Improving Continual Learning of Acoustic Scene Classification via Mutual Information Optimization

Muqiao Yang*, **Umberto Cappallazzo†**, **Xiang Li***, **Bhiksha Raj***

*Cargegie Mellon University, †University of Trento

Introduction

Continual Learning: aims to address catastrophic forgetting that makes the model have tendency to abruptly erase past knowledge while learning new tasks; to incrementally accumulate knowledge over time like human perception.

The objective requires the model to absorb both task-specific and task-agnostic knowledge to adapt to different domains. Focusing on acoustic scene classification, we demonstrates that mutual information can help the feature extractor learn task-agnostic knowledge, while helping the classifier learn task-specific knowledge.

(1) For the feature extractor part, we first present that it is theoretically sound to learn **task-agnostic** knowledge by maximizing the MI between the feature representations of the original input and an augmented acoustic scene of the same input.

(2) For the classifier part, we show that by selecting the memory samples with a combination of surprise and learnability criteria, the samples are expected to be both **representative and informative** to boost the continual learning performance of the acoustic scene classification model.

Background

Problem Statement

Class-incremental learning (CIL): new classes of acoustic scenes may keep appearing in continuous streams of data. Compared to another category of continual learning, i.e., task-incremental learning, CIL does not have access to task identities during inference time. Therefore, its objective is to build a holistic classifier among all of the seen classes by making use of the label information only.

Notations and Augmentations

Our mutual information optimization relies on the comparisons between different augmented representations of acoustic scenes, which are also called pseudo-labeled samples.

Augmentations may include: add Gaussian noise, apply band-stop filtering, or invert along the time axis, etc.

X/X': original/augmented input

Z/Z': feature representation of original/augmented input

Y: prediction logits

I(•,•): Mutual Information

H(•,•): Shannon/conditional Entropy

Method

Feature extractor

We would like to guarantee that the encoded representations can preserve sufficient information from the original inputs regardless of their classes. Therefore, as shown in the equation, maximizing the MI between Z and Z′ is equivalent to maximizing the lower bound of the MI between input X and the encoded features Z. The MI is further estimated through the infoNCE (noise contrastive estimation) loss.

$$\begin{aligned} I(X;Z) &= H(Z) - H(Z|X) \\ &= H(Z) - H(Z|X, Z') \\ &\geq H(Z) - H(Z|Z') = I(Z;Z') \\ &\geq \underbrace{\frac{1}{N}\sum_{i=1}^{N} \log \frac{(f(z_i, z_i')/\tau)}{\sum_{j=1}^{N}(f(z_i, z_j')/\tau)}}_{\triangleq \mathcal{L}_{\mathrm{NCE}}(Z,Z')} \end{aligned}$$

Intuition: The feature extractor would like to extract task-agnostic knowledge such that the mutual information between the original inputs and encoded feature can be maximized.

Methods cont'd

Classifier

We sample from memory to recall past knowledge. We expect the selected samples only bring extra information but also make sure the new information can be effectively learned by the model.

$$\begin{aligned} &\mathcal{L}_{\mathrm{NCE}}(Z, \{Z'\}, Y) \\ &= \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{1}{\sum_{k=1}^{N} \mathbb{1}(y_k = y_i)} \sum_{y_k = y_i} \Big(\sum_{\hat{z_i} \in \mathcal{S}_{z_i}} \log(f(z_i, \hat{z_i})/\tau) \\ &\quad - \log \sum_{j=1}^{N} \Big(\sum_{\hat{z_j} \in \mathcal{S}_{z_j}} f(z_j, \hat{z_j})/\tau \Big)\Big)\Big] \end{aligned}$$

$$\begin{aligned} \mathrm{score}_t(Y, Z) = &-\mathcal{L}_{\mathrm{NCE}}(Z_{t-1}, \{Z'_{t-1}\}, Y_{t-1}) \\ &+ \mathcal{L}_{\mathrm{NCE}}(Z_t, \{Z'_t\}, Y_t) \end{aligned}$$

Intuition: Two criteria for sample selection, reflected in the scoring equation.

Surprise (representative) is to favor samples that brings more surprise from past knowledge to the current model

Learnability (informative) is to favor samples with higher learnability, since they maximize the MI between Z and Z' given Y by the current model, which aligns with our objective function.

Experiments & Results

Experimental setting

We compare our mutual information based methods with other continual learning methods including Random sampling, Herding sampling, Gradient-based sample selection (GSS), and uncertainty-based sampling. Fine-tune means fline training without any continual learning approaches performed, which is the lower bound of our performance.

Evaluation Metric

We use average Acc, backward transfer (BWT) and forward transfer (FWT) to show that our method helps not only learn task-agnostic knowledge, but also preserve the task-specific knowledge.

Method	Memory Size	Acc ↑	BWT ↑	FWT ↑	Memory Size	Acc ↑	BWT ↑	FWT ↑
fine-tune	-	19.1	-58.7	0.0	-	20.4	-56.0	0.0
Random	0.2k	22.5	-52.5	26.6	0.2k	42.8	-28.5	49.8
	0.5k	24.6	-49.7	27.3	0.5k	49.8	-27.8	54.3
	1k	26.2	-47.6	29.7	1k	52.6	-27.0	59.2
Herding [24]	0.2k	47.5	-30.8	49.3	0.2k	51.6	-26.9	56.0
	0.5k	49.3	-28.7	50.6	0.5k	54.3	-26.3	63.3
	1k	50.8	-27.9	52.2	1k	56.2	-24.8	65.2
GSS [25]	0.2k	48.8	-30.3	49.8	0.2k	51.9	-25.3	56.5
	0.5k	49.6	-29.3	50.8	0.5k	54.6	-25.8	62.9
	1k	50.3	-28.2	51.9	1k	56.1	-24.6	63.7
Uncertainty [27]	0.2k	50.9	-28.9	51.6	0.2k	55.9	-24.5	63.8
	0.5k	51.8	-27.6	53.1	0.5k	57.6	-23.7	67.5
	1k	52.9	-27.1	53.9	1k	58.9	-22.8	69.0
MIO (Ours)	0.2k	52.1	-28.5	53.4	0.2k	58.0	-23.5	64.7
	0.5k	53.7	-27.4	55.9	0.5k	60.7	-22.9	69.1
	1k	**55.3**	**-26.5**	**57.3**	1k	**64.1**	**-22.5**	**74.8**

Table 1. Quantitative results for continual learning on TAU Urban Acoustic Scenes and Environmental Sound Classification-50 with different memory selection methods and size.

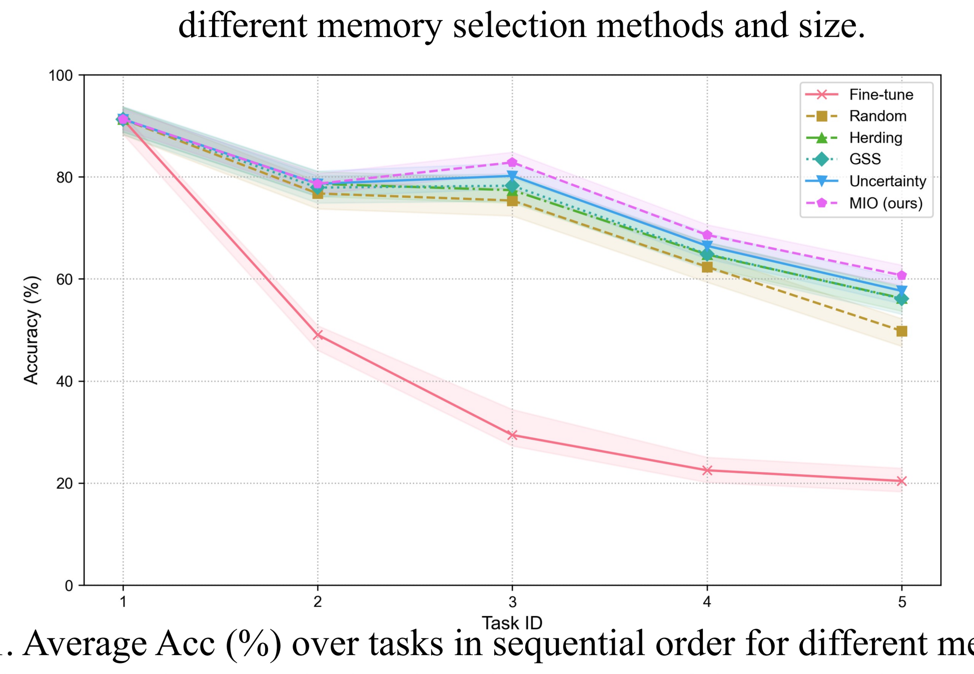

Fig. 1. Average Acc (%) over tasks in sequential order for different methods. The accuracies are calculated on the test sets of the seen tasks so far.

Conclusion

We propose to optimize different levels of the model to learn task-agnostic and task-specific knowledge from the perspective of mutual information, and select samples from the memory buffer that are both representative and informative..