Life Long Learning for Handwritten Digits Recognition

Shen Ruoyue
20M38216

Handwritten Digits Dataset

- 4 volunteers to write in a square box with 7 different virtual pen types: [Pencil, GelPen, Paintbrush, Pigment, Pen, WaterColorBrush, Crayon].

- Every writer need to write one digit using 7 different kinds of pen for 3 times, so the dataset has in total $4 \times 3 \times 10 \times 7 = 840$ images annotated with digit class, pen type and writer name, with a size of 410×410.

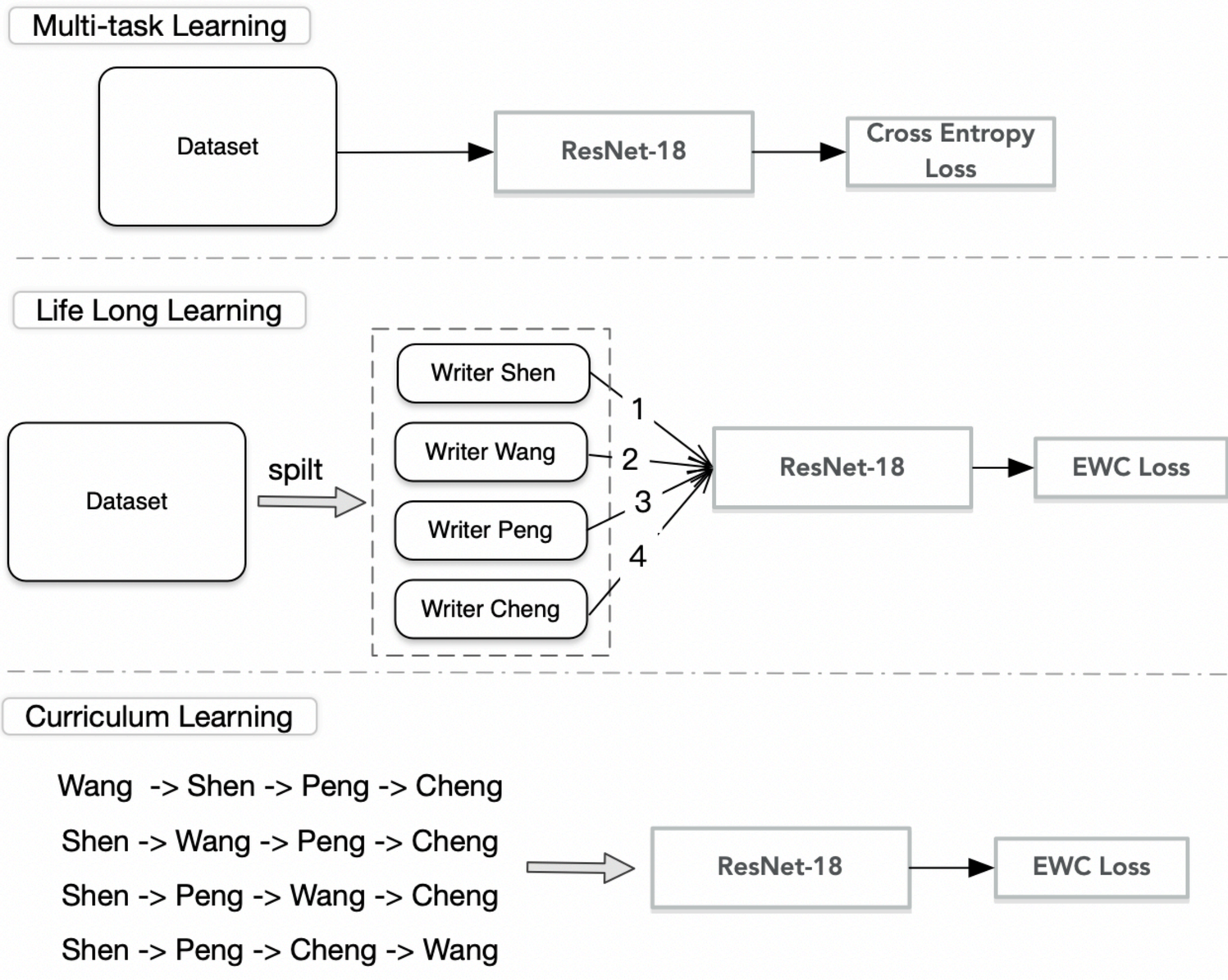
```
FileName,PenName,Writer
U0030/U0030_00001.png,Pencil,Shen
U0030/U0030_00002.png,Pencil,Shen
U0030/U0030_00003.png,Pencil,Shen
U0030/U0030_00004.png,Pen,Shen
U0030/U0030_00005.png,Pen,Shen
U0030/U0030_00006.png,Pen,Shen
U0030/U0030_00007.png,GelPen,Shen
U0030/U0030_00008.png,GelPen,Shen
U0030/U0030_00009.png,GelPen,Shen
U0030/U0030_00010.png,Paintbrush,Shen
U0030/U0030_00011.png,Paintbrush,Shen
U0030/U0030_00012.png,Paintbrush,Shen
```

Life Long Learning

- Life long learning wants to use a unified network structure to train separately on different tasks, and the network can also be competent for all tasks.

Curriculum Learning

- The order of tasks will affect the effect of learning, focus on the relationship between each task to find the proper learning order.



Life Long Learning

ResNet-18 + Elastic Weight Consolidation (EWC)

- **ResNet**

$$H(x) = F(x) + x$$

- **EWC**

$$L'(\theta) = L(\theta) + \lambda \sum_i b_i \left(\theta_i - \theta_i^b\right)^2$$

$$b_i = \frac{1}{|D|} \sum_{d \in D} \frac{\partial^2 L(d, \theta)}{\partial \theta^2}$$

Experiment

Baseline Method

- **Multi-task Learning:** Train a multi-task model using ResNet with the whole dataset as upper bound.
- **Basic ResNet:** Use four sub-tasks to train the model continuously and see the performance on different test set.
- **ResNet with EWC:** Replace the original loss function with the EWC loss function to implement life long learning.

Metrics

- Matrix: Item $R_{i,j}$ means the performance on test set of task j after training on task i.
- Average Accuracy: $\mathrm{ACC} = \frac{1}{T}\sum_{i=1}^{T} R_{T,i}$
- Backward Transfer: $\mathrm{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1} R_{T,i} - R_{i,i}$
- Forward Transfer: $\mathrm{FWT} = \frac{1}{T-1}\sum_{i=2}^{T} R_{i-1,i} - \bar{b}_i$

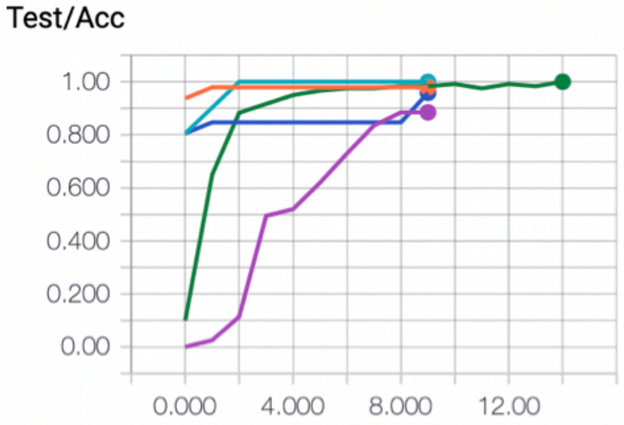
Experiment

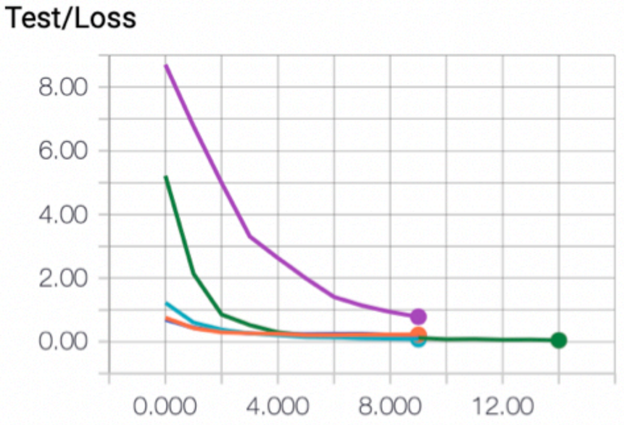
Setting

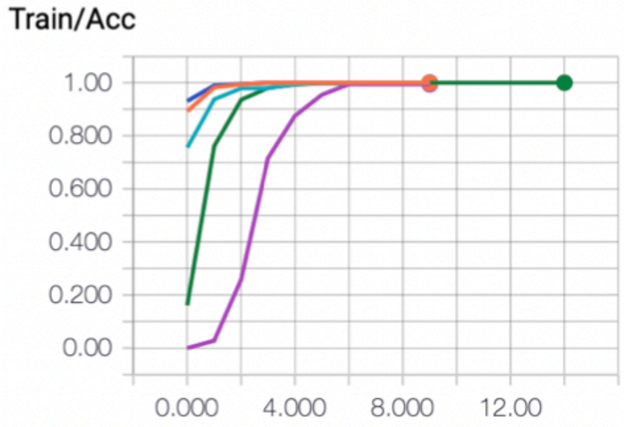
- Split dataset into sub-tasks according to the author and get 4 sub-tasks.
- Resized images to 256 × 256, and center cropped to 224 × 224.
- Adam with the fixed learning rate 1e-5 and no weight decay.
- Mini-batch size is 8 and every task is trained for 10 epochs.
- Basic loss function for classification is cross entropy loss, for life long learning, replace it with the EWC loss function.
- Repeat the experiment for each setting for 5 times, and get the average values and range.

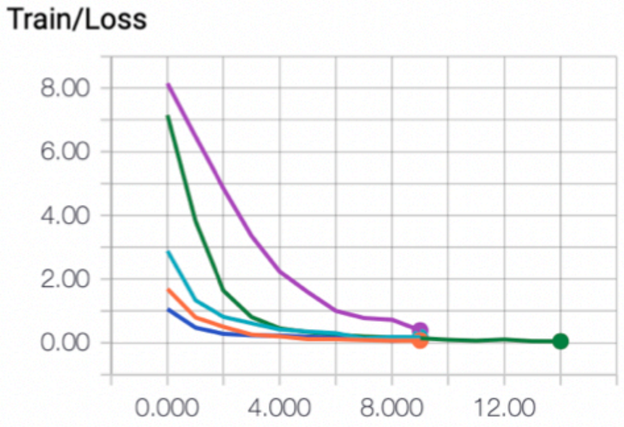
Experiment

Life Long Learning vs Multi-task Learning





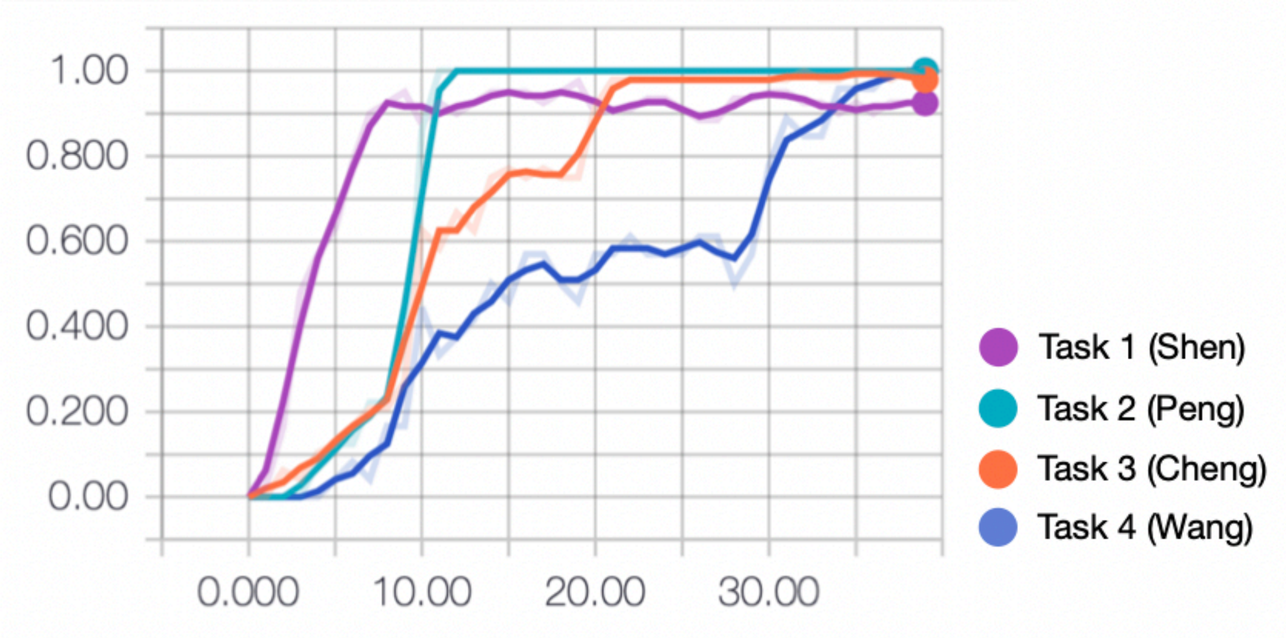




Life Long Learning: converge slower for 1st task; for the subsequent tasks, the accuracy is high and the loss is small, indicating that the knowledge learned from the first task can be well transferred to the subsequent tasks.

Experiment

Variation of the accuracy using (EWC)



Training on previous task can improve the performance on future tasks, and training on later task doesn’t affect performance on previous tasks.

Experiment

Table 1. Experiment result

Model	Task Order	Average Accuracy	Backward Transfer	Forward Transfer
ResNet with EWC	Wang_Shen_Peng_Cheng	0.9598 ± 0.0046	0.0594 ± 0.0122	0.7536 ± 0.0115
	Shen_Wang_Peng_Cheng	0.9439 ± 0.0098	-0.0122 ± 0.0325	0.6755 ± 0.0120
	Shen_Peng_Wang_Cheng	0.9549 ± 0.0132	-0.0156 ± 0.0404	0.5671 ± 0.0278
	Shen_Peng_Cheng_Wang	0.9694 ± 0.0181	0.0055 ± 0.0148	0.5505 ± 0.0389
Basic ResNet	Wang_Shen_Peng_Cheng	0.9565 ± 0.0215	0.0623 ± 0.0145	0.7649 ± 0.0103
	Shen_Wang_Peng_Cheng	0.9389 ± 0.0076	0.0232 ± 0.0381	0.6833 ± 0.0097
	Shen_Peng_Wang_Cheng	0.9509 ± 0.0105	-0.0091 ± 0.0456	0.5722 ± 0.0088
	Shen_Peng_Cheng_Wang	0.9613 ± 0.0167	0.0041 ± 0.0174	0.5324 ± 0.0347
Multi-task Learning	Whole Dataset	0.9958 ± 0.0042	-	-

- **Life Long Learning**: EWC indeed improves ACC for all the tasks. But for BWT and FWT, EWC methods only outperforms the basic one in the last task order, which means that EWC is not so efficient on handling the forgetting problem and knowledge transferring in this dataset.

Experiment

Table 1. Experiment result

Model	Task Order	Average Accuracy	Backward Transfer	Forward Transfer
ResNet with EWC	Wang_Shen_Peng_Cheng	0.9598 ± 0.0046	0.0594 ± 0.0122	0.7536 ± 0.0115
	Shen_Wang_Peng_Cheng	0.9439 ± 0.0098	-0.0122 ± 0.0325	0.6755 ± 0.0120
	Shen_Peng_Wang_Cheng	0.9549 ± 0.0132	-0.0156 ± 0.0404	0.5671 ± 0.0278
	Shen_Peng_Cheng_Wang	0.9694 ± 0.0181	0.0055 ± 0.0148	0.5505 ± 0.0389
Basic ResNet	Wang_Shen_Peng_Cheng	0.9565 ± 0.0215	0.0623 ± 0.0145	0.7649 ± 0.0103
	Shen_Wang_Peng_Cheng	0.9389 ± 0.0076	0.0232 ± 0.0381	0.6833 ± 0.0097
	Shen_Peng_Wang_Cheng	0.9509 ± 0.0105	-0.0091 ± 0.0456	0.5722 ± 0.0088
	Shen_Peng_Cheng_Wang	0.9613 ± 0.0167	0.0041 ± 0.0174	0.5324 ± 0.0347
Multi-task Learning	Whole Dataset	0.9958 ± 0.0042	-	-

- **Curriculum Learning**: Train the dataset with meaningful order (Shen-Peng-Cheng-Wang), EWC method can handle the forgetting problem and knowledge transfer- ring to some extent. Besides, this order achieved the best results in average accuracy.

Thank you for your attention!
Q & A

Shen Ruoyue
20M38216