Life Long Learning for Handwritten Digits Recognition

Shen Ruoyue 20M38216
Tokyo Institute of Technology
ruoyue@ks.c.titech.ac.jp

Abstract

Handwritten digit recognition is a fundamental problem in Machine Learning. The appearance of various advanced models has made the performance on this task perfect. We proposed a new dataset in which 4 volunteers write digits with 7 different pens. It contains 840 images and has three types of annotations: class label, pen type and writer name. We believe these annotations can promote research on tasks other than simple handwritten digit recognition. We then present the task of life long learning on this dataset, which requires the model to continuously learn new tasks without performance degradation. Comprehensive experiments and analyses demonstrate the effectiveness of life long learning and curriculum learning and how the annotations in our dataset can contribute to other new tasks.

1. Introduction

Handwritten digit recognition is a fundamental problem in the field of Machine Learning (ML) and Computer Vision (CV). This task uses a picture of a handwritten digit as input and outputs the class which the digit belongs to from 0-9. In 1998, Yann LeCun et al. reviewed various methods applied to handwritten character recognition in [5], whose results showed that Convolutional Neural Networks (CNN) outperformed all other models. Since then, neural networks and the MNIST dataset [5] have become the baseline for handwritten digit recognition.

With the rapid development of deep learning, the accuracy of handwritten digit recognition was pushed very high, even almost reaching 100%. It is difficult to make a breakthrough in recognizing categories of handwritten digits, but there are still many topics related to handwritten digit recognition that can be studied. For example, infer a person's personality through this person's writing habits, use other annotations except the category to do some more complex and novel researches, and so on.

Therefore, we proposed a new dataset in which 4 volunteers write digits with 7 different pens. In addition to the category annotations, the dataset also has pen types and writer name annotations. We believe that these annotations can promote research on tasks other than handwritten digit recognition.

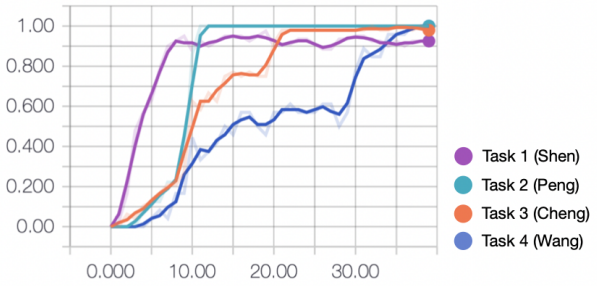

Figure 1. Variation of accuracy rate. It's obvious that training on previous task can improve the performance on future tasks, and training on later task doesn't affect performance on previous tasks.

With such annotated dataset, we then propose the task of life long learning on handwritten digits. Dividing the dataset according to different writers, we can obtain 4 different sub-tasks. By continuously training the model on each task and observing its performance on the test set of all tasks, we can measure the effectiveness of the life long learning and curriculum learning. Fig. 1 shows the variation of the accuracy rate when using the elastic weight consolidation(EWC) [3] method for life long learning.

To summarize, our contributions are:

1. We collect a new handwritten digits dataset, containing 721 training samples, 121 testing samples. It not only has labels for classification, but also provides 7 different pen types and 4 different writers' annotation information, which can be conveniently used for other tasks.

2. We implement life long learning in practice on our proposed dataset, using the elastic weight consolidation method (EWC [3]), as well as designing sufficient experiments to demonstrate the effectiveness and importance of curriculum learning.

2. Related Work

2.1. Image Classification

Image classification is a core task in computer vision, which takes an image as input, and output the class of this image from a predefined category set. MNIST dataset [5] is a typical supervised dataset for image classification, which contains 60,000 training images and 10,000 testing images, with size 28×28. Since the introduction of AlexNet [4] in 2012, deep convolutional neural networks (DCNN) have led to a series of breakthroughs for of image classification. Since then, various new architectures have been proposed, like VGG [8], GoogLeNet [9]. But deeper neural networks are more difficult to train because of the degradation problem. ResNet [2] imported identity mapping to solve this problem properly and lift the performance on CV tasks to a new level.

2.2. Life Long Learning

Life Long Learning With rapid development of machine learning(ML), there are many powerful network structures which can solve many basic tasks like image classification, semantic segmentation and so on. However, a certain network can only handle a given task, where is still a long distance from the real artificial intelligence (AI) as human expected. 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. Elastic weight consolidation(EWC) [3] is a typical solution for the catastrophic forgetting problem in life long learning. It's basic idea is that only change the parameters unimportant to previous tasks, and protect the important ones from being changed by regularizing the loss function.

Multi-task Learning Life long learning only uses the data of the current task for training; while multi-task learning requires the data of all previous tasks, which may cause problems with increased data storage and calculations with the growth of task amount. So what life long learning actually wants to solve is how to let the model learn every different tasks well without using multi-task learning. Experiment shows that using multi-task learning allows the model to learn all tasks and achieve good results, so the results of multi-task learning are often used as the upper bound of life-long learning.

Transfer Learning Both transfer learning and life long learning want to take advantage of knowledge learned on previous tasks to improve the performance of future tasks. Although transfer learning only focuses on the performance on the current task, life long learning needs to consider the performance on all tasks, requiring the model not forgetting the previous tasks. In other words, life long learning includes transfer learning to achieve the transformation of knowledge, but life long learning also wants to achieve the retention of knowledge.

Figure 2. Some examples from our dataset

2.3. Curriculum Learning

The main idea of Curriculum Learning [1] is that like the learning process of human and animals, the order of tasks will indeed affect the effect of learning. If we organize the tasks with a meaningful order, it's easier for the model to find a better local optimum, and at the same time speed up the training.

Since the order of training tasks will have a great impact on the final result, curriculum learning is a problem that needs to be considered for life long learning. A recent work Taskonomy(task+taxonomy) [10] focus on the relationship between each task to find the proper learning order.

3. Dataset

In this section, we describe our proposed dataset, a new handwritten digit dataset not only has labels for classification, but also provides seven different pen types and four different writers' annotation information. Therefore, it can be conveniently used for other tasks like pen recognition, writer recognition and so on.

3.1. Data Collection

To collect the data, we asked four people to write in a square box with seven different pen types: [Pencil, GelPen, Paintbrush, Pigment, Pen, WaterColorBrush, Crayon]. For the ease of data collection and the quality of samples, we ask the writer to write on iPad with Apple Pencil using the App called Sketches, and unified the transparency and size parameters in the software for different writers.

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. On this basis, we simply divided

the dataset into a training set with 721 samples and a test set with 121 samples at a ratio of 1:4. When splitting, preserve the percentage of samples for each class. Besides, all the images have a size of 410×410.

3.2. Dataset Analysis

As mentioned above, this dataset has three kinds of annotations: class label, pen type and writer name. So it's reasonable to support different tasks like pen type recognition, writer recognition and so on.

Although this dataset is still small in size, we ensure the diversity in the process of data collection. As shown in Fig. 2, samples in the same row is written by the same person, and we can see the difference between different types of pens. Not only the same digit written by different people differ greatly, sometimes the same digit written by the same person is different. What is more interesting is the last column, although the same parameters are controlled when using the software, different people have different writing habits and stress, the effects of writing with the same pen are also different.

4. Methods

In order to apply life long learning to this dataset, first we need to get a base network for image classification. Here we just use ResNet [2] pretrained on ImageNet [7] as our classification network. For life long learning, we utilize elastic weight consolidation method (EWC [3]), a typical solution to avoid the catastrophic forgetting problem by only changing the parameters unimportant to previous tasks.

4.1. ResNet

When it comes to a deep neural network, due to the chain rule of the derivative, multiplying gradients less than 1 will make the gradient smaller and smaller, eventually leading to a gradient of 0 in some layer. The parameters of the first few layers will no longer be updated.

In order to solve this problem, ResNet [2] introduces residual connections, which adding an identity mapping in the network structure to change the feature into $H(x) = F(x) + x$. This formulation can be realized by feedforward neural networks with shortcut connections as in Fig. 3.

There are 5 different depth structures for ResNet in [2], which is 18, 34, 50, 101, and 152. For our task, the amount of data is small, and does not require a very complex model to achieve good results. For computational simplicity, we used ResNet-18 for the classification network. Besides, we also pretrain it on ImageNet [7], because without pretraining the model is very unstable.

4.2. EWC

Elastic weight consolidation(EWC) [3] is a typical solution for the catastrophic forgetting problem in life long learning, whose basic idea is that only change the parameters unimportant to previous tasks, and protect the important ones from being changed by regularizing the loss function. So the loss function we need to minimize is:

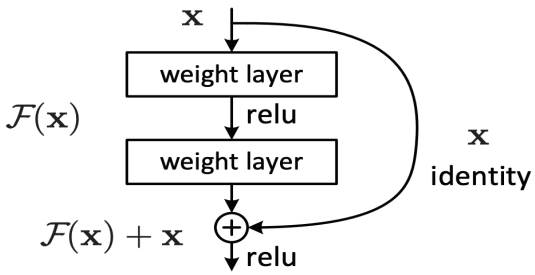

Figure 3. Shortcut connections

$$L'(\theta) = L(\theta) + \lambda \sum_i b_i \left(\theta_i - \theta_i^b\right)^2 \tag{1}$$

where $L(\theta)$ is the loss function for the current task, the latter summation term is used to limit the change of parameters, like a regular item. λ is the weight showing how important the old task is compared to the new one, while b_i is a guard, indicating the importance of parameter θ_i to the previous task. The larger the b_i, the more important θ_i is to the previous task, the smaller it changes in learning new tasks.

In EWC method, the importance of parameters is evaluated using second-order derivatives:

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

where D is the dataset, L is the loss function.

5. Experiments

5.1. Dataset

We initialize the ResNet-18 weights from a pretrained model on ImageNet [7] released by Pytorch, because without pretraining the model is very unstable.

For training and testing, we only use the dataset we proposed, containing 721 training samples, 121 testing samples.

For life long learning and curriculum learning, we need different sub-tasks. In order to measure the effect of different models fairly, when dividing the data into different sub-tasks, we use the same training set and test set divided in multi-task learning. Then divide it into different tasks according to the writer name annotation. In this dataset, we have four writers: [Wang, Shen, Peng, Cheng], so we can get four sub-tasks for life long learning. Task 1(Writer Shen) has 173 training samples and 37 testing samples, Task

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	-	-

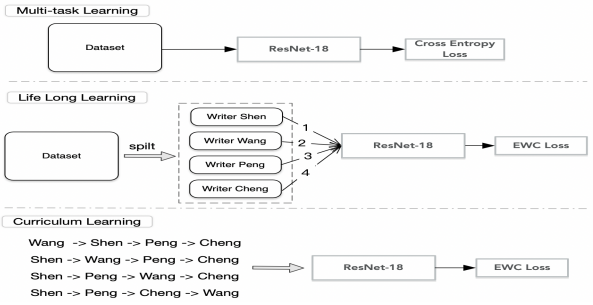

Figure 4. Baseline Models Training

2 (Writer Peng) has 188 training samples and 22 testing samples, Task 3 (Writer Cheng) has 168 training samples and 42 testing samples, Task 4 (Writer Wang) has 191 training samples and 19 testing samples.

3.

5.2. Baseline Method

The training process of different baseline models is shown in Fig. 4.

Multi-task Learning As mentioned before, the results of multi-task learning are often used as the upper bound of life-long learning. So here we train a multi-task model using ResNet with the whole dataset.

Basic ResNet In this setting, we use four sub-tasks to train the model continuously and see the performance on different test set for 4 tasks after training each task.

ResNet with EWC In this setting, we replace the original loss function with the EWC loss function to implement life long learning.

5.3. Experiment Setting

Our implementation follows the practice in [2]. The image is resized to 256 × 256, and center croped to 224 × 224. We use Adam with the fixed learning rate 1e-5 and without weight decay. The mini-batch size is 8 and every task is trained for 10 epochs. The basic loss function for classification is cross entropy loss, for life long learning, we replace it with the EWC loss function.

In order to get the accurate results, we repeat the experiment for each setting for 5 times, and get the average values and range.

5.4. Metrics

Life long learning is still a novel field so different papers proposed their own evaluation metrics. Here we use the metrics from [6].

In order to measure the ability of life long learning, we need to construct a matrix. The rows represent the model after training on task T_i (The first row is random initiation without training), and the columns represent what task the test set is from. For example, the item $R_{i,j}$ means the performance on test set of task j after training on task i.

According to this matrix, we can get three metrics:

Average Accuracy Measures the overall performance after learning all tasks on the test set.

$$\mathrm{ACC} = \frac{1}{T}\sum_{i=1}^{T} R_{T,i} \tag{3}$$

Backward Transfer Measures the degree of forgetting on previous task after training on the last task T. It is usually negative, and the larger the better, meaning that the model has forgotten very little about the previously learned knowledge. Besides, positive backward transfer means the knowledge learned in the future is a boost to previous task.

$$\mathrm{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1} R_{T,i} - R_{i,i} \tag{4}$$

Forward Transfer Measures how the learned knowledge on previous tasks affects the performance on task T_i before learning task T_i. It is usually positive, and the larger the better, meaning that the model can transfer knowledge to future tasks.

$$\mathrm{FWT} = \frac{1}{T-1}\sum_{i=2}^{T} R_{i-1,i} - \bar{b}_i \tag{5}$$

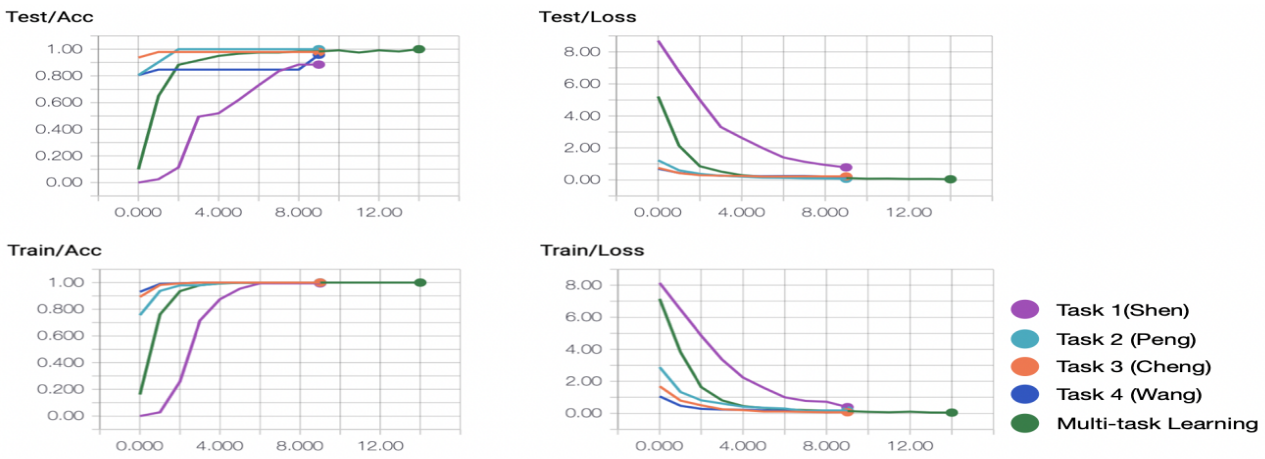

Figure 5. Training process for life long learning and multi-task learning

5.5. Results

The experiment result is shown in Table 1. The Model column uses different model to verify the efficiency of life long learning method, the Task Order column uses different task order to verify that meaningful sequences can indeed improve model performance.

5.5.1 Life Long Learning vs Multi-task Learning

The variation of accuracy and loss on the training and test sets are shown in Fig. 5. When multi-task learning is performed using the whole dataset, the variation curve of accuracy and loss proves the validity of the dataset. When using life long learning, for the first task, the convergence is slower because of the small amount of data, while for the subsequent tasks, the accuracy is high and the loss is small at the beginning of the training, indicating that the knowledge learned from the first task can be well transferred to the subsequent tasks.

In Table 1, comparing the result between basic ResNet model and the one added EWC method for life long learning, it's obvious that using EWC indeed improves the average accuracy for all the tasks. But for backward transfer and forward transfer, EWC methods only outperforms the basic one in the Shen-Peng-Cheng-Wang task order, which means that EWC is not so efficient on handling the forgetting problem and knowledge transferring in this dataset.

5.5.2 Curriculum Learning

As mentioned in Section 5.5.1, if we train the dataset with meaningful order (Shen-Peng-Cheng-Wang), EWC method can handle the forgetting problem and knowledge transferring to some extent. Besides, this order achieved the best results in average accuracy.

The experimental results on the four sequences in Table 1 show that the best results are obtained by putting writer "Wang" at the end of the training. By analyzing the dataset, it is not difficult to find that "Shen", "Peng", and "Cheng" write more neatly and similarly, while "Wang" scribbles and has a different style from the others (the bottom row in Fig. 2). Therefore, the possible reason for this result is that letting the model learn simpler tasks first and then more complex tasks can improve its accuracy.

An interesting phenomenon is that "Shen", "Peng", and "Cheng" are from the south of China, and "Wang" is from the north of China. It may be the influence of the environment that has caused the difference in writing habits. Writing habits of people from the same country and region may have similarities, while differ from those of people from other countries and regions, which might be an interesting potential direction for future works.

This results inspires us to consider the impact of task order on performance when doing life long learning, and adopting a suitable task order can improve the performance of the model to some extent.

6. Disscussion and Conclusion

We collected a new handwritten digits dataset, containing 840 samples, with three kinds of annotation: class label, pen type, writer name. We further implement the ResNet and EWC method on this dataset for life long learning. We believe this dataset can do more interesting tasks except life long learning. Experiments show that EWC indeed help improve life long learning performance, show the effectiveness and importance of curriculum learning, as well as providing explainable results. However, there is still much left for future improvement. We hope this novel handwritten digit dataset and the baselines will encourage the community to develop interesting tasks and models in future work.

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[5] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[6] David Lopez-Paz and Marc' Aurelio Ranzato. Gradient episodic memory for continual learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6467–6476. Curran Associates, Inc., 2017.

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[10] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.