Overcoming Catastrophic Forgetting for Continual Learning via Feature Propagation

Xuejun Han[1], Yuhong Guo[1,2]

[1] Carleton University, Ottawa, Canada [2] Canada CIFAR AI Chair, Amii, Canada

We propose a general feature-propagation based contrastive continual learning method in an online fashion:

- *feature propagation* to align the current and previous representation spaces;
- *contrastive representation learning* to bridge the domain shifts among distinct tasks;
- *supervised contrastive learning* to further mitigate the class-wise shifts in the feature space

The extensive experiments are implemented in multiple *image classification* tasks.

Problem Setting

- Consider a data stream of unknown distributions $\{D_1, \ldots, D_T\}$.
- In the continual learning (CL), the model g is expected to learn the tasks sequentially - it can only learn one task at a time without forgetting what has learned in the past.
- In the online CL, the model can only experience the current dataset D_t once.
- For rehearsal-based CL, a small memory set M storing a few past examples is reserved and can be revisited multiple times along the learning of new tasks.
- Thus, the objective of the continual learner g at task t is to minimize the following loss:

$$\mathcal{L}_{er} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t\cup\mathcal{M}}\, l(g(\mathbf{x}), y)$$

which is also known as *experience replay.*

Methodology

Experience Replay with Feature Propagation

- Suppose the model consists of a feature extractor ψ and classifier f.
- Denote ψ_o is the model state before learning current task.
- Apply feature propagation procedure to training data $D_t \cup M$ in the feature space.
- The example embeddings derived from ψ are amended by fusing a weighted sum of all example embeddings come from ψ_o:

$$\tilde{\psi}(\mathbf{x}_i) = (1-w)\cdot\psi(\mathbf{x}_i) + w\cdot\sum_{\mathbf{x}_j\in\mathcal{D}_t\cup\mathcal{M}} A_{ij}\,\psi_o(\mathbf{x}_j)$$

- The propagation weight A_{ij} for examples x_i and x_j is set as:

$$A_{ij} = \frac{\exp(-d(\psi(\mathbf{x}_i), \psi_o(\mathbf{x}_j))\cdot\eta)}{\sum_{\mathbf{x}_{j'}\in\mathcal{D}_t\cup\mathcal{M}}\exp(-d(\psi(\mathbf{x}_i), \psi_o(\mathbf{x}_{j'}))\cdot\eta)}$$

  where $d(\cdot,\cdot)$ is the Euclidean distance.

- The fusion of the two terms is motivated by the exponential moving average which can move forward without forgetting the past.
- In consequence, the amended experience replay loss is

$$\mathcal{L}_{er'} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t\cup\mathcal{M}}\, l(f(\tilde{\psi}(\mathbf{x})), y)$$

Contrastive Representation Rehearsal

- Enforce the example embeddings to stay near the previous corresponding ones;
- Protect the representation space from dramatically changing after being retrained on new tasks.
- The proposed contrastive loss is defined as follows,

$$\mathcal{L}_{cl} = -\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_t\cup\mathcal{M}} \log \frac{\exp(-d(\psi(\mathbf{x}), \psi_o(\mathbf{x}))\cdot\tau)}{\sum_{\mathbf{x}_j\in\mathcal{D}_t\cup\mathcal{M}}\exp(-d(\psi(\mathbf{x}), \psi_o(\mathbf{x}_j))\cdot\tau)}$$

❖ *Up to this point, the contrastive continual learning with feature propagation (**CCL-FP**) is define as:*

$$\mathcal{L}_{ccl-fp} = \mathcal{L}_{er'} + \alpha\mathcal{L}_{cl}$$

Supervised Contrastive Replay

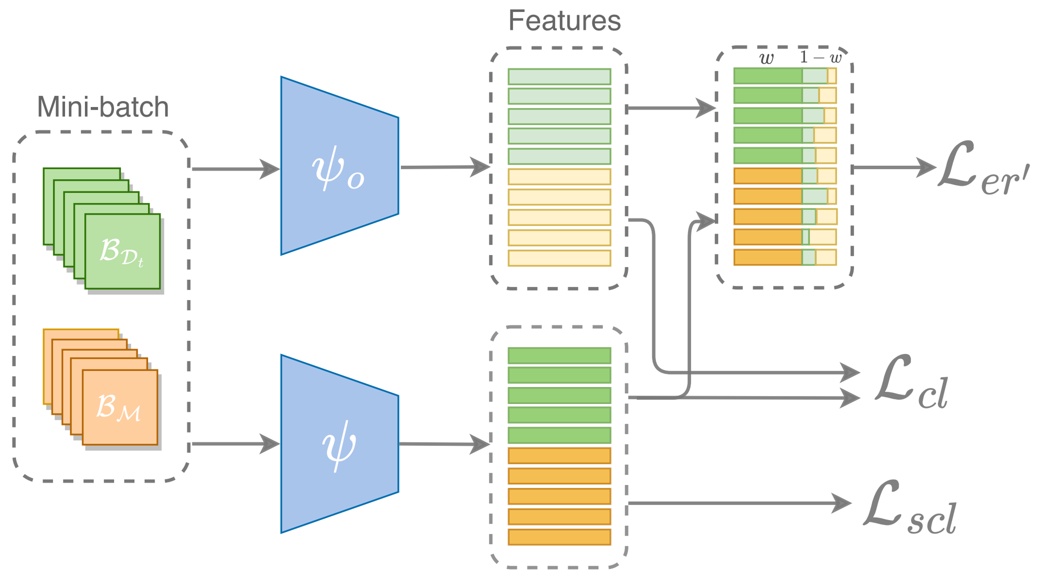
- To further improve the feature space by a supervised contrastive loss.
- Try to eliminate the domain and class shifts between tasks.
- The supervised contrastive loss is defined as follows,

$$\mathcal{L}_{scl} = -\mathbb{E}_{\mathbf{x}_i\sim\mathcal{D}_t\cup\mathcal{M}}\mathbb{E}_{\mathbf{x}_k\sim\mathcal{S}_i} \log \frac{\exp(-d(\psi(\mathbf{x}_i), \psi(\mathbf{x}_k))\cdot\tau)}{\sum_{\mathbf{x}_j\in(\mathcal{D}_t\cup\mathcal{M})\setminus\mathbf{x}_i}\exp(-d(\psi(\mathbf{x}_i), \psi(\mathbf{x}_j))\cdot\tau)}$$

❖ *By integrating the supervised contrastive loss into the overall objective, we obtain our intensified model named by **CCL-FP+**:*

$$\mathcal{L}_{ccl-fp+} = \mathcal{L}_{er'} + \alpha\mathcal{L}_{cl} + \beta\mathcal{L}_{scl}$$

➢ An overview of the proposed method:



Experiments

Datasets

We compare our methods with several state-of-the-art continual learning methods on Split MNIST, Permuted MNIST, Rotated MNIST, Split CIFAR-10, Split CIFAR-100 and Split Tiny ImageNet.

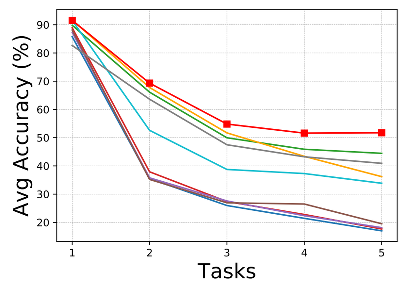
Evaluation metric

We use *average accuracy* (↑) to evaluate the overall performance of models on test data of all seen tasks, defined as follows,

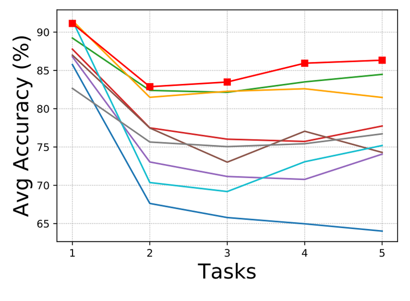
$$ACC_t = \frac{1}{t}\sum_{k=1}^{t} R_{t,k}$$

which indicates the average accuracy on test data of task 1 to t after the model has learned continually up till task t.

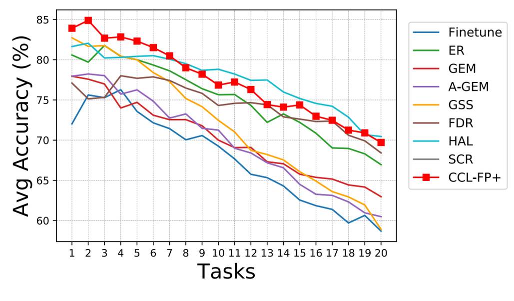
Experimental Results



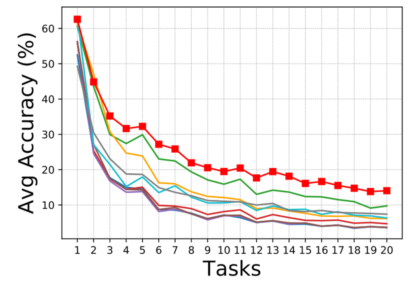
(a) S-CIFAR-10 | Class-IL



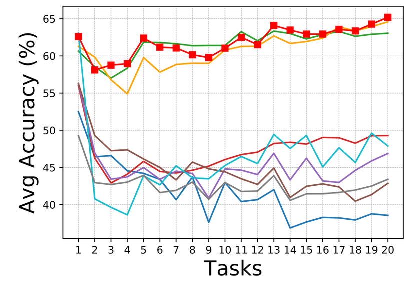
(b) S-CIFAR-10 | Task-IL



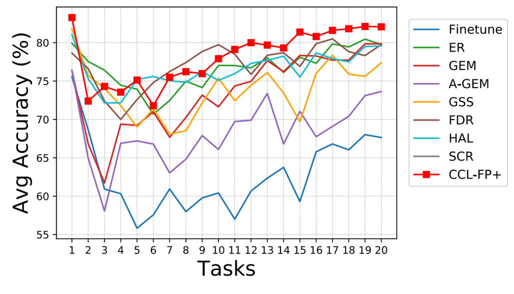
(c) P-MNIST | Domain-IL



(d) S-CIFAR-100 | Class-IL



(e) S-CIFAR-100 | Task-IL



(f) R-MNIST | Domain-IL

Model	S-MNIST		S-CIFAR-10		P-MNIST
	Class-IL	Task-IL	Class-IL	Task-IL	Domain-IL
Joint	**95.59** ± **0.31**	**99.33** ± **0.17**	**58.89** ± **3.26**	**87.58** ± **1.85**	**77.65** ± **1.09**
Finetune	19.62 ± 0.12	95.25 ± 1.66	17.00 ± 1.20	64.02 ± 3.53	58.68 ± 0.46
ER-Reservior	76.43 ± 3.08	98.77 ± 0.14	44.45 ± 3.69	84.42 ± 1.15	66.95 ± 1.40
GEM	80.79 ± 1.47	97.68 ± 0.32	18.66 ± 0.91	77.74 ± 2.60	62.96 ± 1.14
A-GEM	45.69 ± 3.77	98.66 ± 0.16	18.13 ± 0.27	74.07 ± 0.76	60.48 ± 2.04
GSS	71.19 ± 1.25	98.45 ± 0.51	36.19 ± 4.38	81.47 ± 1.74	58.91 ± 0.96
FDR	81.03 ± 2.23	98.66 ± 0.52	19.51 ± 1.04	74.29 ± 3.49	68.41 ± 2.72
HAL	79.15 ± 2.03	98.81 ± 0.18	33.86 ± 1.73	75.19 ± 2.57	**70.83** ± **1.86**
SCR	–	–	40.91 ± 1.07	76.72 ± 2.28	–
CCL-FP (ours)	88.67 ± 0.97	**99.15** ± **0.37**	50.11 ± 3.69	85.44 ± 2.03	66.91 ± 0.95
CCL-FP+ (ours)	**89.16** ± 1.14	99.14 ± 0.05	**51.74** ± **2.41**	**86.33** ± **1.47**	69.22 ± 1.07

Model	S-CIFAR-100		S-Tiny-ImageNet		R-MNIST
	Class-IL	Task-IL	Class-IL	Task-IL	Domain-IL
Joint	**19.60** ± **2.14**	**69.80** ± **2.17**	**14.21** ± **0.66**	**43.89** ± **0.88**	**84.12** ± **0.61**
Finetune	3.58 ± 0.13	39.55 ± 4.42	4.77 ± 0.23	26.93 ± 1.59	67.64 ± 2.17
ER-Reservoir	9.74 ± 0.98	63.05 ± 0.82	7.21 ± 0.29	36.75 ± 0.79	79.77 ± 0.86
GEM	4.69 ± 0.41	49.29 ± 0.73	6.76 ± 0.45	29.05 ± 0.74	79.85 ± 2.17
A-GEM	3.67 ± 0.10	46.88 ± 1.81	5.43 ± 0.11	29.67 ± 0.91	73.64 ± 3.68
GSS	6.15 ± 0.49	64.58 ± 2.29	6.41 ± 0.42	39.71 ± 0.72	77.37 ± 3.55
FDR	3.65 ± 0.10	42.87 ± 2.62	4.83 ± 0.43	26.97 ± 2.69	79.75 ± 3.12
HAL	6.31 ± 0.71	47.88 ± 2.76	3.85 ± 0.32	21.70 ± 1.12	78.65 ± 1.57
SCR	7.38 ± 0.61	43.41 ± 1.31	2.61 ± 0.28	13.03 ± 0.85	–
CCL-FP (ours)	13.64 ± 1.04	**65.19** ± **0.65**	**10.52** ± **0.28**	39.44 ± 0.48	80.68 ± 1.74
CCL-FP+ (ours)	**14.05** ± **0.85**	65.19 ± 1.88	10.13 ± 0.44	**39.99** ± **0.59**	**82.06** ± **1.29**

↑ *Table above:* The average accuracy ± standard deviation (%) by the end of training for baselines and our models across 5 runs with different random seeds. The results for joint training, i.e. the upper bound, and the best accuracies for CL models on each benchmark are marked in bold.

			S-MNIST		S-CIFAR-10		R-MNIST
$w\neq 0$	$\alpha\neq 0$	$\beta\neq 0$	Class-IL	Task-IL	Class-IL	Task-IL	Domain-IL
			76.43	98.77	44.45	84.42	79.77
✓			83.48	98.72	50.45	85.23	79.38
	✓		78.04	99.13	44.82	84.63	79.89
		✓	77.37	98.67	46.41	85.55	81.01
✓	✓		88.67	99.15	50.11	85.44	80.68
✓		✓	82.95	98.73	50.36	84.73	81.78
	✓	✓	78.31	98.89	47.52	85.91	81.79
✓	✓	✓	89.16	99.14	51.74	86.33	82.06

↑ *Table above:* The The ablation study on S-MNIST, S-CIFAR-10 and R-MNIST datasets. All results are the average accuracy across 5 runs with different random seeds.

← *Figure Left:* The evolution curves of average accuracy on test data of all seen tasks as new tasks are learned, on all datasets in task-il, class-il and domain-il settings.