Incremental Learning Based on Angle Constraints

Yang Zhou\textsuperscript{1,2}, Yunbai Qin\textsuperscript{1}, F. Jiang\textsuperscript{1,2,*}, Kunkun Zheng\textsuperscript{3} and Mingcan Cen\textsuperscript{1}

\textsuperscript{1}College of Electronics and Engineering, Guangxi Normal University, Guilin, 541004, China
\textsuperscript{2}Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, 541004, China
\textsuperscript{3}Guangdong Aohong Technology Co., LTD, Zhongshan, 528437, China

*Corresponding author: fjiang@mailbox.gxnu.edu.cn

Abstract. With the rapid growth of the Internet, it has become easy to obtain new data for many application domains. However, when adding new data to the current system of artificial neural networks (ANNs) to learn, it can cause the network to completely forget what it has learned before, which is called catastrophic forgetting. The main reason for these problems is the inability of ANNs to balance new classes with old ones. Therefore, to address the challenge of learning new knowledge while not suffering from catastrophic forgetting, some incremental learning algorithms have been proposed to alleviate. This paper proposes features that balance new classes with old classes by using angular distillation. And some exemplars from the old classes are retained to improve the performance on the old data. The effectiveness of our algorithm is demonstrated on CIFAR-100 dataset.

1. Introduction

During the grow-up of human beings, people learn and receive new things every day. Through some learning skills, humans usually do not forget what they have already learned. Therefore, we can imitate the human learning style and design sustainable machine learning [1] models i.e. incremental learning models. Incremental learning is very similar to the human's own learning model, where the learned knowledge is defined based on accumulation over time. However, much of the machine learning that has been developed in recent years has focused on batch learning. All training samples are available at once prior to training, and after learning these samples, the learning process terminates, and no new knowledge is learned. If new data is added directly to the model, most of the models suffer from catastrophic forgetting [2, 3] and a dramatic drop in performance for the old classes. In practical applications, however, training samples are not all available at once, but are obtained gradually over time. And the model needs to be retrained after new samples arrive, which consumes a lot of space and time. Only incremental learning algorithms can continuously learn new knowledge from new samples and can preserve most of the knowledge already learned without having to relearn it for all the data.

To overcome forgetting problems, a number of methods based on knowledge distillation [4] have been proposed recently. They use the network to separate the critical distribution information from the original data at very high-temperature parameter T. The temperature parameters are then recovered and the two are weighted to fully integrate. Although distillation can maintain the classification performance on the old class, it still cannot balance the new class well with the old classes. As the incremental step
proceeds, the number of classes gradually increases, the old classes become less and less influential, and being able to accurately identify them is a big challenge.

In this paper, we propose a model mainly to target the problem of imbalance between new and old classes. The main contributions: (1) In order to maintain the performance of the old class, we not only used knowledge distillation but also some examples from the old classes. (2) We use cosine regularization [5] to reduce the features into a three-dimensional sphere. In this way, the network can use more angular information to represent the features and reduce the bias caused by the amplitude of the features. In this work, we divided the CIFAR-100 [6] dataset into 5, 10, 20, and 50 incremental steps, and verified the effectiveness of the method in the image classifier.

2. Relate Work
Within the field of machine learning [7, 8], incremental learning has been one of the hot spots for research. Due to the rapid development of deep learning, more and more researchers have been attracted to study incremental learning. Here we mainly describe two types of methods that rely on regularization or rely on knowledge distillation.

Using regularization method: To alleviate forgetting, these methods [9] control and reduce or prevent changes of parameters on networks with fixed learning capacity. Therefore, for the selection of network parameters, it is important to prioritize the usage of the parameters according to the importance of the weights. Elastic weight consolidation (EWC) [2] was first proposed to add penalties for changes of important parameters in a new training task. Zenke et al. [10] showed that the importance of weights needs to be evaluated online during training tasks, which can increase flexibility. Serrae et al. [11] proposed to add a small portion of weights to the base network, trained together with the main model, so that past task information is retained without affecting the present task learning.

Using distillation method: As described in [4], knowledge distillation is an effective method for transferring knowledge from one network to another. It was first used in Learning without Forgetting (LwF) [12] to accomplish the task of incremental learning, where knowledge is retained in the original model using modified cross-entropy loss. Hou et al. [13] proposed to select parts of old task data based on knowledge distillation for training to adapt to new tasks. Rebuffi et al. [14] proposed an incremental classifier and representation learning method. Through the recent sample mean classification rules, knowledge distillation and prototype exercises for feature learning. Aljundi et al. [15] proposed to use an automatic coding gate to identify the dedicated network required for the new task, thus selecting the old task with the highest similarity to the new task for further training.

Overall, the incremental learning approach of regularization is more concerned with parameter changes but is limited by the number of tasks easily. The knowledge distillation approach is focused on data to refine knowledge but may lead to forgetting the previous tasks as the incremental steps increase.

3. Our Approach
In this paper, as shown in Figure 1, we introduce an additional regularization term in the distillation loss function to consolidate the prior knowledge data while learning new knowledge. Although similar to the way [14, 16] retains the exemplars from old classes, we target different ways to improve the representation of the model to the data. Typically, Neural networks use a softmax to generate the probability distribution of classes. However, to balance the new classes with the old ones, we add regularization to the classification layer.
3.1. Angle constraint
In a typical neural network with multiple classifications, the probability distribution of sample $x$ is usually output using softmax. This is shown as follows:

$$P_k(x) = \frac{e^{W_k x + b_i}}{\sum_j e^{W_j x + b_j}}$$  \hspace{1cm} (1)

where $x$ is the feature vector, $W$ is the weight vector, and $b$ is the deviation vector. These parameters constitute the probability prediction of the output layer. We add cosine regularization \cite{5} to the final classification layer, dividing the norm of the input weight vector and the norm of the input vector. The final output probability is shown in equation (3).

$$f_W(x) = \cos \theta = \frac{W \cdot X}{||W|| ||X||}$$  \hspace{1cm} (2)

$$P_k(x) = \frac{e^{f_W(x) + b_i}}{\sum_j e^{f_W(x) + b_j}}$$  \hspace{1cm} (3)

The range of the output features of the network is constrained by $f_W(x)$ to $[-1,1]$, and then the probability distribution is derived by softmax. Although cosine regularization is widely used in other vision tasks, it is only used to constrain new and old classes in this work.

3.2. Knowledge distillation method
In order to maintain the performance of the old classes, we use exemplars of the old classes and knowledge distillation \cite{4}. Knowledge distillation is used to generate softening labels and softening probabilities using the parameter $T$, as shown in equation (4) and equation (5) respectively. $T$ is a temperature scalar that separates critical information from the original dataset when the temperature parameter is set very high.

$$q_k(x) = \frac{e^{f_W(z_k/T)}}{\sum_{i=1}^m e^{f_W(z_i/T)}}$$  \hspace{1cm} (4)

$$p_k(x) = \frac{e^{f_W(z_k/T)}}{\sum_{i=1}^m e^{f_W(z_i/T)}}$$  \hspace{1cm} (5)

The network is initially trained without distillation loss, using only the cross-entropy \cite{17} loss function. When training examples with old classes, the network will use the distillation loss function as shown in Equation (6) and the cross-entropy loss function as shown in Equation (7). We set the temperature $T$ to 2.
\[ L_D = \sum_{x \in X_m \cup X_n} \sum_{k=1}^{m} -q_k(x) \log[p(x)] \]  \hspace{1cm} (6) \\
\[ L_{CE} = \sum_{x \in X_m \cup X_n} \sum_{k=1}^{n+m} -y_k(x) \log[P_k(x)] \]  \hspace{1cm} (7) 

The final integration of the two-loss functions, where \( \gamma \) is set to 0.5 in order to balance the two, given as:

\[ L_{AK} = (1 - \gamma) \times L_{CE} + T^2 \times \gamma \times L_D \]  \hspace{1cm} (8) 

4. Experiment

We compare our approach with LwF and baseline convolutional neural networks (CNN-Base).

4.1. Experiment details

**CIFAR-100**: The dataset has a total of 100 classes. Each class has 600 colored images of size 32 × 32, including 500 for the training set and 100 for the testing set. We divided the 100 classes into 5, 10, 50, and 100 incremental batches for the experiments.

We build a 32-layer ResNet [18] on the framework of Pytorch for the experiments. For the CIFAR-100 dataset, each incremental batch is trained in 70 epochs. The size of each batch is 64. The starting value of the learning rate is set to 2 at 48, 60, and 68 epochs divided by 5, respectively. During the experiment, we use SGD [17] to optimize the loss function and use L2 regularization to reduce model overfitting with a weight decay parameter of 0.00001. Images for the training process are randomly flipped and cropped, but the dataset is not expanded.

To make the comparison experiments fair, we use the same neural network, keep the same number of old sample examples and follow the same rules to split the classes. Each incremental batch is evaluated on a test set after training. Multiple experiments are performed on a fixed order of classes, and the experimental accuracy is compared by taking the average value.

4.2. Analysis of experimental results

In this section, we focus on analyzing the classification accuracy of different methods at the same incremental steps. We compare with LwF and CNN-Base. As shown in Figure 2, (a), (b), (c), and (d) depict the incremental learning results of dividing CIFAR-100 into 5, 10, 20, and 50 incremental batches, respectively. The LwF method does not use old class examples, but only combines knowledge distillation and fine-tuning methods to optimize network parameters. CNN-Base is a convolutional neural network (CNNs) that uses only old class examples and knowledge distillation.

![Figure 2](image-url)
In Table 1 we summarize the average accuracy of each incremental batch finished. The first row of the table shows the number of classes in each incremental step. The other rows show the accuracy values for the different methods. The best results are shown in bold. LwF does not add examples from its own old classes compared to our model. CNN-Base has fewer angle constraints than our model. We compare LwF and CNN-Base and find that preserving the old class examples does improve the performance of the network in increments. The comparison of Our-CNN and CNN-Base shows the effectiveness of the angle constraint.

| # classes | 5     | 10    | 20    | 50    |
|-----------|-------|-------|-------|-------|
| Our-CNN   | 59.75 | 61.60 | 63.60 | 67.00 |
| CNN-Base  | 58.05 | 60.30 | 62.20 | 65.00 |
| LwF       | 17.32 | 28.10 | 53.80 | 63.00 |

Figure 3 shows the comparison between the confusion evidence of CNN-Base and Our-CNN, which can further illustrate the capability of these two models. The classifier of CNN-Base is similar to our model when the increments are in relatively small steps. However, as the incremental steps increase, CNN-Base has a significant decay due to poor performance on old classes. The confusion matrix of Our-CNN shows that our model has better robustness and has a stronger ability to represent.
5. Conclusion
In this paper, we propose a method that trains CNNs by using angle-constrained output features. Under the angle constraint, the model is optimized by combining cross-entropy and distillation loss functions so that the accuracy of the network decays more slowly as incremental learning proceeds gradually. Experimental results on CIFAR-100 draw conclusions in this paper. Exemplars from old classes can be good for improving the performance of old classes of the network. The angle constraint also balances the new class and the old class well. In future work, we plan to change the representability of the network. By using more sensitive features to discriminate classes, a more human-like approach might perform well on incremental learning.

Acknowledgments
This work was supported by National Natural Science Foundation of China (61762018). This work was supported in part by the Innovation special project of Zhongshan Science and Technology Bureau under Grant 2019AG001, the Colleges and Universities Key Laboratory of Intelligent Integrated Automation, Guilin University of Electronic Technology, China (GXZDSY2016-03), the research funding of Guangxi Key Lab of Multi-source Information Mining & Security (18-A-02-02), Natural Science Foundation of Guangxi (2018GXNSFAA281310), the Guangxi Key Research and Development Funding (2019AB35004).

Reference
[1] P. Fuangkhon and T. Tanprasert, "An incremental learning algorithm for supervised neural network with contour preserving classification," in 2009 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2009, vol. 2: IEEE, pp. 740-743.
[2] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, and A. Grabska-Barwinska, "Overcoming catastrophic forgetting in neural networks," Proc Natl Acad Sci U S A, 114, no. 13, pp. 3521-3526, 2016.
[3] R. Kemker, M. Mcclure, A. Abitino, T. Hayes, and C. Kanan, "Measuring Catastrophic Forgetting in Neural Networks," arXiv preprint arXiv:1708.02072, 2017.
[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," Computer, 14, no. 7, pp. 38-39, 2015.
[5] C. Luo, J. Zhan, X. Xue, L. Wang, R. Ren, and Q. Yang, "Cosine Normalization: Using Cosine Similarity Instead of Dot Product in Neural Networks," in International Conference on Artificial Neural Networks, 2018: Springer, pp. 382-391.
[6] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Computer Science Department, University of Toronto, Tech. Rep., 1, 01/01 2009.
[7] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Distance-based image classification: Generalizing to new classes at near-zero cost," IEEE transactions on pattern analysis and machine intelligence, 35, no. 11, pp. 2624-2637, 2013.
[8] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," Advances in neural information processing systems, 13, pp. 409-415, 2000.
[9] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 139-154.
[10] F. Zenke, B. Poole, and S. Ganguli, "Continual Learning Through Synaptic Intelligence," Proceedings of machine learning research, 70, p. 3987, 2017.
[11] J. Serra, D. Suris, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," in Proceedings of Machine Learning Research (PMLR), 2018, vol. 80, pp. 4548 – 4557.
[12] Z. Li and D. Hoiem, "Learning without Forgetting," IEEE Transactions on Pattern Analysis & Machine Intelligence, 40, no. 12, pp. 2935-2947, 2017.
[13] S. Hou, X. Pan, C. Change Loy, Z. Wang, and D. Lin, "Lifelong learning via progressive distillation and retrospection," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 437-452.

[14] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental Classifier and Representation Learning," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 2001-2010.

[15] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3366-3375.

[16] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-End Incremental Learning," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 233-248.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.