Continual Learning for Activity Recognition

Ramesh Kumar Sah[1], Seyed Iman Mirzadeh[1], and Hassan Ghasemzadeh[2]

***Abstract*— The recent success of deep neural networks in prediction tasks on wearable sensor data is evident. However, in more practical online learning scenarios, where new data arrive sequentially, neural networks suffer severely from the "catastrophic forgetting" problem. In real-world settings, given a pre-trained model on the old data, when we collect new data, it is practically infeasible to re-train the model on both old and new data because the computational costs will increase dramatically as more and more data arrive in time. However, if we fine-tune the model only with the new data because the new data might be different from the old data, the neural network parameters will change to fit the new data. As a result, the new parameters are no longer suitable for the old data. This phenomenon is known as *catastrophic forgetting*, and *continual learning* research aims to overcome this problem with minimal computational costs. While most of the continual learning research focuses on computer vision tasks, implications of catastrophic forgetting in wearable computing research and potential avenues to address this problem have remained unexplored. To address this knowledge gap, we study continual learning for activity recognition using wearable sensor data. We show that the catastrophic forgetting problem is a critical challenge for the real-world deployment of machine learning models for wearable sensor data. Moreover, we show that the catastrophic forgetting problem can be alleviated by employing various training techniques.**

I. Introduction

In recent years, Deep Neural Networks (DNN) have demonstrated superior performance in various domains, from natural language processing to computer vision and signal processing. The utility of deep learning in medical field is also well understood [1]. However, deep neural networks do not learn "continually" as we humans do. In contrast, deep neural networks learn in an "isolated" manner. For instance, a typical activity recognition system will be trained on a single dataset of multiple subjects, performing a specific number of activities (i.e., classes). As soon as the training is completed, the model will perform predictions on the sensor data that have already been seen during the training time. In contrast, humans are "lifelong" learners, and we accumulate and retain knowledge from our previous experiences. The goal of continual learning (lifelong learning) [2] is to mimic this learning experience of humans for machines by modeling the learning problem as a "sequence" of tasks or "stream" of data that will arrive in time. Currently, deep neural networks do not perform well in continual learning scenarios, and this is a major obstacle toward reaching Artificial General Intelligence (AGI) [3].

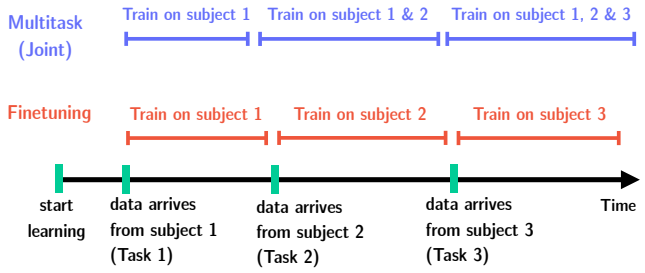

Fig. 1. Comparison of two different learning approaches to sequential learning: multitask (joint) and fine-tuning.

Continual learning is a challenging problem since neural networks suffer from the "catastrophic forgetting" problem [4]. Catastrophic forgetting happens because in continual learning, the model is trained in an online manner, and the training data distribution changes over time. For instance, the sensor readings from two different subjects could have different distributions even for the same activity. Consequently, the previous parameters the model learned for previous subjects will adapt to the new subject, but the new parameters are no longer effective for the previous subjects. Even though this is a significant challenge for the AI community, we still do not understand the catastrophic problem clearly.

In this paper, we study continual learning in the context of wearable sensor systems and in particular for activity recognition in non-stationary environments due to sequential nature of learning across users. In Section II, we discuss approaches that one can take for continual learning of human activities, and demonstrate the adverse impacts of catastrophic forgetting on the activity recognition performance. In Section III, we study the continual learning and the catastrophic forgetting problem, followed by proposed algorithms that can alleviate catastrophic forgetting in Section IV. Finally, in Section V we study the effectiveness of various continual learning algorithms in the context of wearable systems. Our contributions can be summarized as follows:

1) We study the continual learning problem motivated by the practical challenges of deploying prediction models on sensor data.
2) We study the effectiveness of state-of-the-art continual learning algorithms for human activity recognition and show that while these algorithms improve the perfor-

[1]Ramesh Sah is a Computer Science graduate student at Washington State University, Pullman, WA, USA. ramesh.sah@wsu.edu. Seyed Iman Mirzadeh is also a Computer Science graduate student at Washington State University, Pullman, WA, USA. seyediman.mirzadeh@wsu.edu

[2]Dr. Hassan Ghasemzadeh is an associate professor of Biomedical Informatics in the College of Health Solutions at Arizona State University, Phoenix, AZ, USA. hassan.ghasemzadeh@asu.edu

mance, they still have a significant gap to reach the joint training accuracy.

II. Multitask Training vs. Fine-Tuning

Figure 1 demonstrates the continual learning setup in human activity recognition systems. Assume a scenario where we continually collect activity data from one subject at a time, each performing the same set of activities. As stated previously, one practical setup in continual learning is to assume that data arrive as a sequence of tasks. Here, we model the data collected from each subject as an individual task. First, the model will be trained on the activities we collect with the first subject. For the second subject, we can pursue two different approaches:

1) **Multitask (Joint)**: In multitask or joint training, the model will be re-trained jointly on the old tasks (i.e., previous subjects) and new task (i.e., current subject). This is shown in the top row of Figure 1.
2) **Fine-tuning**: In fine-tuning, the model will not be re-trained on the old tasks (i.e., previous subjects) and will only be fine-tuned using the data received for the current task (i.e., current subject).

Let us discuss the implications of these two learning approaches by comparing their performance and training time. First, we compare the performance of multitask (MTL) and fine-tuning (FT) on the PAMAP2 dataset [5], which includes eight subjects[1]. We measure the performance using the average validation accuracy of the model on all the tasks trained up to each time. For instance, the validation accuracy for task 1 is measured on the validation (test) set of subject 1, the accuracy for task 2 is measured on validation sets of subject 1 and subject 2, and so on.

Figure 2 illustrates the catastrophic forgetting problem in continual learning of human activity recognition. If we fine-tune the model only on the new data, the performance drops significantly as more and more tasks (subjects) arrive. The reason is that at each time, we train the model only on the latest subject, and the model fits the new data by minimizing the classification loss on the data for that specific subject. Consequently, if the distribution of the new training data (i.e., latest subject) is different from the old data (i.e., previous subjects), the new parameters change so much that they can no longer perform prediction on the data from previous subjects. We refer to this problem as the "distribution shift" problem, which is responsible for catastrophic forgetting.

In contrast to fine-tuning, we can approach the sequential learning of multiple tasks using multitask (joint) training. In this scenario, the distribution shift problem does not exist because, at each time, the model will be trained on the old data (i.e., previous subjects) and new data simultaneously. Hence, the catastrophic forgetting does not happen, and as we can see in Figure 2, the average accuracy of joint training does not drop as opposed to fine-tuning approach. However, joint training has a significant problem: it is practically infeasible in a real-world setting. The reason is that when

[1]Experimental details are discussed in Section V.

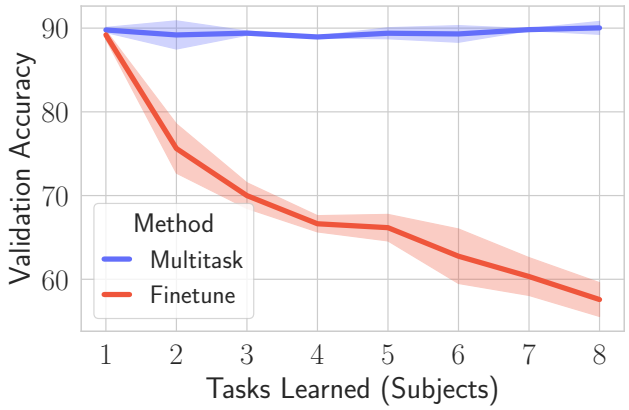

Fig. 2. Evolution of the average accuracy in the continual learning experience for multitask and joint training.

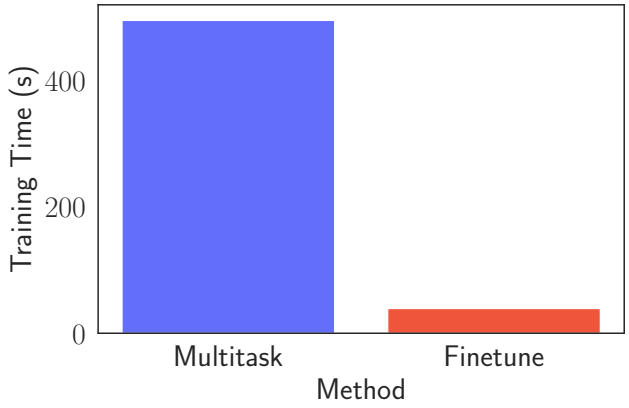

Fig. 3. Comparison of the training time of the multitask and joint training of the networks in Figure 2.

the model sees data for subject T, it needs to be trained on the data for all previous subjects t where $1 \leq t \leq T$ and hence the computation cost grows rapidly as more and more data arrive. Figure 3 compares the computation cost of the models in Figure 2 where the average training time of joint training is about 12 times of fine-tuning.

The ultimate aim of continual learning is to achieve the performance of multitask learning, with the training setting of fine-tuning where at each time, the model does not have access to the training data of previous tasks.

III. Continual Learning

A standard metric for success in Artificial Intelligence (AI) is the ability to mimic human learning. For instance, we typically measure different human skills, such as driving a car or recognizing an image to develop an AI system that can match these abilities given enough training data. However, this measurement emphasizes the result (e.g., classification accuracy) and overlooks the fact that human learning has a critical characteristic: learning through time. Humans can gain new experiences while maintaining the knowledge gained from previous experiences. This may be

the consequence of time moving only forward, and the world is constantly changing. Hence, humans have evolved to learn in such a non-stationary environment. Continual learning aims to mimic this human ability to learn in non-stationary environments. In this work, we follow a popular categorization of non-stationary classification environments in continual learning [6], [7]:

1) **New Instances (NI)**: In this scenario, all classes are shown in the first task while subsequent instances of known classes become available over time.
2) **New Classes (NC)**: Where new classes are available so that the model should deal with the learning of new classes without forgetting previously learned classes. In the literature, this scenario is also known as Class Incremental Learning.
3) **New Instances & Classes (NIC)**: In this scenario, each task can include both new instances of known classes or completely new classes.

We note that there are other categorizations of continual learning setups, such as dividing the environments into task-based and task-agnostic depending on whether or not boundaries between tasks are defined [8]. However, in the context of human activity recognition using sensor data, we use the more popular categorization of continual learning scenarios.

IV. Overcoming Catastrophic Forgetting

With the significance of continual learning being known to the AI community, the research on overcoming catastrophic forgetting has rapidly accelerated in recent years. Continual learning algorithms can be categorized into three types [9].

1) Regularization: These methods explicitly apply regularization techniques to ensure parameters do not change too much. For instance, Elastic Weight Consolidation (EWC) [10] uses an estimate of second-order curvature of the minima (i.e., neural network parameters after training on each task) and include that information in the loss function as a regularizer. However, regularization-based methods do not perform as well as experience replay methods in practice.

2) Experience Replay: Experience Replay (ER) methods build and store a memory of the knowledge learned so far. This knowledge is mainly referred to as replay buffer or episodic memory. In its simplest form, this knowledge could be storing a few examples from previous tasks and replaying/rehearsing them in addition to the new data at each time. ER-Ringbuffer [11] is a popular approach that stores a fixed amount of data points for each task, adds them to the new data obtained for the current task, and trains the model on the new dataset. Averaged Gradient Episodic Memory proposes another form of storing knowledge (A-GEM) [12]. Instead of storing raw examples, A-GEM stores the gradient directions for previous tasks and uses it to modify the gradient updates for the current task to ensure that the new gradient is not in a direction that increases the forgetting.

3) Parameter Isolation: Parameter isolation methods allocate different subsets of the parameters to each task. The intuition is that if we can delegate each task's learning to a specific part of the model, we can avoid catastrophic forgetting. Another perspective to these methods is by gating mechanisms that improve the stability and control the plasticity by activating different gates (subsets of the model parameters) for each task [13]. [14] proposes a bio-inspired approach for a context-dependent gating for any specific task.

We note that a continual learning algorithm does not necessarily belong to only one of the above categories. For example, Mode Connectivity SGD (MC-SGD) [15] exploits an important characteristic of multitask solution based on low-loss linear paths in the parameter space. Moreover, MC-SGD uses episodic memory to estimate any point loss in the parameter space for previous tasks and adds this information as a regularization term to the loss function.

V. Experiments

In this section, we first introduce our experimental setup, and in the second part, we study the effectiveness of state-of-the-art continual learning algorithms for human activity recognition.

A. Experimental Setup

1) Benchmark: We use the PAMAP2 [5] human activity recognition dataset includes sensor readings of 8 subjects, each performing 12 daily activities. In addition, three IMU sensors were placed on the hand, chest, and ankle. We have used only the IMUS sensors' 3D acceleration and gyroscope readings for our analysis. Moreover, we have divided the time series into the 2.56s windows with 50% overlap. In our experiments, we use the data collected from each subject as a separate task. Each subject has different physical characteristics that for the same activity, the sensor readings are different enough for catastrophic forgetting to happen. Finally, we used 25% of the data to measure test (validation) accuracy and the rest for training.

B. Training Details

We used a standard neural network architecture of our prediction model that includes two 1D convolutional layers, each with 64 filters of length 5. After two convolutional layers, we used one max-pooling layer of length 4 and one fully connected layer with 200 neurons for classification. All layers in the network use the ReLU activation function. We use the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.01 and a momentum value of 0.8. We use CL-Gym [16] continual learning library for implementation and training of our experiment on a 2.7 GHz Dual-Core Intel Core i5 on a MacBook Pro. We note that while we only report results using the convolutional network, catastrophic forgetting still happens regardless of the choice of architecture. Finally, all the reported results are the average of five runs using different random initialization.

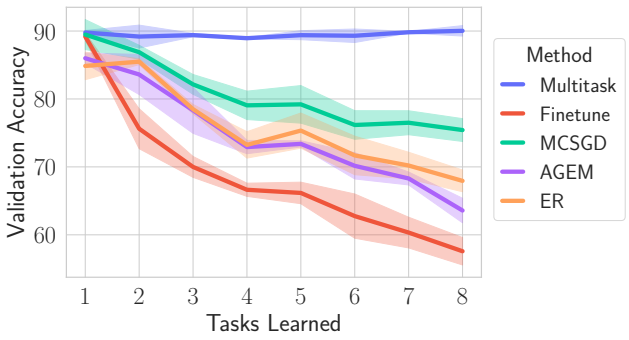

Fig. 4. Evolution of the average validation accuracy.

1) Evaluation: We evaluate each algorithm using two common metrics used repeatedly in CL research [11], [12], [15], [17]:

- **Average Accuracy**: defined as

$$\frac{1}{T}\sum_{j=1}^{T} a_{T,j} \tag{1}$$

Where T represents the number of tasks, and $a_{i,j}$ denotes the test accuracy on task j after the algorithm has finished learning task i. The average accuracy measures how well the network is able to predict the data from all the tasks it has learned up to each time.

- **Average Maximum Forgetting**: defined as

$$\frac{1}{T-1}\sum_{j=1}^{T-1} \max_{l\in\{1,\dots,T-1\}} (a_{l,j} - a_{T,j}) \tag{2}$$

Where T represents the number of tasks, and $a_{i,j}$ denotes the test accuracy on task j after the algorithm has finished learning task i. The average forgetting shows the decrease in performance for each of the tasks between their peak accuracy and their accuracy after the learning experience is finished.

C. Comparison of Continual Learning Algorithms

We compare the effectiveness of state-of-the-art continual learning algorithms in Figure 4 and Table I. Figure 4 shows the evolution of the average accuracy for each algorithm throughout the continual learning experience, similar to Figure 2 introduced in Section I. While all algorithms can outperform fine-tuning, we can see that they still have a visible performance gap to the multitask training. However, the best method (MC-SGD) can decrease the forgetting by nearly 29% while with only double training time, which is significantly lower than the multitask (joint) training time. It is worth mentioning that MC-SGD is also the current state-of-the-art continual learning algorithm on computer vision tasks, which might hint that the advances in continual learning in other domains can also improve the continual performance time-series domain.

TABLE I

COMPARISON OF CONTINUAL LEARNING ALGORITHMS.

Method	Average Accuracy(%)	Average Forgetting(%)	Average Training Time (s)
Finetune	57.6 (± 2.02)	34.0 (± 3.09)	39.1
AGEM [12]	63.6 (± 1.81)	21.4 (± 1.89)	87.9
ER-Ring [11]	67.9 (± 1.6)	20.7 (± 1.76)	50.4
MCSGD [15]	75.4 (± 1.58)	5.0 (± 0.62)	81.1
Multitask	90.1 (± 0.77)	0.0	495.5

VI. CONCLUSION

Efficient deployment of machine learning models is a fundamental problem for wearable systems. Besides the challenges arising from limited energy and computation [18], continual learning can also be viewed as an important research problem that arises from a practical scenario where the data arrives sequentially and re-training the model is computationally expensive. We note that the implications of continual learning problems go beyond wearable systems, and catastrophic forgetting is a significant obstacle towards reaching artificial general intelligence.

In this work, we have studied continual learning and the catastrophic forgetting problem in the context of sensor systems. We showed that while current continual learning methods can alleviate the catastrophic forgetting, there is still a visible gap between them and multitask training that calls for further research on the intersection of continual learning and wearable systems.

REFERENCES

[1] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 4–21, 2016.

[2] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robotics and Autonomous Systems*, vol. 15, no. 1, pp. 25 – 46, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/092188909500004Y

[3] D. Hassabis, D. Kumaran, C. Summerfield, and M. M. Botvinick, "Neuroscience-inspired artificial intelligence," *Neuron*, vol. 95, pp. 245–258, 2017.

[4] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," ser. Psychology of Learning and Motivation, G. H. Bower, Ed. Academic Press, 1989, vol. 24, pp. 109 – 165. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0079742108605368

[5] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," *2012 16th International Symposium on Wearable Computers*, pp. 108–109, 2012.

[6] V. Lomonaco and D. Maltoni, "Core50: a new dataset and benchmark for continuous object recognition," in *CoRL*, 2017.

[7] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural networks : the official journal of the International Neural Network Society*, vol. 113, pp. 54–71, 2019.

[8] C. Zeno, I. Golan, E. Hoffer, and D. Soudry, "Task agnostic continual learning using online variational bayes with fixed-point updates," *ArXiv*, vol. abs/2010.00373, 2020.

[9] M. D. Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks." *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2021.

[10] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 13, pp. 3521–3526, 2017.

[11] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. S. Torr, and M. Ranzato, "On tiny episodic memories in continual learning," *arXiv preprint arXiv:1902.10486*, 2019.

[12] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with a-gem," in *International Conference on Learning Representations*, 2018.

[13] S. I. Mirzadeh, M. Farajtabar, and H. Ghasemzadeh, "Dropout as an implicit gating mechanism for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 945–951.

[14] N. Y. Masse, G. D. Grant, and D. Freedman, "Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization," *Proceedings of the National Academy of Sciences*, vol. 115, pp. E10 467 – E10 475, 2018.

[15] S. I. Mirzadeh, M. Farajtabar, D. Gorur, R. Pascanu, and H. Ghasemzadeh, "Linear mode connectivity in multitask and continual learning," in *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

[16] S. I. Mirzadeh and H. Ghasemzadeh, "Cl-gym: Full-featured pytorch library for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 3621–3627.

[17] S. I. Mirzadeh, M. Farajtabar, R. Pascanu, and H. Ghasemzadeh, "Understanding the role of training regimes in continual learning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[18] S. I. Mirzadeh and H. Ghasemzadeh, "Optimal policy for deployment of machine learning models on energy-bounded systems," in *IJCAI*, 2020.