Task-agnostic Continual Learning with Hybrid Probabilistic Models

Polina Kirichenko [1] Mehrdad Farajtabar [2] Dushyant Rao [2]
Balaji Lakshminarayanan [3] Nir Levine [2] Ang Li [2]
Huiyi Hu [2] Andrew Gordon Wilson [1] Razvan Pascanu [2]

[1]New York University

[2]DeepMind

[3]Google Brain

August 17, 2021

Existing Approaches

- re-sample the data or design specific loss functions that better facilitate learning with imbalanced data
- enhance recognition performance of the tail classes by transferring knowledge from the head classes

Our Contribution

- Hybrid Continual Learning (HCL) - a normalizing flow-based approach.
- Generative replay and a novel functional regularization are employed to alleviate forgetting. The functional regularization is shown to be better than generative replay.
- HCL achieves strong performance on *split MNIST*, *split CIFAR*, *SVHN-MNIST* and *MNIST-SVHN* datasets.
- HCL can detect task boundaries and identify new as well as recurring tasks.

Continual Learning (CL)

- A CL model $g_\theta : \mathcal{X} \to \mathcal{Y}$.
- A sequence of τ supervised tasks: $T_{t_1}, T_{t_2}, \ldots, T_{t_\tau}$. τ is not known in advance.
- Each task $T_i = \{(x_j^i, y_j^i)\}_{j=1}^{N_i}$, where $x_j^i \in \mathcal{X}^i$ and $y_j^i \in \mathcal{Y}^i = \{1, \ldots, K\}$.
- The corresponding data distribution of task T_i is $p_i(x, y)$.
- **Constraint**: While training on a task T_i the model cannot access to the data from previous $T_1, \ldots, T_{i-1}$ or future tasks $T_{i+1}, \ldots, T_\tau$.
- **Objective**: Minimize $\sum_{i=1}^{M} \mathbb{E}_{x,y\sim p_i(\cdot,\cdot)} l(g_\theta(x), y)$ for some risk function $l(\cdot, \cdot)$ and generalize well on all tasks after training.

Task-agnostic CL

In this work, the *task-agnostic* setting is considered, where the task index is not provided and the model has to infer it from data.

An illustration of the HCL Framework

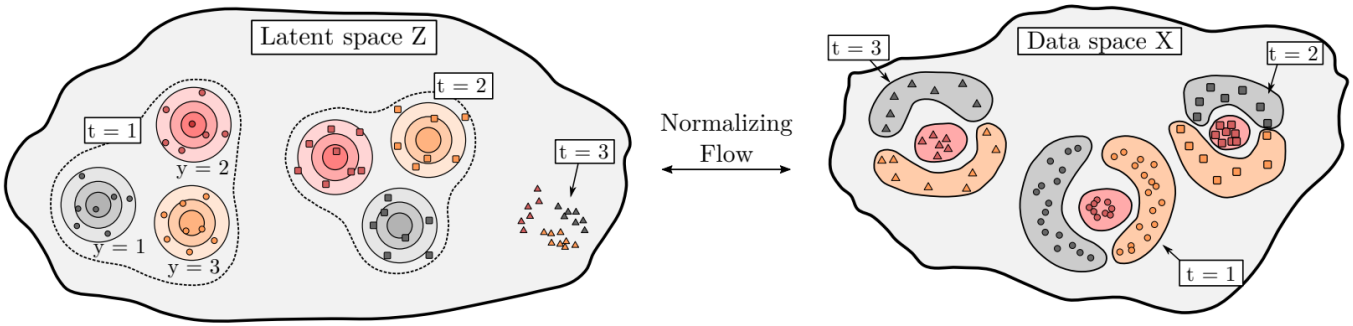


Figure 1. An illustration of the proposed Hybrid Continual Learning (HCL) framework. HCL models the distribution of each class in each task as a latent Gaussian distribution transformed by a normalizing flow. We show the Gaussian mixtures corresponding to the two tasks t_1 and t_2 in the latent space on the left, and the corresponding data distributions on the right. If a new task $t = 3$ appears, HCL identifies it using the typicality of the flow's statistics, and initializes the Gaussian mixture for a new task.

Modeling the Data Distribution

- $p_t(x, y)$: the joint distribution of the data x and the class label y conditioned on a task t.

$$p_t(x, y) \approx \hat{p}(x, y|t) = \hat{p}_X(x|y, t)\hat{p}(y|t)$$

- $\hat{p}_X(x|y, t)$ is modeled by a normalizing flow f_θ with a base distribution $\hat{p}_Z = \mathcal{N}(\mu_{y,t}, I)$.

$$\hat{p}_X(x|y, t) = f_\theta^{-1}\left(\mathcal{N}(\mu_{y,t}, I)\right)$$

- $\mu_{y,t}$ is the mean of the latent distribution corresponding to the class y and task t.
- $\hat{p}(y|t)$ is assumed to be a uniform distribution over the classes for each task: $\hat{p}(y|t) = 1/K$.

Task Identification

Three statistics on data batches B.

- log-likelihood

$$S_1(B, t) = \sum_{(x_j, y_j) \in B} \hat{p}_X(x_j | y_j, t)$$

- log-likelihood of the latent variable

$$S_2(B, t) = \sum_{(x_j, y_j) \in B} \hat{p}_Z(f_\theta(x_j) | y_j, t)$$

- log-determinant of the Jacobian

$$S_3(B, t) = S_1(B, t) - S_2(B, t)$$

Task Identification

- For each task t, keep track of the mean μ_S^t and standard deviation σ_S^t for the statistics over a window of the last l batches of data.
- If all the statistics $|S(B, t') - \mu_S^{t'}| < \lambda \sigma_S^{t'}$ for one of the previous tasks, identify a switch to the task t'.
- Otherwise, switch to a new task and add new Gaussians for this task in the latent space.

Generative Replay (HCL-GR)

- Store a single snapshot of the HCL model $\hat{p}_X^{(k)}(x|y,t)$ with weights $\theta^{(k)}$.
- Generate and replay data from old tasks using the snapshot: $x_{GR} \sim \hat{p}_X^{(k)}(x|y,t)$, where $y \sim U\{1,\ldots,K\}$ and $t \sim U\{t_1,\ldots,t_k\}$.
- Maximize the likelihood $\mathcal{L}_{GR} = \log \hat{p}_X(x_{GR}|y,t)$ under the current model $\hat{p}_X(\cdot)$.
- The resulting loss in generative replay training is $\mathcal{L}_{ll} + \mathcal{L}_{GR}$, where $\mathcal{L}_{ll}$ is the log-likelihood of the data on the current task.
- Update the snapshot with new weights $\theta^{(k+1)}$ after detecting the task change $T_{t_{k+1}} \rightarrow T_{t_{k+2}}$.

Functional Regularization (HCL-FR)

Enforce the flow to map samples from previous tasks to the same latent representations as a snapshot model.

- Store a single snapshot of the model $\hat{p}_X^{(k)}(\cdot)$ and produce samples $x_{FR} \sim \hat{p}_X^{(k)}(x|y,t)$ for $y \sim U\{1,\dots,K\}$, $t \sim U\{t_1,\dots,t_k\}$.
- $\mathcal{L}_{FR} = \|f_\theta(x_{FR}) - f_{\theta^{(k)}}(x_{FR})\|^2$, where f_θ is the current flow mapping and $f_{\theta^{(k)}}$ is the snapshot model.
- The resulting loss in functional regularization is $\mathcal{L}_{ll} + \alpha\mathcal{L}_{FR}$.

Compared Methods

- Adam: Train the model with Adam optimizer without any extra steps for preventing forgetting.
- Multi-Task Learning (MTL): When training on T_{t_i}, it has access to all previous tasks $T_{t_1}, \ldots, T_{t_{i-1}}$.
- Experience Replay (ER): Reserve a buffer with a fixed size of 1000 samples for each task and randomly select samples to add to that buffer during training on each task.
- CURL: CURL incorporated a generative model (VAE) with an expanding Gaussian mixture in latent space and likelihood-based task-change detection.

Datasets

- Split MNIST: Split the dataset into 5 binary classification tasks. The epoch is set to 30 for each task. The *Glow* architecture is used to model the data distribution.
- MNIST-SVHN and SVHN-MNIST: The size of the MNIST images is upscaled to $32 \times 32 \times 3$. The *RealNVP* architecture is used to model the data distribution. The epoch is set to 90 for each task.
- Split CIFAR: Each task corresponds to 2 classes of CIFAR-10 and 10 classes of CIFAR-100 respectively. The image embeddings are used, which are extracted by an EfficientNet model pre-trained on ImageNet. The epoch is set to 15 for each task. The *RealNVP* architecture is used.

Metrics

$a_{i,j}$: The accuracy of the model on task i after training on j tasks.

- Final accuracy on each task: $a_{i,\tau}$, $i \in \{1, \ldots, \tau\}$.
- Average final accuracy across tasks: $\frac{1}{\tau} \sum_{i=1}^{\tau} a_{i,\tau}$.
- Average forgetting: $\frac{1}{\tau-1} \sum_{i=1}^{\tau-1} (a_{i,i} - a_{i,\tau})$.
- Overall accuracy: the final accuracy on $(K \times \tau)$ - way classification.

Experiment Results

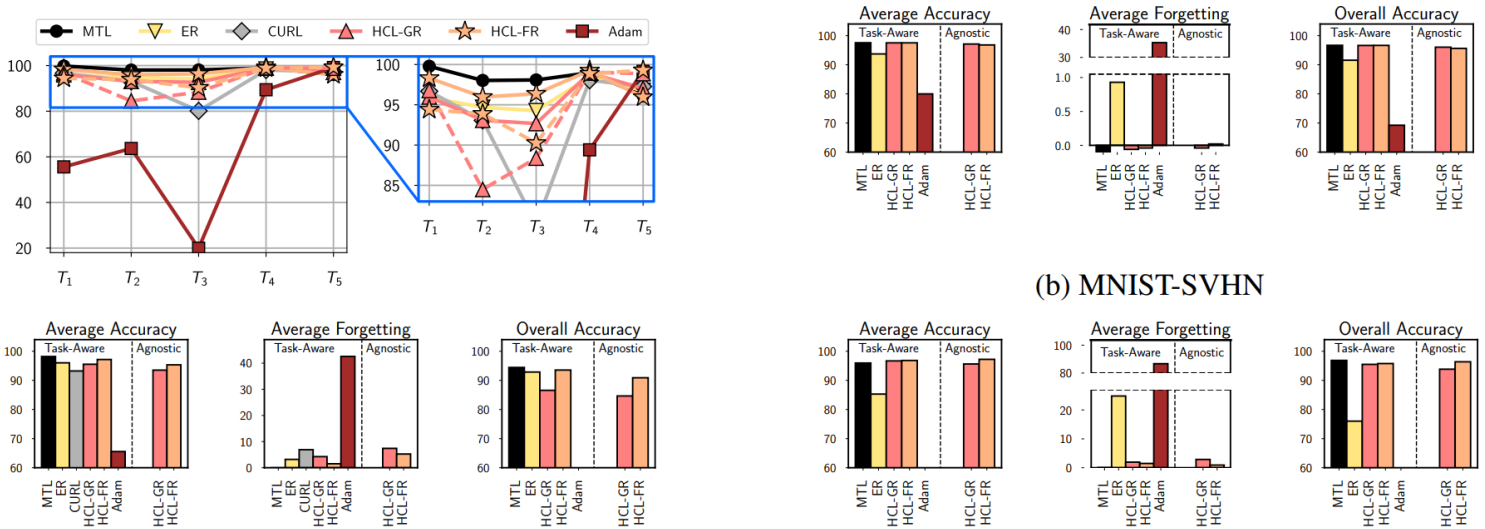


Figure 2. Results on **(a)** Split MNIST, **(b)** MNIST-SVHN and **(c)** SVHN-MNIST image datasets. For Split MNIST, in the top panel we show the performance of each method on each of the tasks in the end of training; for HCL we show the results in the task-agnostic setting with dashed lines. We also show average accuracy, forgetting and overall accuracy for each of the datasets and methods. HCL provides strong performance, especially on SVHN-MNIST where it achieves almost zero forgetting and significantly outperforms ER.

Experiment Results

Table 1. Results of the experiments on **split MNIST** dataset with MTL (multitask learning), Adam (regular training without alleviating forgetting), ER (standard experience or data buffer replay with the capacity of 1000 samples per task), HCL-GR (generative replay), HCL-FR (functional regularization). The dataset with 10 classes is split into 5 binary classification tasks, as well as task-agnostic versions of HCL-FR and HCL-GR.

TASK #	1	2	3	4	5	ACC AVG	FORGET AVG	FULL ACC
MTL	99.78 ±0.15	98.02 ±1.50	98.08 ±0.96	98.98 ±0.33	96.15 ±1.87	98.20 ±0.88	−1.32 ±1.03	94.44 ±1.06
ADAM	55.59 ±4.74	63.66 ±3.10	19.96 ±7.03	89.41 ±7.15	99.16 ±0.41	65.56 ±1.33	42.57 ±1.49	19.66 ±0.08
ER	95.92 ±5.00	94.69 ±1.98	94.27 ±2.20	98.44 ±0.41	96.77 ±1.00	96.02 ±1.31	3.19 ±1.92	92.86 ±1.92
CURL	96.67 ±0.64	93.06 ±1.42	80.14 ±5.70	98.05 ±0.86	97.27 ±0.41	93.23 ±1.06	6.90 ±1.47	–
HCL-FR	98.31 ±1.03	95.97 ±0.81	96.37 ±1.06	99.24 ±0.08	95.95 ±3.04	97.17 ±0.65	1.53 ±0.91	93.55 ±1.40
HCL-GR	95.97 ±3.65	93.08 ±4.17	92.67 ±2.43	98.94 ±0.66	97.02 ±1.69	95.54 ±1.21	4.25 ±1.99	86.58 ±7.37
HCL-FR (TA)	94.41 ±3.42	93.88 ±0.37	90.25 ±4.11	98.86 ±0.23	99.28 ±0.23	95.33 ±0.67	5.24 ±0.95	90.89 ±0.96
HCL-GR (TA)	96.78 ±0.98	84.49 ±5.57	88.38 ±4.79	99.04 ±0.53	98.87 ±0.06	93.52 ±1.84	7.41 ±2.18	84.65 ±3.46

Experiment Results

Table 2. Results of the experiments on **SVHN-MNIST** and **MNIST-SVHN** datasets with MTL (multitask learning), Adam (regular training without alleviating forgetting), ER (standard experience or data buffer replay with the capacity of 1000 samples per task), HCL-GR (generative replay), HCL-FR (functional regularization), as well as task-agnostic versions of HCL-FR and HCL-GR.. Each dataset contains two 10-way classification tasks corresponding to MNIST and SVHN.

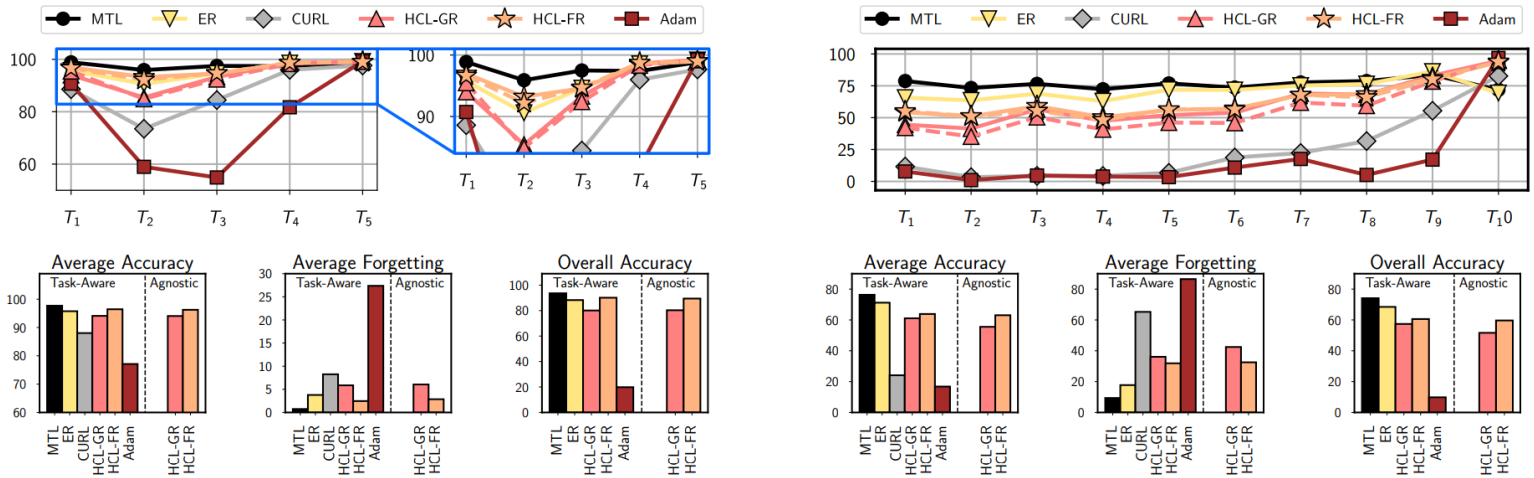
SVHN-MIST

TASK #	1	2	ACC AVG	FORGET AVG	FULL ACC
MTL	95.96	99.18	97.57	0.01	96.86
ADAM	9.28	99.18	54.23	86.69	30.74
ER	71.14	99.45	85.295	24.83	76.00
HCL-FR	94.48	99.32	96.90	1.49	95.78
HCL-GR	94.03	99.35	96.69	1.94	95.5
HCL-FR (TA)	95.19	99.26	97.23	0.92	96.38
HCL-GR (TA)	91.76	99.50	95.63	2.87	93.84

MNIST-SVHN

TASK #	1	2	ACC AVG	FORGET AVG	FULL ACC
MTL	99.58	95.56	97.57	-0.11	96.68
ADAM	64.16	95.82	79.99	35.31	69.23
ER	98.54	88.89	93.72	0.93	91.56
HCL-FR	99.51	95.55	97.53	-0.04	96.65
HCL-GR	99.53	95.52	97.53	-0.06	96.63
HCL-FR (TA)	99.47	94.14	96.81	0.02	95.62
HCL-GR (TA)	99.56	94.68	97.12	-0.04	96.04

Experiment Results on Split CIFAR



(a) Split CIFAR-10 (b) Split CIFAR-100

Figure 3. Results on Split CIFAR embedding datasets. We use embeddings extracted by an EfficientNet model (Tan & Le, 2019) pre-trained on ImageNet. In the top panels we show the performance of each method on each of the tasks in the end of training; for HCL we show the results in the task-agnostic setting with dashed lines. At the bottom, we show average accuracy, forgetting and overall accuracy for each of the methods. HCL outperforms CURL and Adam and performs on par with experience replay with a large replay buffer. HCL-FR provides better performance than HCL-GR.

Experiment Results on Split CIFAR-10

Table 3. Results of the experiments on **split CIFAR-10** embeddings dataset extracted using EfficientNet model pretrained on ImageNet. The dataset with 10 classes is split into 5 binary classification tasks. The methods used are MTL (multitask learning) setting, Adam (regular training without alleviating forgetting), ER (standard data buffer replay with the capacity of 1000 samples per task), CURL (Rao et al., 2019), HCL-GR (generative replay), HCL-FR (functional regularization), as well as task-agnostic versions of HCL-FR and HCL-GR.

15 EPOCHS PER TASK

TASK #	1	2	3	4	5	ACC AVG	FORGET AVG	FULL ACC
MTL	98.87 ±0.08	95.90 ±0.29	97.48 ±0.12	97.40 ±0.25	98.82 ±0.13	97.69 ±0.05	0.75 ±0.14	93.61 ±0.14
ADAM	90.73 ±1.05	58.97 ±4.88	54.90 ±3.82	81.70 ±3.76	99.25 ±0.04	77.11 ±1.33	27.38 ±1.63	19.85 ±0.01
ER	95.75 ±0.43	90.60 ±1.06	94.62 ±0.28	98.57 ±0.15	99.20 ±0.11	95.75 ±0.35	3.78 ±0.46	88.27 ±0.52
CURL	88.59 ±3.85	73.48 ±6.00	84.46 ±2.82	95.98 ±0.64	97.65 ±0.41	88.03 ±2.15	12.05 ±2.79	–
HCL-FR	96.95 ±0.44	93.22 ±0.59	94.58 ±0.19	98.50 ±0.08	98.97 ±0.19	**96.44** ±0.05	**2.47** ±0.09	**90.12** ±0.35
HCL-GR	93.98 ±0.27	85.43 ±0.25	93.28 ±0.92	98.63 ±0.16	99.20 ±0.08	94.11 ±0.21	5.86 ±0.24	80.10 ±1.21
HCL-FR (TA)	96.63 ±0.33	92.18 ±1.33	94.70 ±0.19	98.73 ±0.25	98.98 ±0.10	96.25 ±0.17	2.86 ±0.28	89.44 ±0.80
HCL-GR (TA)	95.47 ±0.73	84.88 ±1.08	92.40 ±0.16	98.32 ±0.22	99.23 ±0.05	94.06 ±0.25	6.05 ±0.35	80.29 ±0.81

Experiment Results on Split CIFAR-10

SINGLE-PASS (1 EPOCH PER TASK)

TASK #	1	2	3	4	5	ACC AVG	FORGET AVG	FULL ACC
MTL	98.92 ±0.10	96.67 ±0.17	97.25 ±0.00	97.20 ±0.23	98.18 ±0.26	97.64 ±0.05	−0.05 ±0.13	93.69 ±0.09
ADAM	92.22 ±1.20	53.35 ±0.53	62.53 ±3.52	78.92 ±6.34	99.13 ±0.15	77.23 ±1.27	26.87 ±1.54	19.83 ±0.03
ER	98.32 ±0.12	94.08 ±0.71	97.12 ±0.14	97.73 ±0.08	98.55 ±0.11	97.16 ±0.09	1.32 ±0.17	91.85 ±0.03
CURL	95.54 ±1.16	80.97 ±4.83	80.55 ±7.16	94.61 ±1.99	96.25 ±0.41	89.58 ±0.92	8.25 ±1.10	–
HCL-FR	95.27 ±0.46	88.03 ±0.40	93.35 ±0.73	98.28 ±0.31	98.92 ±0.06	94.77 ±0.09	4.68 ±0.13	85.94 ±0.01
HCL-GR	93.68 ±0.55	85.82 ±0.27	93.28 ±0.41	98.52 ±0.17	99.10 ±0.04	94.08 ±0.08	5.73 ±0.06	82.85 ±0.40
HCL-FR (TA)	95.35 ±0.16	87.12 ±0.81	93.40 ±0.32	98.27 ±0.16	98.90 ±0.08	94.61 ±0.24	4.85 ±0.32	85.72 ±0.37
HCL-GR (TA)	93.37 ±0.27	82.28 ±1.94	92.33 ±0.65	98.23 ±0.17	99.17 ±0.18	93.08 ±0.54	7.05 ±0.56	79.93 ±0.84

Experiment Results on Split CIFAR-100

Table 4. Results of the experiments on **split CIFAR-100** embeddings dataset extracted using EfficientNet model pretrained on ImageNet. The dataset with 100 classes is split into ten 10-way classification tasks. The methods used are MTL (multitask learning) setting, Adam (regular training without alleviating forgetting), ER (standard data buffer replay with the capacity of 1000 samples per task), CURL (Rao et al., 2019), HCL-GR (generative replay), HCL-FR (functional regularization), as well as task-agnostic versions of HCL-FR and HCL-GR.

15 EPOCHS PER TASK

TASK #	1	2	3	4	5	6	7	8	9	10	ACC AVG	FORGET AVG	FULL ACC
MTL	78.77 ±0.39	73.30 ±1.27	76.53 ±1.30	72.33 ±0.25	76.87 ±0.78	73.63 ±2.36	77.63 ±1.60	78.73 ±0.12	83.53 ±0.80	71.87 ±0.97	76.32 ±0.52	9.37 ±0.34	74.11 ±0.55
ADAM	7.77 ±0.25	1.00 ±0.29	4.63 ±0.60	3.80 ±1.26	3.33 ±1.15	10.83 ±0.46	17.60 ±1.34	5.07 ±0.05	17.10 ±0.80	96.83 ±0.25	16.80 ±0.27	86.49 ±0.33	9.84 ±0.04
ER	65.70 ±1.23	63.57 ±1.60	68.83 ±0.78	62.97 ±1.32	71.93 ±0.76	71.73 ±0.24	75.20 ±0.71	76.40 ±1.08	86.20 ±0.57	69.37 ±1.72	71.19 ±0.22	17.74 ±0.32	68.46 ±0.27
CURL	11.62 ±2.31	3.34 ±1.07	4.34 ±2.01	4.26 ±2.02	6.72 ±3.23	18.76 ±1.25	22.28 ±2.41	31.66 ±2.02	55.44 ±1.19	82.68 ±0.52	24.11 ±0.72	65.24 ±0.68	–
HCL-FR	54.27 ±1.68	51.00 ±1.87	59.10 ±1.85	50.30 ±1.82	56.10 ±0.29	57.10 ±1.08	67.67 ±2.05	68.33 ±0.87	81.20 ±0.75	93.47 ±0.21	63.85 ±0.80	31.89 ±0.99	60.58 ±0.78
HCL-GR	44.53 ±1.18	41.43 ±1.75	56.57 ±0.25	47.53 ±0.59	52.00 ±1.77	53.87 ±1.24	68.93 ±1.60	68.10 ±0.43	82.93 ±0.24	94.67 ±0.19	61.06 ±0.43	36.10 ±0.56	57.39 ±0.60
HCL-FR (TA)	55.03 ±0.57	51.13 ±1.89	55.33 ±4.23	48.17 ±0.63	56.50 ±1.56	56.53 ±1.02	67.87 ±0.66	65.97 ±2.59	80.17 ±1.36	93.93 ±0.33	63.06 ±0.64	32.52 ±0.88	59.66 ±0.70
HCL-GR (TA)	42.10 ±0.75	35.17 ±2.32	50.37 ±0.86	40.77 ±0.88	46.13 ±0.63	45.83 ±1.01	61.87 ±0.87	59.33 ±0.48	78.47 ±0.68	95.30 ±0.22	55.53 ±0.23	42.46 ±0.25	51.64 ±0.14

Experiment Results on Split CIFAR-100

SINGLE-PASS (1 EPOCH PER TASK)

TASK #	1	2	3	4	5	6	7	8	9	10	ACC AVG	FORGET AVG	FULL ACC
MTL	86.23 ±0.19	81.43 ±0.54	83.77 ±0.73	79.80 ±0.36	78.53 ±0.37	72.87 ±1.64	73.47 ±0.54	63.13 ±1.54	60.80 ±0.70	35.93 ±2.52	71.60 ±0.38	−12.18 ±0.13	68.80 ±0.52
ADAM	8.47 ±1.15	1.13 ±0.42	5.03 ±0.37	2.30 ±0.65	2.90 ±0.45	9.83 ±0.66	19.07 ±0.52	12.00 ±1.28	22.33 ±2.18	95.87 ±0.17	17.89 ±0.18	83.64 ±0.28	11.29 ±0.31
ER	78.23 ±0.79	76.77 ±1.22	80.40 ±0.42	80.10 ±0.45	80.63 ±0.26	74.03 ±1.17	72.07 ±0.12	62.87 ±1.22	58.27 ±1.70	17.97 ±0.76	68.13 ±0.36	−27.10 ±0.25	65.13 ±0.25
CURL	12.82 ±4.73	4.94 ±1.73	9.88 ±2.95	9.38 ±2.89	11.52 ±2.67	14.44 ±4.05	22.44 ±3.18	15.64 ±1.87	24.38 ±6.32	59.98 ±9.29	18.54 ±1.54	44.00 ±2.66	–
HCL-FR	50.30 ±1.16	13.80 ±2.27	48.37 ±1.05	38.63 ±0.57	37.90 ±2.14	39.77 ±0.25	51.47 ±5.70	53.33 ±2.26	73.77 ±0.31	94.03 ±0.76	50.14 ±0.41	43.31 ±0.17	45.76 ±0.34
HCL-GR	56.10 ±1.22	42.67 ±0.94	58.93 ±1.84	39.63 ±1.68	41.53 ±2.49	36.13 ±1.60	45.83 ±4.90	46.00 ±2.25	51.40 ±6.68	90.87 ±1.40	50.91 ±0.82	40.40 ±0.85	46.10 ±0.99
HCL-FR (TA)	49.50 ±2.40	16.27 ±2.23	48.73 ±2.19	37.73 ±1.14	38.87 ±1.90	37.80 ±3.19	51.77 ±1.65	50.53 ±2.09	75.07 ±1.61	93.63 ±0.25	49.99 ±0.86	43.04 ±0.58	45.64 ±0.97
HCL-GR (TA)	35.20 ±0.91	30.20 ±1.99	41.17 ±1.92	27.30 ±5.62	35.57 ±1.14	33.13 ±2.83	43.70 ±0.43	45.23 ±1.27	68.03 ±1.00	95.50 ±0.43	45.50 ±0.46	52.08 ±0.57	40.87 ±0.57

Comparison of FR and GR

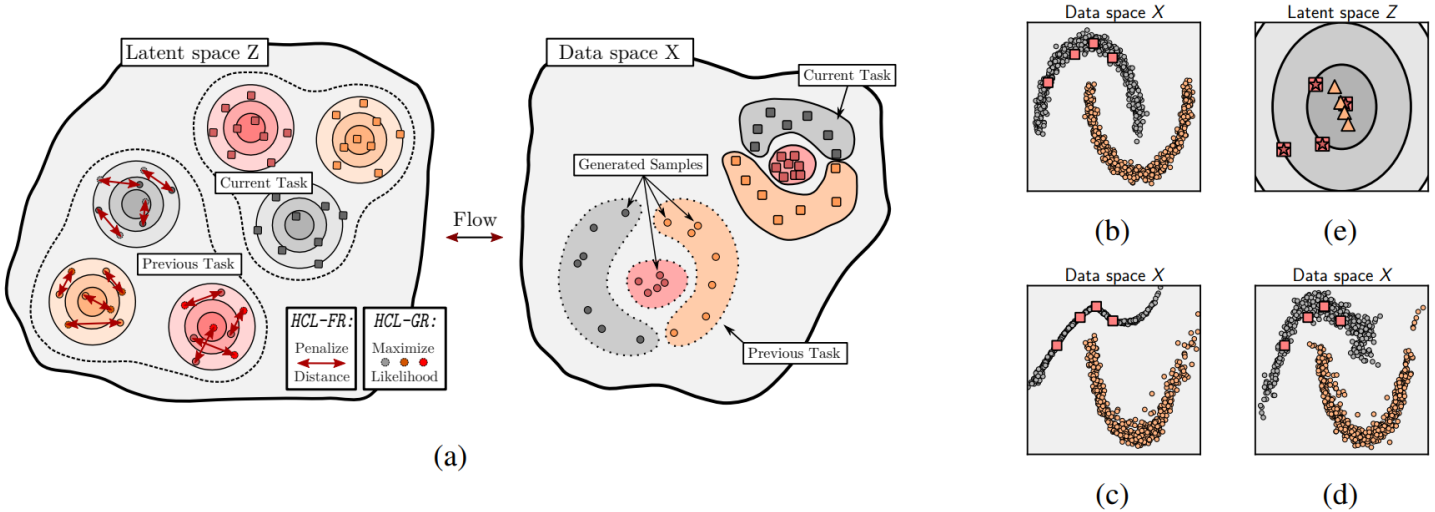


Figure 4. Comparison of functional regularization and generative replay. **(a)**: A visualization of HCL-FR and HCL-GR; HCL-GR forces the model to maintain high likelihood of the replay data, while HCL-FR penalizes the distance between the locations of the latent representations for the sampled data for the current and snapshot models. **(b)**: Two moons dataset; data from the first and second tasks is shown with grey and orange circles, and coral squares show the replay samples. **(c)**: Learned distribution after training on the second task with HCL-GR, and **(d)** HCL-FR. **(e)**: Locations of images of the replay data in the latent space for the model trained on the first task (**squares**), HCL-GR (**triangles**) and HCL-FR (**stars**). HCL-FR restricts the model more than GR: the locations of replay samples in the latent space coincide for HCL-FR and the model trained on the first task. Consequently, HCL-FR preserves more information about the structure of the first task.

Thanks!