# Overcoming Catastrophic Forgetting for Continual Learning via Model Adaption

Wenpeng Hu[1,2,*], Zhou Lin[1,*] , Bing Liu[3,*], Chongyang Tao[2], Zhengwei Tao[2], Dongyan Zhao[2], Jinwen Ma[1], and Rui Yan[2,†]

[1]Department of Information Science, School of Mathematical Sciences, Peking University
[2]ICST, Peking University, Beijing, China
[3]Department of Computer Science, University of Illinois at Chicago

## Introduction

- **Learning multiple tasks sequentially is important for the development of AI and lifelong learning systems.**
  - For continual learning and knowledge accumulation
  - For dealing with catastrophic forgetting problem in neural networks.
- **We propose Parameter Generation and Model Adaptation (PGMA) to deal with the problem. The proposed approach learns to build a model, called the solver, with two sets of parameters. The first set is shared by all tasks learned so far and the second set is dynamically generated to adapt the solver to suit each test instance in order to classify it.**

## Advantages

***What can our model do?***

- **Ease the problem of accuracy deterioration.**
- **No parameter increase or network expansion is needed to learn new tasks.**
- **No previous data need to be stored to enable the system to remember the previously learned models or knowledge.**
- **The proposed approach PGMA works well for different scenarios and different types of datasets, and outperforms the existing baselines markedly.**

## Existing Work

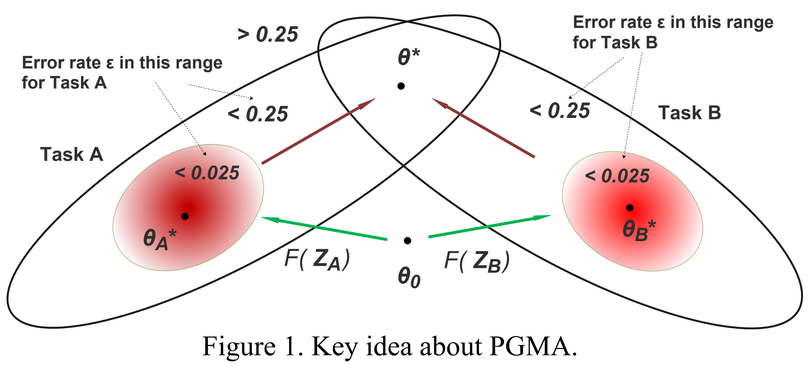


Figure 1. Key idea about PGMA.

### Shortcomings:

Existing joint parameterization methods, learning shared parameters for all tasks, suffer from accuracy deterioration. Dynamic changing methods require constant increase of model parameters and thus can result in a huge and complex model.

## Proposed Framework

### We propose a new framework and a model, which includes:

➢ Data Generator (DG) ➢ Solver ➢ Dynamic Parameter Generator (DPG)

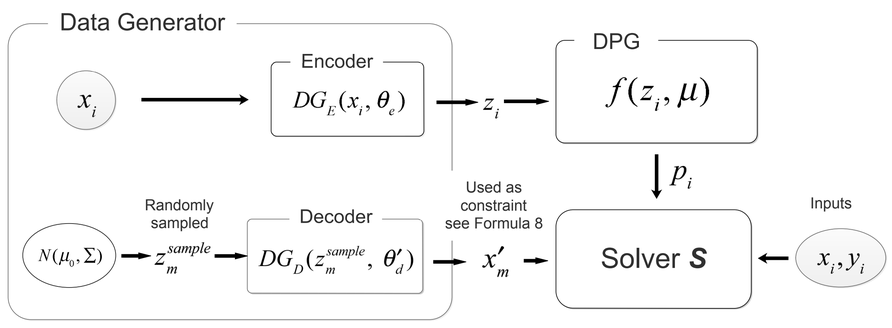


Figure 2. Sequential training in the PGMA framework.

Given a test instance $x_i$, $DE_G$ first generates its embedding $z_i$, which is fed to $f(\cdot)$ to generate a set of parameters $p_i$. Solver S then takes $x_i$ as input and uses the trained/learned shared parameters $\theta_0$ and $p_i$ to classify $x_i$. $\theta_0$ contains the common features of all tasks learned so far. $p_i$ simply adapts S for $x_i$ in order to classify $x_i$.

***Combination Method:*** $\theta_i^* = \text{combine}(\theta_0, \mathbf{p}_i) = \{[\mathbf{w}_{i,k}^*]\}_{k=1}^K = \{[\theta_{0,k}; \mathbf{p}_{i,k}]\}_{k=1}^K$

***Constraints:***

$$\min \sum_{m=1}^{M} \sum_{k=1}^{K} ||\mathbf{w}_{i,k}^* \mathbf{x}'_{m,k} - \mathbf{w}_{i-1,k}^* \mathbf{x}'_{m,k}||$$

$$\min \sum_{m=1}^{M} ||\mathbf{z}_m^{sample} - DG_E(x'_m, \theta_e)||$$

$$\min \sum_{m=1}^{M} ||DG_D(\mathbf{z}_m^{sample}, \theta_d) - x'_m||$$

## Experimental Results

### Datasets

➢ Experiments conducted using two image datasets (MNIST and CIFAR-10) and two text datasets (DBPedia ontology (Lehmann et al., 2015) and THUCNews (Li et al., 2006)).

### Results

| Model | shuffled MNIST (3 tasks) | disjoint MNIST (2 tasks) | disjoint CIFAR-10 (2 tasks) | THUCNews (2 tasks) | DBPedia (2 tasks) | DBPedia (3 tasks) |
|---|---|---|---|---|---|---|
| Adam | 91.46 | 48.64 | 41.99 | 46.78 | 47.66 | 41.10 |
| EWC | 96.70 | 48.96 | 37.75 | 45.02 | 53.95 | 35.89 |
| GR | 97.57 | 89.96 | 65.11 | 81.55 | 88.41 | 82.14 |
| IMM | 97.92 | 94.12 | 62.98 | 80.32 | 92.65 | 85.31 |
| Our PGMA | **98.14** | **96.53** | **69.51** | **85.12** | **94.70** | **88.06** |

Table 1. Average accuracy over all tasks in a sequence after the tasks have all been learned.

| Model | shuffled MNIST | disjoint MNIST | DBPedia | CIFAR-10 | THUCNews |
|---|---|---|---|---|---|
| GR | 94.54 | 75.47 | 63.71 | 31.09 | 47.35 |
| IMM | 96.09 | 67.25 | 64.04 | 32.36 | 46.61 |
| Our PGMA | **96.77** | **81.70** | **69.68** | **40.47** | **52.93** |

Table 2. Average accuracy over 5 tasks in a sequence after the tasks have all been learned.

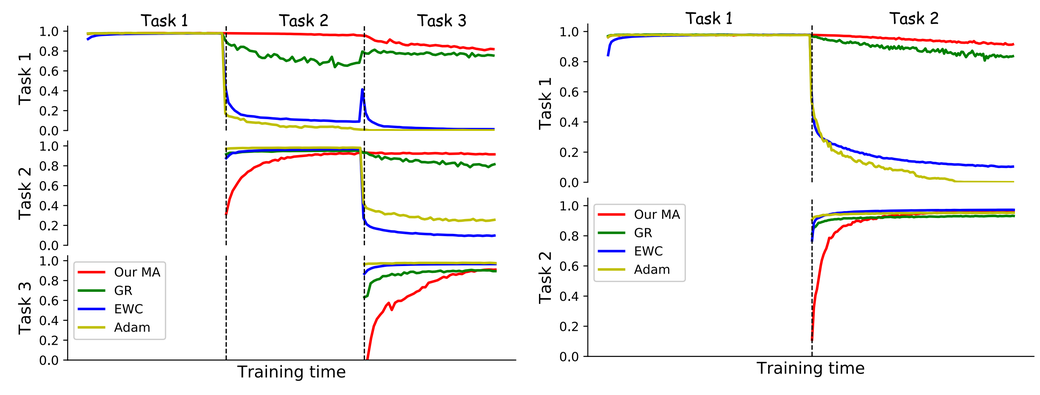


Figure 3. Experimental results on DBPedia dataset.

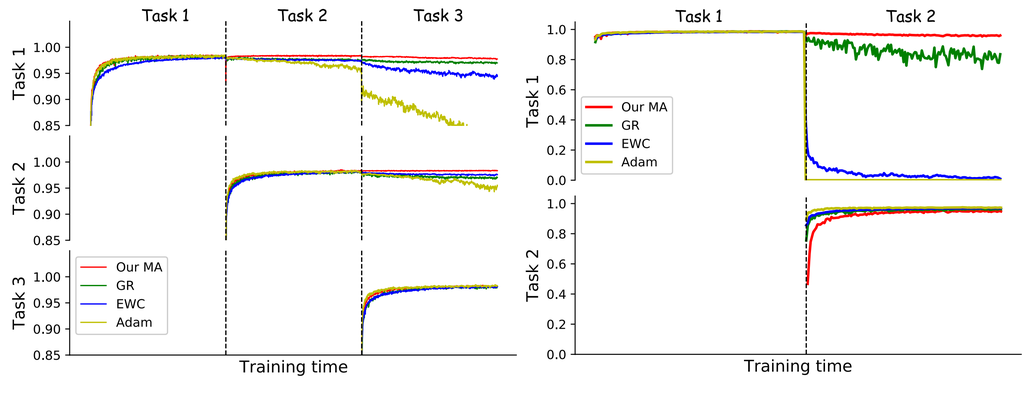


Figure 4. Experimental results on MNIST dataset.