A Simple and Effective Approach to Continual Learning for Image Classification

Overview of the Winning Entry for the CVPR 2020 CLVISION Challenge

Zheda Mai[1], Hyunwoo Kim[2], Jihwan Jeong[1], Scott Sanner[1]

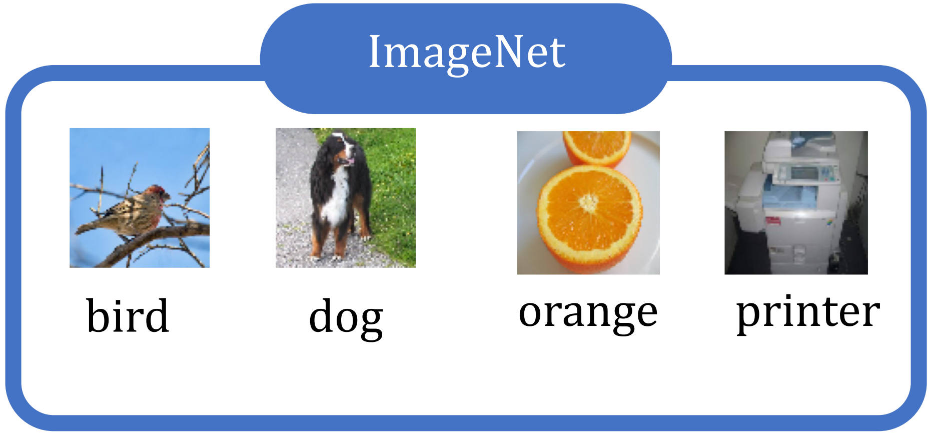
University of Toronto[1], LG Sciencepark[2]

Agenda

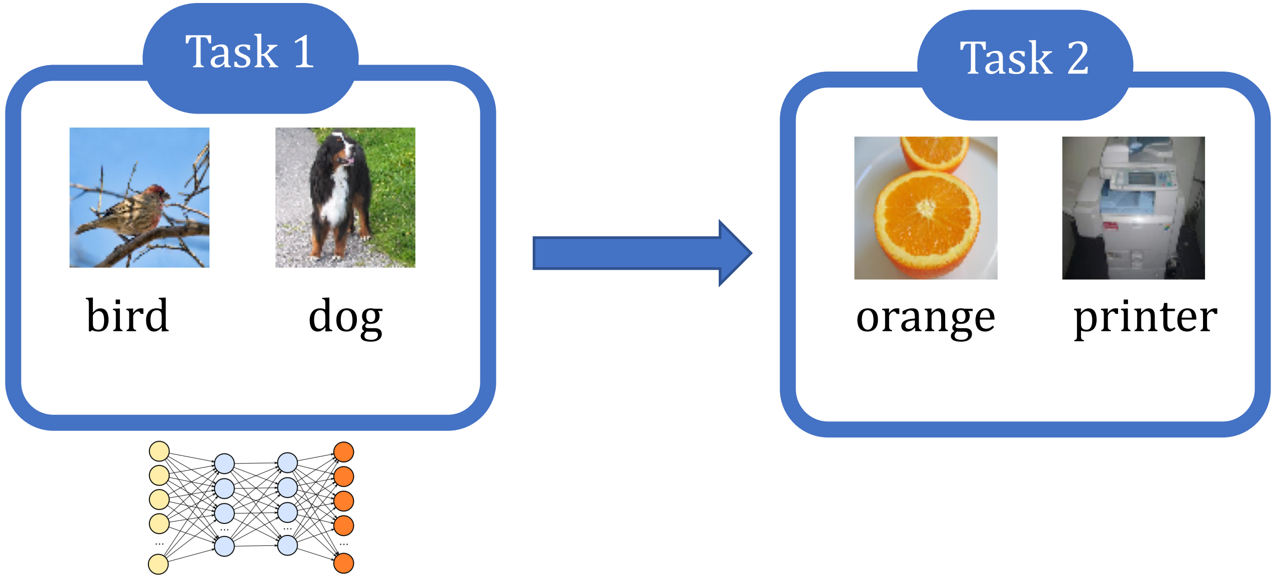
- Introduction to Continual Learning
- CLVISION challenge at CVPR2020
- Winning solutions presentation

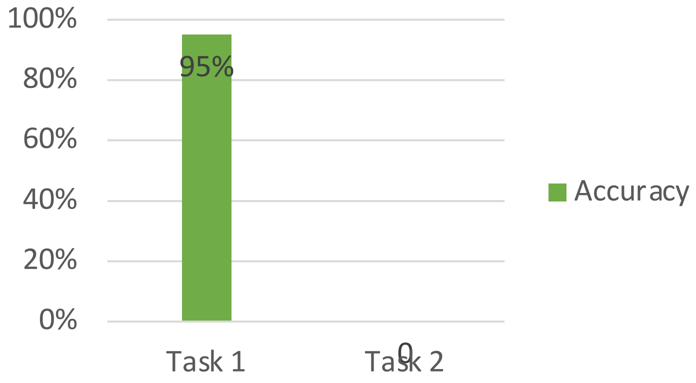
Neural Network can’t learn continuously

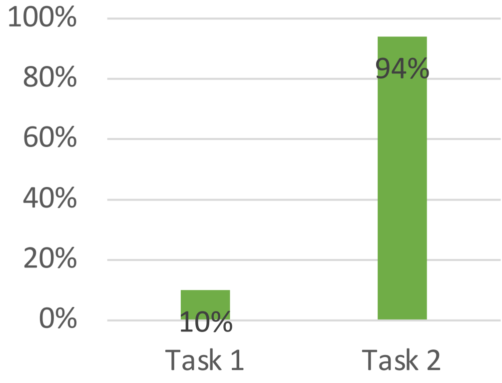
- Conventional deep learning
  - Mini-batches are iid-sampled from the whole dataset
  - Example: ImageNet classification



Neural Network can’t learn continuously







- When incrementally learning from non-stationary data with SGD, neural networks suffer from *catastrophic forgetting*.
- Continual Learning(CL) attempts to teach neural networks how to learn continuously.
- Two main challenges
  - Avoid forgetting from old tasks
  - Improve current task learning

CLVISION Challenge at CVPR2020

Based on CORe50 ((**C**)ontinual (**O**)bject (**Re**)cognition) dataset with 50 classes

- Each column represents a category
- Each row shows objects with non-stationarity
  - holding hand (left or right)
  - background environments
  - illumination
  - occlusion

CLVISION Challenge - Evaluation

- Final test accuracy
- Average validation accuracy over time
- Total training & test time
- Ram usage
- Disk usage

Final aggregation metric

CL_score: weighted average of all the metrics (0.3, 0.1, 0.15, 0.125, 0.125)

Three challenge tracks

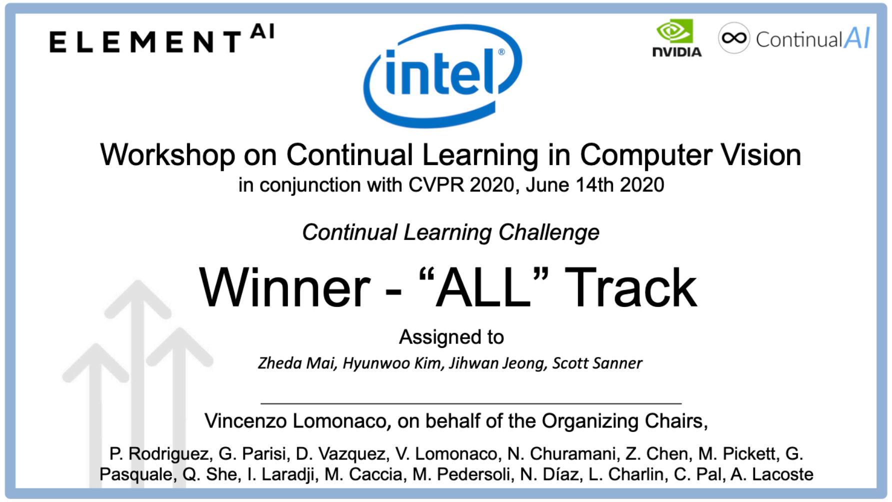
- New instances(NI)
  - 8 tasks of the same 50 classes,
  - Each task has images collected in different environmental conditions
  - No task label is given

- New instances & classes (NIC)
  - 391 tasks, each one has 300 images of the same class
  - The class can be seen or completely new
  - No task label is given

- Multi-Task New classes(NC)
  - 9 tasks, 10 classes in the first one and 5 classes in the other 8 tasks
  - **Task label is given**

Final Ranking

TEAM NAME	TEST ACC (%)	VAL ACC$_{avg}$ (%)	RUN$_{time}$ (M)	RAM$_{avg}$ (MB)	RAM$_{max}$ (MB)	DISK$_{avg}$ (MB)	DISK$_{max}$ (MB)	CL_{score}
UT_LG	0.92	0.68	68.67	10643.25	11624.87	0	0	0.694359483
JODELET	0.88	0.64	6.59	15758.62	18169.32	0	0	0.680821395
AR1	0.80	0.58	20.46	8040.47	10092.72	0	0	0.663760006
YC14600	0.91	0.65	64.88	16425.64	19800.48	0	0	0.653114358
ICT_VIPL	0.95	0.68	76.73	2459.31	2459.68	392.1875	562.5	0.61726439
SOONY	0.88	0.63	120.33	14533.97	15763.60	0	0	0.612231922
REHEARSAL	0.75	0.52	22.87	19056.77	23174.11	0	0	0.570829566
JIMIB	0.91	0.74	242.12	17995.61	23765.51	0	0	0.542653619
NOOBMASTER	0.76	0.53	147.59	24714.06	30266.62	0	0	0.464365891
NAÏVE	0.23	0.24	5.16	15763.46	18158.02	0	0	0.32735254
AVG	0.80	0.59	77.54	14539.12	17327.49	39.22	56.25	0.58



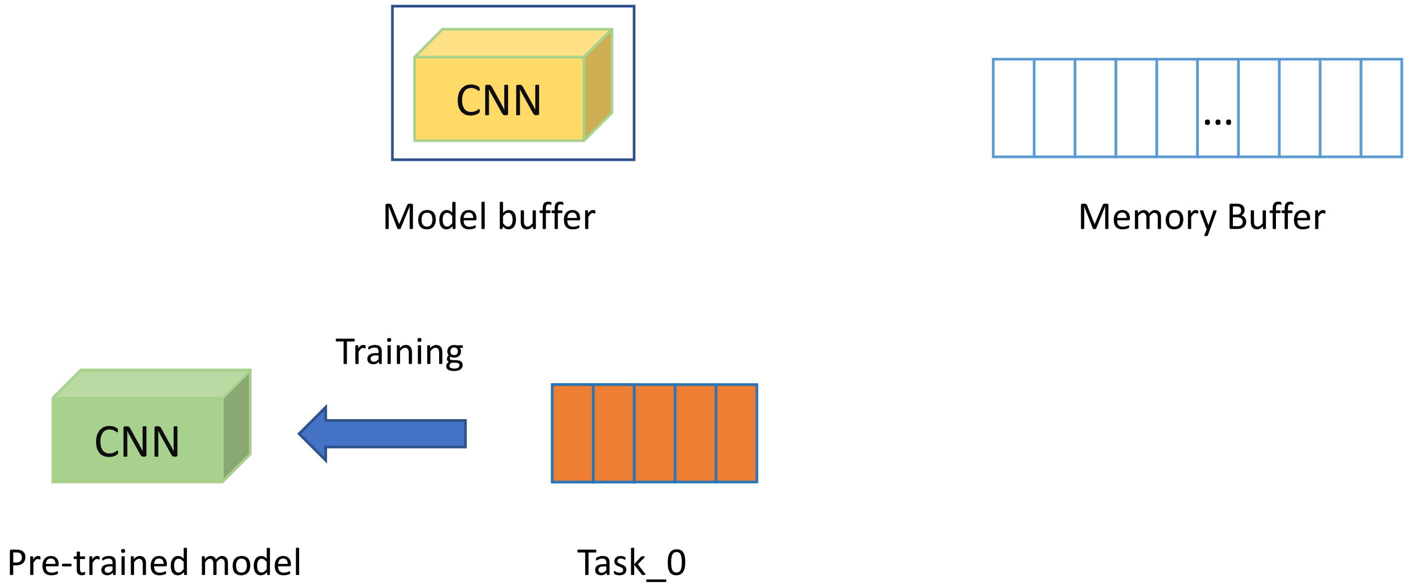
Winning solutions

- Team UT_LG
  - University of Toronto
  - LG Sciencepark
- Team ICT_VIPL
  - Institute of Computing Technology
  - University of Chinese Academy of Sciences
- Team YC14600
  - University of Bristol
  - Amazon

Team UT_LG

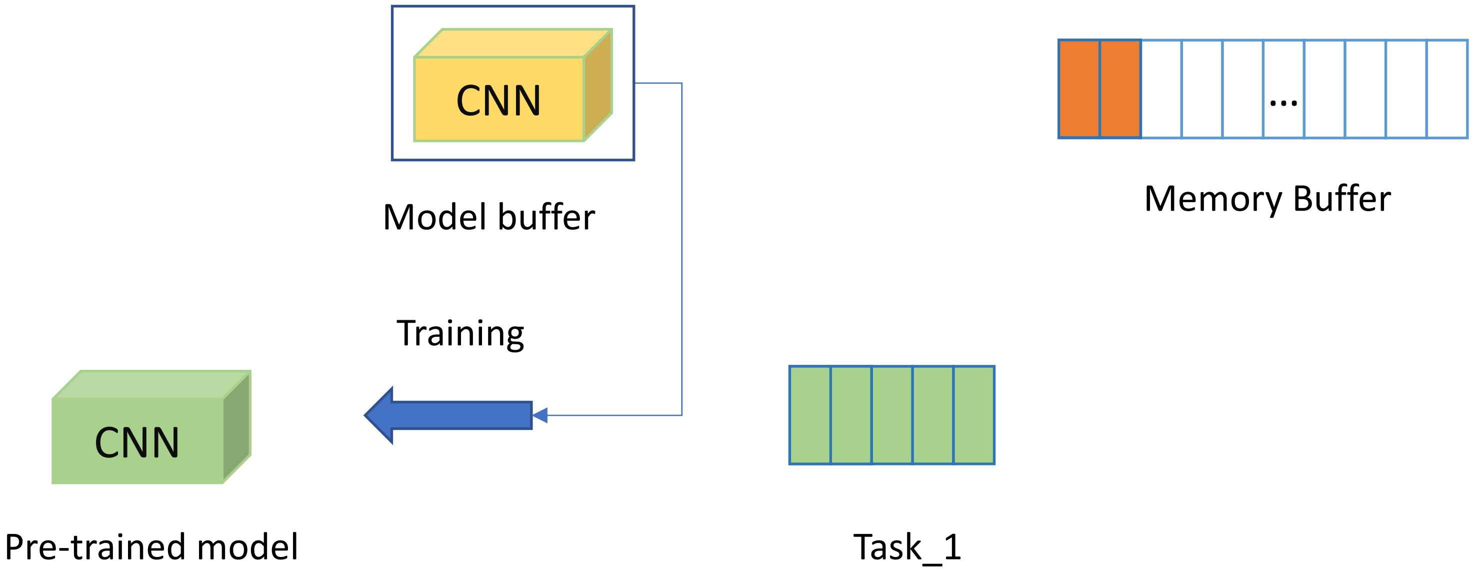
Batch-level Experience Replay with Review for Continual Learning

- Used in NI & NIC track, where no task label is given

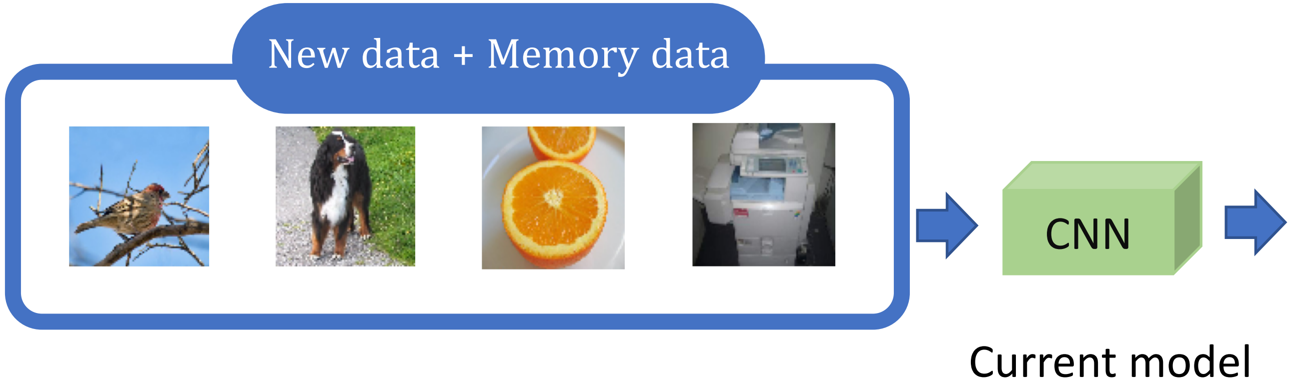


Team UT_LG

Batch-level Experience Replay with Review for Continual Learning

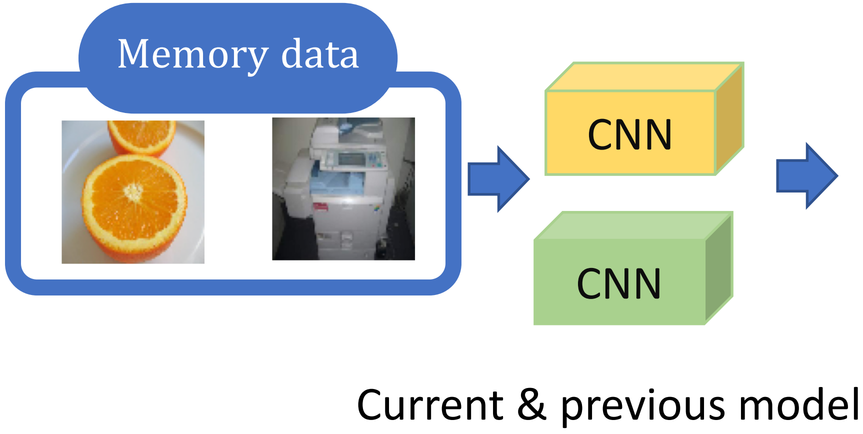


Batch-level Experience Replay with Review for Continual Learning



$$\mathcal{L}_{CE}(\mathbf{x}, y) = \sum_{c=1}^{D_{new}+D_{mem}} -\delta_{c=y} \log\left(p_c(\mathbf{x})\right)$$

cross-entropy loss for new data and memory data

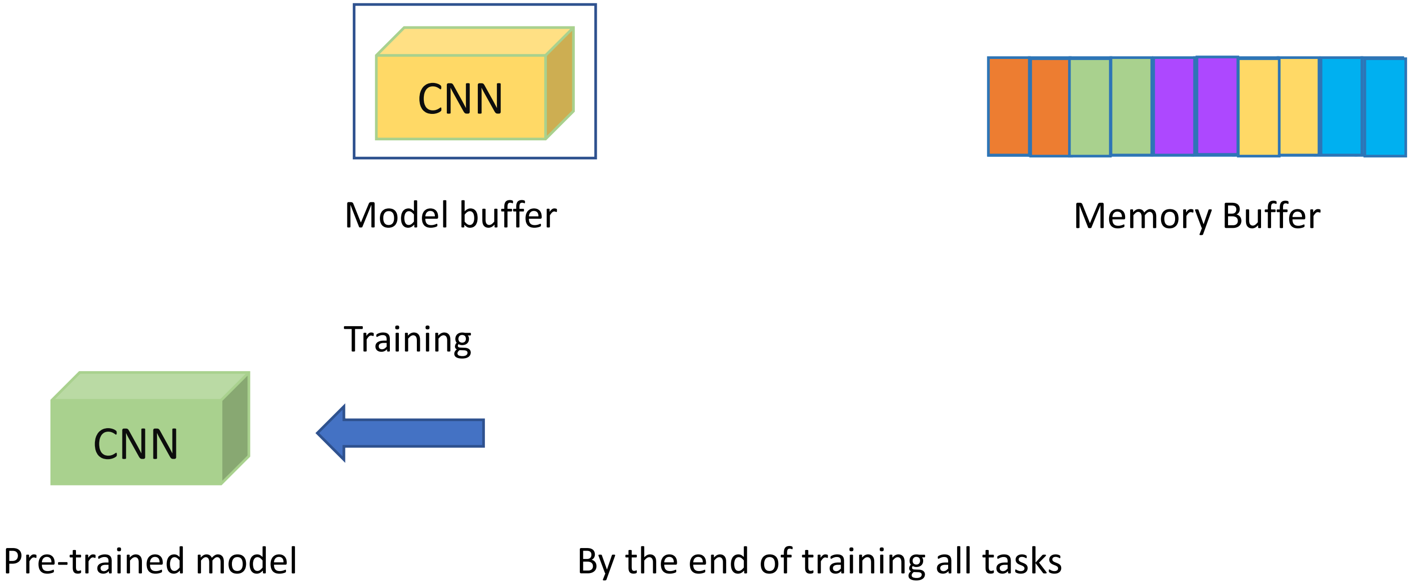


$$\mathcal{L}_{KD}(\mathbf{x}) = \sum_{c=1}^{C_{mem}} -\hat{q}_c(\mathbf{x}) \log\left(q_c(\mathbf{x})\right)$$

knowledge distillation loss for memory data

Total Loss $\quad \mathcal{L}(\mathbf{x}, y) = \mathcal{L}_{CE}(\mathbf{x}, y) + \lambda \mathcal{L}_{KD}(\mathbf{x}) + L_2$

Batch-level Experience Replay with **Review** for Continual Learning



By the end of training all tasks

Smaller learning rate for review

NC – task label is given

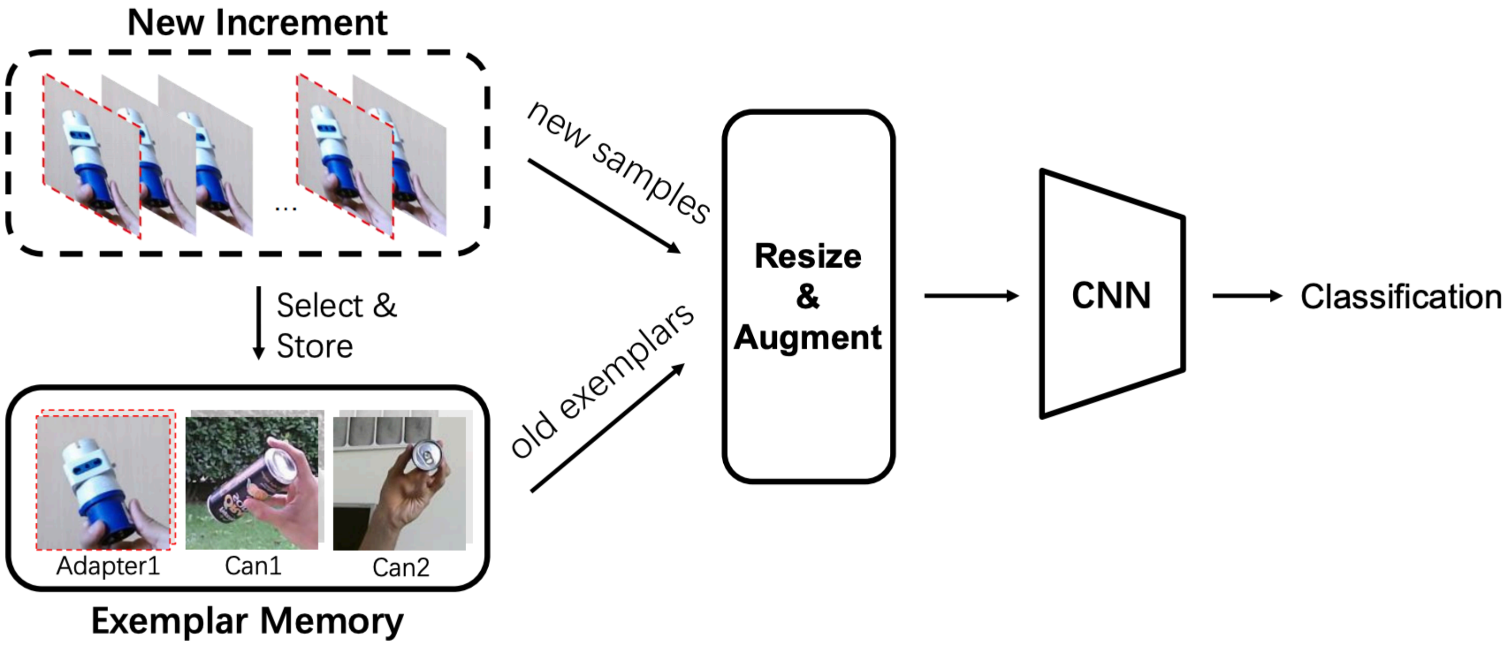
Training

Test

Training Details

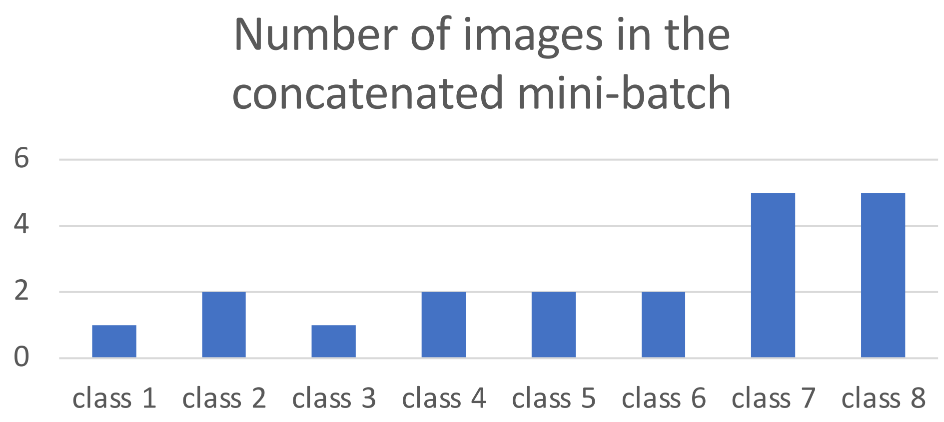
- Architecture
  - A pre-trained DenseNet-161 is used for all scenarios
  - Freeze the the first two blocks and tune the other two blocks
- Pre-processing
  - Center-crop the images with size 100x100
  - Resize the images to 224x224
  - Pixel-level and spatial-level data augmentation
- Memory buffer strategy
  - Update: Reservoir sampling
  - Retrieve: Random

ICT_VIPL Team



- For every incoming mini-batch, retrieve another mini-batch from memory buffer
- Concatenate them to create a new mini-batch
- Resize and data augmentation

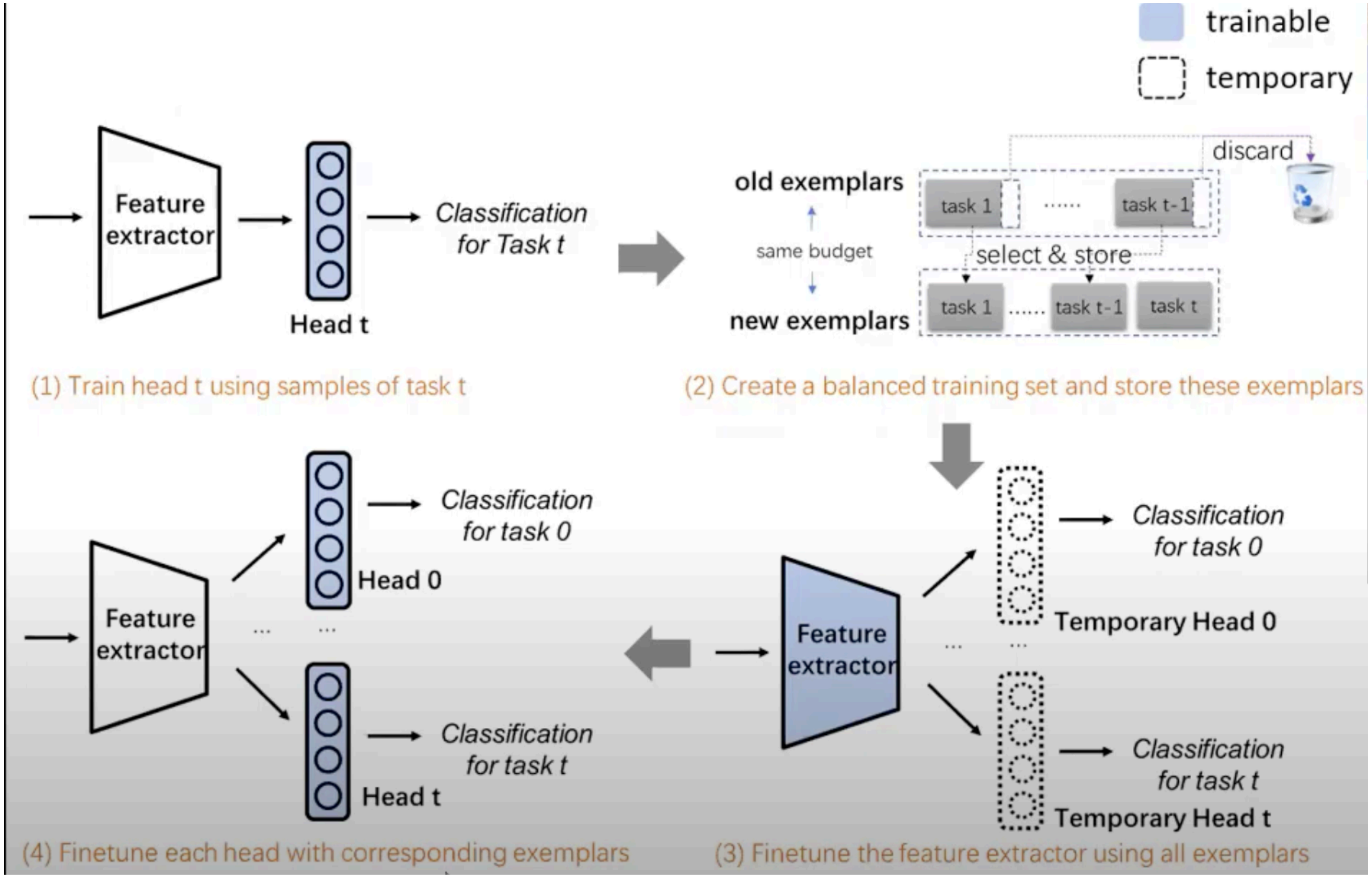
ICT_VIPL Team



- Problem
  - Incoming mini-batch(size 10) contains class 7 and class 8
  - Memory mini-batch(size 10) contains class 1-6
  - There is imbalance in the concatenated mini-batch
  - More severe when the model sees more classes

- Their solution
  - Divide the softmax output by the class prior estimated by the ratio of the corresponding class samples in the current training set.
  - This is a popular strategy to tackle class imbalance problem

NC - task label is given



Team YC14600

Proposed Discriminative Representation Loss

- Minimize the similarities of representations between samples from different classes
- Maximize the similarities of representations between samples from the same class

$$\mathcal{L} = \mathcal{L}_{clf} + \lambda \mathcal{L}_{DR}, \quad \lambda > 0, \quad \text{where} \quad \mathcal{L}_{DR} = \min_{\Theta}(\mathcal{L}_{bt} - \mathcal{L}_{wi})$$

- L_{clf} is the cross-entropy loss for the classification task
- L_{bt} is the similarities of representations between samples from different classes
- L_{wi} is the similarities of representations between samples from a same class

Thank you!
Q&A

zheda.mai@mail.utoronto.ca