ECCV 2018 European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany

Semi-Supervised Deep Learning with Memory

Yanbei Chen[1] Xiatian Zhu[2] Shaogang Gong[1]
yanbei.chen@qmul.ac.uk eddy@visionsemantics.com s.gong@qmul.ac.uk
[1]Queen Mary University of London [2]Vision Semantics Ltd., London, UK

Queen Mary University of London

Introduction

Problem & Motivation

- Semi-supervised multi-class classification is a task of learning from sparse labelled and abundant unlabelled training data.
- The goal of semi-supervised learning is to boost the model performance by utilising the large amount of unlabelled data when only a limited amount of labelled data is available.

Key Contribution

- A novel **Memory-Assisted Deep Neural Network** characterised by a memory mechanism that permits the deep model to additionally learn from its memory (*assimilation*) and adjust itself to fit optimally the incoming training data (*accommodation*) incrementally.

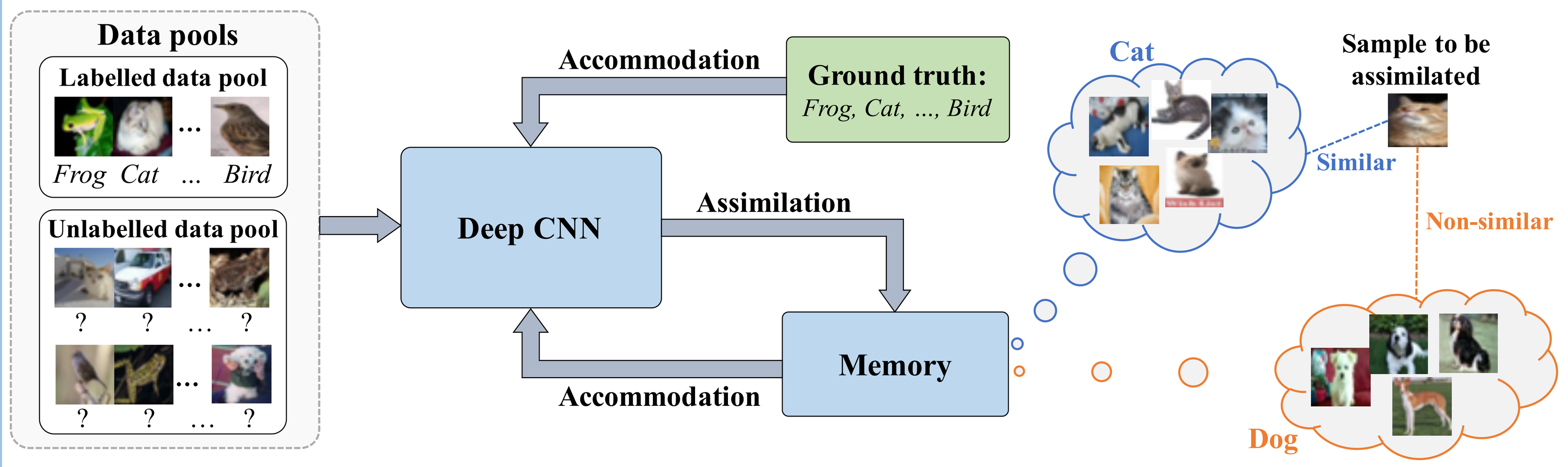

Figure 1. Illustration of the Memory-Assisted Deep Neural Network, which is inspired by the spirit of Piaget's theory on human's ability of continual learning.

Main Ideas

- **Memory of model learning:**
  - (1) class-level feature representation (key embedding): represent the cluster centroid dynamically evolving in the feature space
  - (2) class-level predictive uncertainty (value embedding): encode model inference uncertainty accumulatively revealed by the preceding training iterations
- **Memory loss:**
  - (1) penalise class distribution overlap
  - (2) encourage network predictions to be consistent with confident memory predictions

Methodology

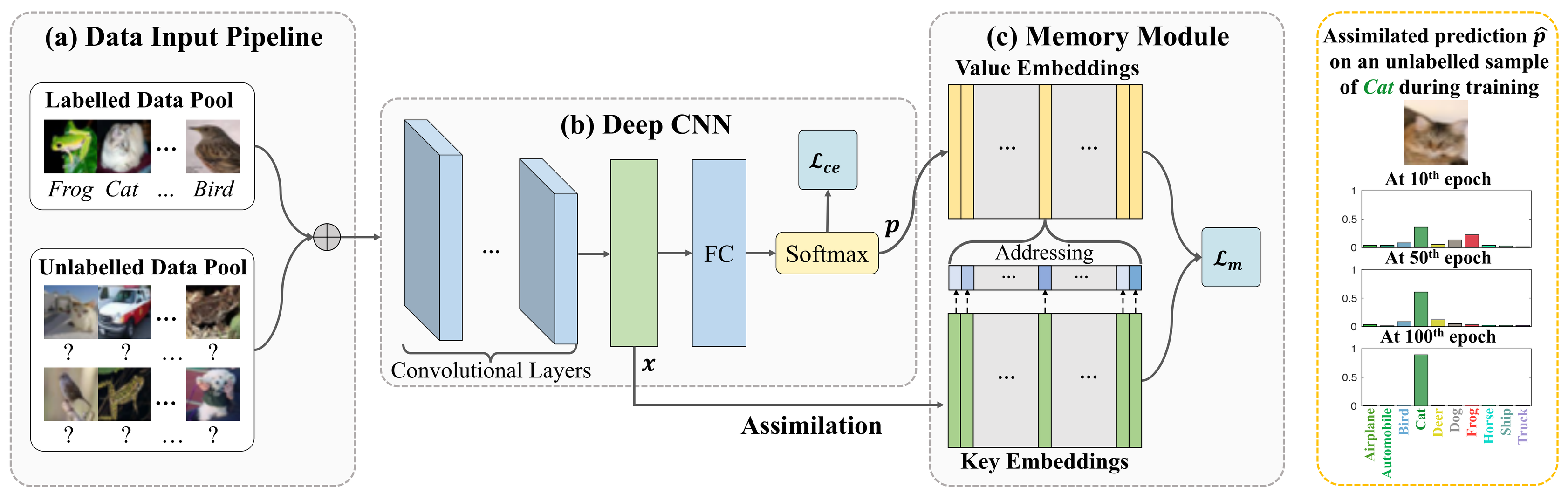

Figure 2. An Overview of the Memory-Assisted Deep Neural Network.

Main components

- Conventional Deep Neural Network
- Memory Module

  $(key, value)$ pairs: iteratively updated by gradients

$$\begin{cases} \boldsymbol{k}_j \leftarrow \boldsymbol{k}_j - \eta \nabla \boldsymbol{k}_j \\ \boldsymbol{v}_j \leftarrow \dfrac{\boldsymbol{v}_j - \eta \nabla \boldsymbol{v}_j}{\sum_{i=1}^{K} (\boldsymbol{v}_{j,i} - \eta \nabla \boldsymbol{v}_{j,i})} \end{cases} \text{with} \begin{cases} \nabla \boldsymbol{k}_j = \dfrac{\sum_{i=1}^{n_j} (\boldsymbol{k}_j - \boldsymbol{x}_i)}{1 + n_j} \\ \nabla \boldsymbol{v}_j = \dfrac{\sum_{i=1}^{n_j} (\boldsymbol{v}_j - \boldsymbol{p}_i)}{1 + n_j} \end{cases}$$

- Assimilation-Accomodation Interaction

(1) Memory Assimilation: derive memory prediction

key addressing & value reading

$$w(\boldsymbol{m}_i|\boldsymbol{I}) = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases} \qquad w(\boldsymbol{m}_i|\boldsymbol{I}) = \frac{e^{-\mathrm{dist}(\boldsymbol{x}, \boldsymbol{k}_i)}}{\sum_{j=1}^{K} e^{-\mathrm{dist}(\boldsymbol{x}, \boldsymbol{k}_j)}} \qquad \hat{\boldsymbol{p}} = \sum_{i=1}^{K} w(\boldsymbol{m}_i|\boldsymbol{I})\, \boldsymbol{v}_i$$

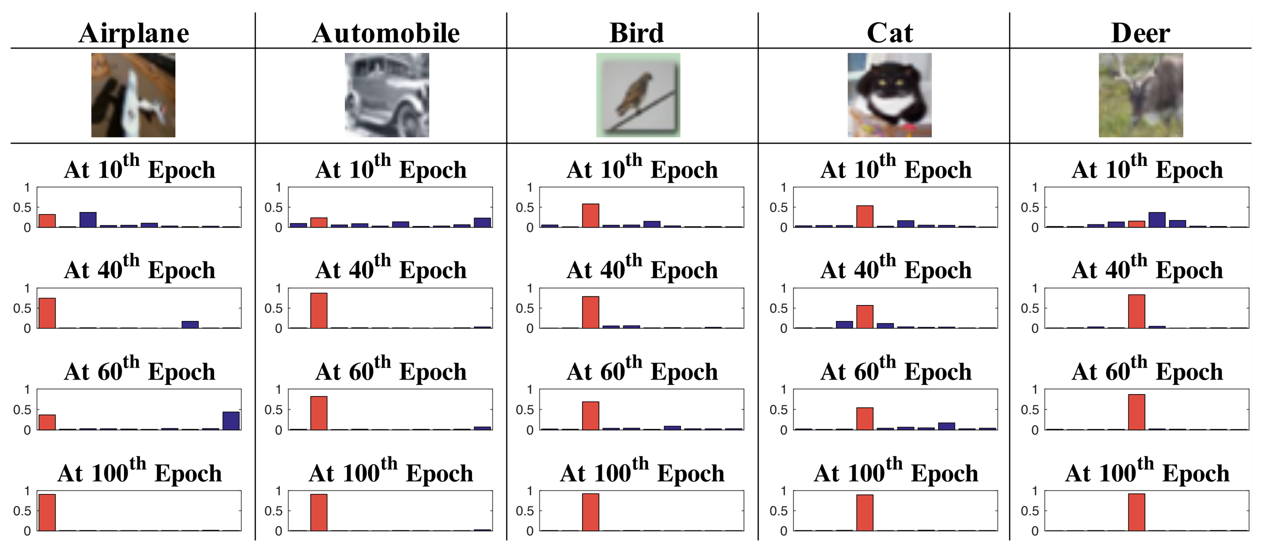

Figure 3. Evolution of memory predictions on the unlabelled samples.

(2) Accommodation: derive memory loss

Model Entropy (ME) + Memory-Network Divergence (MND)

$$H(\hat{\boldsymbol{p}}) = -\sum_{j=1}^{K} \hat{\boldsymbol{p}}(j) \log\, \hat{\boldsymbol{p}}(j) \qquad D_{\mathrm{KL}}(\boldsymbol{p}||\hat{\boldsymbol{p}}) = \sum_{j=1}^{K} \boldsymbol{p}(j) \log \frac{\boldsymbol{p}(j)}{\hat{\boldsymbol{p}}(j)}$$

Final semi-supervised learning objective

$$\mathcal{L}_m = H(\hat{\boldsymbol{p}}) + \max(\hat{\boldsymbol{p}}) D_{\mathrm{KL}}(\boldsymbol{p}||\hat{\boldsymbol{p}})$$

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_m$$

Experiments & Ablation Studies

Experiments on three benchmark datasets

Methods	SVHN [20]	CIFAR10 [13]	CIFAR100 [13]
DGM* [12]	36.02 ± 0.10	–	–
Γ-model [24]	–	20.40 ± 0.47	–
CatGAN* [30]	–	19.58 ± 0.58	–
VAT [19]	24.63	–	–
ADGM* [16]	22.86	–	–
SDGM* [16]	16.61 ± 0.24	–	–
ImpGAN* [27]	8.11 ± 1.3	18.63 ± 2.32	–
ALI* [5]	7.42 ± 0.65	17.99 ±1.62	–
Π-model [14]	4.82 ± 0.17	12.36 ± 0.31	39.19 ± 0.36
Temporal Ensembling [14]	4.42 ± 0.16	12.16 ± 0.24	37.34 ± 0.44
Mean Teacher [32]	**3.95 ± 0.19**	12.31 ± 0.28	–
MA-DNN (Ours)	4.21 ± 0.12	**11.91 ± 0.22**	**34.51 ± 0.61**

Table 1. Comparison with the state-of-the-art SSL methods.

Evolution of the memory module

- **key embeddings:** capture the global manifold structure for deriving probabilistic assignments based on *cluster assumption*

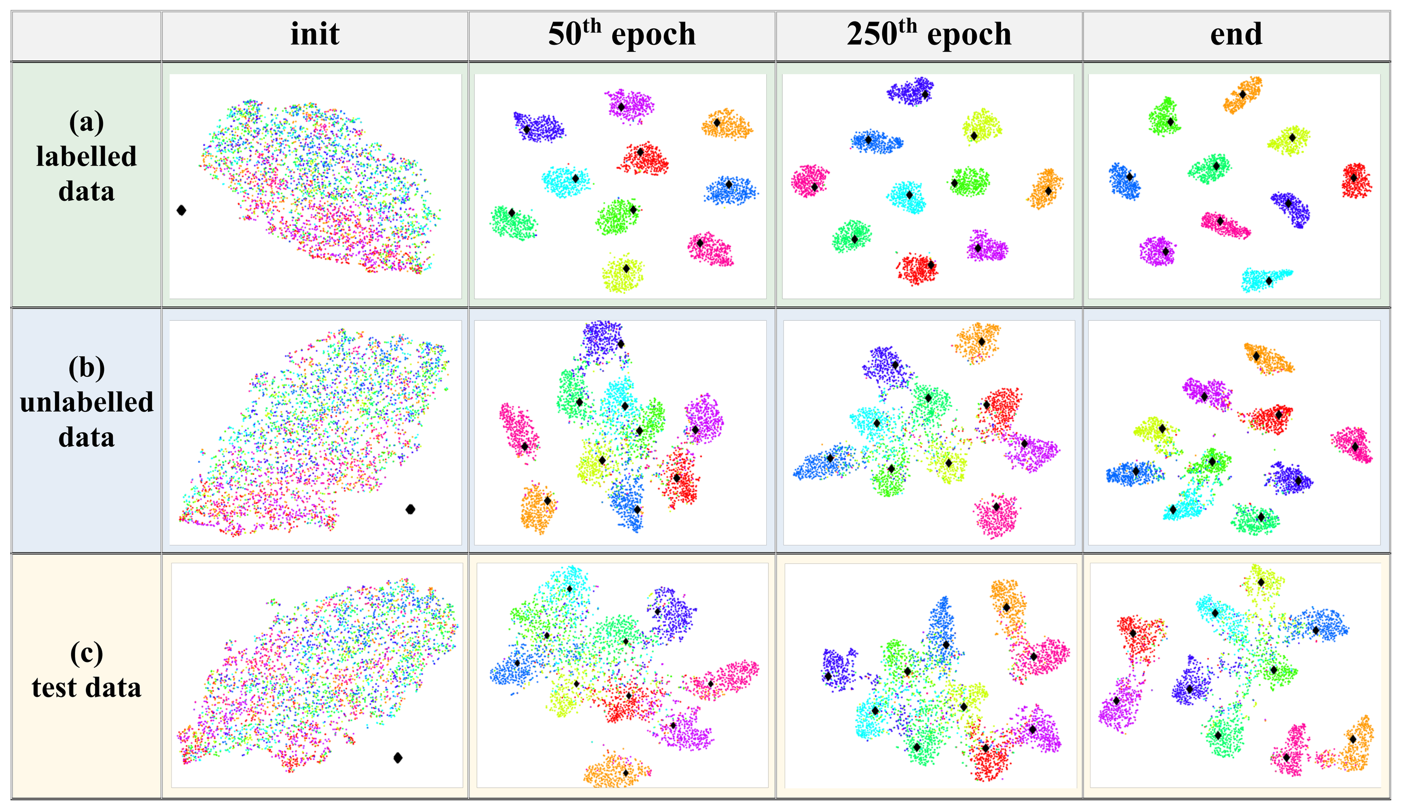

Figure 4. Evolution of the key embeddings.

- **value embeddings:** capture the model inference uncertainty to smooth the memory predictions with uncertainty

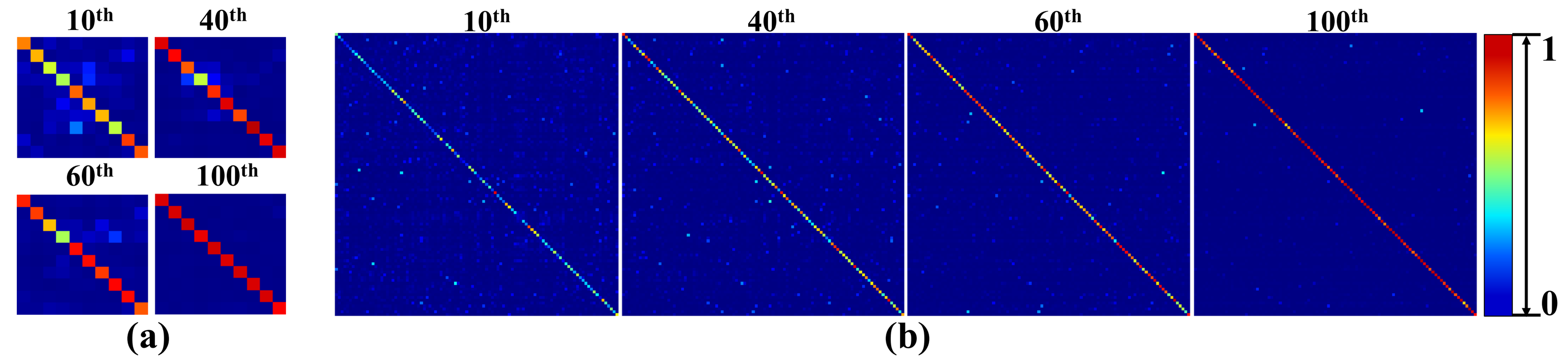

Figure 5. Evolution of the value embeddings.

Reference

[1] Chapelle, O., Schlkopf, B., Zien, A.: Semi-supervised learning. The MIT Press (2010)
[2] Kaiser,L.,Nachum,O.,Roy,A.,Bengio,S.: Learning to remember rare events. In: International Conference on Learning Representation (2017)
[3] Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representation (2017)