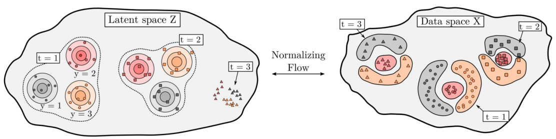
Task-agnostic Continual Learning with Hybrid Probabilistic Models

Polina Kirichenko Mehrdad Farajtabar Dushyant Rao Balaji Lakshminarayanan Nir Levine Ang Li Huiyi Hu Andrew Gordon Wilson Razvan Pascanu

Overview

We propose **HCL**, a **H**ybrid generative-discriminative approach to **C**ontinual **L**earning for classification

- Model each task and each class with a normalizing flow
- Same flow is used to learn data distribution, classify data, identify task changes and avoid forgetting
- Strong performance on a range of problems in task-aware and task-agnostic settings



Task-agnostic continual learning

Task 1 $\left(x : \right.$  $, \; y \in \{0, \ldots, 9\} \; , \; t = 1 \left. \right)$

Task 2 $\left(x : \right.$  $, \; y \in \{0, \ldots, 9\} \; , \; t = 2 \left. \right)$

- Sequence of tasks, each with the same set of classes
- We need to avoid forgetting old tasks when training on new tasks
- **Task agnostic:** we do not have task IDs *t*, model has to detect task boundaries

Normalizing flows

- Deep generative models based on invertible neural networks
- We can compute density of the data exactly via change of variables

$$z \sim p_Z \qquad x = f_\theta^{-1}(z) \qquad p(x) = p_Z(f_\theta(x)) \cdot \left| \frac{\partial f_\theta}{\partial x} \right|$$

Modeling the data distribution

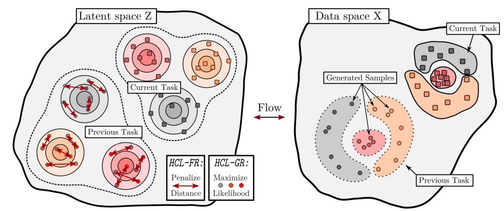
$$p_t(x, y) \approx \hat{p}(x, y|t) = \hat{p}_X(x|y, t)\hat{p}(y|t) \qquad \hat{p}(x|y, t) = f_\theta^{-1}(\mathcal{N}(\mu_{y,t}, I))$$

HCL approximates the data distribution with a single normalizing flow, with each class-task pair *(y, t)* corresponding to a unique Gaussian in the latent space

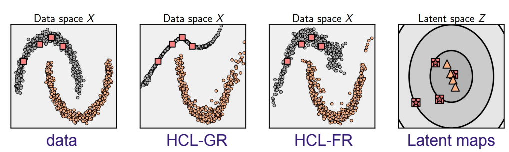
- Train via maximum likelihood
- Make predictions via Bayes rule:

$$\hat{y} = \arg\max_y \sum_{t=1}^{\tau} \hat{p}_X(x|y, t)$$

Avoiding forgetting



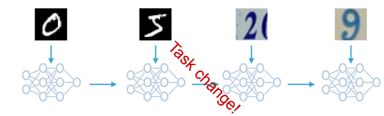
- Save a snapshot $\hat{p}_X^{(k)}$ of the model after detecting task *k*
- Generate data $x_R \sim \hat{p}_X^{(k)}(x|y, t)$
- **Generative replay:** maximize $\log \hat{p}_X(x_R|y, t)$ or
- **Functional Regularization:** minimize $\|f_\theta(x_R) - f_{\theta^{(k)}}(x_R)\|^2$



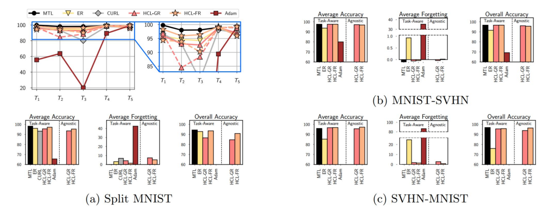
HCL-FR restricts the model more than GR: the locations of replay samples in the latent space coincide for HCL-FR and the model trained on the first task.

Task boundary identification

HCL uses a method based on Density of States Estimation (DoSE; Morningstar et al.): check that the statistics extracted by the flow model are within the typical set



Results

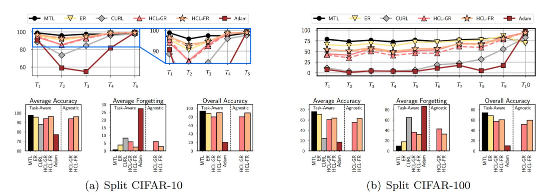


(b) MNIST-SVHN

(a) Split MNIST

(c) SVHN-MNIST

HCL provides strong performance, especially on SVHN-MNIST where it achieves almost zero forgetting and significantly outperforms ER.



(a) Split CIFAR-10

(b) Split CIFAR-100

On CIFAR, we train the models on EfficientNet embeddings. HCL outperforms CURL (Rao et al.) and Adam and performs on par with experience replay with a large replay buffer.

HCL-FR provides better results than HCL-GR overall.