NeurIPS 2024
The 38th Annual Conference on Neural Information Processing Systems

DASH: Warm-Starting Neural Network Training in Stationary Settings without Loss of Plasticity

Baekrok Shin*, Junsoo Oh*, Hanseul Cho, Chulhee Yun

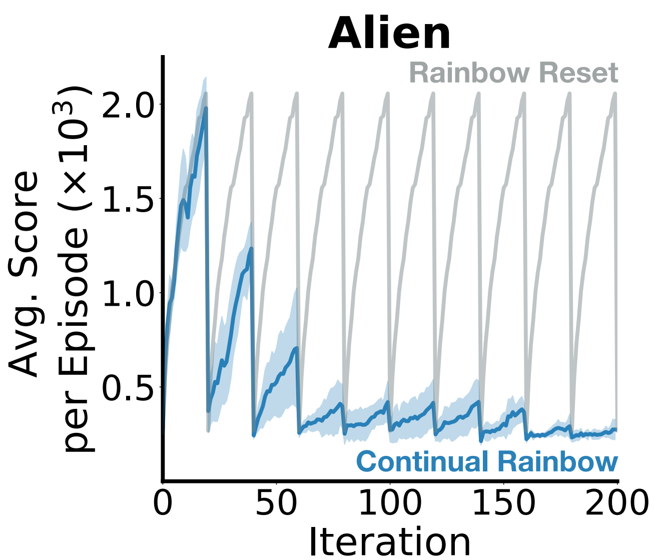
KAIST AI

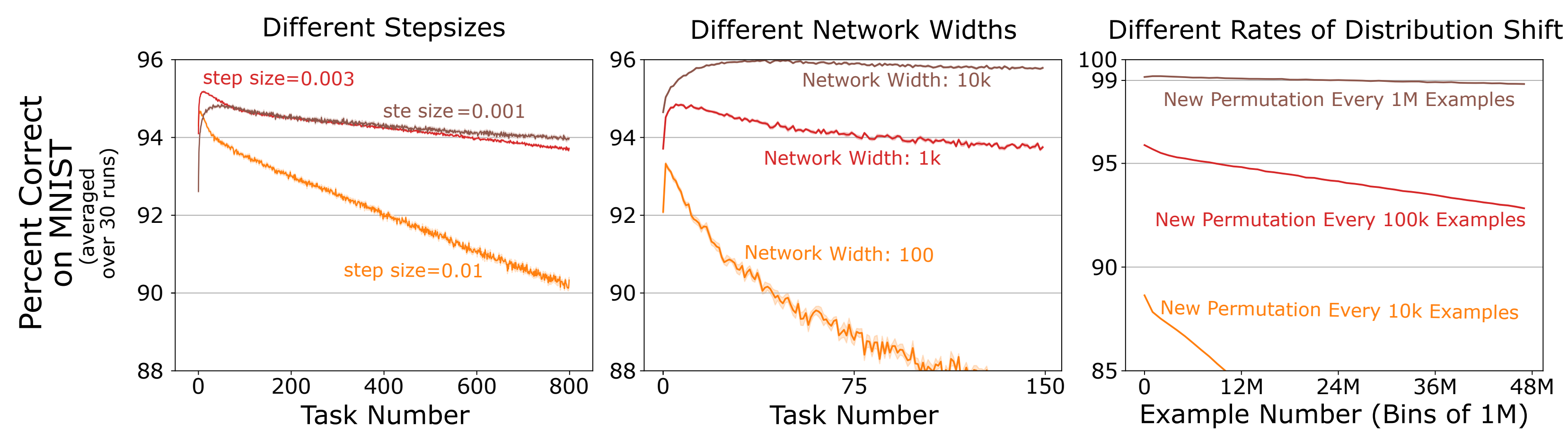
OptiML

KAIST AI
Graduate School of AI

Plasticity of Neural Networks

Under Non-Stationary Data Distribution



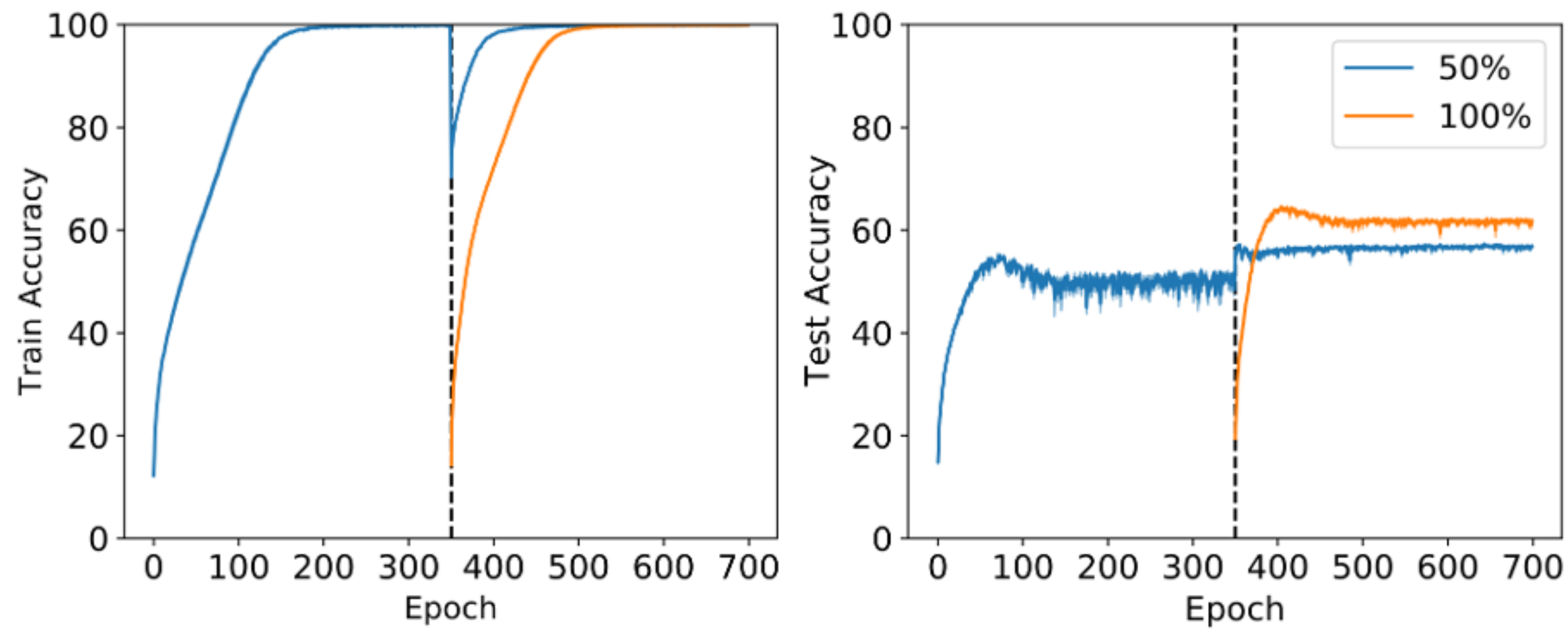


- **Plasticity**: Ability of the model to adapt to new information
- Plasticity loss is often observed in Reinforcement Learning and Continual Learning, where the data distribution is *non-stationary*.

Figures from Abbas et al., (2023), Dohare et al., (2023)

Plasticity of Neural Networks

Under Stationary Data Distribution

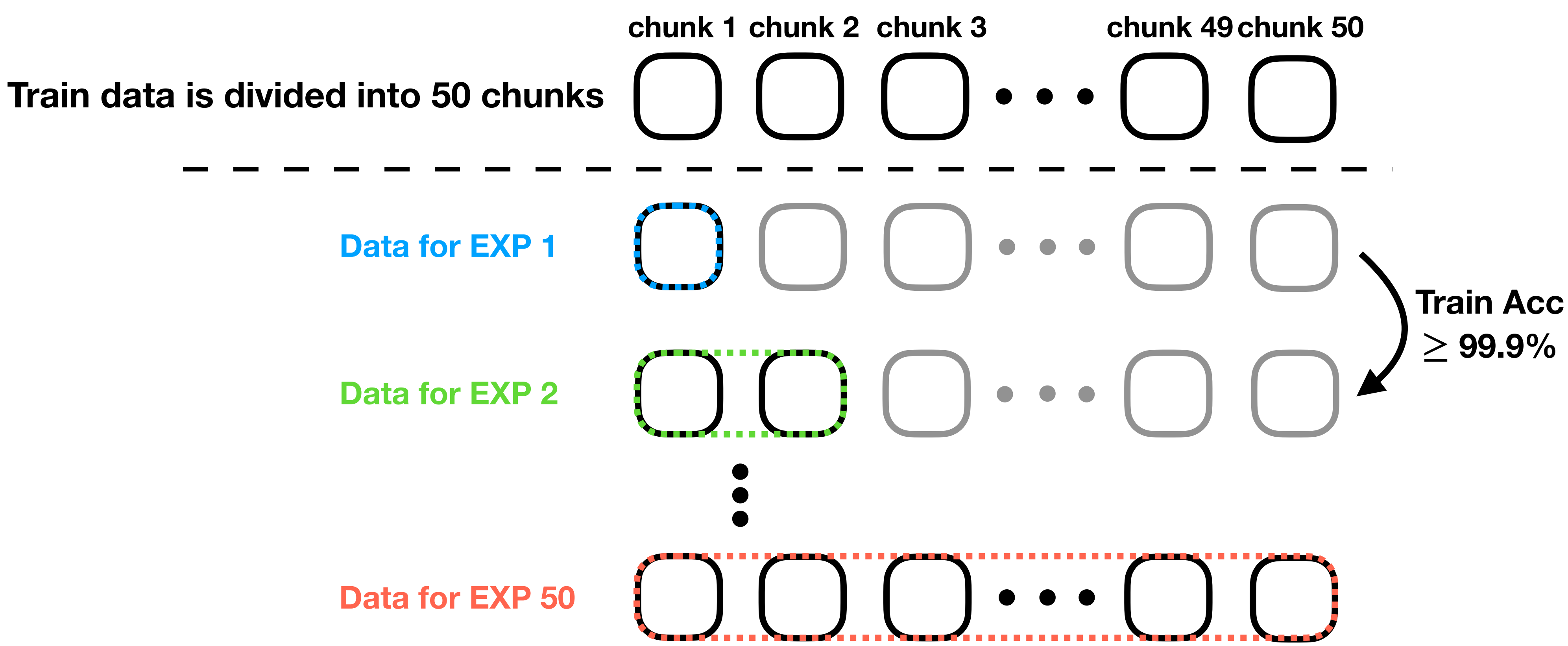


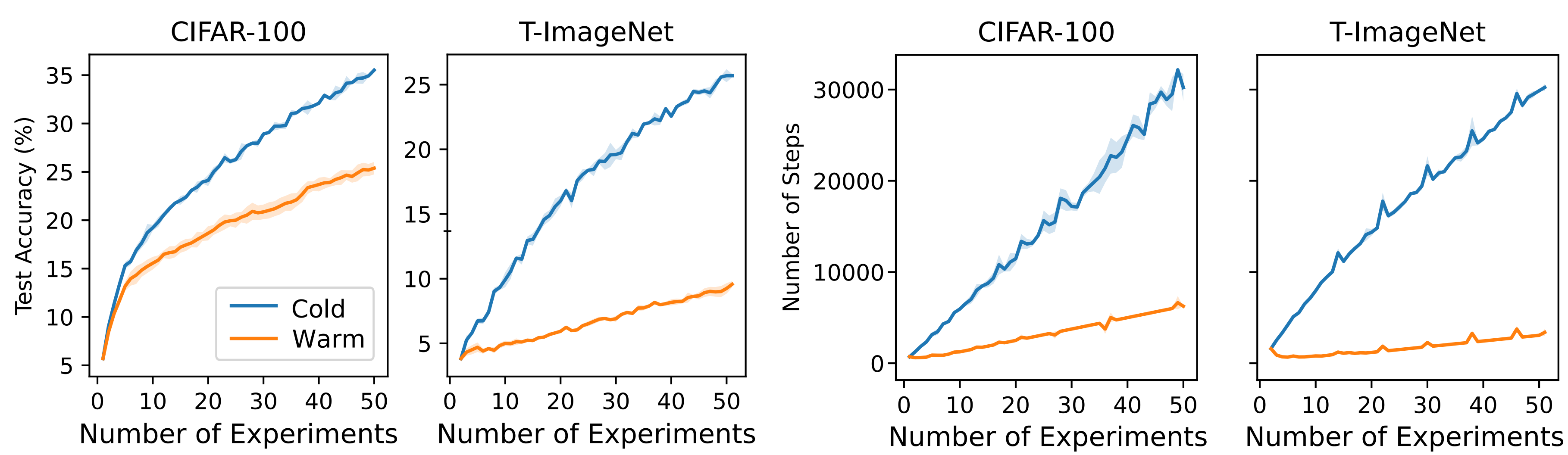
- Surprisingly, models pre-trained on a portion of a dataset and then trained on the full dataset (*warm-start*) tend to generalize worse than models trained from scratch on the full dataset (*cold-start*).

Figure from Ash & Adams, (2020)

Plasticity of Neural Networks

Under Stationary Data Distribution





Warm-Starting vs. Cold-Starting

Theoretical Framework

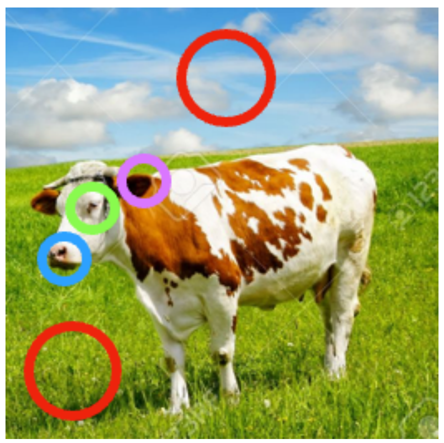


Image data consists of label dependent *features* and label independent *noises*

Features: ears, eyes, mouth, …

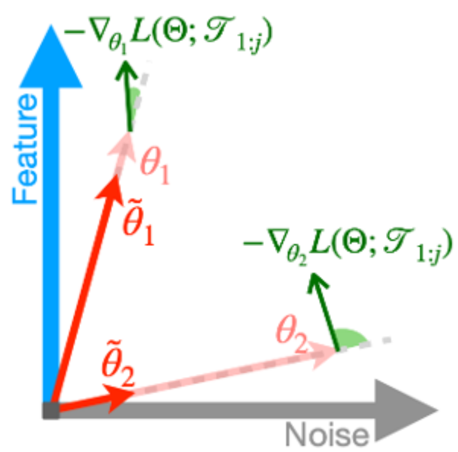
Noises: grass, sky, …

Theoretical Results (Informal)

- When warm-starting, the model cannot learn many features **due to noise memorization** and achieves poor generalization performance.
- When cold-starting, the model forgets the memorized noise, allowing it to learn more features, but it requires longer training time.
- If the model can retain the learned features while forgetting the memorized noise (*ideal method*), it can learn more features while converging faster compared to cold-starting.

DASH: Direction-Aware SHrinking

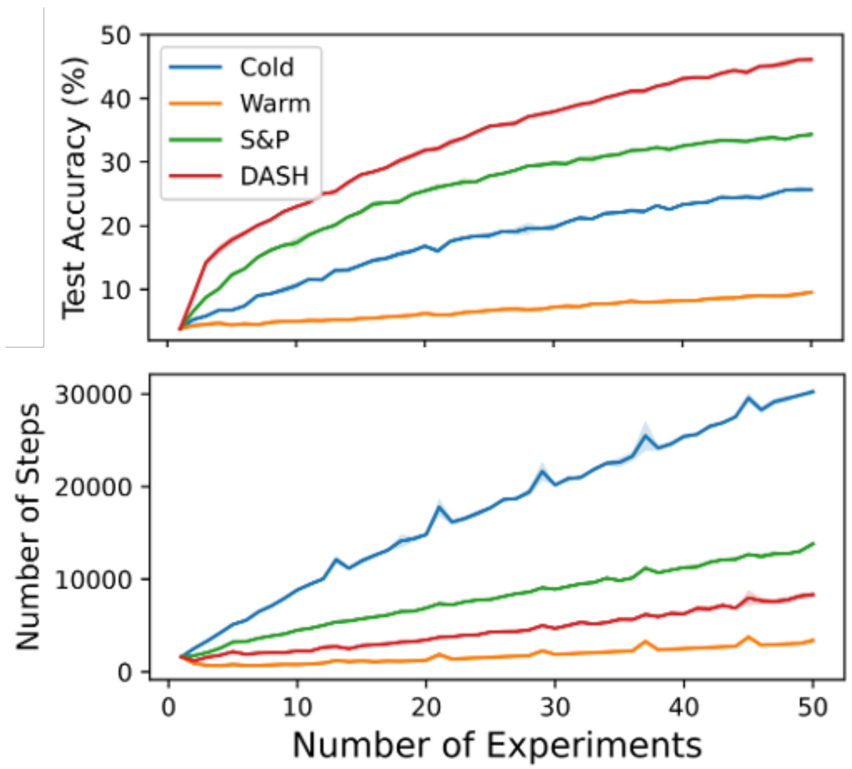
Q. How can this ideal method can be implemented in **real-world neural net training?**



- When new train data $\mathscr{T}_j$ comes in, DASH calculates negative gradient of the loss calculated with train data $\mathscr{T}_{1:j}$
- Then, shrink the weights proportionally to the cosine similarity between the current weight θ and $-\nabla_\theta L$

- Neurons that *learned features*:
  - Show high cosine similarity with new data's negative gradient
  - **Are retained** by not shrinking, preserving learned features
- Neurons that *memorized noise*:
  - Show low cosine similarity with new data's negative gradient
  - **Are shrunk** to forget memorized noise, and this effectively redirects the weight towards feature learning

Experimental Results



ResNet-18	Test Acc at Last Experiment		Number of Steps at Last Experiment		AVG of Test Acc across All Experiments		AVG of Number of Steps across All Experiments	
T-ImageNet	SGD	SAM	SGD	SAM	SGD	SAM	SGD	SAM
Random Init	25.69 (0.13)	31.30 (0.09)	30237 (368)	40142 (368)	17.37 (0.06)	21.95 (0.11)	17503 (53)	22513 (74)
Warm Init	9.57 (0.24)	13.94 (0.37)	3388 (368)	5474 (0)	6.70 (0.04)	9.88 (0.21)	1785 (5)	2773 (7)
S&P	34.34 (0.48)	37.39 (0.18)	13815 (368)	26066 (1606)	25.43 (0.02)	28.47 (0.08)	7940 (15)	13172 (182)
DASH	**46.11 (0.34)**	**49.57 (0.36)**	**8341 (368)**	**12251 (368)**	**33.06 (0.15)**	**35.93 (0.17)**	**4439 (48)**	**7900 (136)**
CIFAR-10								
Random Init	67.32 (0.51)	75.68 (0.39)	**5161 (156)**	17125 (292)	57.66 (0.11)	66.27 (0.13)	2916 (37)	**8121 (26)**
Warm Init	63.53 (0.56)	70.99 (0.59)	1173 (0)	3910 (247)	54.87 (0.18)	63.27 (0.55)	665 (11)	2153 (23)
S&P	81.25 (0.14)	85.53 (0.22)	5395 (625)	32649 (978)	71.74 (0.16)	76.19 (0.04)	**2766 (53)**	15552 (1558)
DASH	**84.08 (0.52)**	**86.75 (0.53)**	6490 (399)	**11886 (2771)**	**75.21 (0.33)**	**77.59 (0.69)**	3454 (55)	8689 (527)
CIFAR-100								
Random Init	35.52 (0.14)	40.27 (0.31)	10557 (247)	14310 (191)	25.72 (0.11)	29.90 (0.06)	5803 (79)	7588 (54)
Warm Init	25.12 (0.59)	32.02 (0.31)	1173 (0)	2346 (0)	19.18 (0.52)	24.01 (0.33)	854 (23)	1294 (12)
S&P	50.08 (0.23)	52.95 (0.36)	4926 (191)	12277 (1226)	37.32 (0.14)	40.36 (0.18)	2929 (27)	**5954 (187)**
DASH	**57.99 (0.28)**	**60.88 (0.29)**	**3519 (0)**	**11730 (1211)**	**43.99 (0.14)**	**46.15 (0.58)**	**2041 (51)**	6675 (797)
SVHN								
Random Init	86.27 (0.46)	89.84 (0.24)	5552 (156)	**10869 (156)**	78.01 (0.10)	83.31 (0.14)	3099 (15)	**5546 (44)**
Warm Init	84.01 (0.41)	88.85 (0.29)	938 (191)	1329 (191)	75.37 (0.50)	81.16 (0.54)	642 (18)	993 (15)
S&P	92.67 (0.17)	94.27 (0.07)	**3597 (156)**	1573 (191)	87.35 (0.14)	89.35 (0.05)	**1858 (12)**	5548 (94)
DASH	**93.67 (0.13)**	**95.19 (0.09)**	5161 (672)	14467 (989)	**89.59 (0.07)**	**91.67 (0.03)**	2619 (68)	8613 (728)

DASH outperforms other baselines in terms of test accuracy while converging faster!