Paper Reading

2023.01.19

Guan Yunyi

Paper - ICCV 2021

FLAR: A Unified Prototype Framework for Few-sample Lifelong Active Recognition

Lei Fan, Peixi Xiong, Wei Wei and Ying Wu

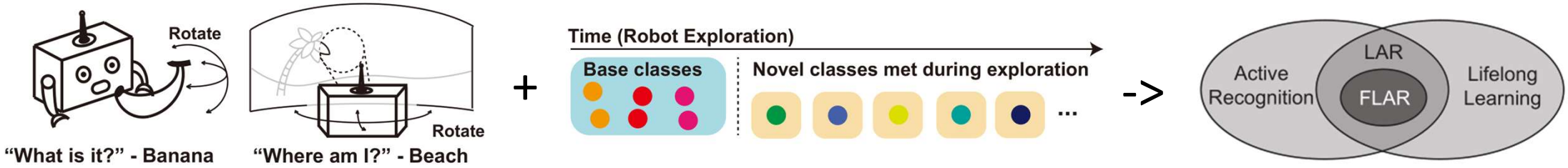
Northwestern University, 2145 Sheridan Road, Evanston, IL, USA

01

Introduction

1.1 Introduction

- **Previous AR research:** fixed-category setting
- recognition can only be made for the samples from the trained categories
- massive training data are usually required
- **Main idea:** propose a unified framework towards Few-sample Lifelong Active Recognition (FLAR), which aims at performing active recognition on progressively arising novel categories that only have few training samples



active recognition lifelong learning

1.2 Requirements and challenges

- **F: few-shot learning,** learn new concepts from limited training samples

  -> overfitting

  -> **prototype to represent each category**

- **L: Lifelong (incremental learning),** adapt recognition learned from old classes to new concepts while avoiding training from scratch

  -> catastrophic forgetting issue

  -> store limited instances in **memory** & **knowledge distillation**

- **AR: active recognition,** making decisions to explore the most informative viewpoints based on the current stage with few training samples

  -> will impede the success of the policy training

  -> a newly designed **reward**

02

Methods

2. Settings instance x

- **AR:** exploration+ aggregating observations + classification

current view

- viewpoint: **discrete**, M azimuths ×N elevations $p_t = (m, n)$ $v_t = \mathcal{P}(x, p_t)$
- Observations: $\mathcal{X}_t = h(v_t, p_{t-1,t}, t)$, where $h(\cdot)$ is a fuse operation
- Representation for object: $q_x = f(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T)$

- **Lifelong:**

- class-incremental task stream $X^{base}, X^1, X^2, \ldots, X^y, \ldots$ $X^y = \{x_1^y, \ldots, x_k^y\}$
- Since the high expense of collecting training samples for new categories, limit k to 3,5,10

2.1 Prototype-guided active recognition - Prototype representation

- **F:** few-shot learning -> overfitting

  -> **prototype** to represent each category by **averaging the aggregations of the observations for each sample**

- Representation for instance x (feature): $q_x = f(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T)$
- Features from the same class: $Q^y = \{q_1^y, q_2^y, \ldots\}$
- **Prototype**: $\mu^y = \frac{1}{|Q^y|} \sum_{q \in Q^y} q$ a weight vector from the final linear layer
- label: $\hat{y} = \mathbf{argmin}_y ||q_x - \mu^y|| = \mathbf{argmax}_y \mu^y q_x$

  Euclidean distances
- category loss: $\mathcal{L}_{category} = -\sum_i F_{softmax}(\hat{y}^i, y^i),$

  i denotes the corresponding training sample

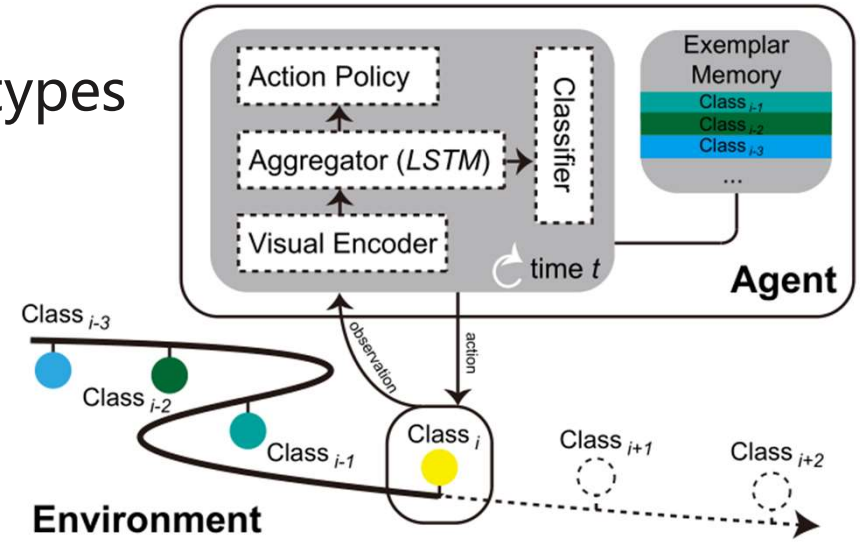
2.1 Prototype-guided active recognition – AR system

- **Visual encoder** (simple 3-layer network) + **LSTM**
- **Policy**:

- Reward: growth of predicted probability of correct category $R(\hat{y}_t, \hat{y}_{t+1}) = 1$
- Policy loss: REINFORCE

$$\mathcal{L}_{policy} = \sum_i \sum_{t=1}^{T-1} \log \pi(a_t^i | \mathcal{X}_{t-1}, \theta) R(\hat{y}_t, \hat{y}_{t+1})^i.$$

- **Classifier**: a linear layer with weight of prototypes



2.1 Prototype-guided active recognition – Other losses

- **Entropy loss**: $\mathcal{L}_{entropy}$

on the action distribution, to promote more exploratory behavior of agent and prevent policy collapse

- **Forecast loss**:

$$\mathcal{L}_{forecast} = \sum_{i} \sum_{t=2}^{T} D(\hat{\mathcal{X}}_t^i, \mathcal{X}_t^i | \mathcal{X}_{t-1}^i, a_{t-1}),$$

the cosine distance between observations inferred using aggregated information and observations obtained using a prediction module (a simple network)

2.2 Lifelong learning on novel classes - Agent memory

- One way to handle catastrophic forgetting in lifelong learning: entangle weight in the classifier with the representation learning process
- Weight in classifier = prototypes -> change along with learning

  -> data distribution of previous classes should be introduced to current training

  -> **memory** to store object instances that best describe the current category
- Save limited m instances $M^y = \{x_1, x_2, \ldots, x_m\}$

- instance is selected if its prototype (the average feature vector) best approximate the overall prototype of training data
- Save in the form of view-grids of m instances
- Selection processes can be done one time for each category

2.2 Lifelong learning on novel classes - Distillation loss

- How to achieve lifelong?

-> **Encourage reproducing the same output of saved instances**

- Before training on new classes, implement recognition episodes on saved instances to get their classifier outputs z

$$\mathcal{L}_{distillation} = -\sum_{i} \sum_{y \in C^{known}} F_{BCE}(z_i^y, f(x_i^y)),$$

F_{BCE} denotes the Binary Cross Entropy function

$$\mathcal{L} = \mathcal{L}_{category} + \mathcal{L}_{policy} + \mathcal{L}_{entropy} + \mathcal{L}_{forecast} + \mathcal{L}_{distillation}.$$

Algorithm 1: Training on task X^i

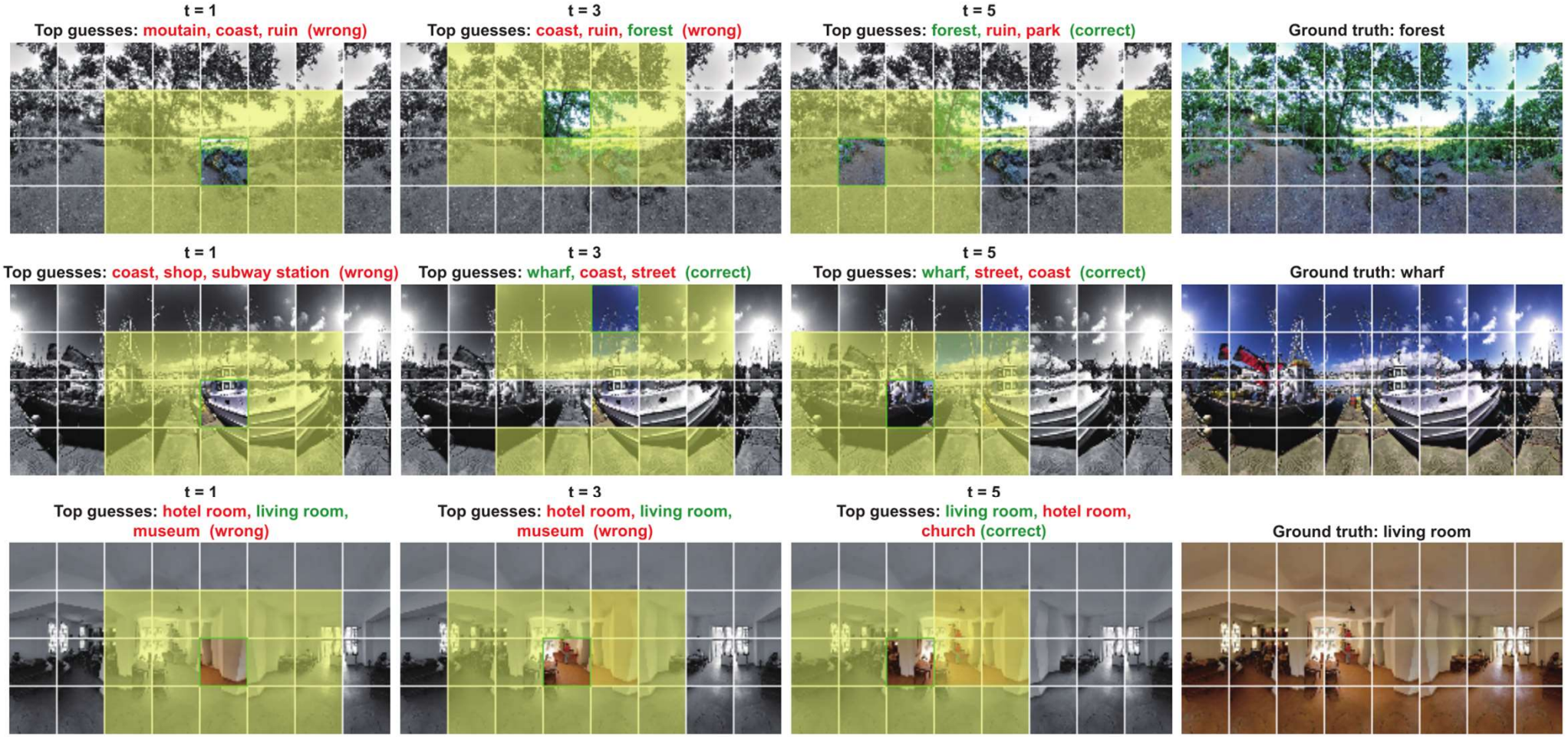
Input: X^i: current task from the stream
Require: f: recurrent embedding module
Require: $Agent$: *AR* system with policy π
Require: $M^{i-1} = \{M^y, y \in C^{known}\}$: memory
$\mathcal{D} = X^i \cup M^{i-1}$: training set
// store network updates for the distillation loss
for $y \in C^{known}$ **do**
  Perform *AR* for all $x_i \in \mathcal{D}$ to get z_i^y
end
// network training
while *epoch reaches maximum* **do**
  Perform *AR* for all $x_i \in \mathcal{D}$
  Back propagate $\mathcal{L}$ defined in Equation 6
end
Update $C^{known} \leftarrow C^{known} \cup y^i$
Update agent memory to M^i

Experiments

3.1 AR result

- Dataset: SUN360

correct wrong guesses within 5 steps



green box: current view, light yellow area: the next movement grid

3.1 AR result

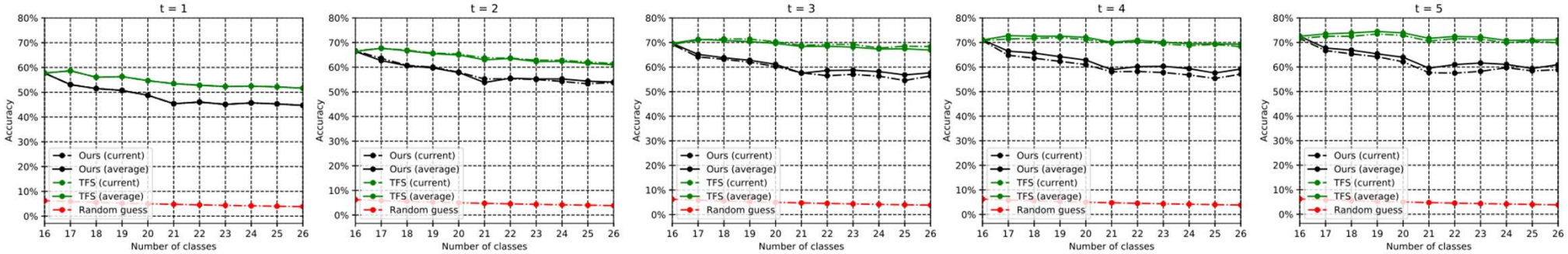
- Dataset: SUN360
- Large improvements denote the advantages of including effective policies during recognition

Method	t = 2 acc.		t = 3 acc.		t = 5 acc.	
	curr.	*avg.*	*curr.*	*avg.*	*curr.*	*avg.*
Single view	51.6	51.6	51.6	51.6	51.6	51.6
Random views	55.6	56.5	57.7	59.1	59.8	62.3
Largest step	54.7	55.7	53.6	56.6	52.4	55.8
Look-Ahead [18]	59.8	60.2	67.8	66.3	69.4	70.6
Ours	**61.5**	**61.0**	**68.4**	**67.0**	**69.9**	**71.1**

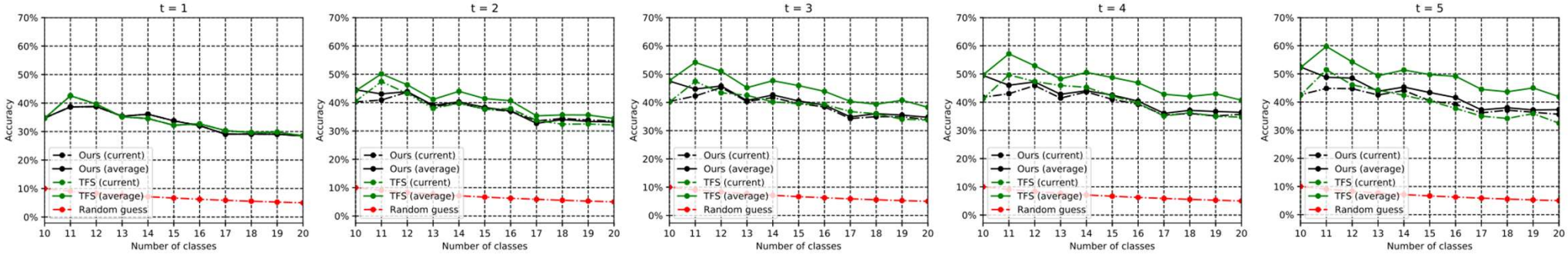
Table 1. Recognition accuracy on the SUN360 dataset [40]. The *curr.* denotes results with current estimates while *avg.* is the average of class likelihoods up to the current step.

3.2 Lifelong learning results

- Better performance on SUN360 than ShapeNet because of larger searching space
- Performance arises with taking more steps



(a) The result of the proposed approach on the SUN360 dataset [40].



(b) The result of the proposed approach on the ShapeNet dataset [10].

Figure 4. Recognition accuracy for both datasets. The method `TFS` is short for Training From Scratch, which could access sufficient data of all categories. The method `Random Guess` defines the lower bound of our performance.

04

Conclusion

4. Concultion

- First research of FLAR: **incrementally** learns **active recognition** on novel categories from **few-shot** samples
- My opinion: AR part basically follows the past research completely

- discrete viewpoint
- discrete action
- only the maximum time step is set