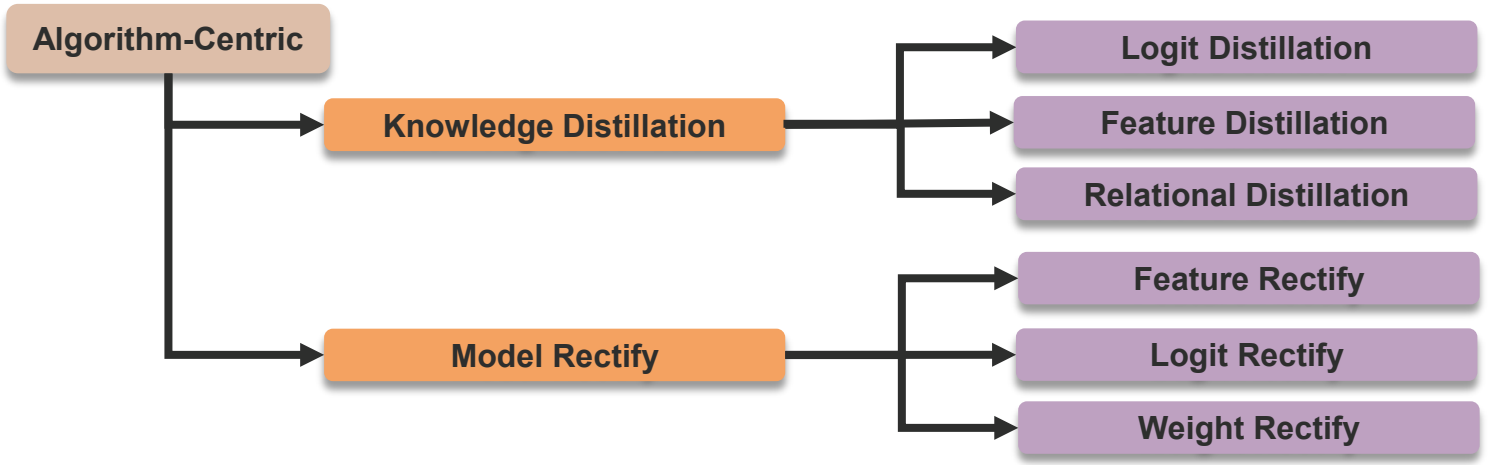
$$\Delta\mu_k^t = \sum_i sim\left(x_i, \mu_k^{t-1}\right) \Delta x_i^t$$

Estimate the drift of old prototypes via that of new class instances

Lu Yu et al. *Semantic drift compensation for class-incremental learning.* CVPR 2020.

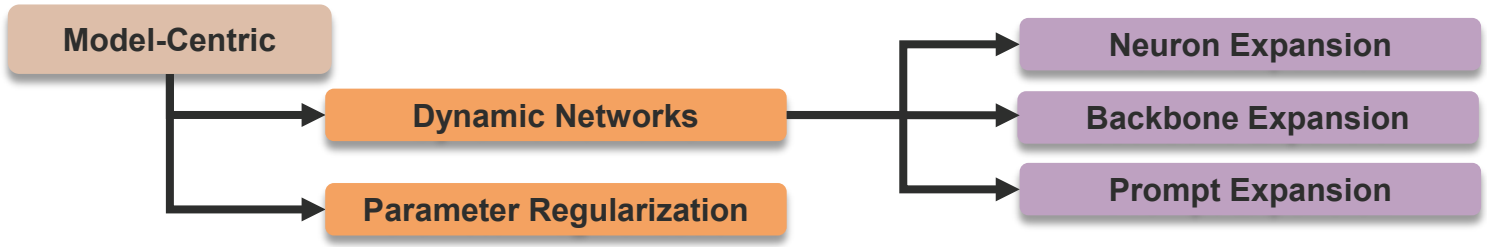
Summary of Algorithm-Centric Methods

- Knowledge distillation requires saving the old model
- Knowledge distillation requires the competitive ability of old model
- There are various phenomena that lead to inductive bias, and all of them can be solved via model rectify
- Model rectify helps understand the "why" of forgetting



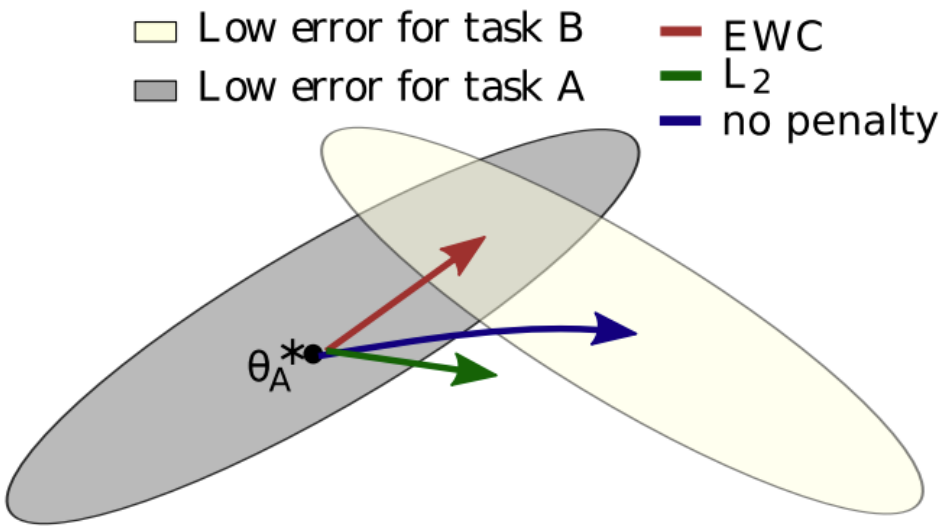
Model-Centric Methods

- **Core Idea**: operate the network structure or estimate parameter importance to resist forgetting
  - Dynamic Networks: adjust the network structure as data evolves to meet the requirements of streaming data
  - Parameter Regularization: regularize the important parameters from being changed when learning new tasks



Parameter Regularization

- We can estimate the importance of each parameter to the task
- Restricting important parameters from being changing



➢ Parameter Regularization [Kirkpatrick et al. PNAS'17]

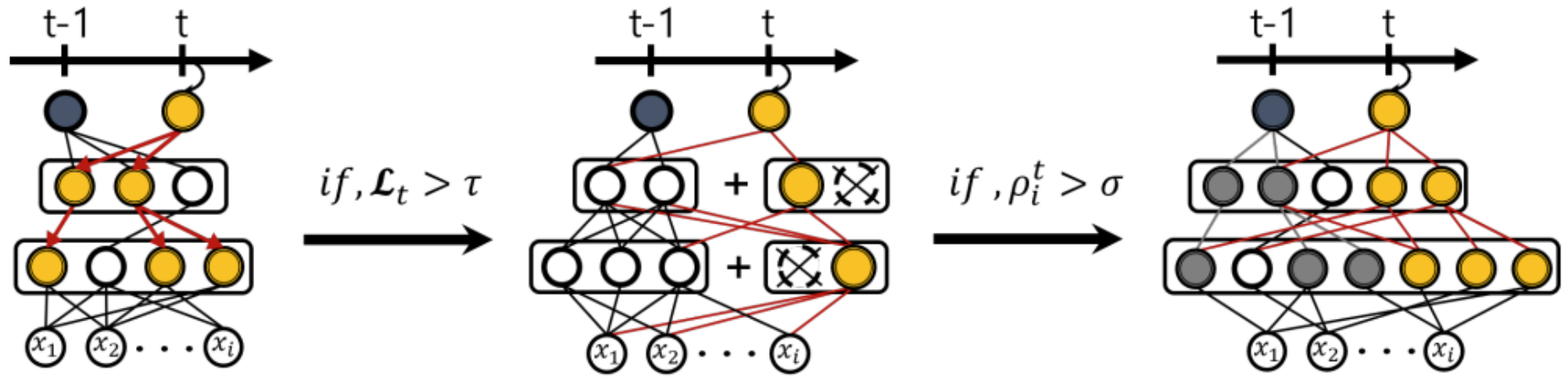
$$\min \ell(f(\mathbf{x}), y) + \frac{1}{2}\lambda \sum_k \Omega_k \left(\theta_k^{b-1} - \theta_k\right)^2$$

Maintain the performance by regularizing important parameters

James Kirkpatrick et al. *Overcoming catastrophic forgetting in neural networks*. PNAS 2017.

Neuron expansion

- If some parameters/neurons are important to the task, we can expand and copy them to maintain the performance [Yoon et al. ICLR'18]

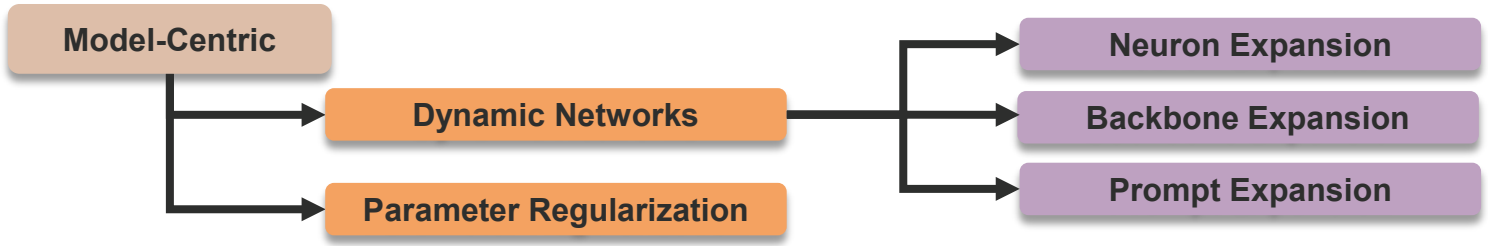


A pipeline of selective retraining, neuron expansion and neuron deletion
It only works under the ***task-IL setting*** to activate corresponding neurons.

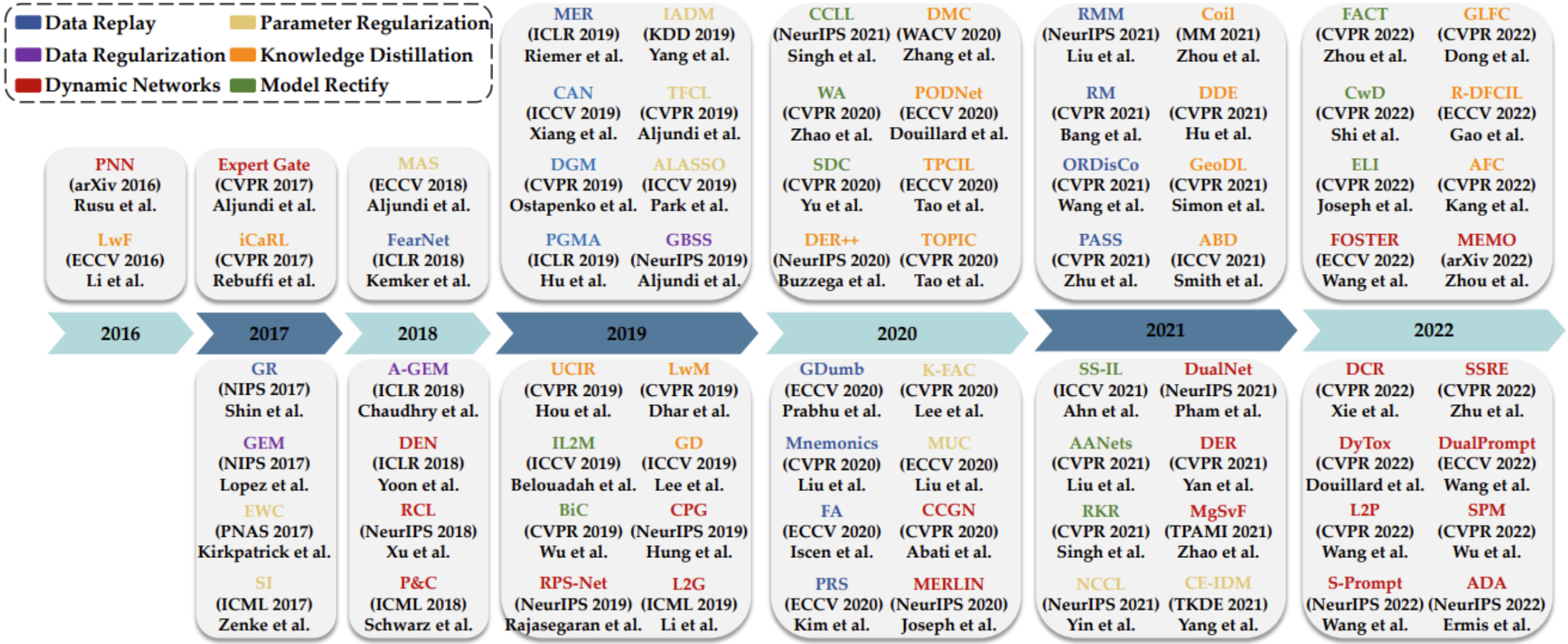
Jaehong Yoon et al. *Lifelong learning with dynamically expandable networks. ICLR* 2018.

Summary of Model-Centric Methods

- Parameter regularization methods weigh the importance of parameters, while they shall conflict at different stages
- Dynamic networks require expanding memory budget as data evolves
- We can strike a balance between backbones and restricted memory
- Adding tokens has shown to be an effective way in the era of ViT



Trends of Continual Learning



Da-Wei Zhou et al. *Deep Class-Incremental Learning: A Survey*. CoRR 2023.

Thanks!
Q & A

https://www.lamda.nju.edu.cn/zhoudw