Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

ICML | 2021

Sharing Less is More: Lifelong Learning in Deep Networks with Selective Layer Transfer

Seungwon Lee
Univ. of Pennsylvania

Sima Behpour
Carnegie Mellon Univ.

Eric Eaton
Univ. of Pennsylvania

Correspondence: {leeswon, eeaton}@seas.upenn.edu

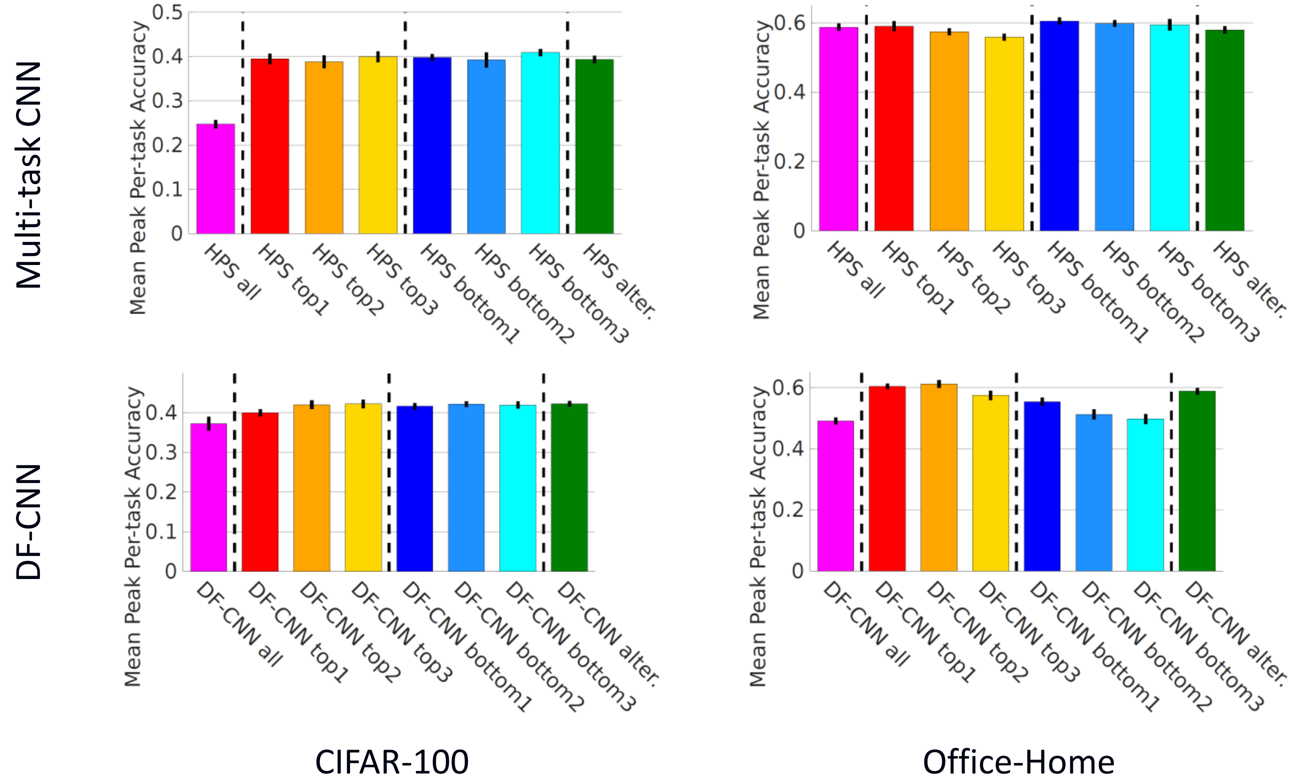
International Conference on Machine Learning 2021

Motivation

- Lifelong/continual ML aims to continually learn, maintain, and reuse knowledge across multiple, consecutive tasks
- Previous work has mainly focused on:
  - Architecture (what / how to transfer)
  - Task relationships (when to transfer)
- Less attention has been given to the granularity of knowledge to transfer (<u>where</u> to transfer)
  - Branching task models in a tree structure
  - Introducing a new learning module per layer between tasks

Motivation

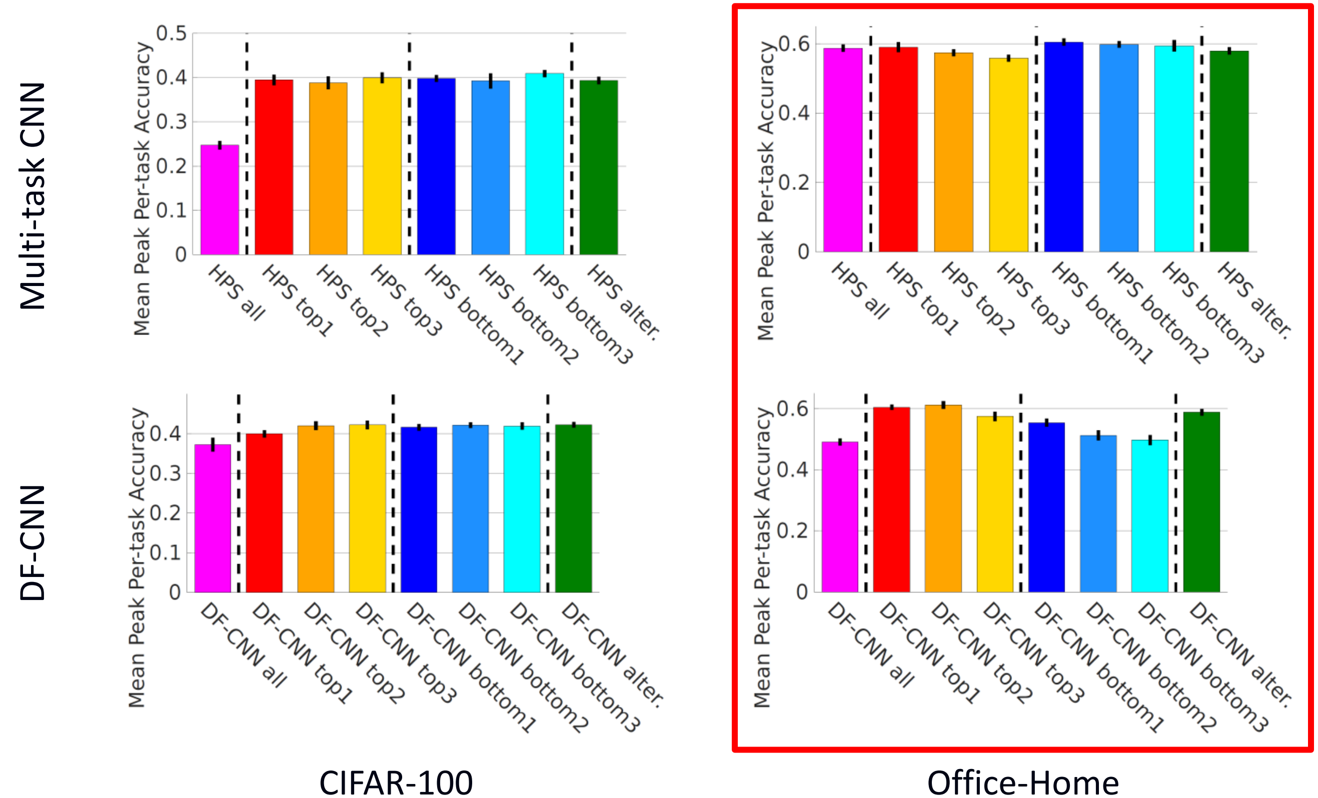
- A simple experiment: evaluation of different architectures



Motivation

Penn Engineering
GRASP LABORATORY
General Robotics, Automation, Sensing & Perception Lab

- A simple experiment: evaluation of different architectures



The optimal transfer configuration varies according to both the architecture and the task relationships

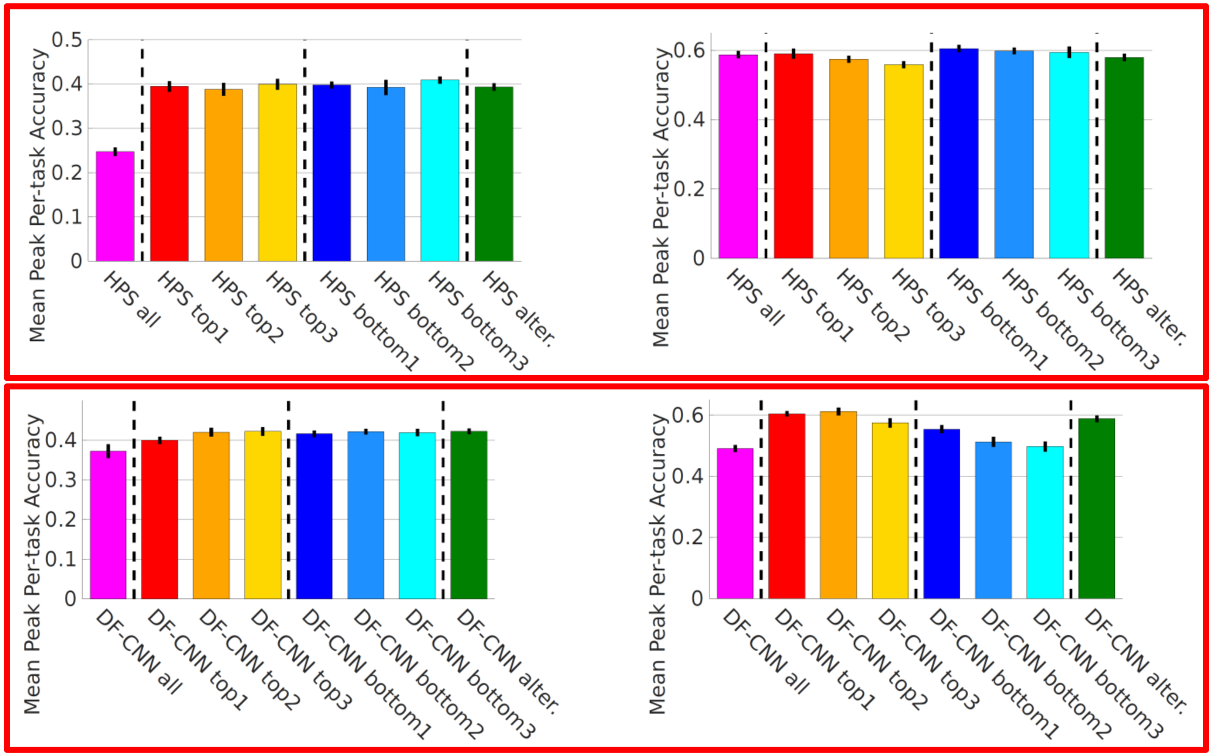
Motivation

Penn Engineering
GRASP LABORATORY
General Robotics, Automation, Sensing & Perception Lab

- A simple experiment: evaluation of different architectures

Multi-task CNN

DF-CNN

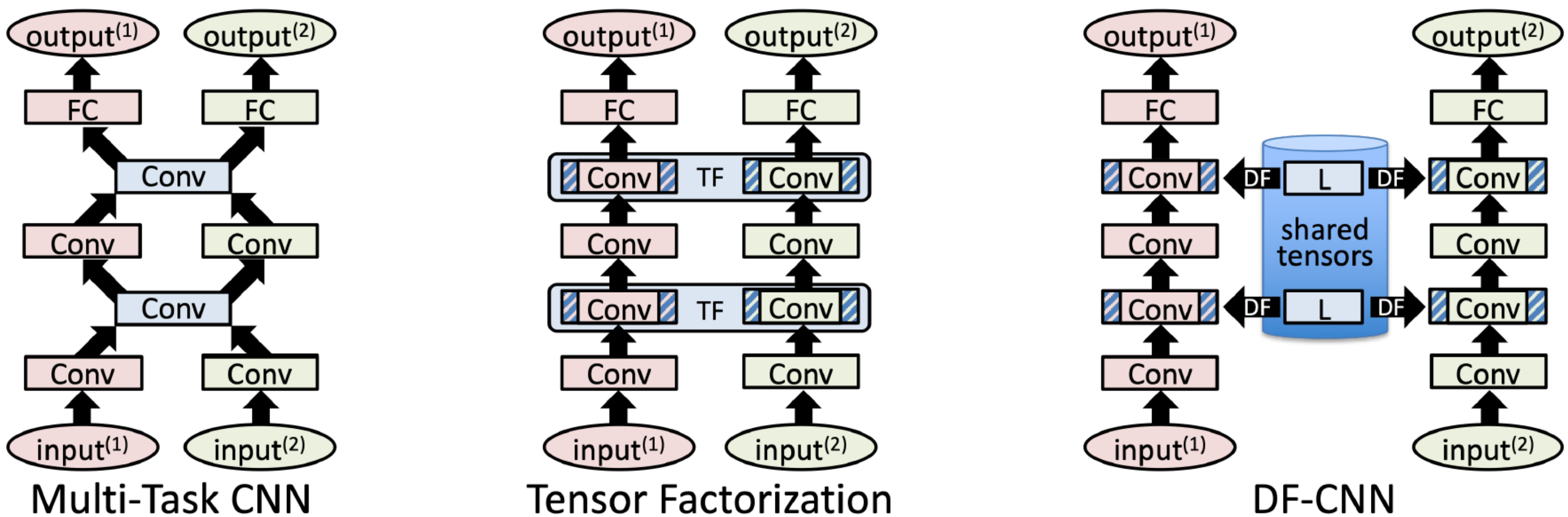


CIFAR-100 Office-Home

The optimal transfer configuration varies according to both the architecture and the task relationships

Lifelong Architecture Search

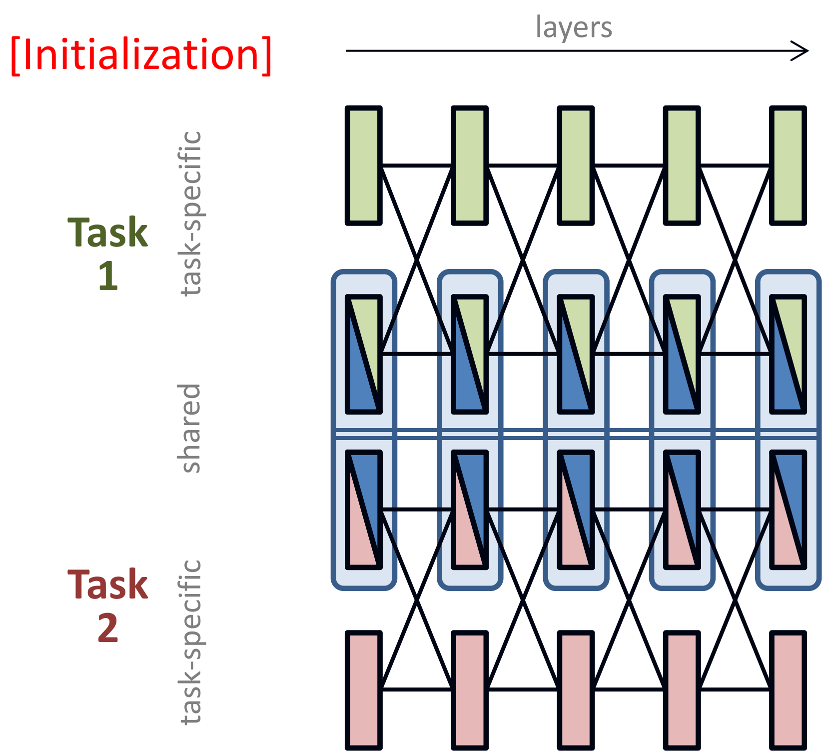
- Difficulties of lifelong architecture search:
  - Size of search space ($T \cdot 2^d$ configurations for d-layer network and T tasks)
  - Dependency on the training of network parameters



Example of an *alternating* transfer configuration for three different learning architectures

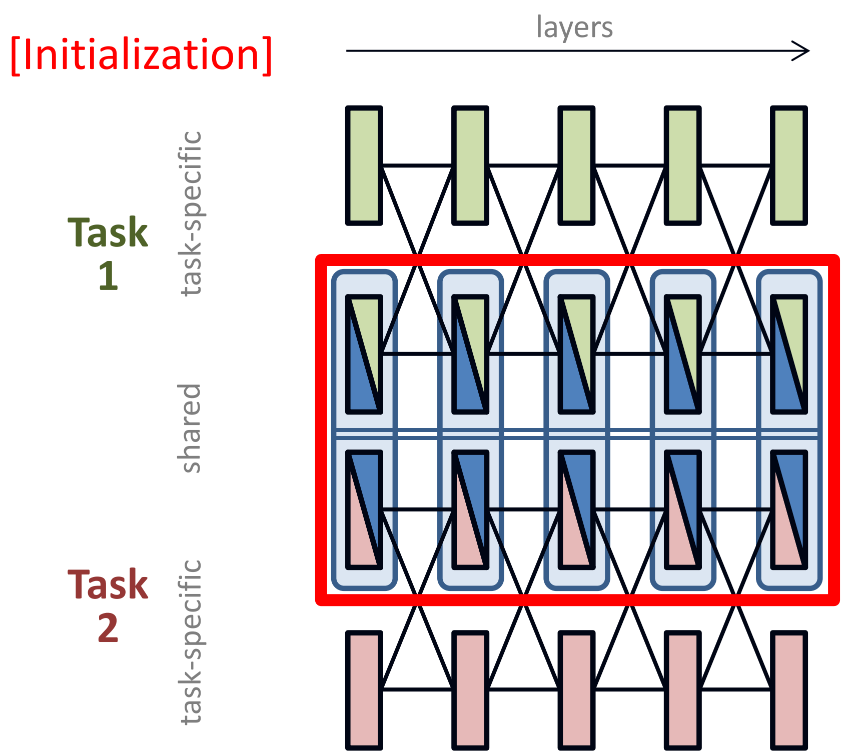
LASEM

- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$



LASEM

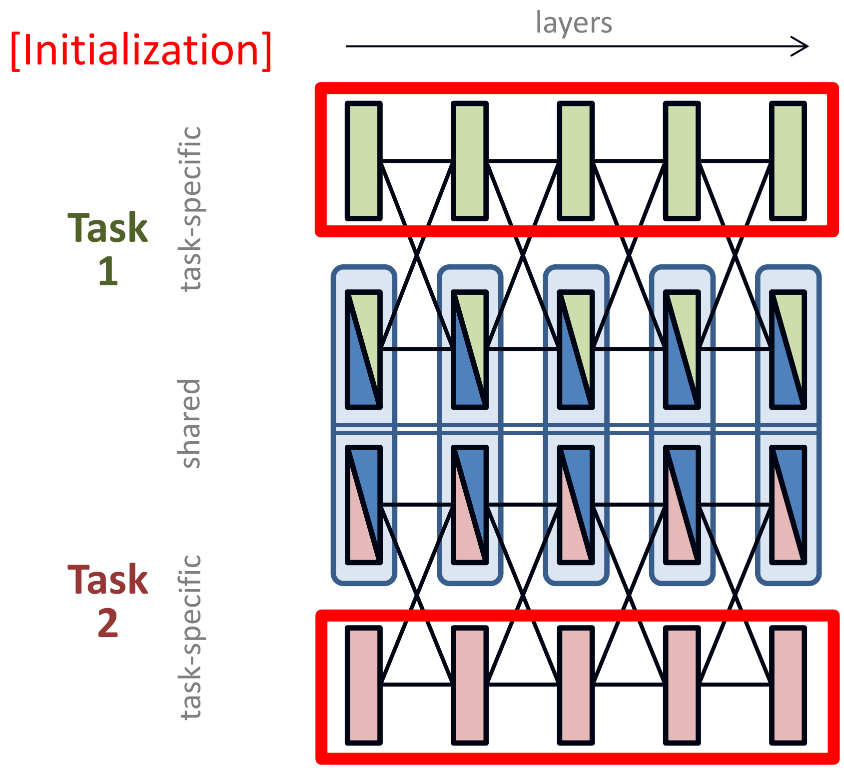
- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$



LASEM

Penn Engineering
GRASP LABORATORY
General Robotics, Automation, Sensing & Perception Lab

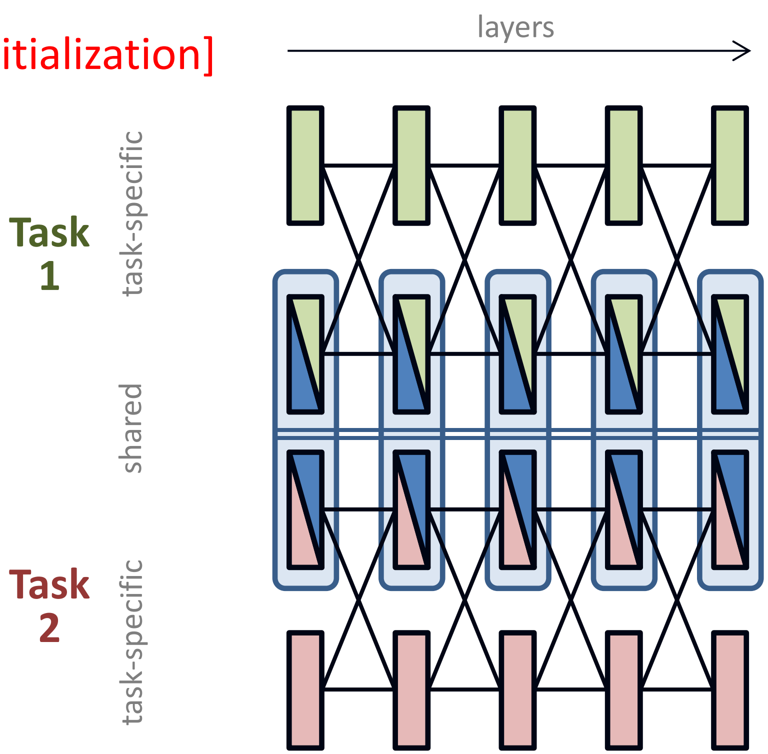
- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$



LASEM

- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$

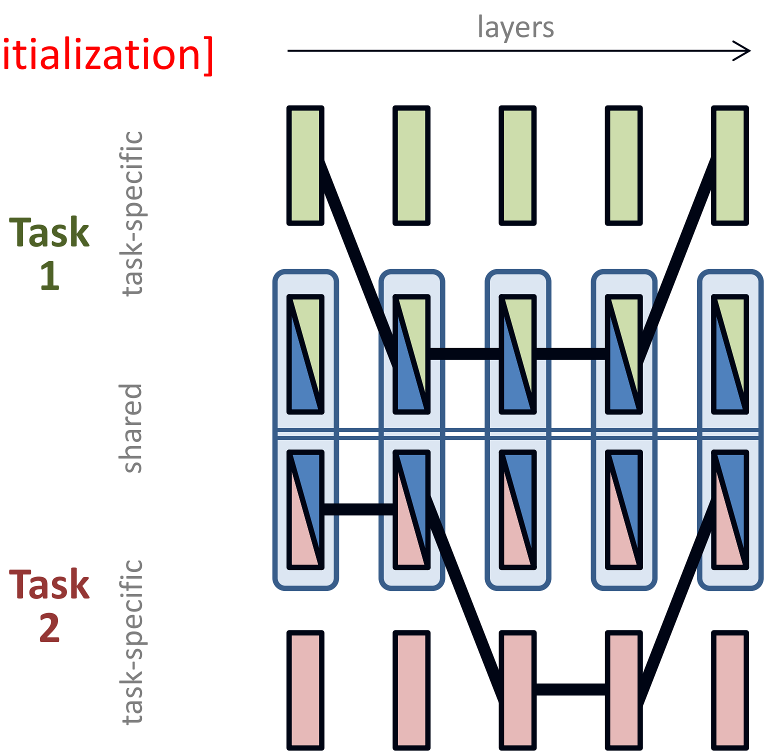
[Initialization]



LASEM

- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$

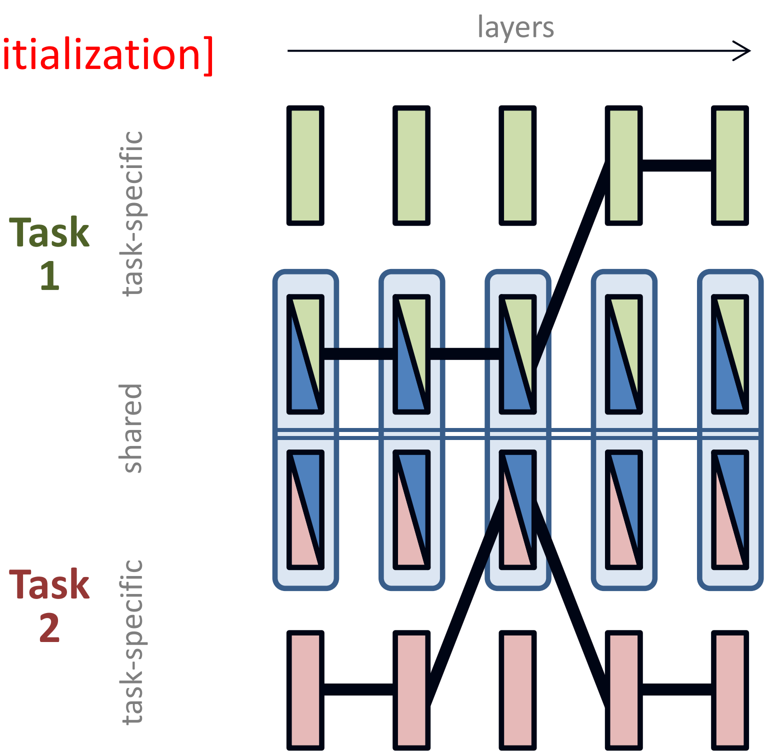
[Initialization]



LASEM

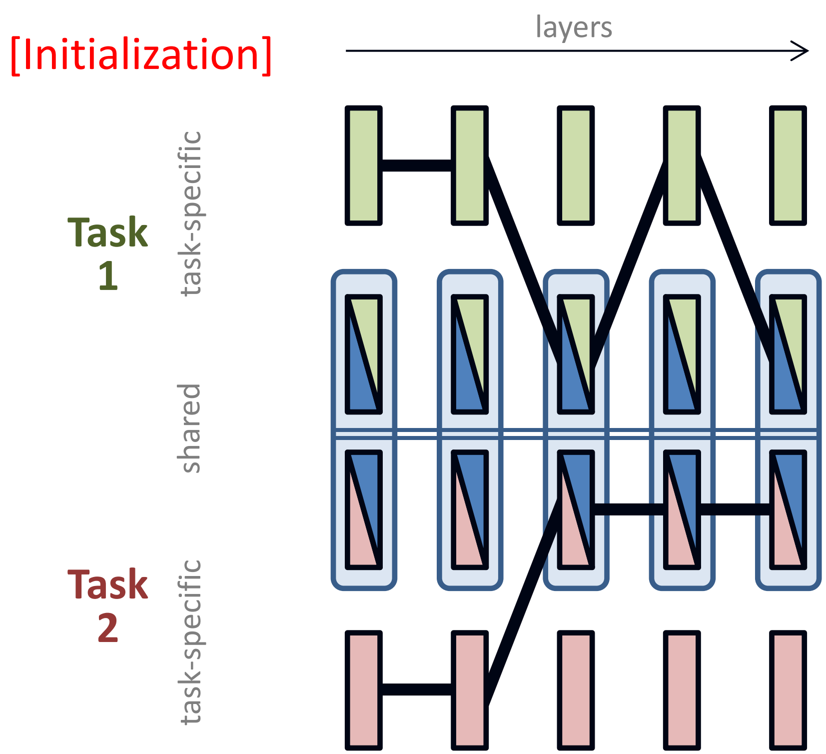
- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$

[Initialization]



LASEM

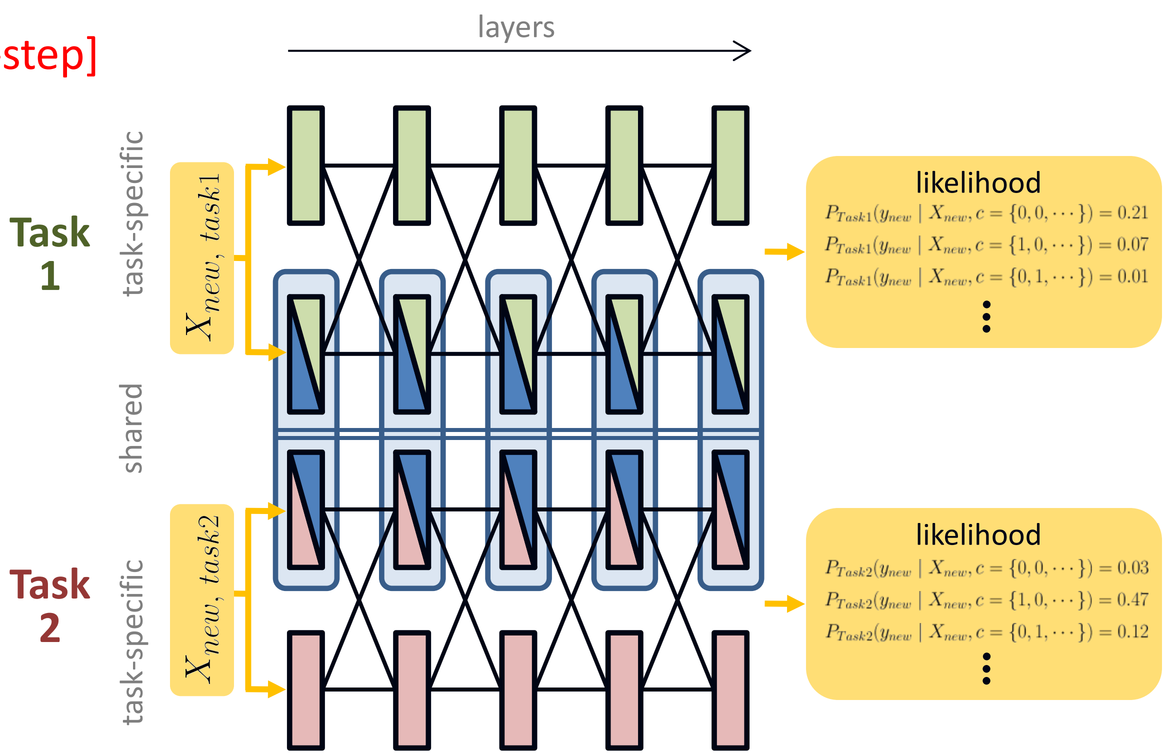
- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$



LASEM

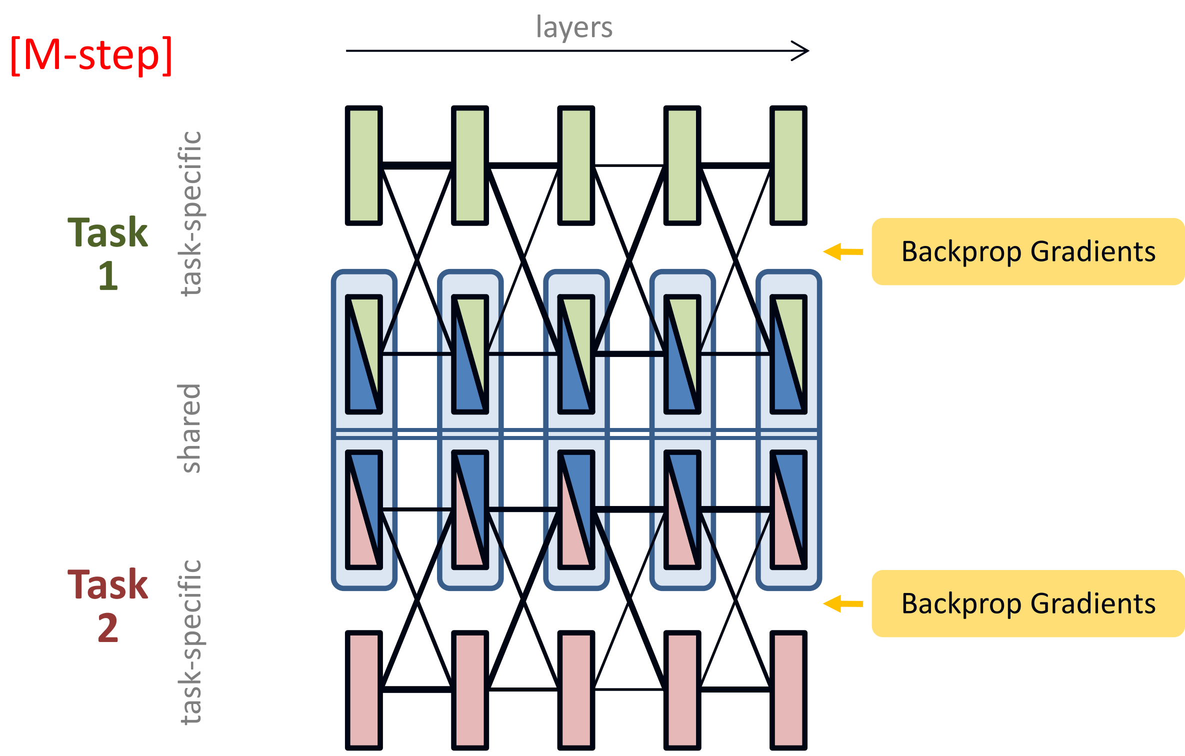
- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$

[E-step]



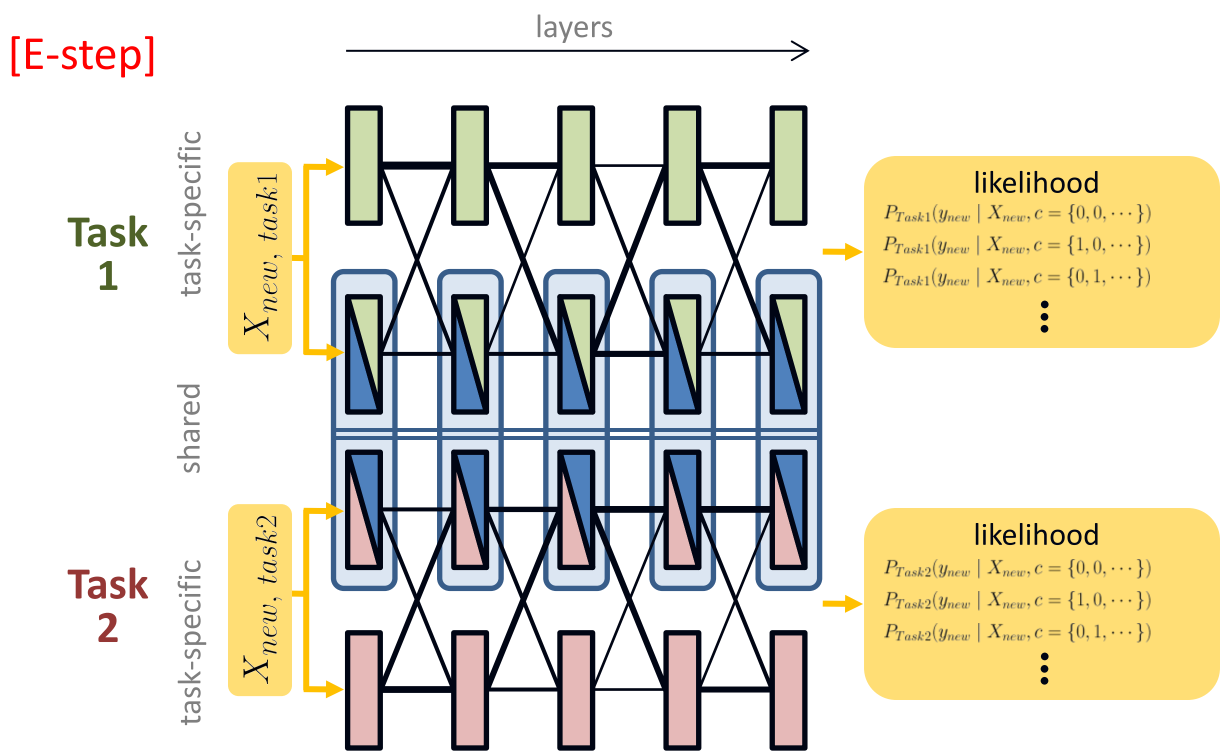
LASEM

- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$



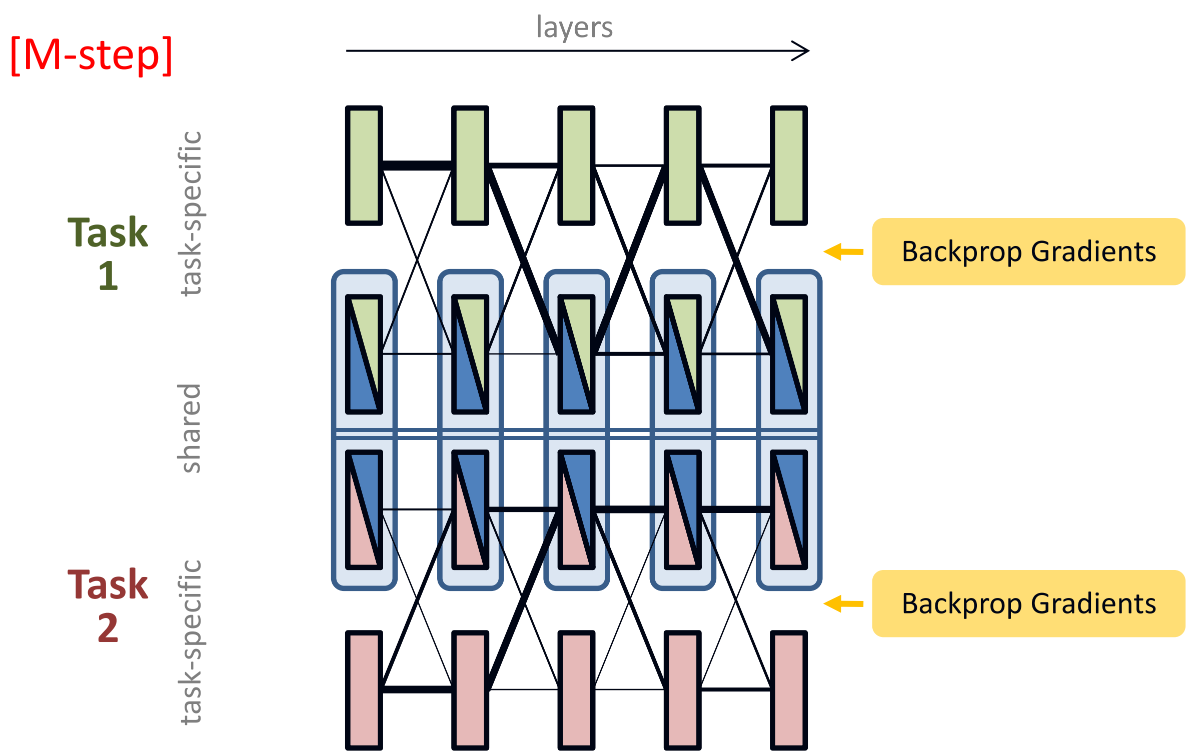
LASEM

- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$



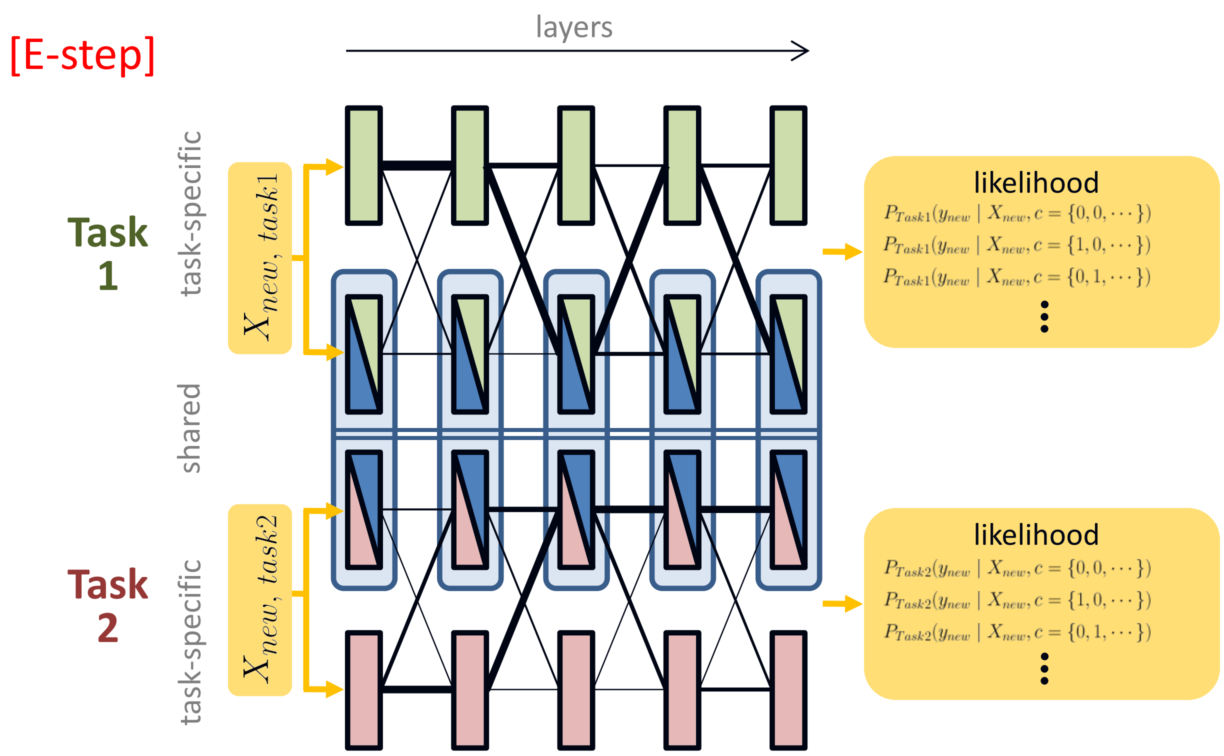
LASEM

- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$



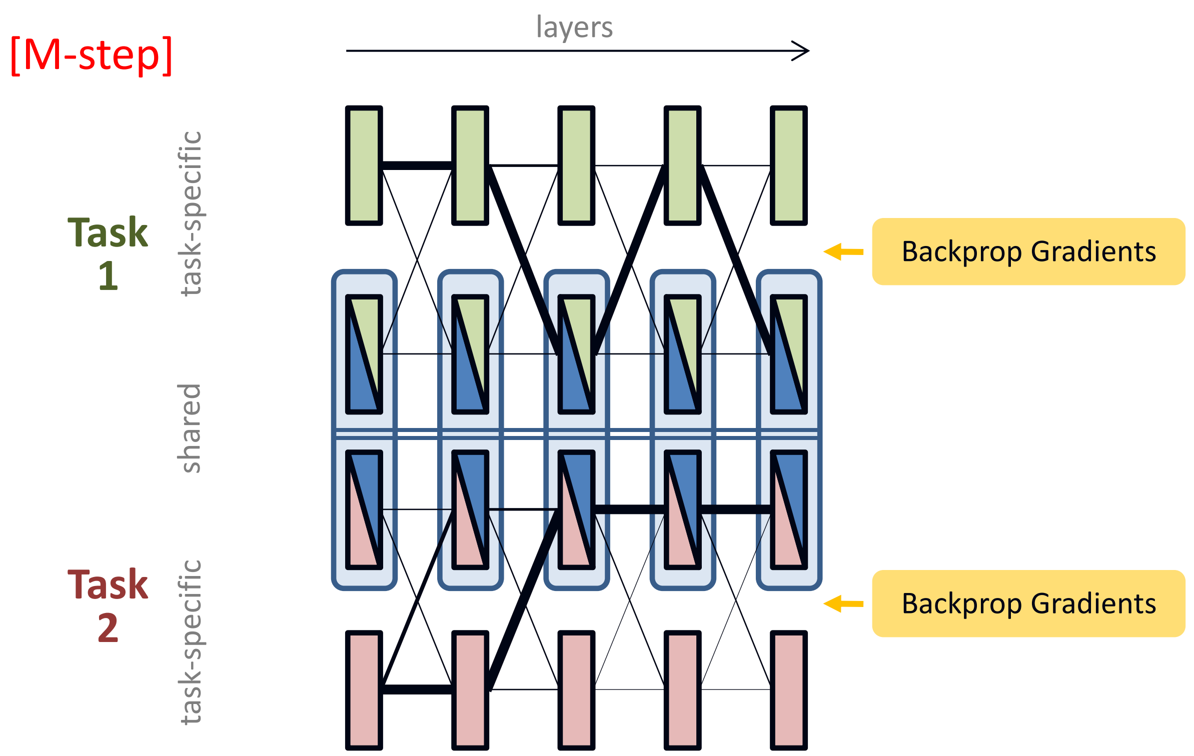
LASEM

- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$



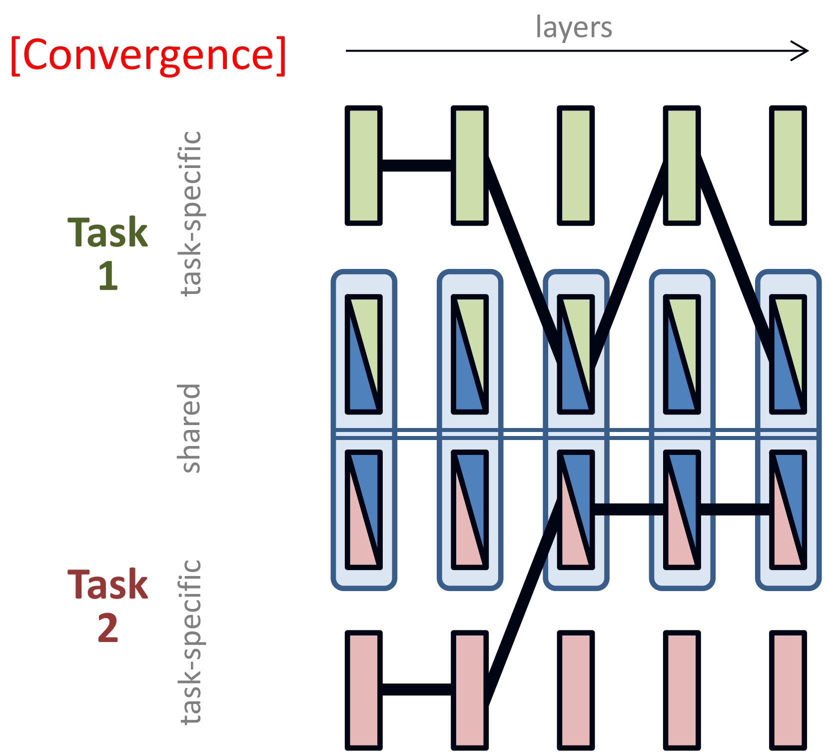
LASEM

- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$



LASEM

- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$



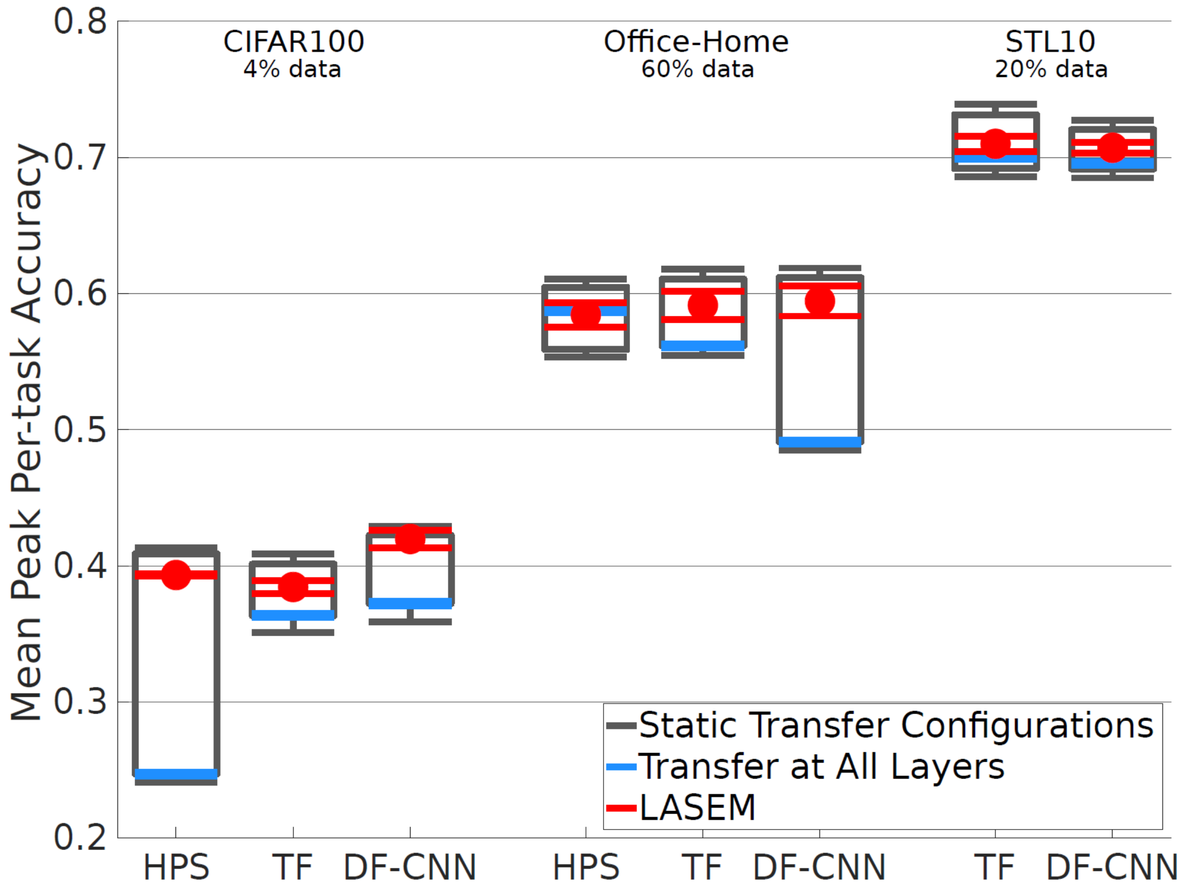
LASEM

- Lifelong Architecture Search via EM algorithm
  - For each new task, initialize transfer-based parameters $\theta_s^{(l)}$ and task-specific parameters $\theta_t^{(l)}$ for layers $l = 1, 2, \cdots, d$
  - (E-step) Estimate posterior probability of transfer configurations
    - prior of configuration $\pi_t(c) = (n_c + 1) / \sum_{\tilde{c}} (n_{\tilde{c}} + 1)$
    - posterior $P(c \mid X_{new}, y_{new}) \propto P(c_{(t)} = c) P(y_{new} \mid X_{new}, c)$
  - (M-step) Update parameters based on the posterior of configurations

$$\theta_s^{(l)} \leftarrow \theta_s^{(l)} + \lambda \sum_{c \in C : c^{(l)} = 1} P(c \mid \mathcal{D}_{new}) \nabla \log \mathcal{L}(\mathcal{D}_{new} \mid c)$$

$$\theta_t^{(l)} \leftarrow \theta_t^{(l)} + \lambda \sum_{c \in C : c^{(l)} = 0} P(c \mid \mathcal{D}_{new}) \nabla \log \mathcal{L}(\mathcal{D}_{new} \mid c)$$

Evaluation: Peak Per-task Accuracy



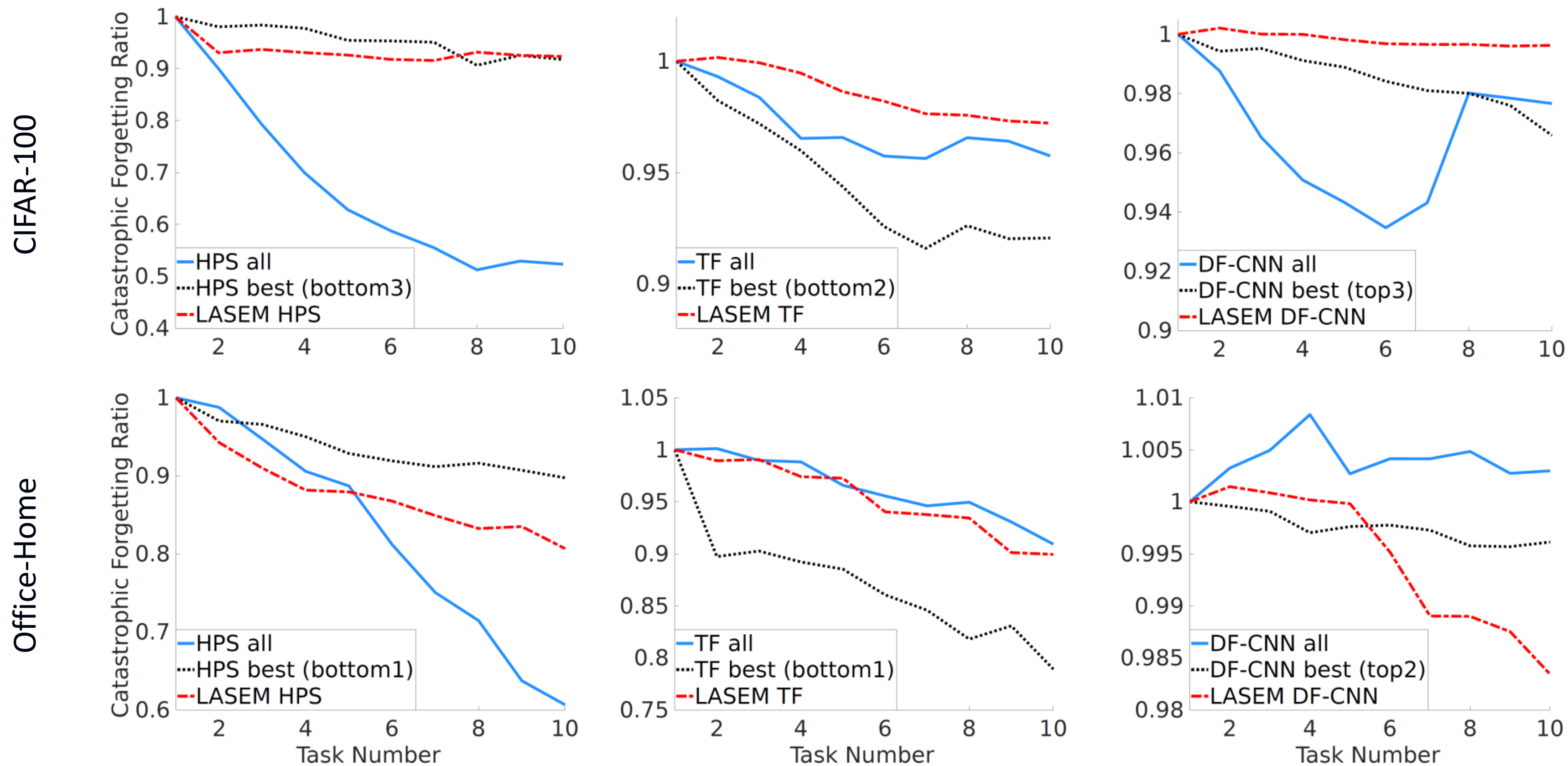
LASEM performs toward the upper range of static transfer configurations

Evaluation: Brute-force Search

Architecture	LASEM	Brute-force Search		Transfer All Layers	
	Accuracy (%)	Accuracy (%)	Relative Time	Accuracy (%)	Relative Time
CIFAR-100 (10 Tasks)					
HPS	39.3 $\pm$ 0.1	40.4 $\pm$ 0.3	6.55	24.7 $\pm$ 0.6	0.78
TF	38.4 $\pm$ 0.5	39.9 $\pm$ 1.1	8.81	36.3 $\pm$ 1.0	0.64
DF-CNN	42.0 $\pm$ 0.6	42.6 $\pm$ 0.7	9.45	36.3 $\pm$ 1.3	0.59
Office-Home (10 Tasks)					
HPS	58.4 $\pm$ 0.9	59.4 $\pm$ 0.2	4.72	54.9 $\pm$ 0.7	0.72
TF	59.1 $\pm$ 1.0	58.7 $\pm$ 0.3	5.22	56.2 $\pm$ 0.7	0.66
DF-CNN	59.5 $\pm$ 1.1	58.8 $\pm$ 0.3	4.04	49.1 $\pm$ 0.6	0.61

LASEM achieves performance of brute-force search 5x – 10x faster

Evaluation: Catastrophic Forgetting



LASEM forgets previous tasks less due to task-specific transfer

Evaluation: Selective Transfer Algos.

Penn Engineering GRASP LABORATORY General Robotics, Automation, Sensing & Perception Lab

Selective Sharing	Accuracy(%)	Forgetting Ratio	Time (k sec)
DEN	48.00 $\pm$ 0.60	0.28 $\pm$ 0.01	55.9
APD-Net	**59.58** $\pm$ **0.45**	0.83 $\pm$ 0.03	21.5
ProgNN	**60.03** $\pm$ **0.45**	1.00 $\pm$ 0.00	96.7
DARTS HPS	45.64 $\pm$ 1.20	0.70 $\pm$ 0.07	43.8
DARTS DF-CNN	56.77 $\pm$ 0.49	0.35 $\pm$ 0.04	33.2
LASEM HPS	58.44 $\pm$ 0.90	0.81 $\pm$ 0.08	70.2
LASEM TF	59.14 $\pm$ 0.80	0.90 $\pm$ 0.04	77.3
LASEM DF-CNN	**59.45** $\pm$ **1.10**	0.98 $\pm$ 0.01	83.2

LASEM achieves high accuracy and low forgetting in comparable time

Evaluation: Scalability

Selective Sharing	Accuracy(%)	Forgetting Ratio	Time (k sec)
CIFAR-100 (10 Tasks)			
ResNet HPS	38.51 $\pm$ 0.53	0.54 $\pm$ 0.03	4.47
LASEM ResNet HPS 4G	39.47 $\pm$ 0.30	0.79 $\pm$ 0.05	11.1
LASEM ResNet HPS 5G	39.07 $\pm$ 1.10	0.79 $\pm$ 0.08	14.4
LASEM ResNet HPS 6G	**40.00** $\pm$ **0.65**	0.75 $\pm$ 0.06	25.1
LASEM ResNet HPS 7G	39.32 $\pm$ 0.33	0.74 $\pm$ 0.07	46.9
CIFAR-100 (40 Tasks)			
ResNet HPS	38.01 $\pm$ 0.27	0.41 $\pm$ 0.02	63.4
LASEM ResNet HPS 4G	**39.89** $\pm$ **0.73**	0.62 $\pm$ 0.03	94.1
LASEM ResNet HPS 5G	38.89 $\pm$ 0.11	0.55 $\pm$ 0.07	109.2
LASEM ResNet HPS 6G	39.17 $\pm$ 0.62	0.56 $\pm$ 0.09	154.1

Group-based LASEM supports deeper nets & longer lifelong scenarios

Summary of Contributions

Penn Engineering
GRASP LABORATORY
General Robotics, Automation, Sensing & Perception Lab

- Investigated the importance of selective layer transfer
- Proposed an EM-based lifelong architecture search algorithm
  - Near-optimal peak per-task accuracy
  - Reduced catastrophic forgetting
  - Enhanced computational efficiency (time/memory)
  - Scalable to deeper architectures and more tasks

Please contact us with questions!


Seungwon Lee
Univ. of Pennsylvania


Eric Eaton
Univ. of Pennsylvania

Correspondence: {leeswon, eeaton}@seas.upenn.edu

Evaluation: Prob. of Selection