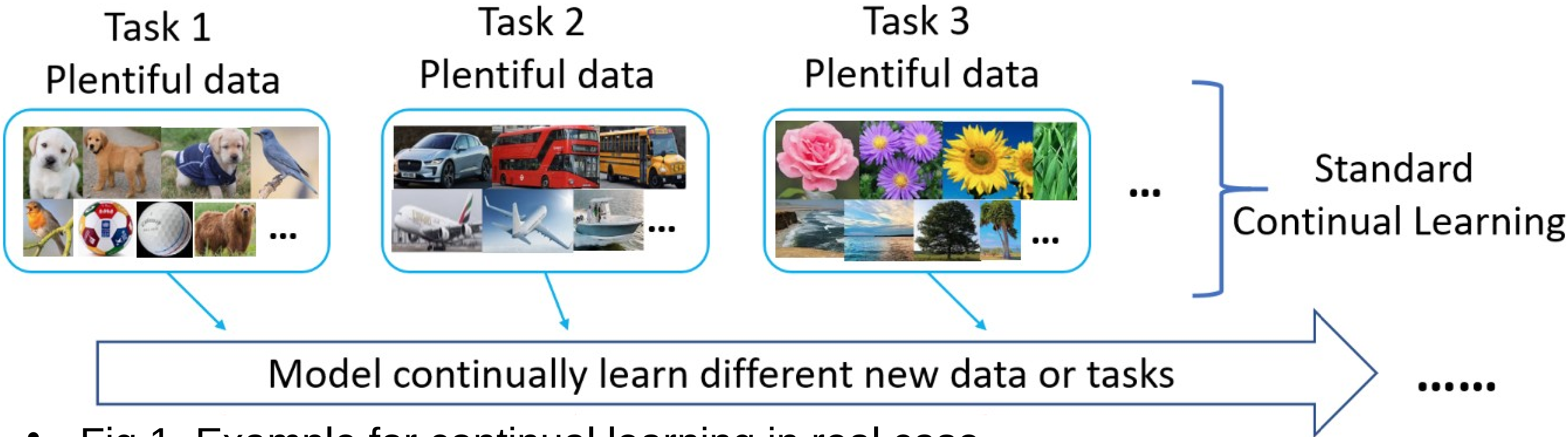
Consideration of compatibility between semi-supervised visual representation learning and incremental learning

20165120 이동건

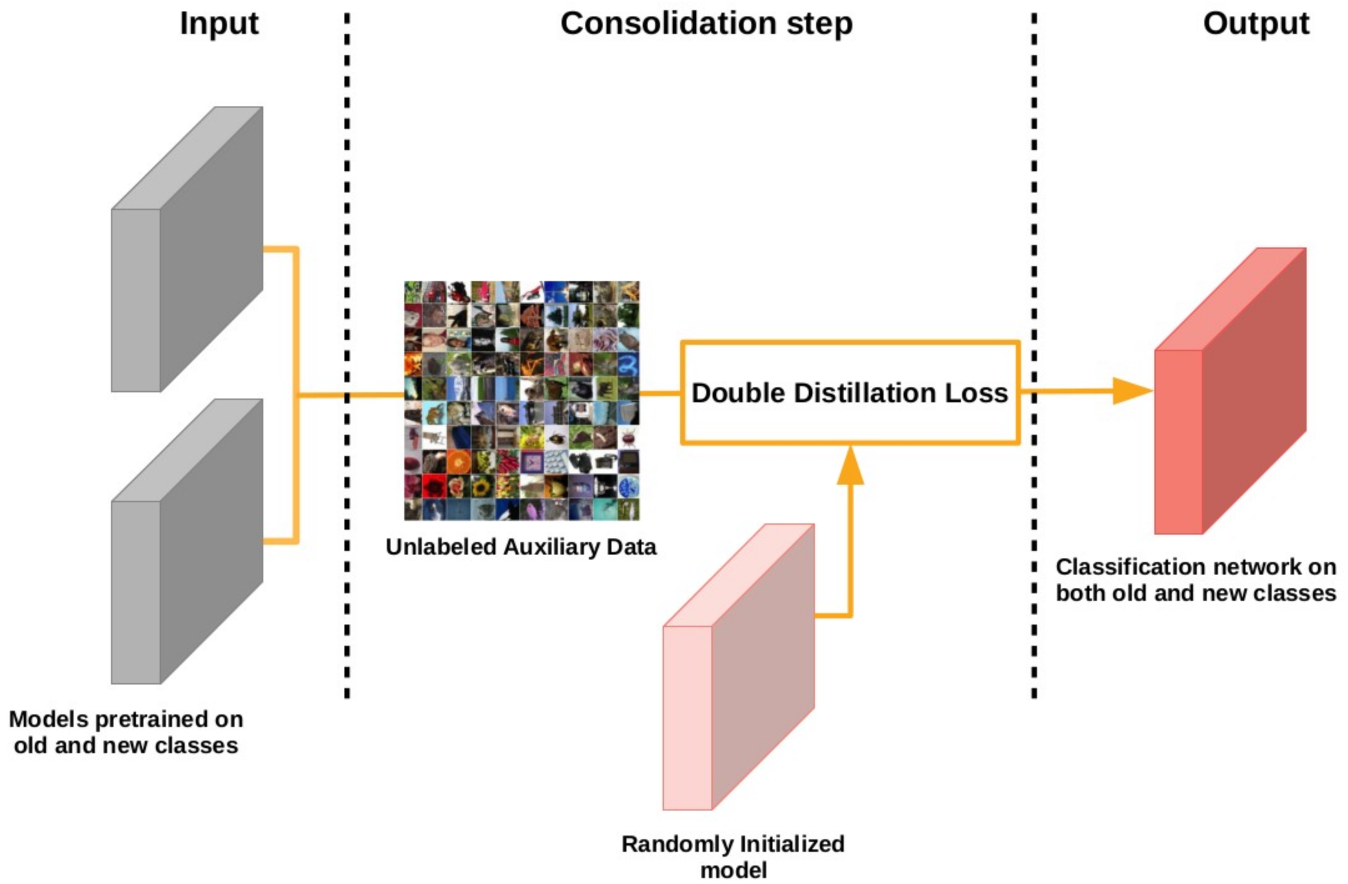
What is continual learning?



- Fig 1. Example for continual learning in real case

- Data stream coming by time

DMC(Deep model consolidation)

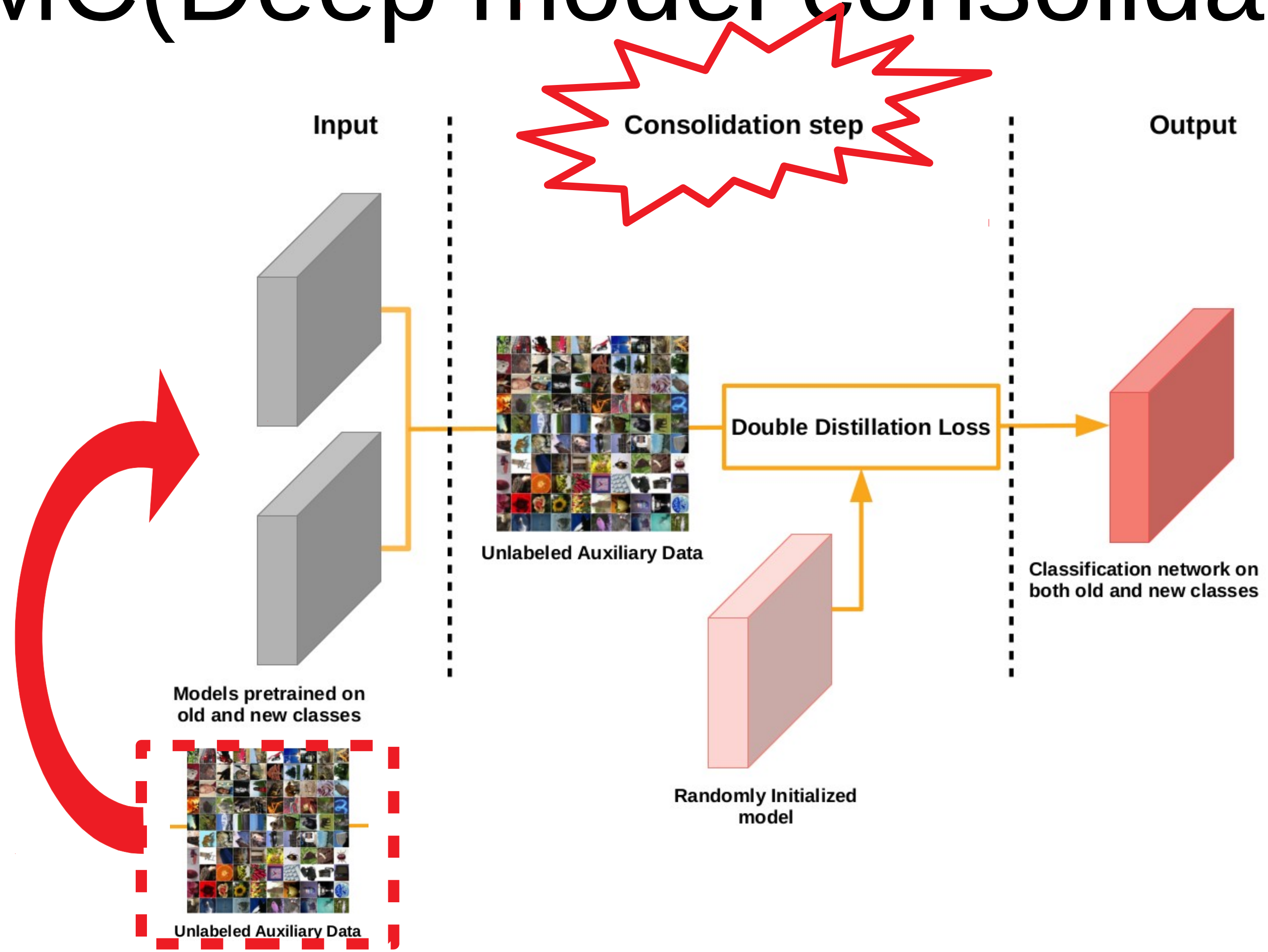


DMC(Deep model consolidation)

Algorithm 1 Consolidation step in DMC

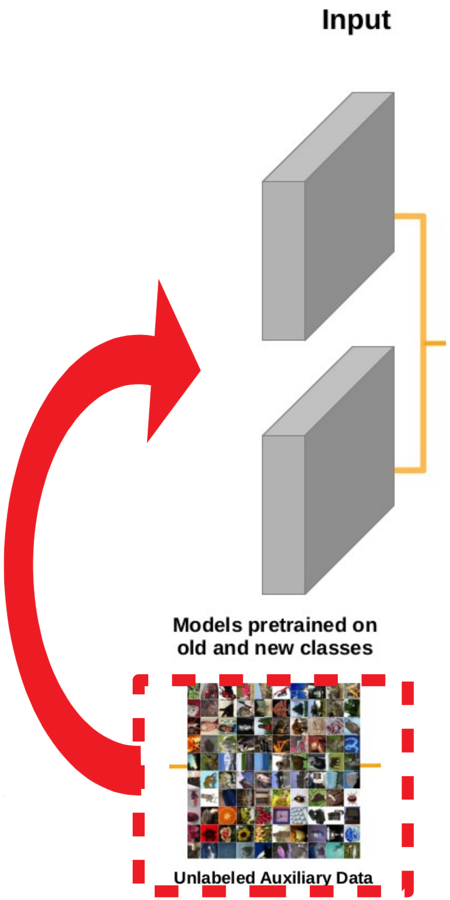
Require: θ_{old} ▷ specified model parameters on old task
Require: θ_{new} ▷ specified model parameters on new task
Require: θ ▷ random initialized model parameters
Require: $D_{unlabel}$ ▷ unlabeled auxiliary data
  for t in [1, epochs] **do**
    for each minibatch B in $D_{unlabel}$ **do**
      $\hat{y}_{old} = f_{old}(x; \theta_{old})$ ▷ logit from old model
      $\hat{y}_{new} = f_{new}(x; \theta_{new})$ ▷ logit from new model
      $\hat{y} = Concatenate(\hat{y}_{old}, \hat{y}_{new})$
      $\dot{y} = Normalize(\hat{y})$ ▷ normalize logit by eq $\boxed{3.2}$
      $y = f(x; \theta)$
      loss = $\frac{1}{|B|} \sum_{i \in B} L_{dd}(\mathbf{y_i}, \dot{\mathbf{y}}_\mathbf{i})$ ▷ loss by eq $\boxed{3.1}$
      update θ using, e.g., ADAM ▷ update network parameters
    end for
  end for
  return θ

DMC(Deep model consolidation)



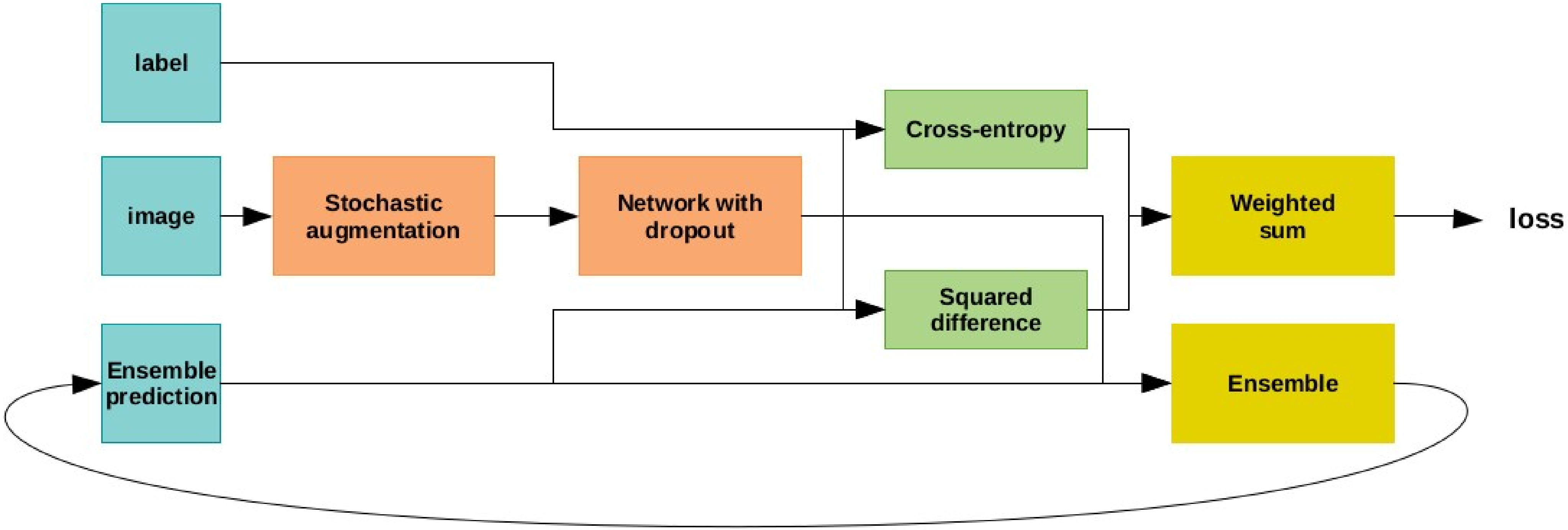
DMC(Deep model consolidation)

How?

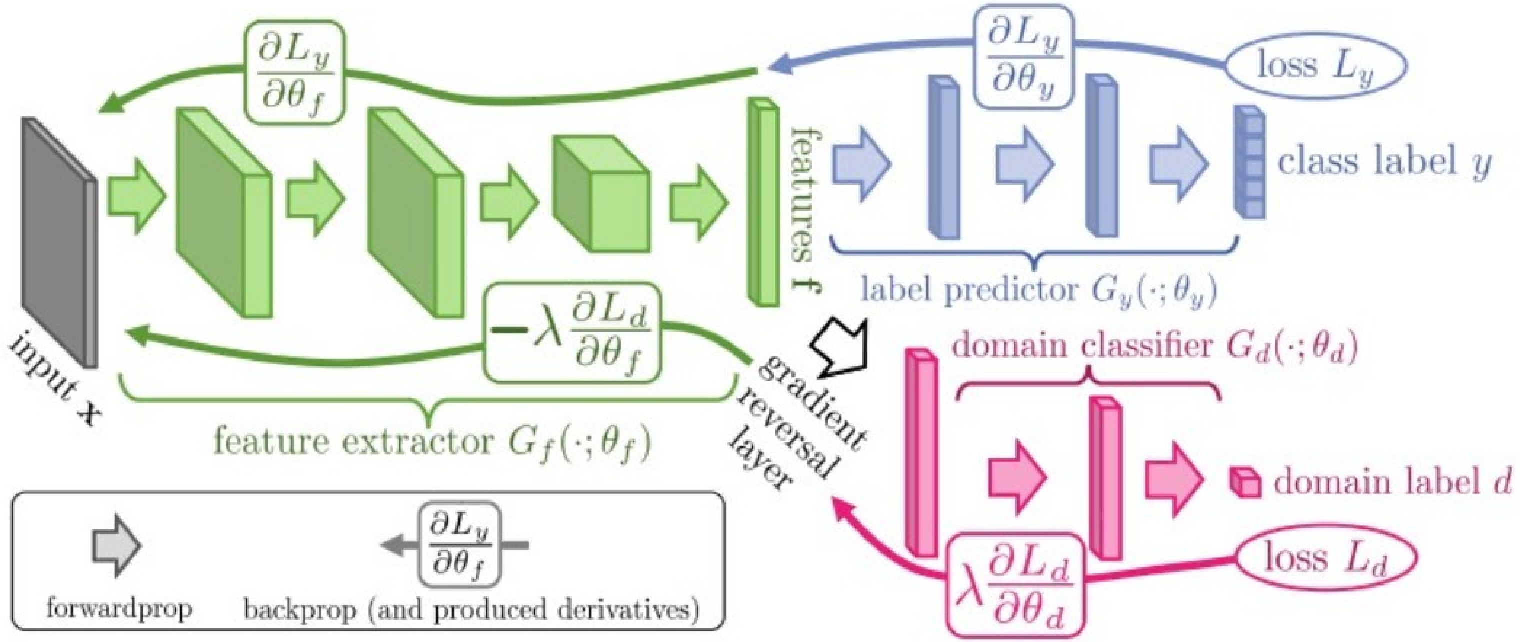


Learning method both labeled data
and unlabeled data
=> **Semi-supervised visual representation Learning**
=> 1. Temporal Ensembling
=> 2. Domain adversarial training

1. DMC with Temporal Ensembling TE-DMC



2. DMC with Domain adversarial training, DAT-DMC



3.DMC with weighting distillation, WD-DMC

Algorithm 1 Consolidation step in DMC

Require: θ_{old}
Require: θ_{new}
Require: θ
Require: $D_{unlabel}$
  for t in [1, epochs] **do**
    for each minibatch B in $D_{unlabel}$ **do**
      $\hat{y}_{old} = f_{old}(x; \theta_{old})$
      $\hat{y}_{new} = f_{new}(x; \theta_{new})$
      $\hat{y} = Concatenate(\hat{y}_{old}, \hat{y}_{new})$
      $\dot{y} = Normalize(\hat{y})$
      $y = f(x; \theta)$
      loss = $\frac{1}{|B|} \sum_{i \in B} L_{dd}(\mathbf{y_i}, \dot{\mathbf{y}}_\mathbf{i})$
      update θ using, e.g., ADAM
    end for
  end for
  return θ

Maximum value of logit

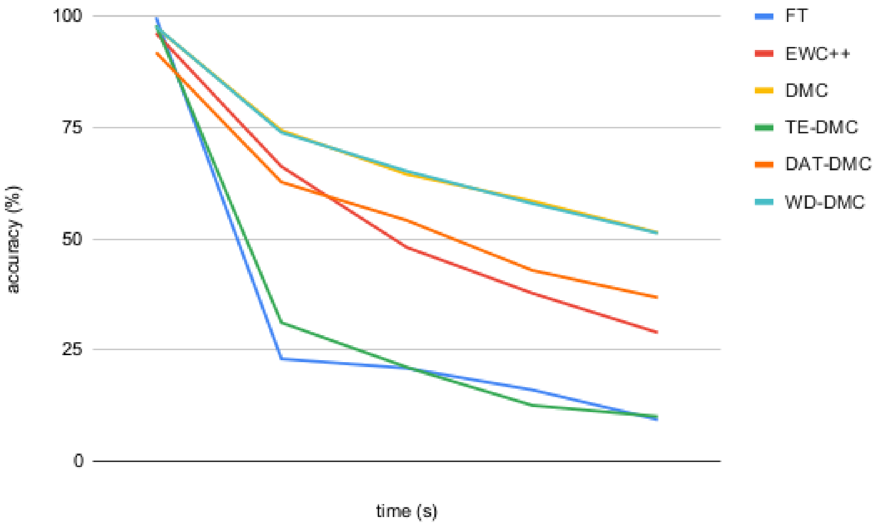
MSE → $\boxed{\text{weighted}}$ sum of squared error

Result

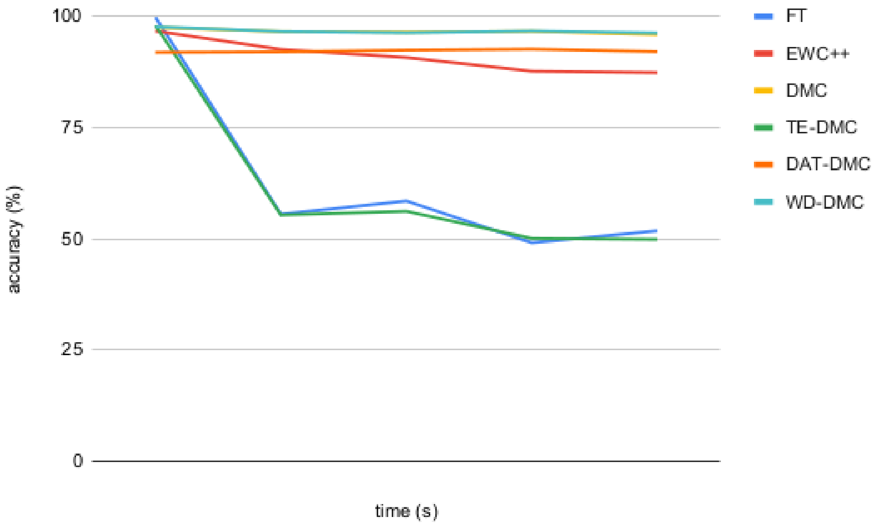
Method	Single-Head Acc. (%)	Multi-Head Acc. (%)
FT	9.29	51.7
EWC++ [4]	28.82	87.34
DMC [38]	51.54	95.94
TE-DMC	10	50.11
DAT-DMC	36.76	92.08
WD-DMC	**51.88**	**96.28**

Table 4.1: **Classification accuracy (single-head and multi-head) on Cifar10.** These are the results of the accuracy of multi-head and single-head, which is evaluated according to whether or not there is prior information about task.

Result



(a) Single-Head Accuracy. (%).



(b) Multi-Head Accuracy. (%)

Figure 4.1: **Test accuracy change along to time.** Except for TE-DMC, all methods perform better than FT.