Lifelong Learning for Dynamic Churn Prediction

1st Fioni Sarnen
School of Computing
Telkom University
Bandung, Indonesia
fionisarnen@student.telkomuniversity.ac.id

2nd Suyanto Suyanto
School of Computing
Telkom University
Bandung, Indonesia
suyanto@telkomuniversity.ac.id

3rd Rita Rismala
School of Computing
Telkom University
Bandung, Indonesia
ritaris@telkomuniversity.ac.id

***Abstract*—Learning continually, accumulating knowledge, and using it to learn new tasks was the characteristic of lifelong learning. Lifelong learning, which is also known as Continual Learning, takes benefit from the one that called the previous (first) task to solve the new task—this schema work by configuring a proper regularization between them. Elastic weight consolidation (EWC) method proposed by Google Deepmind provides a way of calculating the importance (preserving the previously acquired knowledge) of weight and selectively adjusts the plasticity. In this paper, the EWC is exploited to tackle the sequence tasks of predicting churn activity. The tasks involve two distinct datasets from the domain of Telecom. The experimental results show that EWC can elevate the model performance in sequence training. Lifelong learning offers a more flexible way of learning to further research in dynamic learning.**

Keywords—churn prediction, lifelong learning, continual learning, elastic weight consolidation

I. INTRODUCTION

Lifelong learning (LL) or also known as continual learning (CL), can mimic the performance of the human brain to learn continuously. Human biologically has the ability for adapting through acquired knowledge and skills such as necessary information into a more complex form in the concept of transfer of learning and knowledge [1]. The capability of this type of learning going against the 'traditional' static learning, where we were assuming all training data available at the time. Learning in the context of "lifelong" provides the ability to transfer knowledge between a sequential task relying on previously acquired knowledge of particular subjects or domains [2].

However, the problem arises with the habit of neural networks, which tend to overwrite prior acquired knowledge when training with multiple streams of data. It is affected by weight in the previous network (essential to the first task) are change or roughly loss when generalizing for the new task—such activity called catastrophic of forgetting [3] [4] [5] [6].

Furthermore, a comprehensive illustration [7] of lifelong learning scenarios is given to determine what approaches are more suitable for the problems, as illustrated in Fig. 1. These protocols distinguish the kind of a series of tasks into particular difficulty, whether task information was given or not. This boundary will lead to a better choice of both lifelong learning strategy and mechanism.

First scenarios, Task-incremental learning (Task-IL), where problems are always given information about which task needs to be done. Domain-incremental learning (Domain-IL) did not provide task identify during the test (models only solve the problem given). Moreover, the last scenarios, more complicated, are Class-incremental learning (Class-IL), which can learn incrementally of new class along with every task.

Furthermore, numerous approaches to lifelong learning have been widely developed. These methods help pull through the most significant limitation of LL concept that it is catastrophic of forgetting. Parisi et al. (2018) have summarised the problems and existing solutions related to lifelong learning approaches in various cases [8]. PathNet (2017) applying an ensembling mechanism with Genetic algorithm that generates agents to review changes to parameters from the backpropagation algorithm in neural networks [9].

Rehearsal mechanism focuses on revisits previous learning samples to prevent forgetting of previously trained knowledge carried by GeppNet (2016) [10]. Rusu et al. (2016) proposed Progressive Neural Network (PNN) to keep a pool of pre-trained models as knowledge and use lateral connections between them to adapt to the new task [11]. Fixed Expansion layer (FEL) model (2013) extend the hidden layer in multilayer perceptron (MLP) with sparse encoding mechanism to help mitigate catastrophic forgetting of the prior learned representation [12].

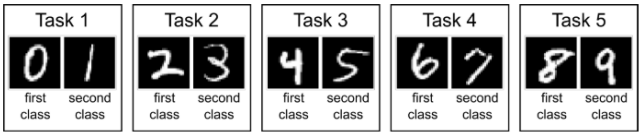

Fig. 1. given schematic MNIST task. Case Example-For Task-IL: given Task 4 and predict whether first or second class. Domain-IL: model needs to predict whether first or second class with task unknown. Class-IL: model need to recognize which digit (incrementally learn all classes) [7].

iCaRL (2017) is a specific model for class-incremental problems [13]. iCaRL monitoring and collecting a various sample that carries the most representative information for each new class. Ultimately, Elastic Weight Consolidation (EWC) proposed by Google Deepmind (2017) uses a regularization-based mechanism to add a constraint on certain weight determined by importance to protect previously learned knowledge [5]. The trick is to lock weight used to solve the first task when training for a new task. However, the tricky part is to know which weight is essential. EWC model uses a Fisher Information Matrix *Fi* [14] to calculate importance from the model's parameter.

In this case, we are applying Elastic Weight Consolidation (EWC) model to sequentially learn multiple tasks and identify the model's impact on our case for mitigating catastrophic forgetting. Weight consolidation method is applied to solve sequential problems which have similarity to first scenarios (Task-IL) that use double-headed output layer to predict Churn activity serially on multiple tasks from the domains of Telecom customers.

A. Elastic Weight Consolidation

The behavior of Elastic Weight Consolidation (EWC) is task-specific synaptic consolidation [15]. EWC counting the importance of weight as the parameter's network for the first task and selectively adjust the plasticity of weight. The term plasticity is the main reason for catastrophic forgetting since the weight learned in the first task can be easily modified

given a new task [16]. Training task (A) with standard neural networks generates a mapping function to expected output targets by finding optimal parameters Θ_A^*, as illustrated in Fig. 2.

Elastic weight consolidation (EWC) adjust parameters from the perspective of probability. Constraining and configuration make it likely there is a solution to solve task B. EWC finds that Θ_B^* is carefully found centered around Θ_A^*. In [5], a substantial weight from a probability perspective is determined using (according to Bayes's rule):

$$log\ p(\Theta|D) = log\ p(D_B|\Theta) + log\ p(\Theta|D_A) - log\ p(D_B) \quad (1)$$

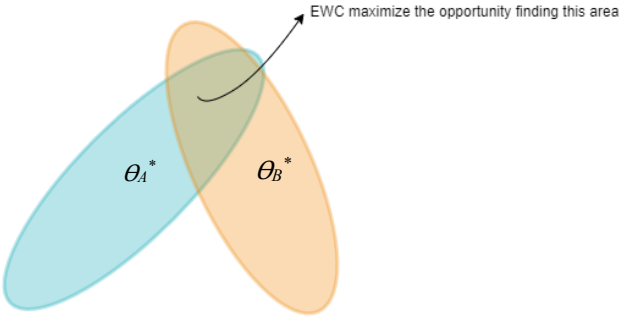

Fig. 2. EWC ensures task A is remembered during train on the new task. Task B. After learning, the first task model finds the optimal parameter Θ_B^* by various configurations. EWC finds a way to constraining important parameters to stay close to their old value.

The perspective can be explained with posterior probability $p(\Theta|D)$, the probability of the data of task B $p(D_B|\Theta)$ as the likelihood function, the posterior probability of task A $p(\Theta|D_A)$, and the last $p(D_B)$ contain loss function for task B. However, the truth of posterior probability is hard to known (intractable). EWC indicate to approximate it with Laplace Propagation [17]. Laplace propagation is employed to find a reasonable distribution approximation to a continuous probability [15].

Fisher information matrix F_i used to determine the importance in EWC by measuring the amount of information carried by random variables to track parameters unknown and the distribution that models the networks. A comprehensive explanation of Fisher information matrix F_i at [14]. EWC no longer uses a loss function that applies initially (base model). Instead, it generates a custom loss $L(\Theta)$ [5] to minimize the loss of new task:

$$L(\Theta) = L_B(\Theta) + \sum \tfrac{1}{2}\lambda\ F_i(\Theta_i - \Theta^*_{A,i})^2 \quad (2)$$

Where Θ_i denote each parameter and $\Theta^*_{A,i}$ denote all parameters significant to the first task (i.e., all accumulate of previous tasks). The loss function for new task L_B and λ denotes how fatal the first task (parameters) to new task.

II. Research Method

A. Dataset

In this case, we are using Churn activity problems. Churn activity described as the loss of customers because they move out to competitors [18]. The most significant factor that causes churn is the correlation between paid subscription status and customer satisfaction. The strategy to be able to predict churn becomes significant for every company. In the real case, data of customers have a massive amount of data and stream over time. Using lifelong learning concepts, we want to show it is very likely to built connection, remembering and preserve the pattern of churn from one task to another.

A lot of churn prediction model adapted e.g., in [18]. This type of models tends to isolate and secluded. However, it is not fair to compare this works to another churn prediction model as it has a whole different focus and goals. The difference lies in how the models treated. The traditional method configures one model to one problem set. Lifelong learning has the ability to use one shared models on different problem sets, by constructing overlapping areas for the solution in all tasks. It is started by acquired knowledge in previous learning (Fig. 2 illustrate the idea).

For the experiments, we use two familiar datasets from Telecom's domains for churn problems as two sequential tasks. As a reference, it is possible to use different data sets or domains or modals to applied using lifelong learning concepts [5]. The first dataset defined as the **First task** is from churn IBM Watson Analytics sample data sets, which has 21 attributes and 7043 records data. The **New task** is the dataset from the UCI Machine Learning Repository dataset, which has 21 attributes and 3333 records. Both data were chosen with consideration having the same attribute variances. Fig. 3 illustrate the proportion of two datasets, and Table 1 describes the attributes of two data sets.

Lifelong learning is developed as an adaptive and dynamic system to observe and predict the pattern of churn. With the flexibility of these concepts, the diversity of the testing method has been developed. Lifelong learning can handle a series of tasks across diverse domains [19], even modality [20], as long as the tasks meet the requirements for not overlapping the category label. Furthermore, what much more significant is how to prepare fair and balanced testing. In this case, lifelong learning help absorb the typical pattern of churn activity from one task to another even from such different data.

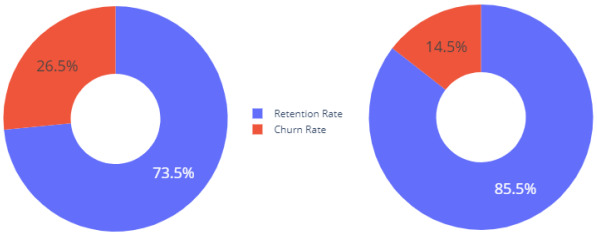

Fig. 3. Churn activity rate from datasets 1 (IBM's) left; Churn rate from datasets 2 (UCI's) right; 2 datasets showing churn to be the minority, having less than 30% of churn activity.

TABLE I. NAME OF THE FEATURES IN EACH TASK

First Task	customerID, gender, SeniorCitizen, Partner Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges **Churn**
New Task	state, account length area code, phone number, international plan