Southwestern University of Finance and Economics (SWUFE)
University of Electronic Science and Technology of China (UESTC)

NEURAL INFORMATION PROCESSING SYSTEMS

Enhancing Knowledge Transfer for Task Incremental Learning with Data-free Subnetwork

Xiaojun Shan
xiaojunshan@std.uestc.edu.cn

Qiang Gao*, Xiaojun Shan*, Yuchen Zhang, Fan Zhou.
the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)

Background

◆ Catastrophic Forgetting & Knowledge Transfer

- ✓ Neuron-wise mask
- ✓ Data-free memory reply

◆ Networks are usually over-parameterized

- ✓ Lottery Ticket Hypothesis
- ✓ Sub-networks

Related Work

NEURAL INFORMATION PROCESSING SYSTEMS

◆ Continual learning

- Regularization-based approaches
- Rehearsal-based approaches
- Architecture-based approaches

◆ Knowledge Transfer

- Bayes model and regression methods
- Mask-based methods
- Few-shot replay methods

Motivations

- **Discover compact subnetworks for (task) incremental learning**
- **Lottery Ticket Hypothesis: a randomly-initialized neural network contains a subnetwork such that, when trained in isolation, can match the performance of the original network.**

➢ Neuron-wise mask

determines which neurons and their corresponding weights should be used for a new coming task

➢ Data-free memory reply

- measure the mask similarity scores
- craft the impressions of the most similar task via data-free memory replay

DSN : Enhancing Knowledge Transfer for Task Incremental Learning with Data-free Subnetwork

Challenges

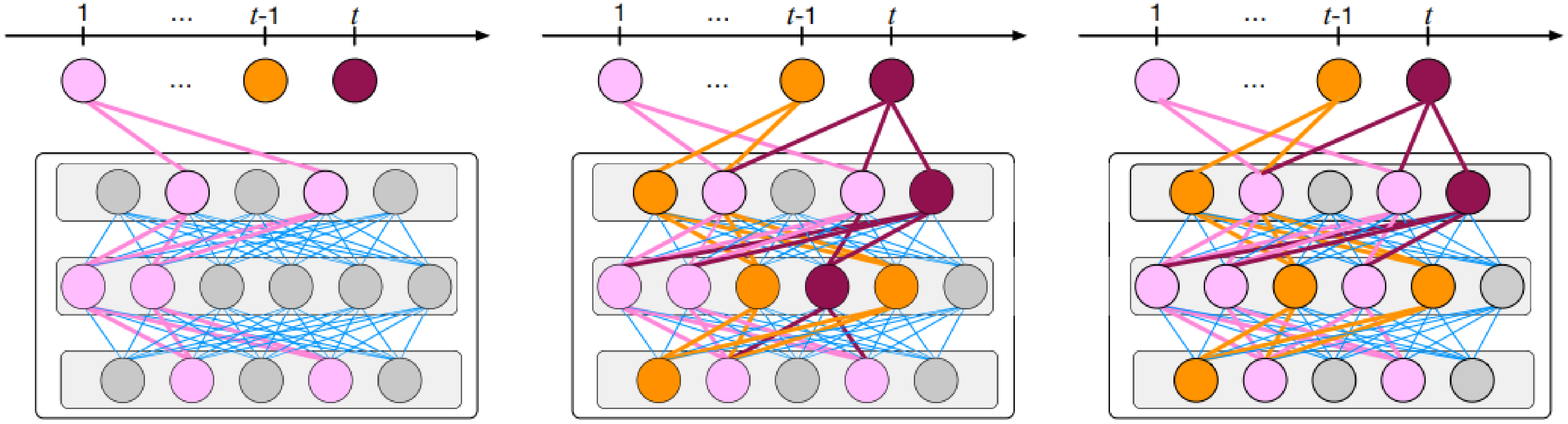
- Catastrophic Forgetting
- Fail to obtain a subnetwork for each corresponding task
- Backward knowledge transfer is not considered

DSN

NEURAL INFORMATION PROCESSING SYSTEMS

Neuron-wise Mask:

- **Layer mask** $\boldsymbol{m}_t^l \in \boldsymbol{m}_t$: $\boldsymbol{m}_t^l = \sigma(\gamma \cdot \boldsymbol{e}_t^l),$
- **Forward** : $\boldsymbol{h}_t^l = \boldsymbol{h}_t^l \odot \boldsymbol{m}_t^l,$
- **Backward:** $\theta_{lij} = \theta_{lij} - \frac{\partial \mathcal{L}}{\partial \theta_{lij}} \odot \max(m_t^{l,i}, m_t^{l-1,j}),$



(a) Dense Initialization (b) Selection for Task t (c) Consolidation for Task 1

Data-free Replay:

Insights

- **A class similarity matrix Mt describe the correlation between different classes.**
- **Model outputs representation sampled form Dirichlet distribution**

for $c = 1 : C_{argmax(S_t)}$ **do**
  Set the concentration parameter $\boldsymbol{\alpha}^c = M^c_{\text{argmax}(S_t)}$;
  for $b = B_1, B_2, \cdots, B_{C_{argmax(S_t)}}$ **do**
    for $i = 1 : b$ **do**
      Sample $\hat{\boldsymbol{o}}^c_{\text{argmax}(S_t)} \sim Dir(C_{\text{argmax}(S_t)}, \beta_b \times \boldsymbol{\alpha}^c)$;
      Initialize $\hat{x}^{c,i}_{\text{argmax}(S_t)}$ to random noise and craft $\hat{x}^{c,i}_{\text{argmax}(S_t)}$ via Eq.$\boxed{9}$;
      $\mathcal{I}_{\text{argmax}(S_t)} \leftarrow \mathcal{I}_{\text{argmax}(S_t)} \cup \hat{x}^{c,i}_{\text{argmax}(S_t)}$;

$$\hat{x}^{c,i}_t = \text{argmin} \mathcal{L}_{IC}(\mathcal{H}(\cdot, \boldsymbol{\theta}(\boldsymbol{n} \odot \boldsymbol{m}_t), \tau), \hat{\boldsymbol{o}}^c_t), \tag{9}$$

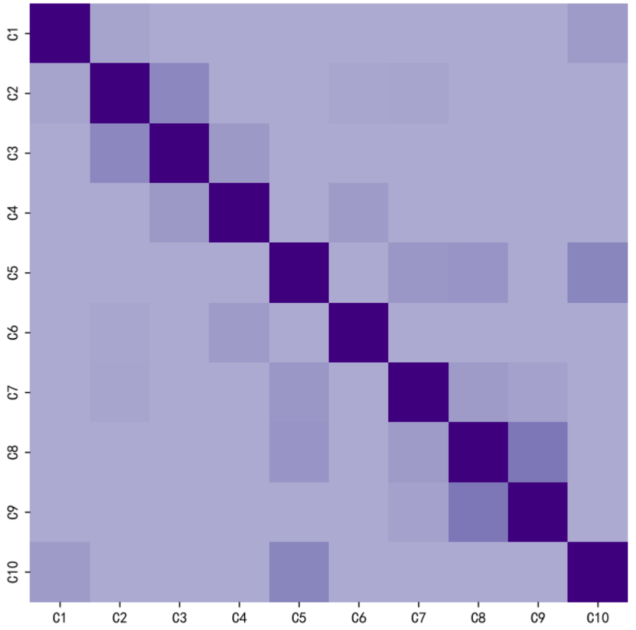


Figure 7: Confusion Matrix of the first task in TinyImageNet.

Evaluation

NEURAL INFORMATION PROCESSING SYSTEMS

Overall Performance

Table 1: Performance comparison of the proposed method and baselines on four datasets.

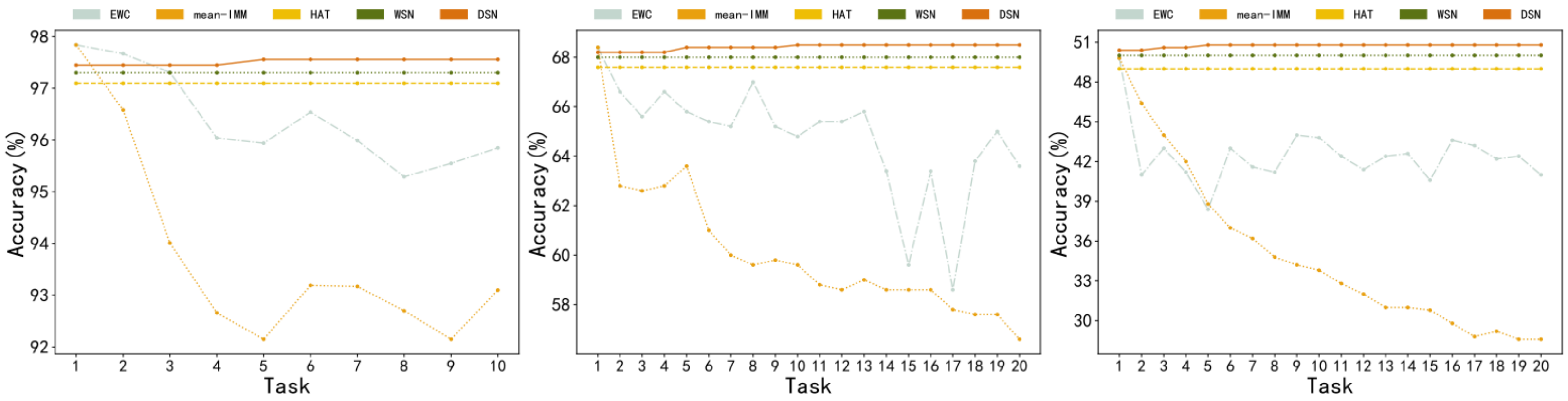
Model	PMNIST			RMNIST			CIFAR-100			TinyImageNet		
	ACC(%)	BWT(%)	Trans(%)	ACC(%)	BWT(%)	Trans(%)	ACC(%)	BWT(%)	Trans(%)	ACC(%)	BWT(%)	Trans(%)
SGD	81.37	-24.52	-17.06	72.83	-25.32	-25.08	59.82	-24.09	-24.02	30.24	-19.12	-19.96
EWC	94.20	-0.32	-4.23	94.86	-0.73	-3.05	67.15	-8.61	-16.69	40.85	-5.24	-9.35
mean-IMM	80.10	-1.13	-18.33	88.81	-0.96	-9.10	56.08	0.23	-27.76	30.10	-3.21	-20.10
mode-IMM	93.13	-4.17	-5.30	89.48	-7.40	-8.43	61.22	-21.49	-22.62	32.26	-19.02	-17.94
PGN	91.89	0.00	-6.54	90.01	0.00	-7.90	53.84	-14.66	-30.00	24.47	-12.12	-25.73
DEN	91.96	-0.41	-6.47	91.53	-0.52	-6.38	59.32	-1.24	-12.79	33.86	-1.30	-3.88
RCL	92.28	0.00	-6.15	93.97	0.00	-3.94	61.77	0.00	-22.07	38.23	0.00	-11.79
HAT	97.10	0.00	-1.33	97.49	0.00	-0.42	71.23	0.00	-12.61	44.51	0.00	-5.69
SupSup	97.02	0.00	-1.41	97.15	0.00	-0.73	71.44	0.00	-12.40	43.22	0.00	-6.98
WSN	97.16	0.00	-1.27	97.32	0.00	-0.59	72.84	0.00	-11.00	45.96	0.00	-4.24
DSN	**98.24**	**0.01**	**-0.19**	**97.73**	**0.02**	**-0.18**	**75.17**	**0.02**	**-8.67**	**46.56**	**0.04**	**-3.64**

- DSN consistently outperforms all baselines regarding *ACC, BWT, and Trans(%)*
- DSN is the first to exceed 0 regarding *BWT*

Evaluation

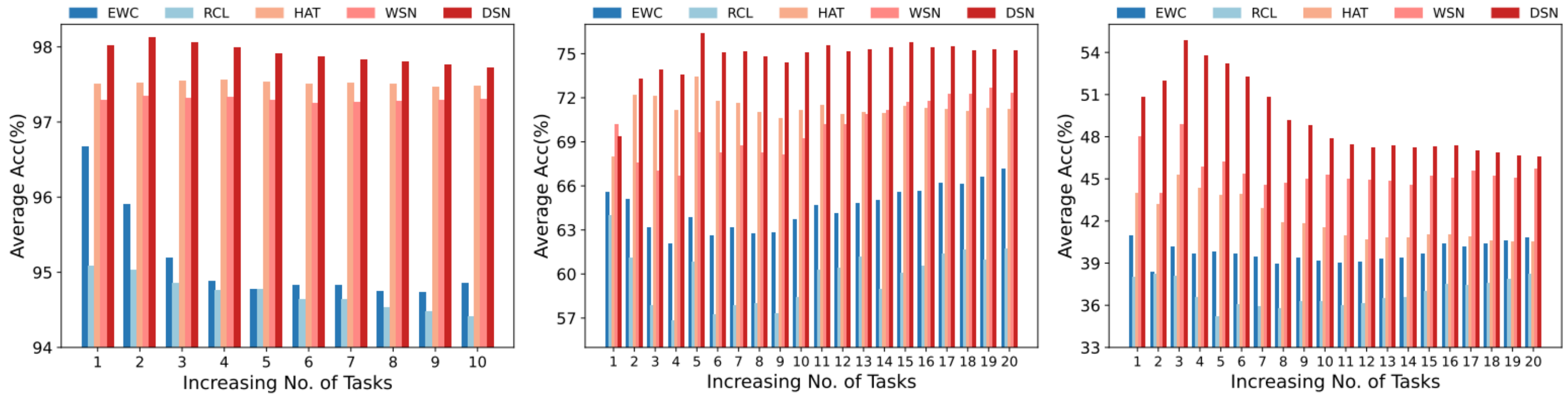
NEURAL INFORMATION PROCESSING SYSTEMS

The accuracy performance of the first task in incremental learning



- When new tasks arrive, DSN is the only one to perform better on the first task

The accuracy performance during entire incremental learning

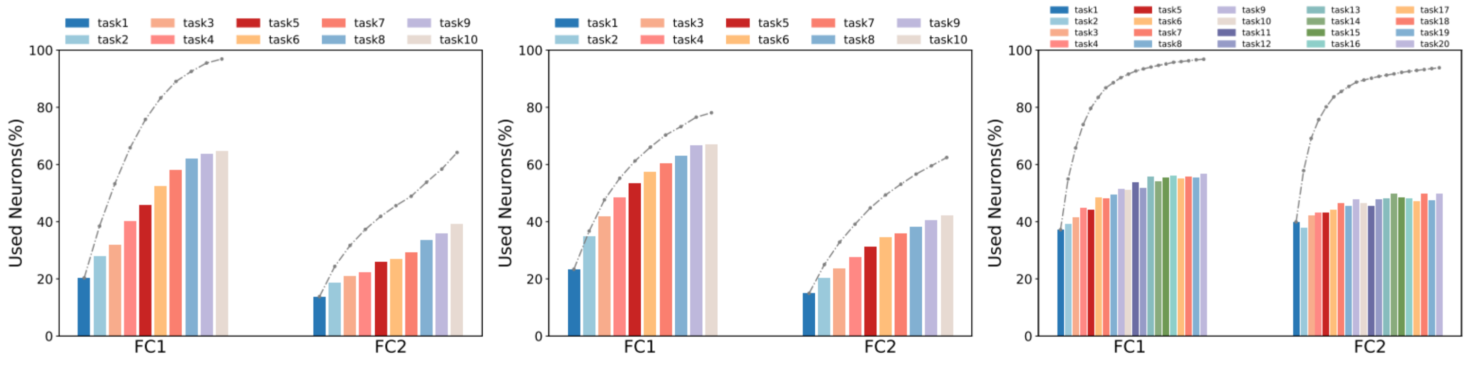


(a) RMNIST (b) CIFAR100 (c) TinyImageNet

- DSN outperforms other baselines during the entire incremental learning process

Evaluation

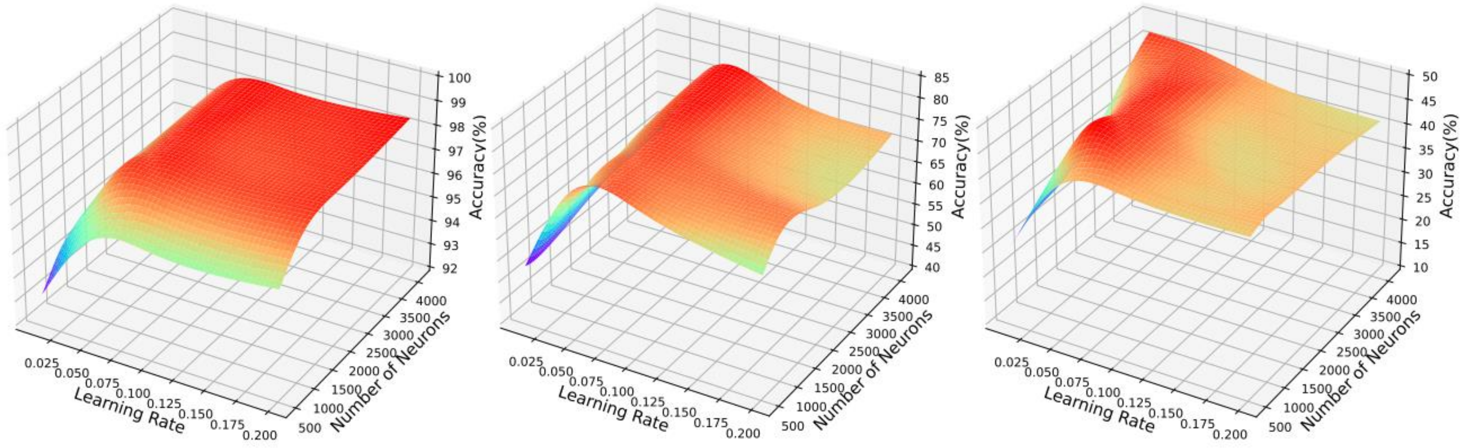
The layer-wise neuron usage in incremental learning



(a) PMNIST (b) RMNIST (c) TinyImageNet

- **DSN prefers to reuse more neurons from earlier tasks when a new task arrives**

Hypernetwork capacity in incremental learning varying different learning rates



(a) RMNIST (b) CIFAR-100 (c) TinyImageNet

Thank you!

Q&A

Xiaojun Shan
xiaojunshan@std.uestc.edu.cn