ADER: Adaptively Distilled Exemplar Replay Towards Continual Learning for Session-based Recommendation

Fei Mi*, Xiaoyu Lin* and Boi Faltings

Motivation

Approaches for session-bases recommendation are developed in an offline manner, in which the recommender is trained on a very large static training set and evaluated on a very restrictive testing set in a *one-time* process. However, a recommender needs to be periodically updated with new data steaming in. In this paper, we study session-based recommendation in a continual learning setup to consider such realistic recommendation scenarios.

Contributions

- The first to study the practical continual learning setting for the session-based recommendation task.
- Propose a method called Adaptively Distilled Exemplar Replay (*ADER*) for this task, and benchmark it with state-of-the-art continual learning techniques.
- Experiment results on two widely used datasets empirically demonstrate the superior performance of *ADER* and its ability to mitigate catastrophic forgetting.

Algorithms

➢ Algorithm to select exemplars at update cycle t

Algorithm 1 *ADER*: ExemplarSelection at cycle t

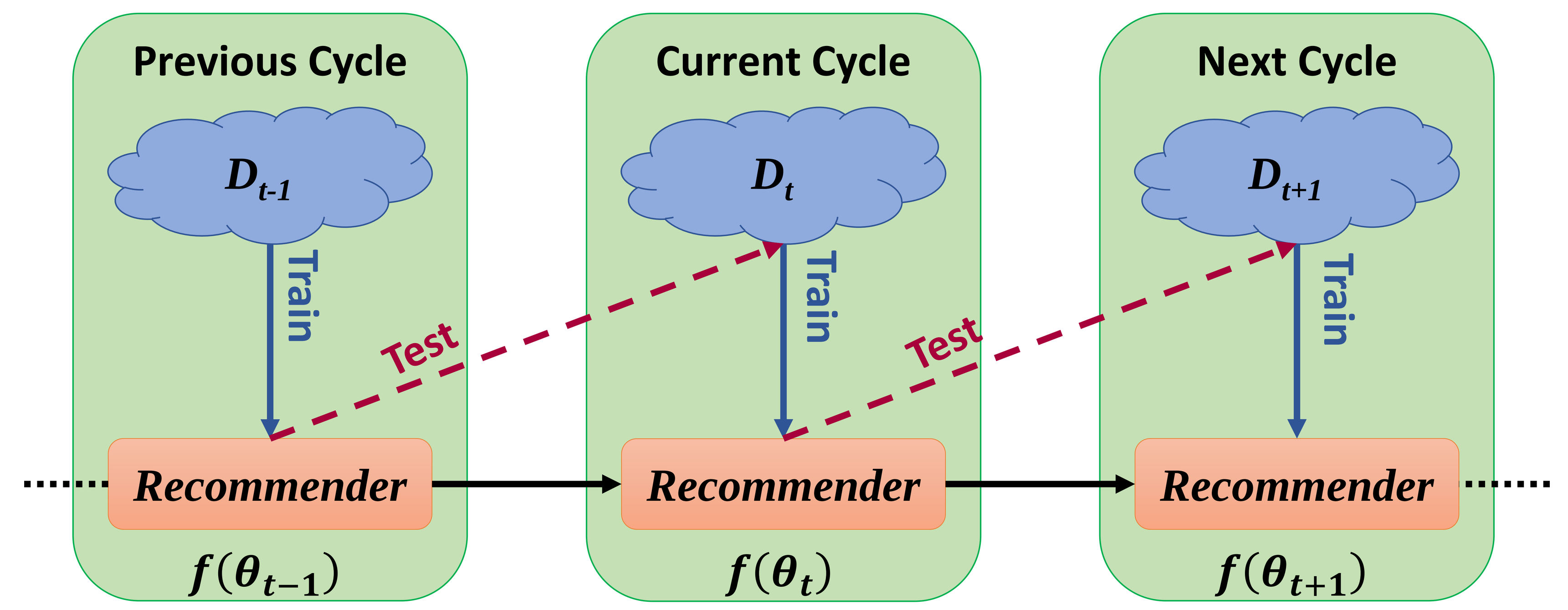
Input: $\mathcal{S} = D_t \cup E_{t-1}$; $M_t = [m_1, m_2, \ldots, m_{|I_t|}]$
  for $y = 1, \ldots, |I_t|$ **do**
    $\mathcal{P}_y \leftarrow \{\mathbf{x} : \forall(\mathbf{x}, y) \in \mathcal{S}\}$
    $\mu \leftarrow \frac{1}{|\mathcal{P}_y|} \sum_{\mathbf{x} \in \mathcal{P}_y} \phi(\mathbf{x})$
    for $k = 1, \ldots, m_y$ **do**
      $\mathbf{x}^k \leftarrow \arg\min_{\mathbf{x} \in \mathcal{P}_y} \|\mu - \frac{1}{k}[\phi(\mathbf{x}) + \sum_{j=1}^{k-1} \phi(\mathbf{x}^j)]\|$
    end for
    $E_y \leftarrow \{(\mathbf{x}^1, y), \ldots, (\mathbf{x}^{m_y}, y)\}$
  end for
Output: exemplar set $E_t = \cup_{y=1}^{|I_t|} E_y$

➢ Algorithm to update model at update cycle t

Algorithm 2 *ADER*: UpdateModel at cycle t

Input: $D_t, E_{t-1}, I_t, I_{t-1}$
  Initialize θ_t with θ_{t-1}
  while θ_t not converged **do**
    Train θ_t with loss in Eq. (2)
  end while
  Compute M_t using Eq. (1)
  Compute E_t using Algorithm 1 with θ_t and M_t
Output: updated θ_t and new exemplar set E_t

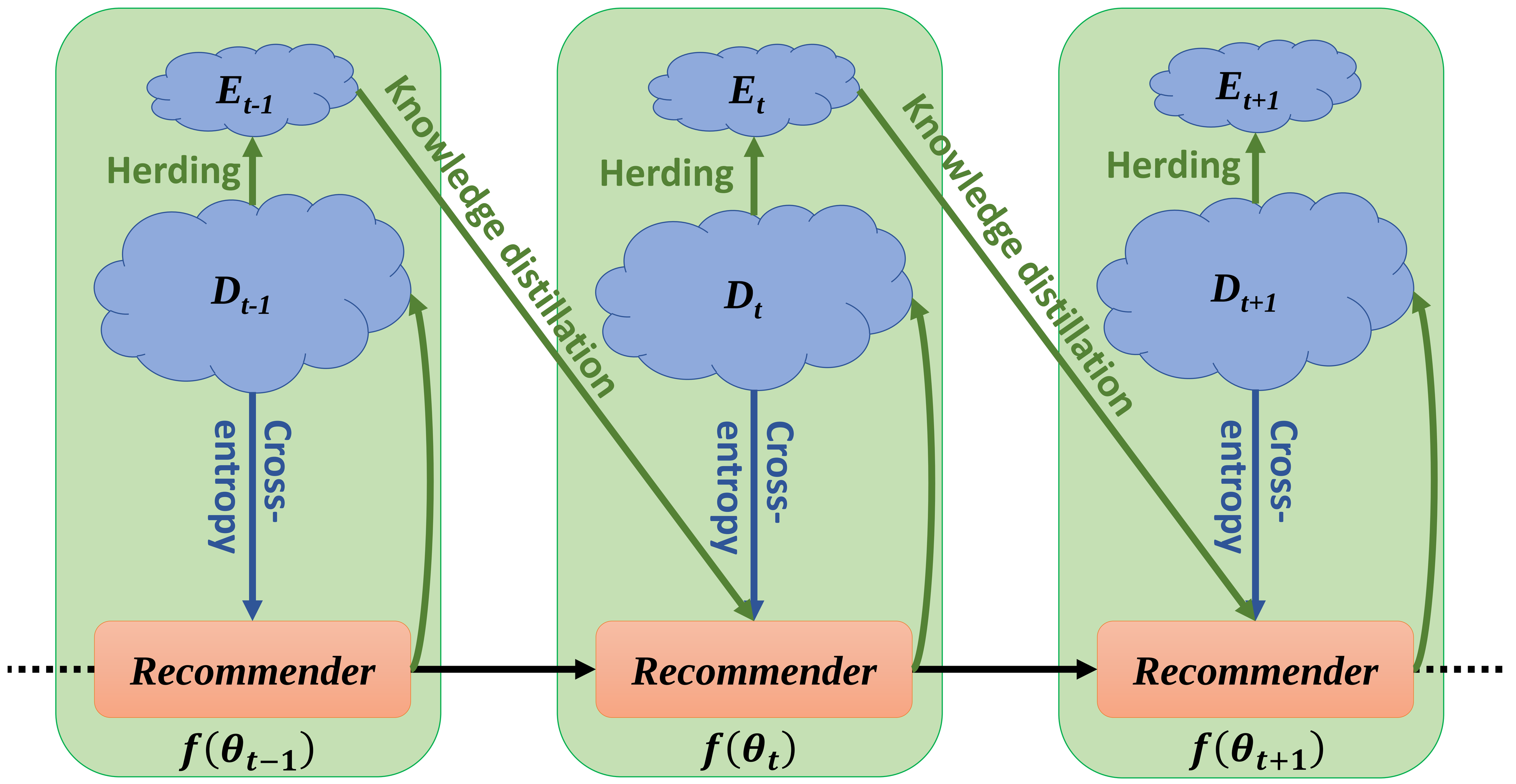
Continual Learning Setup



Notations

- ➢ $f(\theta_{t-1})$: recommendation model obtained until the last update cycle $t-1$.
- ➢ D_t: new incoming data at update cycle t.
- ➢ $f(\theta_t)$: the updated model, after $f(\theta_{t-1})$ is trained on D_t.
- ➢ E_t: exemplar set selected until update cycle t.
- ➢ I_t: the set of appeared items until cycle t.
- ➢ $\phi(\cdot)$: the feature extractor in recommendation model $f(\theta)$.

Proposed Method



Exemplar Selection

➢ **What is the criterion for selecting exemplars of an item/label?**

Herding technique [Welling, 2009; Rebuffi et al., 2017].

➢ **How many exemplars are stored for each item/label?**

The number of exemplars is proportional to its appearance frequency. Suppose N exemplars in total, the number of exemplars $m_{t,i}$ at cycle t for item $i \in I_t$ is:

$$m_{t,i} = N \cdot \frac{|\{x, y = i\} \in D_t \cup E_{t-1}|}{|D_t \cup E_{t-1}|} \qquad \text{Eq. (1)}$$

Proposed Adaptive Distillation Loss

➢ **Knowledge Distillation (KD) loss on exemplars E_{t-1}**

$$L_{KD}(\theta_t) = -\frac{1}{|E_{t-1}|} \sum_{(x,y) \in E_{t-1}} \sum_{i=1}^{|I_{t-1}|} \hat{p}_i \cdot \log(p_i)$$

where $\widehat{P}$ is predicted distribution generated by $f(\theta_{t-1})$, and P is the prediction of $f(\theta_t)$.

➢ **Regular Cross-Entropy (CE) loss computed w.r.t. D_t**

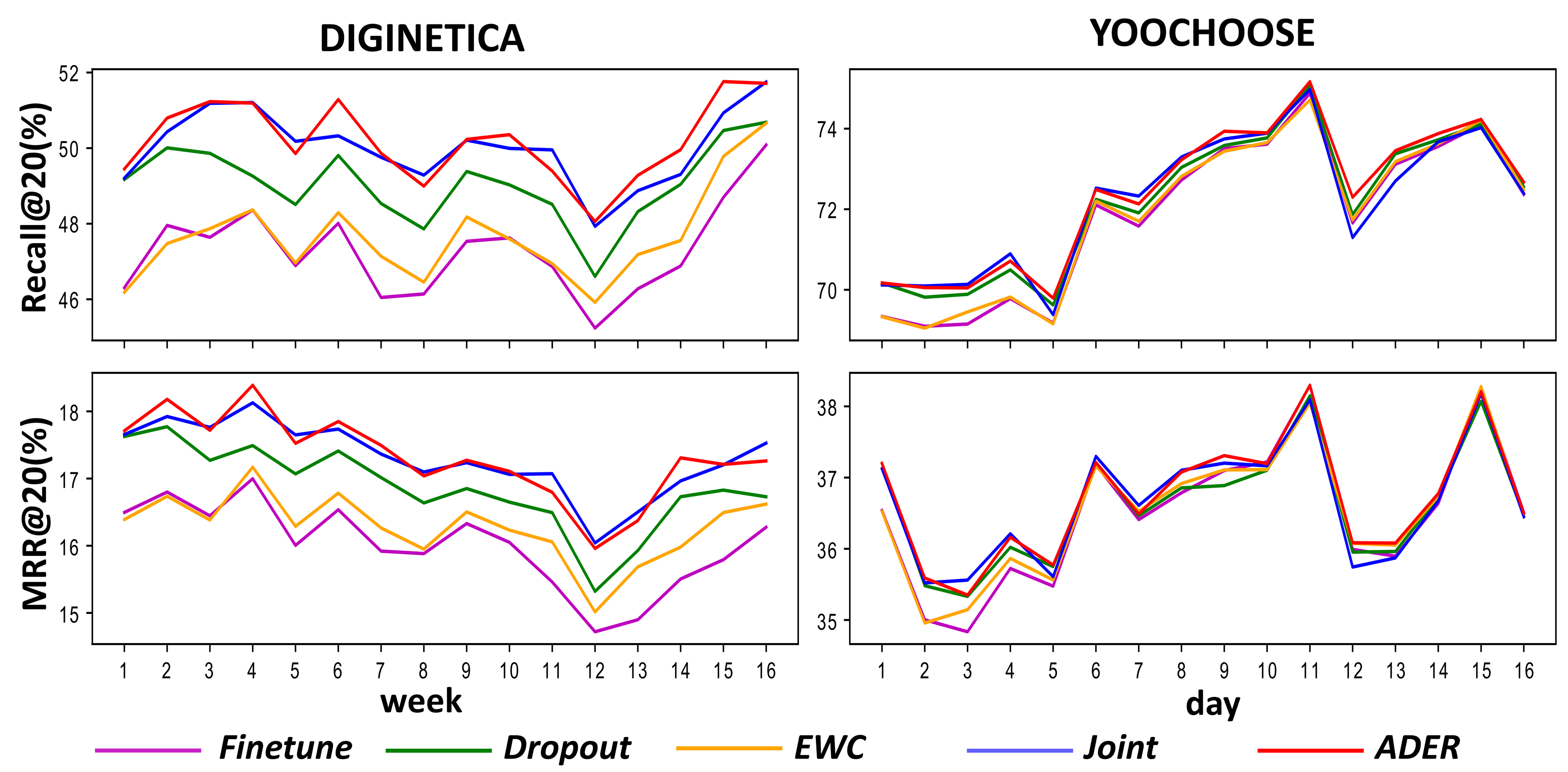
$$L_{CE}(\theta_t) = -\frac{1}{|D_t|} \sum_{(x,y) \in D_t} \sum_{i=1}^{|I_t|} \delta_{i=y} \cdot \log(p_i)$$

➢ **Adaptive Distillation Loss**

$$\lambda_t = \lambda_{base} \cdot \sqrt{{}^{|I_{t-1}|}/_{|I_t|} \cdot {}^{|E_{t-1}|}/_{|D_t|}}, \qquad L_{ADER} = L_{CE} + \lambda_t \cdot L_{KD} \qquad \text{Eq. (2)}$$

Results

	DIGINETICA					YOOCHOOSE				
	Finetune	*Dropout*	*EWC*	*Joint*	*ADER*	*Finetune*	*Dropout*	*EWC*	*Joint*	*ADER*
Recall@20	47.28%	49.07%	47.66%	50.03%	**50.21%**	71.86%	72.20%	71.91%	72.22%	**72.38%**
Recall@10	35.00%	36.53%	35.48%	37.27%	**37.52%**	63.82%	64.15%	63.89%	64.16%	**64.41%**
MRR@20	16.01%	16.86%	16.28%	17.31%	**17.32%**	36.49%	36.60%	36.53%	36.65%	**36.71%**
MRR@10	15.16%	16.00%	15.44%	16.43%	**16.45%**	35.92%	36.03%	35.97%	36.08%	**36.14%**



➢ Ablation Study (on DIGINETICA dataset)

- *$ER_{herding}$* : A vanilla exemplar replay by using L_{CE}, rather than L_{KD}, on exemplars.
- *ER_{random}* : It differs from *$ER_{herding}$* by selecting exemplars of an item at random.
- *ER_{loss}* : It differs from *$ER_{herding}$* by selecting exemplars of an item with smallest L_{CE}.
- *$ADER_{equal}$* : This differs from *ADER* by selecting equal number of exemplars for each item.
- *$ADER_{fix}$* : It differs from *ADER* by not using the adaptive λ_t in Eq.(2), but a fixed λ.

	$ER_{herding}$	*ER_{random}*	*ER_{loss}*	*$ADER_{equal}$*	*$ADER_{fix}$*
Recall@20	49.44%	49.14%	49.31%	49.92%	50.09%
Recall@10	36.88%	36.61%	36.65%	37.21%	37.41%
MRR@20	16.95%	16.79%	16.90%	17.23%	17.29%
MRR@10	16.08%	15.92%	16.02%	16.35%	16.41%

https://dl.acm.org/doi/abs/10.1145/3383313.3412218

fei.mi@epfl.ch