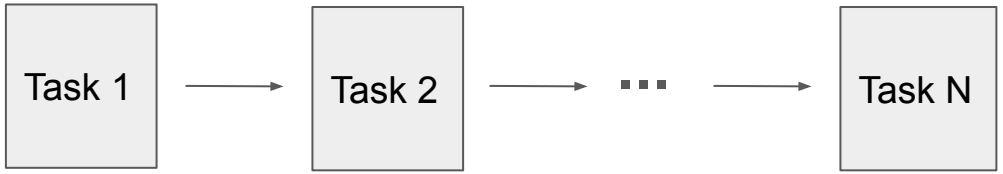
Simple Continual Learning Strategies for Safer Classifiers

Ashish Gaurav, Sachin Vernekar, Jaeyoung Lee, Vahdat Abdelzad, Krzysztof Czarnecki, Sean Sedwards

- Real world systems need to re-train when requirements change
- Use (sequential) **continual learning**



- For **safety** we prefer preserving already validated solutions over learning new solutions (*stability* over *plasticity*)

UNIVERSITY OF WATERLOO

- Generic regularized training objective:

$$\mathcal{L}_i(\boldsymbol{\theta}) + \lambda G_{1:i-1}(\boldsymbol{\theta}, \boldsymbol{\theta}^*_{1:i-1})$$

loss for the new task

stability / plasticity parameter

loss to preserve the configuration that solves the old tasks

- Popular methods (eg. EWC) use an **L2** objective:

$$\mathcal{L}_i(\boldsymbol{\theta}) + \lambda \sum_j F_{jj}(\boldsymbol{\theta}_j - \boldsymbol{\theta}^*_{j,1:i-1})^2$$

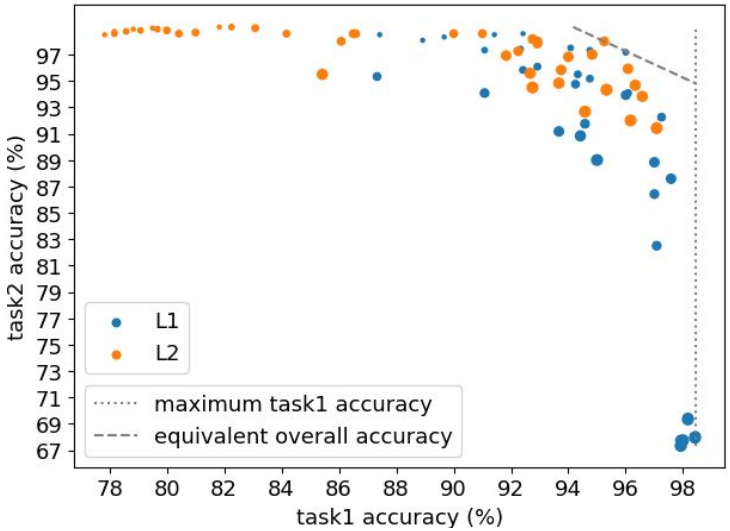
- An alternative is to use an **L1** objective:

$$\mathcal{L}_i(\boldsymbol{\theta}) + \lambda \sum_j C_j^m |\boldsymbol{\theta}_j - \boldsymbol{\theta}^*_{j,1:i-1}|$$

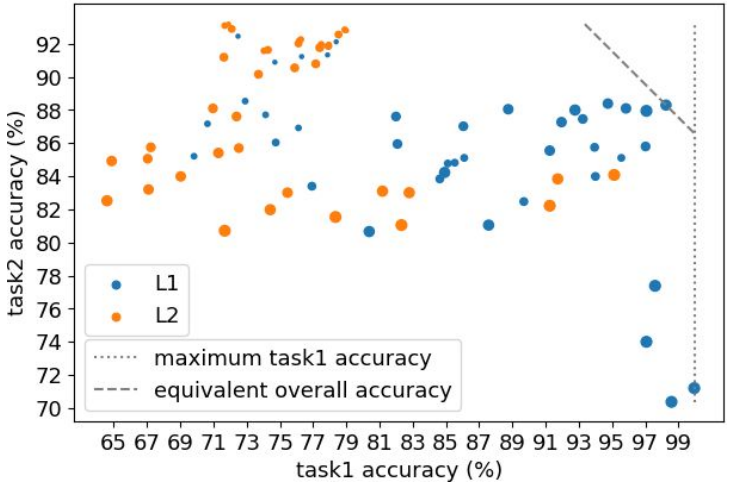
UNIVERSITY OF WATERLOO

L1 vs L2 Objective

L1 objective often leads to solutions with stronger preservation.



2-task, Sim-EMNIST



2-task, Split-MNIST

UNIVERSITY OF WATERLOO

Other Contributions

- We **quantify** amount of forgetting in continual learning for classification
  - we thus propose a per-parameter importance metric for the L1 regularization loss
  - and propose methods that:
    - never forget beyond a certain specified amount (finer control)
    - freeze weights in a more principled manner to achieve higher stability (Fisher freezing)
- We conduct **experiments** on grayscale and real world datasets

Please visit our poster for more details:

Simple Continual Learning Strategies for Safer Classifiers

UNIVERSITY OF WATERLOO