Continual Learning

Christian Geishauser
Dialog Systems and Machine Learning Group
Heinrich-Heine University Düsseldorf

Great sources to get started

- Embracing Change: Continual Learning in Deep Neural Networks[1]
  - Hadsell et. al. 2020
- Continual Lifelong Learning with Neural Networks: A Review
  - Parisi et. al. 2019
- Towards Continual Reinforcement Learning: A Review and Perspectives
  - Khetarpal et. al. 2020
- Continual Lifelong Learning in Natural Language Processing: A Survey
  - Biesialska et. Al. 2020

1 https://www.youtube.com/watch?v=ES1CA9Fi5uc

Examples of Human Learning

Heinrich Heine Universität Düsseldorf

- I learned how to jump for the first time
  - I will reuse that skill to jump over everything I can find!
- I move to a new city
  - I quickly adapt to my new environment!
- Learning a new programming language
  - I could learn the second programming language much faster than the first!
  - I still know how to program in the other language!

Machine Learning

Heinrich Heine Universität Düsseldorf

- Deep Learning is optimised for static, large-scale datasets
  - Supervised learning: learn on fixed dataset with fixed number of classes
  - Reinforcement learning: learn in stationary, self-contained environments
- Gradient-based optimisation assumes that dataset is balanced (i.i.d.)
- Humans dont learn well from randomly sampled data

Defining Continual Learning

- The world is highly non-stationary!
  - Household robot for cleaning needs to learn how to wash dishes
  - Suddenly many news articles about Covid-19 (new vocabulary needed)
  - No one booking a hotel anymore, but many more ordering food in a restaurant

- Continual learning: learning environment is non-stationary, divided into a set of tasks that need to be completed sequentially
  - Compared to multi-task learning, do not see all tasks at once
  - Compared to curriculum learning, learner has no control over task ordering
  - Compared to transfer learning, also previous tasks are important

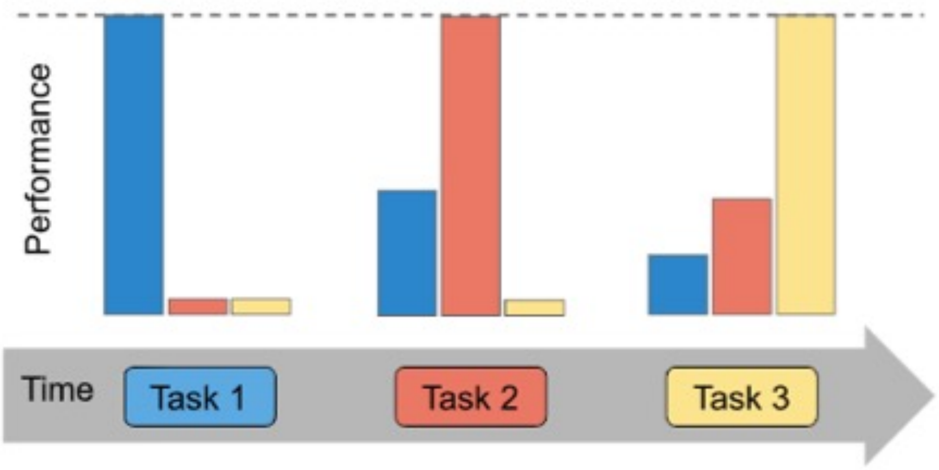
Desiderata of Continual Learning

Continual learning methods involve balancing competing objectives:

- Minimal access to previous tasks
  - The model does not have infinity storage for previous experience
- Minimal increase in model capacity and computation
  - Must be scalable: Should not add a new model for each task
- Fast adaptation and recovery
  - Fast adaptation to novel tasks or domain shifts and of fast recovery

Desiderata of Continual Learning

- Minimizing catrastophic forgetting (CF) and interference
  - Training on new task should not significantly reduce performance of previously learned tasks

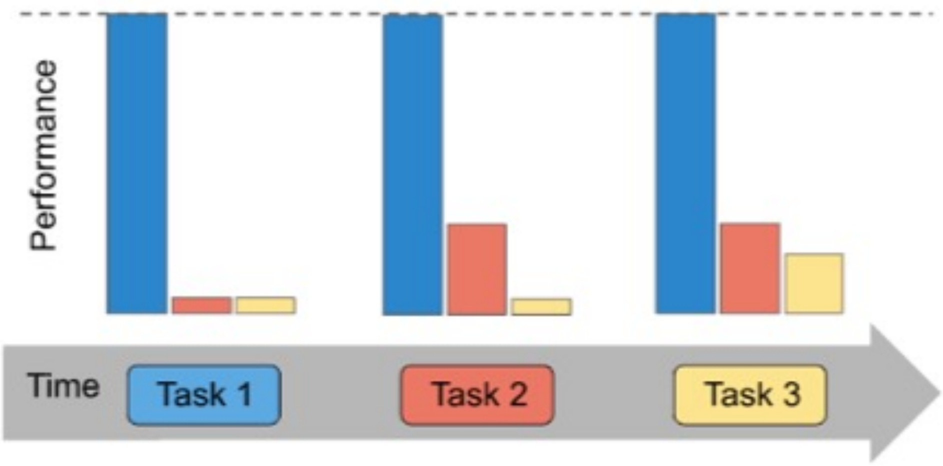


Hadsell et. al. 2020

Desiderata of Continual Learning

hhu
Heinrich Heine
Universität Düsseldorf

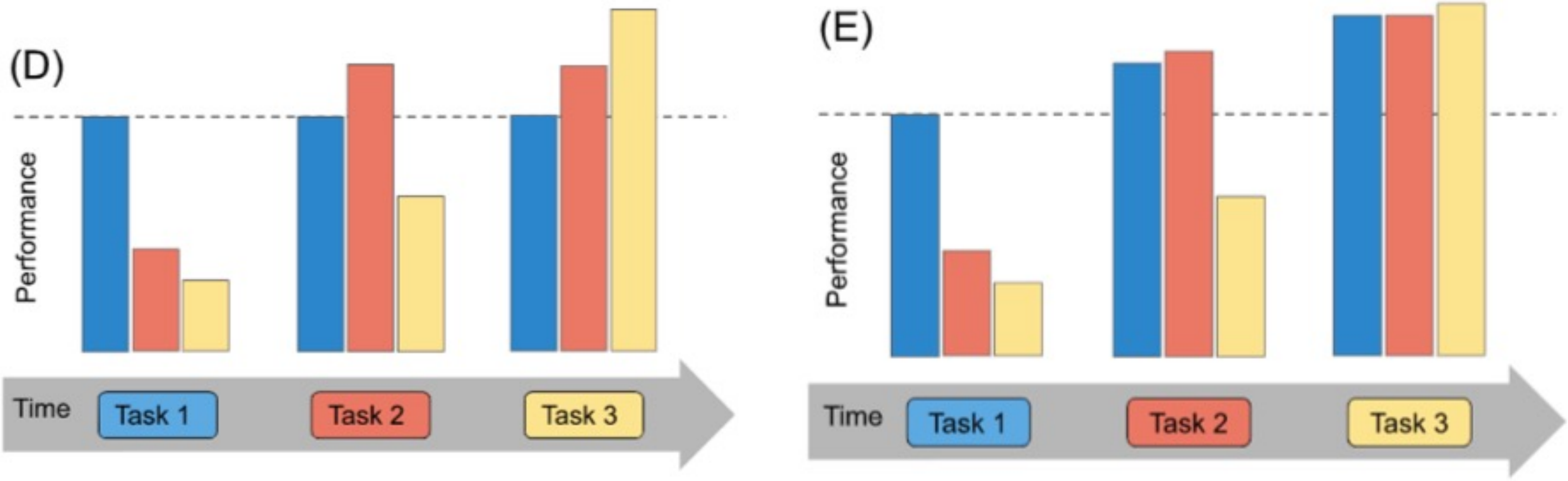
- Maintaining plasticity
  - Model should be able to keep learning effectively as new tasks are observed



Hadsell et. al. 2020

Desiderata of Continual Learning

- Maximizing forward and backward transfer
  - Learning a task should improve related tasks, both past and future



Hadsell et. al. 2020

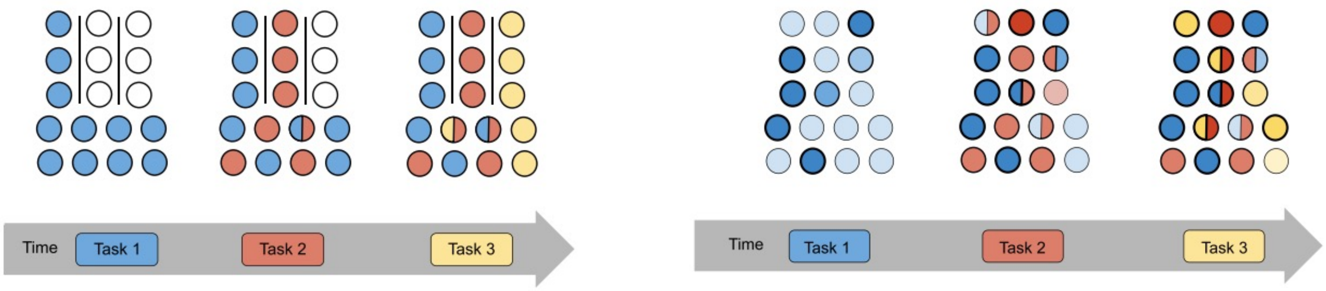
Desiderata of Continual Learning

- These points are competing against each other
- Maintaining perfect recall in a fixed-capacity model is impossible
- Fast adaptation competes with stabilisation (stability-plasticity dilemma in the brain)

How about ..?

- Obvious idea: Use an independent model for every task
- Downsides:
  - Requires significant storage

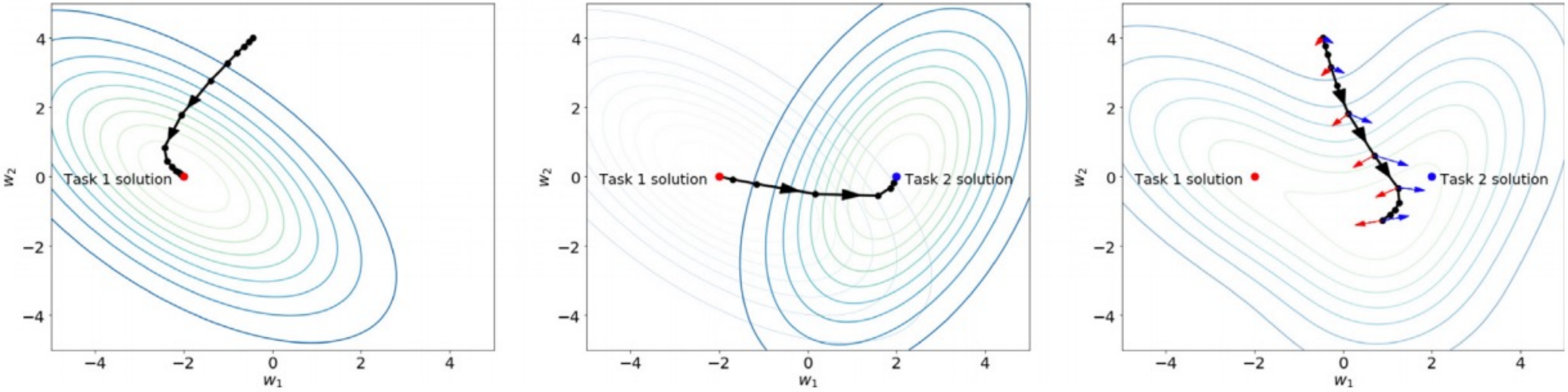
- -> Share parts of the network structure across tasks



Hadsell et. al. 2020

Tug-of-war dynamics

- Sharing parts of the network creates a new challenge: Catastrophic forgetting
  - Straightforwardly learning the new task results in forgetting how to solve old tasks



(A) Training on task 1

(B) Training on task 2

(C) Training on task 1 and 2 simultaneously

Hadsell et. al. 2020

Taxonomy of Continual RL Approaches

- Regularisation-based
- Architectural
- Memory-based
- Learning to learn/Meta-learning
- Learning to explore
- Skill learning
- …

Regularization based

- Approach to deal with catastrophic forgetting

- Regularizes the updates on the current task through
  - Regularising the gradient
  - Regularising the loss
  - Using knowledge distillation
  - …

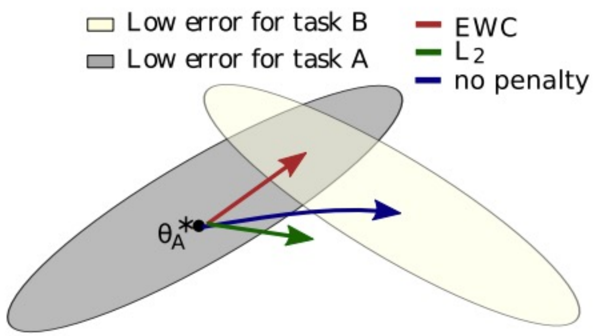
Lopez-Paz et. al. 2017[1]

- Maintains a memory $\mathcal{M}_t$ for every task t
- Updates on new observed sample are constrained to not increase loss on previous tasks for $\mathcal{M}_t$
- Derive equations for gradient-based optimization
- Allows for backward-transfer
- Propose metrics to measure forward and backward transfer

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

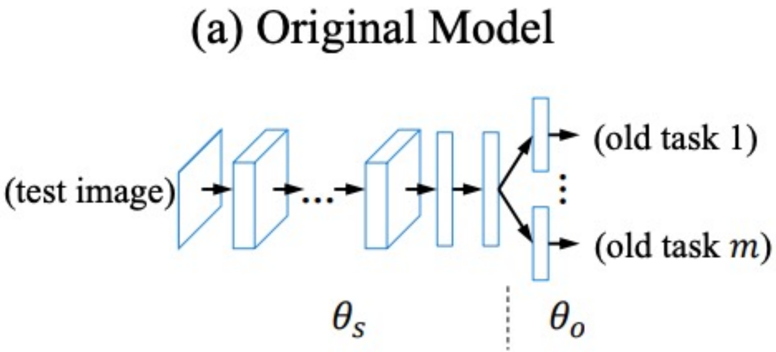
Kirkpatrick et. al. 2017[1]

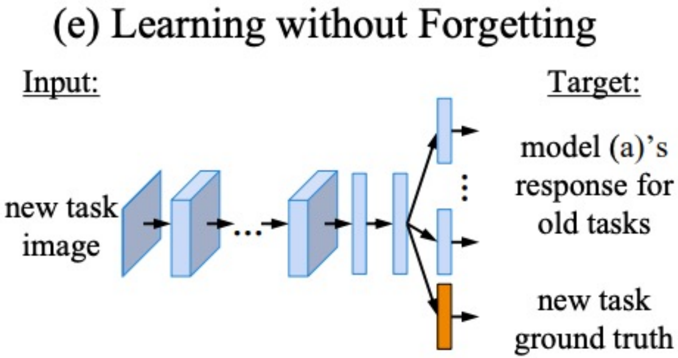
- Elastic Weight Consolidation
  - Inspired by synaptic consolidation in the brain that reduces plasticity of specific synapses
  - Regularizes the loss to
    - Remember old tasks by selectively slowing down learning on weights important for those tasks
  - Relies on Fisher information matrix to measure sensitivity of parameters to each task



Li et. al. 2017[1]

- Output probabilities for each new image should be close to recorded output from original network
- Use knowledge distillation
- No memory needed



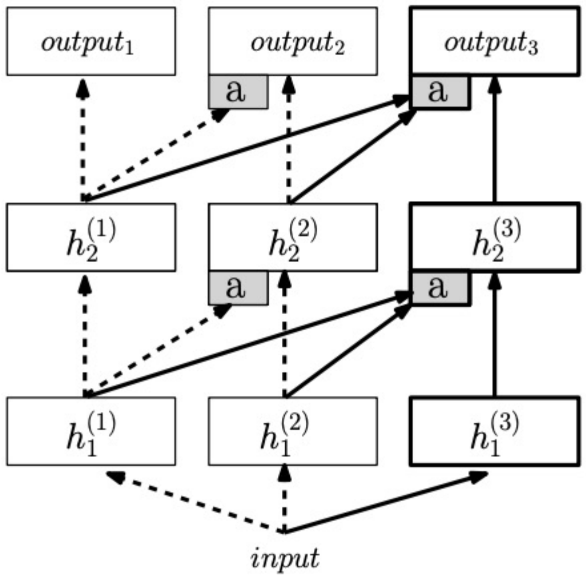


Architectural methods

- Prevent forgetting by applying modular changes to the network architecture
- Typically previous task parameters are kept fixed
- Main drawback: Substantially growing number of parameters

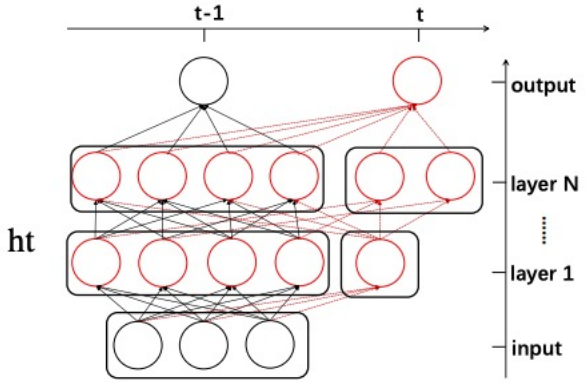
Rusu et. al. 2016[1]

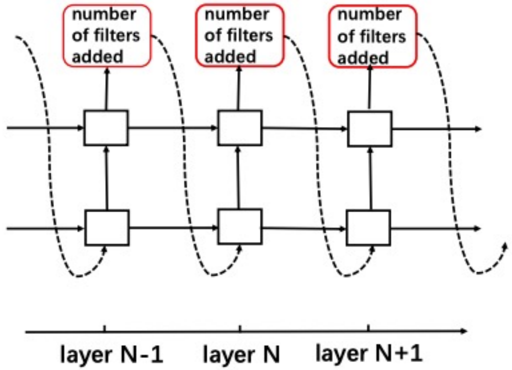
- Immune to forgetting by design (new network for each task)
- Leverage prior knowledge through lateral connections
- Substantial growth of network parameters
- Actually only a fraction of the new capacity is utilized



Xu et. al. 2018[1]

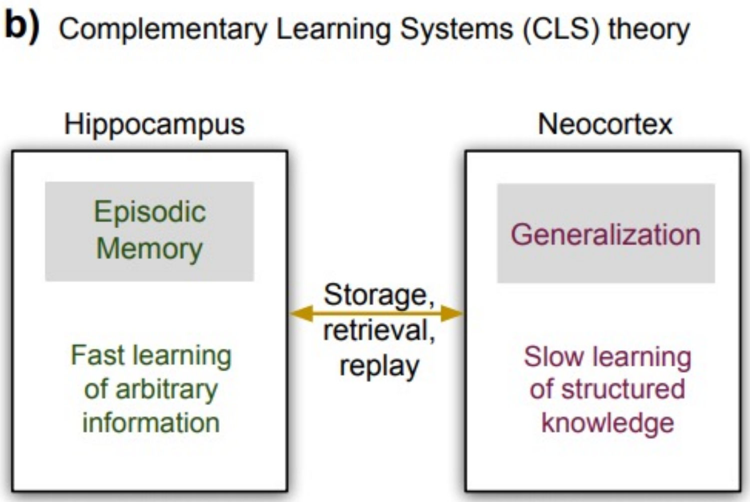
- Deciding optimal number of nodes to add is posed as a reinforcement learning problem
- Reward encodes validation accuracy and network complexity
- Only new parameters are trained





Complementary Learning Systems theory[1]

- Brain learns and memorizes
- Episodic memory stores specific events from the past
- Neocortex for long-term retention
  - Slow learning rate
  - Builds overlapping representations of learned knowledge
- Hippocampal system exhibits short-term adaptation and rapid learning of novel information
  - Encodes sparse representation of events
  - Rapid learning rate
  - Used for replaying memories (also reactived during sleep)



1 adapted from Parisi et. al. 2019

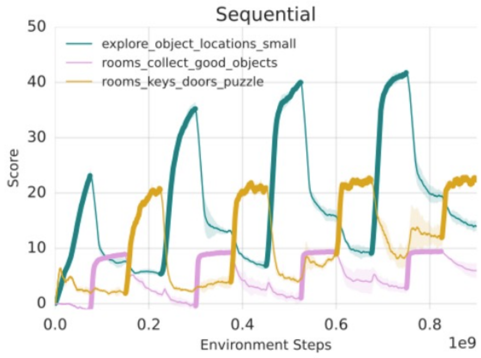
Memory-based methods

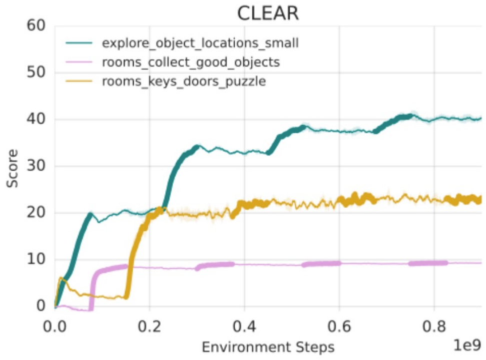
- Rehearsal methods against catastrophic forgetting
  - Store and replay past experiences
- Episodic memory for inference
  - Encoding, storing and recalling knowledge or experience
- Memory grows with number of tasks
  - Use generative memory methods to generate rehearsal data as needed

Rolnick et. al. 2019[1]

- Off-policy learning and behaviour cloning for enhanced stability
- On-policy learning to preserve plasticity (50-50 ratio of on- and off-policy data used)

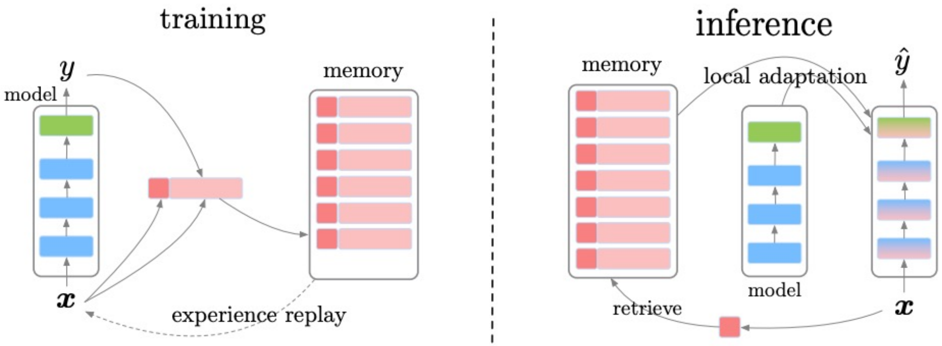
$$L_{\text{policy-cloning}} := \sum_a \mu(a|h_s) \log \frac{\mu(a|h_s)}{\pi_\theta(a|h_s)}, \qquad L_{\text{value-cloning}} := ||V_\theta(h_s) - V_{\text{replay}}(h_s)||_2^2$$





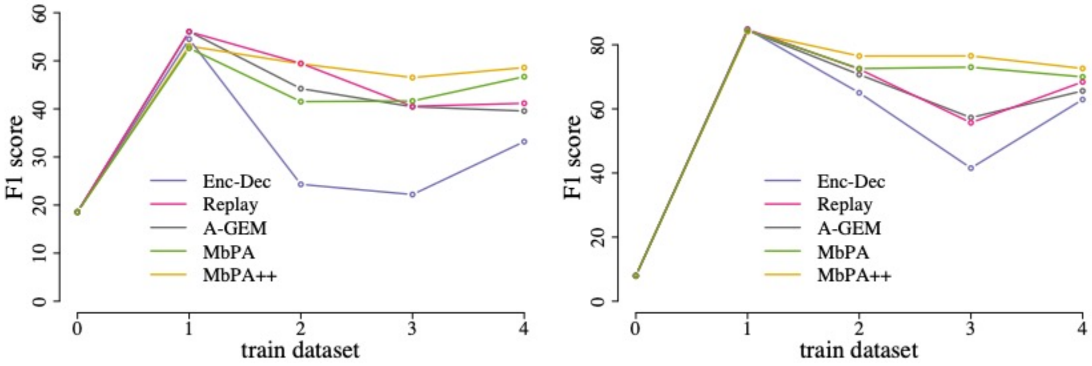
d‘Autume et. al. 2019[1]

- Episodic memory is a key-value memory block
  - Key representation of x obtained using pretrained BERT model
  - Values given by x, y
- During trainig: Use sparse experience replay to seldomly update network
  - Together with training on freshly observed samples
- During testing: Retrieve K nearest neighbors $(x_i, y_i)_{i=1}^{K}$ through key matching and perform local adaptation with it



d'Autume et. al. 2019[1]

- Evaluated on text classification and question answering
  - Evaluated on different datasets but having the same task
  - Question answering: SQuAD 1.1, TriviaQA and QuAC



(b) QA–QuAC

(c) QA–SQuAD

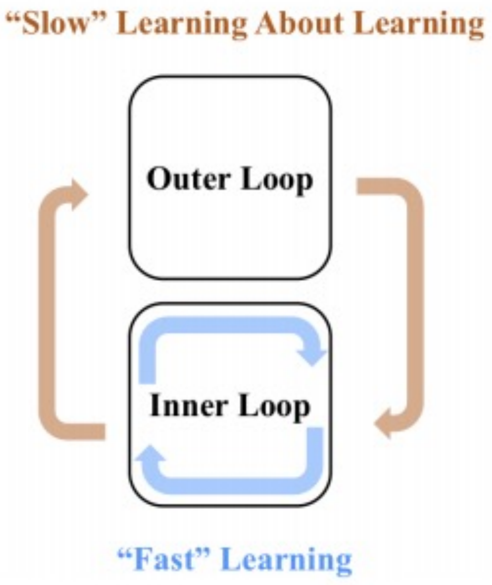
Learning to learn/Meta-learning

- Solutions described so far prescribe hand-engineered mechanisms or architectures
  - Strikes different trade-offs between desiderata
- Can we find better trade-offs by learning a solution from data rather than designing it?
- Can we use meta-learning for rapid learning of new tasks?

Learning to learn/Meta-learning

Heinrich Heine
Universität Düsseldorf

- Meta-learning comprises of two timescales of optimization
  - Inner loop that optimizes specific tasks
  - Outer loop that optimizes performance over multiple inner loops
  - Most prominent example: MAML (Finn et. al. 2017)
  - MAML: Find parameters that can learn a new task quickly after only few update steps



Khetarpal et. al. 2020

Finn et. al. 2019[1]

- Tackles the problem of fast adaptation, becoming faster the more tasks you observed
- Meta-learning usually learns on a set of training tasks in order to rapidly adapt to a new seen task
  - Distinct phases of meta-training and meta-testing/deployment
  - Assume sufficiently large set of tasks for meta-training
  - Tasks come from a fixed distribution
  - In the real world, tasks are likely available only sequentially

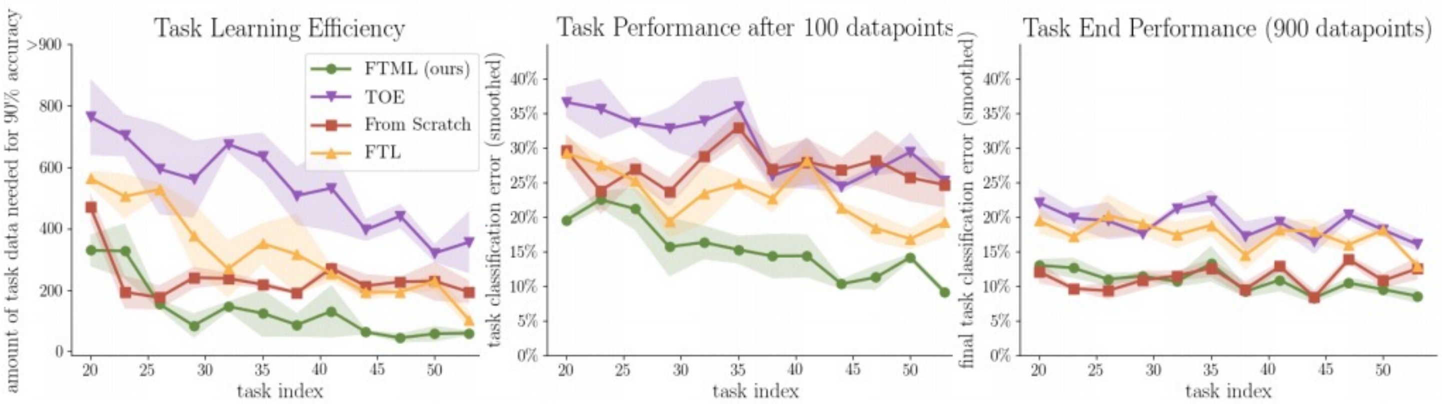
Finn et. al. 2019[1]

- Online Meta-Learning (Finn et. al. 2019)
  - Extends MAML to the sequential learning setting
  - Meta-update uses data for all previously seen tasks
  - Inner-loop update only uses current task data
  - Computationally demanding
  - Only focuses on efficient forward transfer, not tackling catastrophic forgetting

$$\mathbf{g}_t(\mathbf{w}) = \nabla_{\mathbf{w}} \mathbb{E}_{k\sim\nu^t} \mathcal{L}\big(\mathcal{D}_k^{\text{val}}, \boldsymbol{U}_k(\mathbf{w})\big), \quad \text{where}$$
$$\boldsymbol{U}_k(\mathbf{w}) \equiv \mathbf{w} - \alpha \nabla_{\mathbf{w}} \mathcal{L}\big(\mathcal{D}_k^{\text{tr}}, \mathbf{w}\big)$$

Finn et. al. 2019[1]

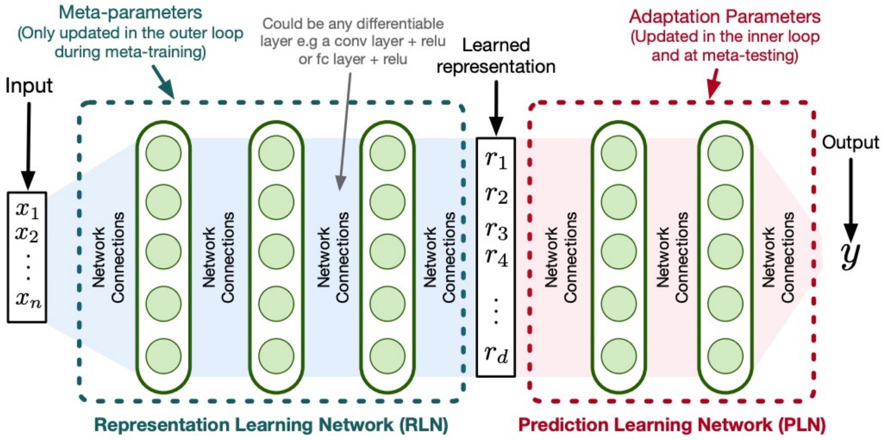
- In particular evaluated on MNIST
  - Tasks created through different backgrounds, rotations, different scaling
  - Evaluated against
    - TOE: Train on everything, i.e. multi-task-learning on all data seen so far
    - FTL: Joint training with fine-tuning, first train on all $t-1$ previous tasks and fine-tune for task t



Javed et. al. 2019[1]

Heinrich Heine
Universität Düsseldorf

- Learn useful representations for online continual learning
  - Inner loop learns on correlated sequences of input, which could lead to catastrophic forgetting
  - Outer loop optimises input representations to reduce forgetting and improve generalization
  - Optimisation leads to sparse input representations even though it was not explicitly trained for it
  - Sparse representations reduce forgetting because each update changes only a small number of weights



Evaluation of CL algorithms

- No established consensus on benchmark datasets/environments and metrics so far
  - Popular datasets for images: Permuted MNIST, splitted CIFAR
  - In RL: ATARI games
    - Not clear how knowledge can be transferred from one to the other
  - Lacking well-suited environments
  - Often designing tasks suitable for a specific question
    - Might result in inherent bias

Evaluation of CL algorithms

- Let $a_{j,i}$ be the performance of task t_i after training on task t_j
- Average accuracy:

$$A_{\mathcal{T}} = \frac{1}{\mathcal{T}} \sum_{i=1}^{\mathcal{T}} a_{\mathcal{T},i}$$

- Forgetting measure:

$$F_{\mathcal{T}} = \frac{1}{\mathcal{T}-1} \sum_{i=1}^{\mathcal{T}-1} f_i^{\mathcal{T}}, \qquad f_i^{\mathcal{T}} = \max_{k \in \{1,\ldots,j-1\}} a_{k,i} - a_{j,i}$$

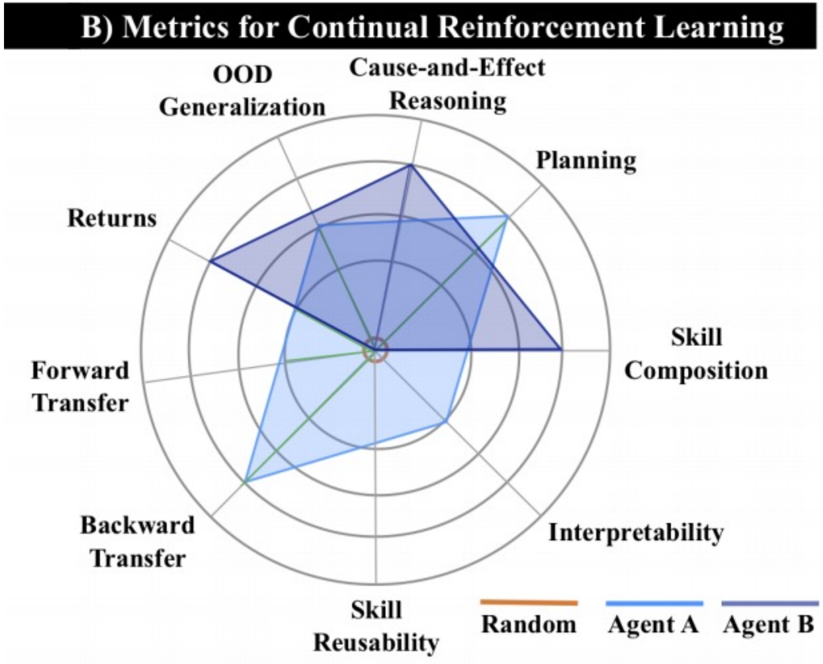
- Forward Transfer:

$$FT_{\mathcal{T}} = \frac{1}{\mathcal{T}-1} \sum_{i=2}^{\mathcal{T}} a_{i-1,i} - b_i$$

  - b_i = test accuracy for task i at random initialization

Evaluation of CL algorithms

- Are skills reused?
- Type of representation or behaviour learned
- Is the agent learning underlying rules of the environment?
- What happens if agent faces situations not in the training distribution?



Khetarpal et. al. 2020

Summary

Heinrich Heine
Universität Düsseldorf

- Continual learning is faced with learning task sequentially
  - In contrast to being exposed to all tasks simulatenously, e.g. multi-task learning
- Creates issues/questions such as
  - Catastropic forgetting
  - How can we leverage past knowledge to learn new tasks quicker
  - How to deal with memory capacity or parameter size
  - Need to focus on multiple objectives of continual learning
- Requires suited datasets and evaluation metrics

Summary

- Regularization methods
  - Gradient episodic memory (GEM)
    - Regularizes the gradients
  - Overcoming catastrophic forgetting in neural networks (EWC)
    - Regularize the loss
  - Learning without forgetting (LwF)
    - Uses knowledge distillation
  - …
- Architectural methods
  - Progressive neural networks (PNN)
  - Reinforced continual learning
  - …

Summary

Heinrich Heine Universität Düsseldorf

- Memory-based approaches
  - Experience replay for continual learning
  - Episodic memory in lifelong language learning
  - …
- Meta learning
  - Online meta-learning
  - Meta-learning representations for continual learning
  - …

Heinrich Heine
Universität Düsseldorf

Thanks!