LLEDA - Lifelong Self-Supervised Domain Adaptation

Mamatha Thota
School of Computer Science
University of Lincoln
Lincoln, LN6 7TS, UK
mthota@lincoln.ac.uk

Dewei Yi
School of Natural and Computing Sciences
University of Aberdeen
Aberdeen, AB24 3UE, UK
dewei.yi@abdn.ac.uk

Georgios Leontidis
Interdisciplinary Centre for Data and AI
School of Natural and Computing Sciences
University of Aberdeen
Aberdeen, AB24 3UE, UK
georgios.leontidis@abdn.ac.uk

Abstract

Lifelong domain adaptation remains a challenging task in machine learning due to the differences among the domains and the unavailability of historical data. The ultimate goal is to learn the distributional shifts while retaining the previously gained knowledge. Inspired by the Complementary Learning Systems (CLS) theory [31], we propose a novel framework called Lifelong Self-Supervised Domain Adaptation (LLEDA). LLEDA addresses catastrophic forgetting by replaying hidden representations rather than raw data pixels and domain-agnostic knowledge transfer using self-supervised learning. LLEDA does not access labels from the source or the target domain and only has access to a single domain at any given time. Extensive experiments demonstrate that the proposed method outperforms several other methods and results in a long-term adaptation, while being less prone to catastrophic forgetting when transferred to new domains.

1. Introduction

Deep neural networks have shown near human level capabilities in many fundamental computer vision tasks [13]. Humans and animals can continuously acquire new information over their lifetime without catastrophically forgetting the prior knowledge learned. This ability to continually learn over time by accommodating new knowledge while retaining the previously learned knowledge is referred to as lifelong or continual learning (in our paper, we will continue to refer to it as lifelong learning). However, artificial neural networks lack these capabilities as new information interferes with previously learned knowledge and sometimes the old knowledge completely gets overwritten by the new one, leading to impaired performance [32]. The root cause of catastrophic forgetting is that learning requires the neural network's weights to change, but the critical change of weights to past learning results in forgetting.

Following the independent and identical distribution (iid) assumption, a deep neural network learned from the training data is ideally expected to perform well on the test data. However, this assumption may not always hold in the real-world applications due to the discrepancy between domain distributions, and applying the trained model to the new dataset may also result in negative performance. In particular, when a model consecutively learns from different visual domains, it tends to forget the past domains in favour of the most recent ones. Domain adaptation (DA) methods based on deep learning have received significant attention in recent years for mitigating the domain shift from the training domain to the inference domain [11, 27, 46, 47].

Current domain adaptation methods operate under the assumption that datasets from both the source and the target domains are accessible at the same time during training, which may not be feasible in practice. In addition, DA algorithms require fully labelled datasets, therefore these algorithms require annotating massive training datasets for newly observed domains, which is a very time-consuming, cumbersome and expensive process. To relax this constraint, we propose a novel framework called LLEDA that can address both catastrophic forgetting and domain-agnostic knowledge transfer without accessing the

labels either from the source or the target domain, and having access to a single domain at any given time.

Motivated by the complementary learning systems [31] (CLS) theory, we propose a lifelong learning framework that reduces catastrophic forgetting, while facilitating domain-agnostic knowledge transfer without using or accessing labelled data or other information from the past domains at any given time. To the best of our knowledge, this is an area of domain adaptation that has not yet been explored. In summary, our work makes the following contributions:

The main contributions of our work are outlined below

1. We propose a novel lifelong learning framework based on the complementary learning systems theory that mimics the workings of the human brain for addressing lifelong domain adaptation with access to multiple sequential domains, all while not using any labels.

2. We propose to overcome catastrophic forgetting by replaying hidden representations rather than raw pixels in the context of Lifelong Domain Adaptation, attempting to maximise generalisation between source and target domains with different distributions.

3. Our proposed self-supervised based approach does not require access to either source or target labels, hence saving time and effort to annotate data and assisting with the labeling bias.

4. Extensive empirical results demonstrate that our method performs competitively across several benchmarks, when compared against other approaches.

2. Related Work

Our work fundamentally lies at the intersection of lifelong learning inspired by CLS theory [31], self-supervised learning, and domain adaptation. [36] categorises catastrophic forgetting mitigation using model regularisation, memory replay or by expanding and training the network. Regularisation methods identify the network weights that contribute significantly to retaining knowledge about a previously learned task and then consolidate them when the model is updated to learn the subsequent tasks [19, 21, 26]. On the other hand, dynamic architectures modify the model's underlying architecture by dynamically accommodating neural resources as it learns new patterns [22, 25, 39]. Similarly, complementary learning systems and replay methods rely on memory replay by storing samples from old distributions and regularly feeding them back to the model to overcome catastrophic forgetting. Alternatively, the model can be expanded progressively to learn the new tasks using added weights that propose ways of constraining the tasks' objectives to avoid forgetting [20, 29, 40]. In this paper, we tackle lifelong learning using Latent Replay by replaying hidden representations rather than raw pixels. We address the problem of lifelong domain adaptation, where the domain sequentially changes, in contrast to the majority of recent research efforts that concentrate on changes with respect to task/class.

Domain Adaptation: Domain adaptation is a special case of transfer learning where the goal is to learn a discriminative model in the presence of domain shift between source and target datasets. Various methods have been introduced to minimise the domain discrepancy in order to learn domain-invariant features. Some involve adversarial methods like DANN [11], ADDA [48] that help align source and target distributions. Other methods propose aligning distributions through minimising divergence using popular methods like maximum mean discrepancy [13, 14, 27, 28, 46, 47], correlation alignment [4, 44], and the Wasserstein metric [8, 24]. MMD was first introduced for the two-sample tests of the hypothesis that two distributions are equally based on observed samples from the two distributions [14], and this is currently the most widely used metric to measure the distance between two feature distributions. The Deep Domain Confusion Network proposed by Tzeng et al. [49] learns both semantically meaningful and domain invariant representations, while Long et al. proposed DAN [27] and JAN [28] which both perform domain matching via multi-kernel MMD (MK-MMD) or a joint MMD (J-MMD) criteria in multiple domain-specific layers across domains.

Self-Supervised Learning: Self-Supervised Learning (SSL) is a paradigm developed to learn visual features from unlabelled data. Recently, SSL approaches have shown significant performance sometimes even surpassing, the performance of supervised baselines [1, 5–7, 10, 16, 17, 53]. These methods use image augmentation techniques to generate multiple views of a given image and learn a model that is invariant to these augmentations. Most recent approaches are divided into two main categories, contrastive and non-contrastive methods. Contrastive methods learn an embedding space where positive pairs are pulled together, whilst negative pairs are pushed away from each other [5, 6, 17]. Non-contrastive methods on the other hand remove the need for explicit negative pairs either by using distillation or by regularisation of the variance and covariance of the embeddings [1, 7, 16, 53]. However, none of these works studied the ability of SSL methods to learn continually and adaptively if they are applied directly. Moreover, very few works have attempted to use SSL in the lifelong domain adaptation setting, e.g. [45] is designed using contrastive learning, so it lacks the capability to adapt using other SSL paradigms. Another example is [43] where their method adapts well only if the domain shift is small between the intermediate domains and is trained using source-labelled data. In this paper, we present a general-purpose framework to incor-

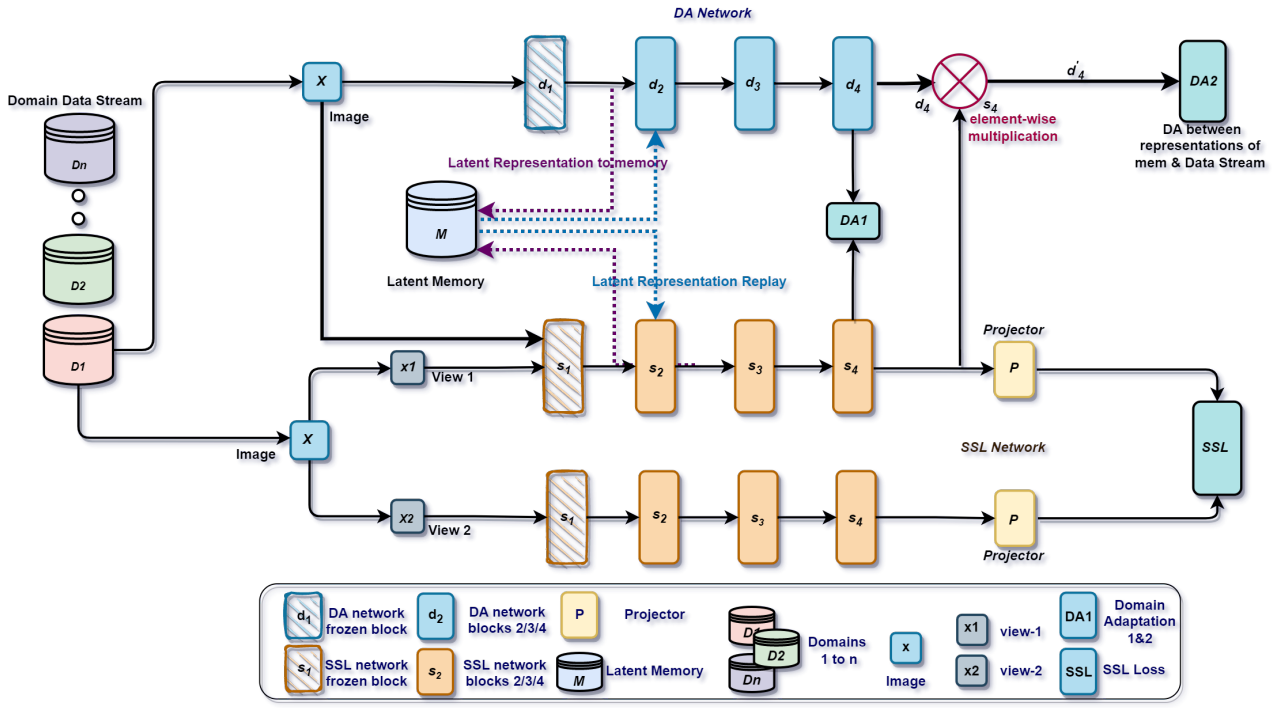

Figure 1. Overview of the proposed LLEDA architechture. LLEDA consists of fast learning **DA network** and slow learning **SSL network**. The **SSL network** learns generic representations using self-supervised learning and **DA network** helps to overcome domain shift by optimising DA loss at two levels, **DA1**- MMD loss between the representations of d_4 and s_4, and **DA2** - MMD loss between memory representations and current data representations.

porate self-supervised learning approaches into the lifelong learning process to extract representations.

Algorithm 1: Pseudocode for the proposed Lifelong Domain Adaptation

Input : Current Domain Data D
Memory Data M
for *sampled minibatch* **do**
  Calculate L_{SSL} loss on D using equation: 5
  Calculate L_{DA1} loss on D using equation: 2
  if *domain* > 1 **then**
    Calculate L_{DA1} loss on M using equation: 2
    Calculate L_{DA2} loss on M and D using equation: 2
  end if
  Add latent representations to memory using algorithm:2
end for

3. Method

Our overall objective is to update a model continually to learn distributional shifts while retaining knowledge about

Algorithm 2: pseudocode for saving random latent representations to memory

$M = \theta$
M_s = memory size
M_c = current memory size
for *each batch* **do**
  $\delta = M_s - M_c$
  S_s = sample size
  S_s = min(S_s, δ)
  R = random sampling of size S_s from batch
  $M = M \cup R$
  M_c = update current memory size
end for

past learnings. Existing domain adaptation and lifelong learning algorithms address the challenge of domain shift, however, they require simultaneous access to the source and the target domains, as well as access to the labelled data or at the very least, the source domain labels. We propose a new lifelong domain adaptation framework (shown in figure 1 and algorithm 1), where we continually update a model to learn the distributional shifts while retaining the prior

knowledge by replaying the hidden latent representations instead of the raw pixels, with access to just a single domain at any given time and no labelled data from either the source domain or the target domain.

LLEDA is inspired by the CLS theory [31], where a the slow-learning SSL network which can be thought of as an analogy of the neocortex gradually acquires structured knowledge using self-supervised learning. In addition, the fast-learning system centred on the hippocampus helps with the rapid learning of specific domain tasks. The fast-learning DA network (hippocampus) addresses domain discrepancy by reducing the distance between the source and target distributions with their mean embeddings in the Reproducing Kernel Hilbert Space. We propose a brain-inspired variant of replay as such we store the internal or hidden latent representations in a buffer which are replayed instead of raw-pixels to overcome catastrophic forgetting. At the same time, the DA network queries the SSL network for quick knowledge acquisition of general representations. The incurred DA losses (DA1 and DA2) are back-propagated through both the learners to reduce the domain shift, interleaved training of the past latent representations from memory with new domain representations consolidate the current learning for long-term retention and to overcome catastrophic forgetting.

3.1. Latent Replay

The mammalian brain has successfully evolved to resist catastrophic forgetting by reactivating, replaying, and recreating the experience preserved in memories [35, 51]. It retains compressed versions of the crucial information from past experiences and reactivates by repeating these neural activity patterns instead of the raw input patterns of prior experiences. Inspired by this, LLEDA uses latent replay for storing the feature representations in the latent memory from a specific layer as a tool to reactivate and replay to overcome catastrophic forgetting.

We freeze the network below the latent replay layer to prevent it from deviating from the feature representations that would have otherwise been generated while feed-forwarding from the input layer, ensuring that these stored latent representations are stable and accurate [37]. As our model does not have access to labels, we follow a simple approach and thereby store a random subset of past latent representations in memory and train the network while interleaving with new domain representations [30]. Following that, we save the latent representations from both the DA and the self-supervised networks for the given random image. During memory consolidation, these memories are interleaved with new latent representations to form a more general representation supporting long-term retention and generalisation when encountering new domain experiences.

As it would be inefficient and impractical to store all past latent representations in the latent buffer, we instead relatively store only a small number of latent representations per domain until the buffer reaches the given number. Thus, at any point in time, the buffer contains a limited size of past random experiences as shown in algorithm 2.

3.2. Domain-specific Representations Learning

Inspired by Dualnet [38], the DA network efficiently utilises the generic representations acquired from the self-supervised network to learn via an adaptation mechanism quickly. The DA network efficiently acquires the low-level general representations from the self-supervised network which can later be used to classify the downstream task alongside reducing the domain shift. It does this in two ways: **Firstly** the DA network calculates the domain shift using representations from block 4 of the Resnet [18] DA1 as shown in figure 2, during both the memory replay and the data source propagation. Maximum mean discrepancy (MMD) is calculated between the representations of DA network and SSL network (DA1) to reduce the domain shift. **Secondly**, MMD is again calculated after the element-wise multiplication between the memory representations and the current data stream propagation (DA2) as shown in figure 2. Finally, we backprapagate these losses through both networks. Again, back propagating the MMD loss at two stages (DA1 and DA2) helps more efficiently in the reduction of domain shift, compared to a single domain adaptation loss.

Let s_4 be the feature representation from the SSL network's residual block and d_4 be the feature representation from the DA network's residual block in ResNet [18] as shown in figure 2, the adapted feature to calculate DA2 loss is obtained as

$$\mathrm{d}_4' = \mathrm{d}_4 \otimes \mathrm{s}_4 \tag{1}$$

where $\otimes$ denotes the element-wise multiplication,the output of the fast DA network d_4, slow SSL network s_4 and the transformed feature $d_4{}'$ all have the same dimension.

The final layer's transformed feature $d_4{}'$ which will be fed into a DA network's head to calculate the DA2 loss using MMD explained later in this section. The fast DA network, therefore, takes advantage of the slow SSL learner's rich feature representations resulting in quick knowledge capture that can be used to reduce domain shift and improve generalisation leading to better identification of classes in the downstream classification task.

MMD defines the distance between the two distributions with their mean embeddings in the Reproducing Kernel Hilbert Space (RKHS). MMD is a two-sample kernel test to determine whether to accept or reject the null hypothesis $p = q$ [14], where p and q are source and target domain probability distributions. In short, the MMD between the distributions of two datasets is equivalent to the distance

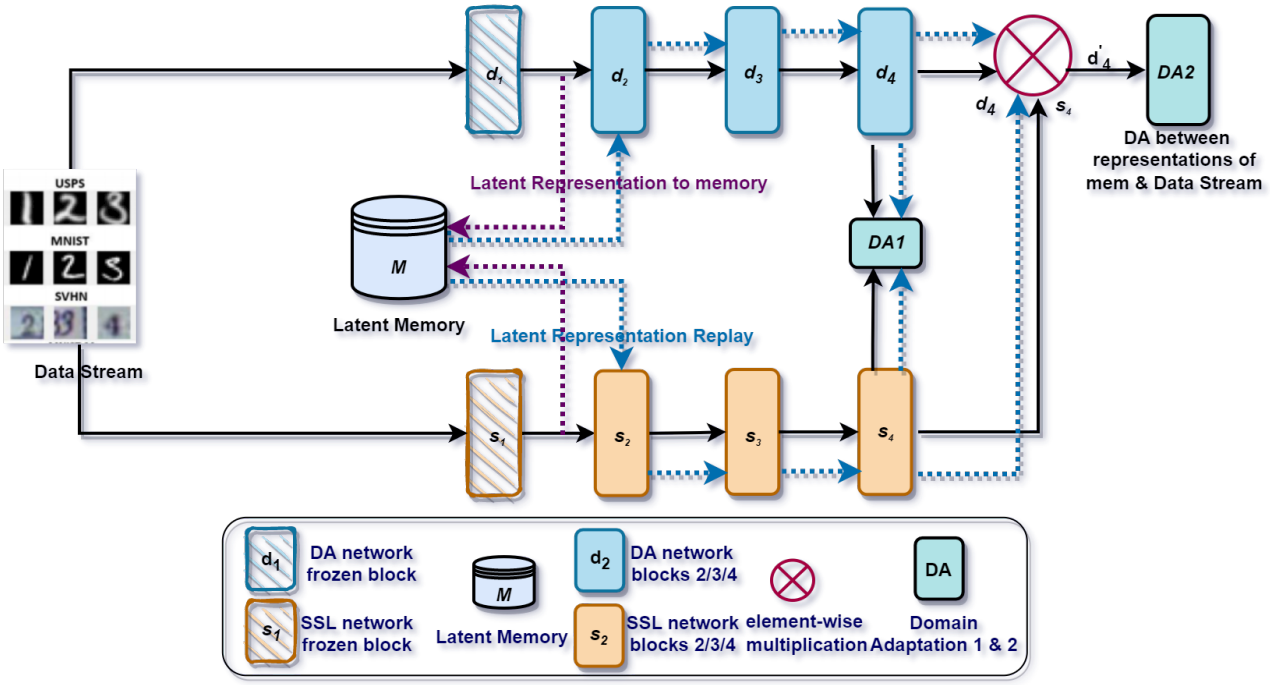

Figure 2. Overview of latent replay. Demonstration of flow of latent representations, **the arrows in blue** show the latent representation flow from memory to the network and **arrows in pink** show the flow of latent representations from network to memory.

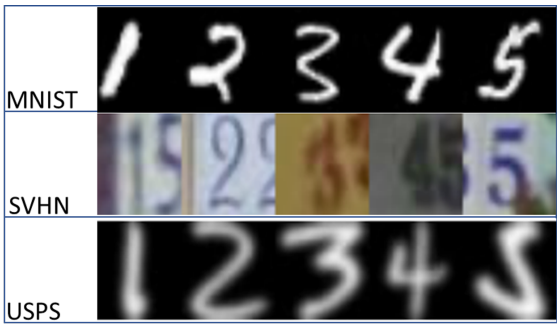

Figure 3. Sample images from Digits Dataset

Figure 4. Sample Images from Office-Home Dataset

between the sample means in a high-dimensional feature space and is computed by the following equation:

$$L_{MMD} = \left\| \frac{1}{N}\sum_{i=1}^{N}\phi(x_i^s) - \frac{1}{M}\sum_{j=1}^{M}\phi(x_j^t) \right\|_H^2 \quad (2)$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}k(x_i^s, x_{i'}^s) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}k(x_i^s, x_j^t) + \frac{1}{M^2}\sum_{j=1}^{M}\sum_{j'=1}^{M}k(x_j^t, x_{j'}^t) \quad (3)$$

where $\phi(.)$ is the mapping to the RKHS H, $k(.,.) = \langle \phi(.), \phi(.) \rangle$ is the universal kernel associated with this mapping, and N, M are the total number of items in the source and target respectively.

3.3. Generalised Feature Learning

The SSL learner is a standard backbone network trained to optimise an SSL loss in order to learn the general representations where labelled data is not available, which is the case with our scenario. Self-supervised methods learn to maximise the similarity between embedding vectors pro-

Figure 5. Sample Images from Office-Caltech Dataset

Dataset	Method	Average
Digits	Baseline	56.7
	DANN	74.5
	DAN	72.9
	CUA	82.1
	GRCL	85.3
	LLEDA-S	**89.0**
	LLEDA	**86.6**

Table 1. Comparison of LLEDA on Digit datasets with state-of-the-art

duced by encoders fed with different views of the same image. As this is a generic step to harness the representations, any SSL method can be applied. However, we have used VICReg [2] to preserve the information content of the representations as it does not require a memory bank, contrastive samples, or a large batch size. VICReg [2] uses a weighted average of invariance, variance and covariance as follows

$$l(z_i, z_j) = \lambda s(z_i, z_j) + \mu[v(z_i) + v(z_j)] + \nu[c(z_i) + c(z_j)] \tag{4}$$

Where λ, μ, ν are the hyper-parameters controlling the importance of each term in the loss. $s(z_i, z_j)$ is the Invariance, $c(z_i)$, $c(z_j)$ is covariance and $v(z_i)$, $v(z_j)$ is variance. The overall objective is given by

$$L = \sum_{I \epsilon D} \sum_{t_i, t_j \sim T} l(z_i, z_j) \tag{5}$$

4. Experiments & Results

4.1. Datasets

We compare and evaluate our method against baseline approaches on a number of benchmark DA datasets, such as Digits, Office-Home [50] and Office-CalTech [12].

Digit Datasets: We consider standard digits datasets broadly adopted by the computer vision community. MNIST [23] and USPS [9] are hand-written grey-scale images, with relatively small domain differences. SVHN [33] includes coloured images of street numbers and contains more than one digit in each image. Sample images of the digit datasets are presented in figure 3. We conducted experiments on two tasks: SVHN $\rightarrow$ MNIST $\rightarrow$ USPS and MNIST $\rightarrow$ SVHN $\rightarrow$ USPS. These two scenarios will allow us to reflect on the performance of lifelong learning scenarios starting from easy datasets, moving to harder ones and vice versa.

Dataset	Method	Average
Office-Home	Baseline	28.7
	DANN	57.6
	DAN	56.3
	CUA	58.6
	LDA-CID	59.4
	LLEDA-S	**60.3**
	LLEDA	**58.2**

Table 2. Comparison of LLEDA on Office-Home datasets with state-of-the-art

Dataset	Method	Average
Office-Caltech	Baseline	52.3
	DANN	81.7
	EWC	84.5
	CUA	84.8
	GRCL	87.2
	LLEDA-S	**87.5**
	LLEDA	**86.1**

Table 3. Comparison of LLEDA on Office-Caltech datasets with state-of-the-art

Office-Home [50]: The office-home data consists of four visual domains: Art (A), Clipart (C), Real World (R), and Product (P) each consisting of images from 65 visual categories totalling 15,500 images in office and home settings leading to the possibility of defining 12 pair-wise binary UDA tasks. Sample images of the office-home datasets are presented in figure 4.

Office-CalTech [12]: This dataset is an extension of the Office-31 [42] with 10 categories shared by Office-31 and the CalTech-256 dataset [15]. This dataset has four domains: Webcam (W), DSLR (D), Amazon (A), and CalTech (C). Sample images of the office-caltech datasets are presented in figure 5.

4.2. Training Methods

We benchmark our LLEDA method against the baseline method which simply finetunes the model as new training

Method	CYCLE-1			CYCLE-2			Avg
	SVHN	USPS	MNIST	MNIST	USPS	SVHN	
VICReg	71.3	93.3	94.1	86.7	85.9	88.7	86.6
SimCLR	73.6	94.8	93.8	78.9	87.2	90.5	86.4
BYOL	70.9	95.5	92.6	86.3	88.9	87.5	86.9

Table 4. Ablation: Comparison of LLEDA using different self-supervised methods.

domains come along, we then compare our LLEDA method with DANN [11] and DAN [27], both of them are classic domain adaptation methods and both these methods have access to source and target data during training. We also compare LLEDA with LDA-CID [41], CUA [3] and GRCL [45] which are continual learning methods with access to source labels. The former is a generative method and the later ones are replay based methods. We also compare LLEDA-S method with all the above methods, this is a supervised version of LLEDA with access to labels.

4.3. Implementation Details

We have 3 stages to our implementation. Firstly, we pretrain the model with Resnet18 as the base encoder on ImageNet without using memory or performing domain adaptation. We then use the pretrained network and train our network as discussed in the methodology section. In the later stage, we train a linear classifier on top of a fixed representation and finally evaluate using the encoder whilst discarding the MMD projection head.

We pretrain our network using the base encoder ResNet-18 [18] on Imagenet. During pretraining, we train LLEDA on two nodes, each consisting of 4 GPUs (Titan Xp GPUs), using LARS optimizer [52] with a batch size of 512 and weight decay of le-6 for a total of 100 epochs. We use the pretrained base encoder trained in the previous step as a base during the training phase and train the datasets by interleaving the stored memory from the given layer and new data source representations. We use SGD optimizer with a batch size of 128 and weight decay of le-4. During finetuning and evaluation, we freeze the trained network (s_1, d_1 as shown in figure 1) and train a linear classifier on top of a fixed representation, whilst discarding the MMD part of the network, which we then use for evaluation. Similar to most self-supervised models [1, 5, 6, 6, 7, 16, 17, 34, 53], we report performance by training a linear classifier on top of a fixed representation to evaluate representations which is a standard benchmark that has been adopted by many papers in the literature.

4.4. Results and Analysis

Baseline: We start by training a basic model M_i on domain Di, we then finetune the model by training on the next available sequential domain $Di + 1$. When this training reaches the end of the cycle, it often performs badly on older domains due to catastrophic forgetting, we treat this as our baseline. In our experiments, we use Resnet18 as our baseline model.

To start with, LLEDA shows increased performance with respect to the baseline 56.7% to 86.6% table 1, but it is a very basic finetuned model. From table ref, we can see that the performance of LLEDA and LLEDA-S on the Digits dataset is significantly better than the other state-of-the-art methods.

Office-Home dataset Again similar to Digits, Office-Home dataset has an increased performance when compared to the baseline method from 28.7% to 58.2% which can be seen in table 2, which is expected. The LDA-CID method has a slight advantage as it has access to labels, so compared to LLEDA, the accuracy is 1.2% higher. On the other hand, LLEDA-S with access to labels has an increased performance of 0.9% compared to the LDA-CID methodology.

Office-Caltech dataset: Similar observation here with respect to the baseline comparison, the performance increased from 52.3% to 86.1% as expected. LLEDA method performed well compared to the other state-of-the-art methods from table 3. GRCL method is 1.1% higher than LLEDA as it has a slight advantage due to its access to labels. But if we compare GRCL with LLEDA-S, LLEDA-S shows a marginal (0.1%) increase in performance compared to the GRCL methodology.

Overall even though LLEDA does not have access to labels or access to source datasets, we can clearly observe from tables-1-3 that the performance of LLEDA i.e, the average accuracy is comparatively the same or better than the other methods.

4.5. Ablation Studies

We analyse the effectiveness of LLEDA using experimental cycles to replicate the lifelong learning scenario. We start by training the image samples from one dataset, and then continue to train on image samples from the next dataset. The two cycles we suggest are as follows SVHN - USPS - MNIST and MNIST - USPS - SVHN, we refer to these as cycle-1 and cycle-2 respectively. We start by

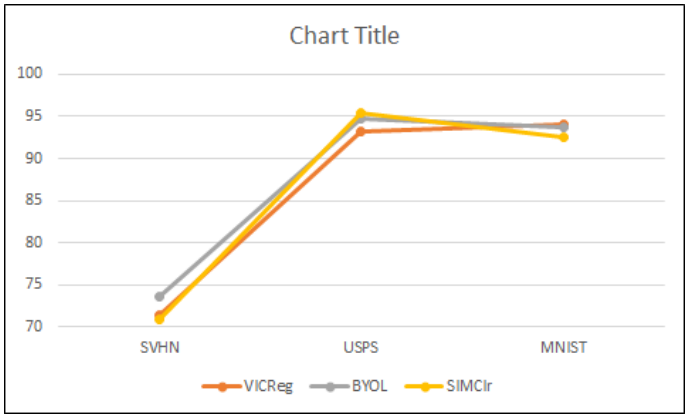

Figure 6. Comparison of SSL methods on the Digits dataset

analysing the LLEDA accuracy with respect to the method of the slow-learning SSL network used. In table 4 and figure 6, we compare three SSL methods: SimCLR, BYOL and VICReg. We chose to compare these 3 methods as the former is a contrastive-based method whereas the latter two are non-contrastive ones. All three methods feature different losses and use different techniques to avoid collapse (e.g. negative samples, redundancy reduction, etc). From table 4, we can see that the average performance of VICReg is robust in comparison to the average performance of contrastive-based SimCLR as the latter requires a require large amounts of contrastive pairs and a higher batch size to converge. The average performance of VICReg slightly underperforms compared to BYOL, but overall, the comparative performance of all three SSL methods is almost similar.

5. Conclusion & Future Work

Inspired by the way the human brain works and the CLS theory, we developed LLEDA which can perform competitively in a continual domain adaptation setting storing and replaying compressed hidden representations rather than the raw pixel data. We have demonstrated that our model can be effectively used for downstream continual domain adaptation tasks without having access to any labelled data. Experimental results show that our proposed LLEDA method significantly outperforms other advanced methods. We hope that the impressive outcomes of our study will inspire future researchers to use source and target unlabeled data in the lifelong domain adaptation setting. In our future work, we would like to work on lossy or lossless compression techniques to further compress the latent representations in order to store and replay in a more efficient manner.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 2, 7

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *CoRR*, abs/2105.04906, 2021. 6

[3] Andreea Bobu, Eric Tzeng, Judy Hoffman, and Trevor Darrell. Adapting to continuously shifting domains. 2018. 7

[4] Chao Chen, Zhihong Chen, Boyuan Jiang, and Xinyu Jin. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3296–3303, 2019. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 7

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 2, 7

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2, 7

[8] Zhihong Chen, Chao Chen, Xinyu Jin, Yifu Liu, and Zhaowei Cheng. Deep joint two-stream wasserstein auto-encoder and selective attention alignment for unsupervised domain adaptation. *Neural computing and applications*, pages 1–14, 2019. 2

[9] John S Denker, WR Gardner, Hans Peter Graf, Donnie Henderson, Richard E Howard, W Hubbard, Lawrence D Jackel, Henry S Baird, and Isabelle Guyon. Neural network recognizer for hand-written zip code digits. In *Advances in neural information processing systems*, pages 323–331. Citeseer, 1989. 6

[10] Aiden Durrant and Georgios Leontidis. Hyperspherically regularized networks for self-supervision. *Image and Vision Computing*, page 104494, 2022. 2

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 1, 2, 7

[12] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2066–2073. IEEE, 2012. 6

[13] Liyun Gong, Mamatha Thota, Miao Yu, Wenting Duan, Mark Swainson, Xujiong Ye, and Stefanos Kollias. A novel unified deep neural networks methodology for use by date recognition in retail food package image. *Signal, Image and Video Processing*, 15(3):449–457, 2021. 1, 2

[14] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012. 2, 4

[15] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 6

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 7

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 7

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 7

[19] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. *CoRR*, abs/1607.00122, 2016. 2

[20] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. *CoRR*, abs/1711.10563, 2017. 2

[21] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016. 2

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[23] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6

[24] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019. 2

[25] Jeongtae Lee, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *CoRR*, abs/1708.01547, 2017. 2

[26] Zhizhong Li and Derek Hoiem. Learning without forgetting. *CoRR*, abs/1606.09282, 2016. 2

[27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 1, 2, 7

[28] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 2

[29] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continuum learning. *CoRR*, abs/1706.08840, 2017. 2

[30] Davide Maltoni and Vincenzo Lomonaco. Continuous learning in single-incremental-task scenarios. *CoRR*, abs/1806.08568, 2018. 4

[31] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995. 1, 2, 4

[32] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 1

[33] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 6

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 7

[35] Joseph O'Neill, Barty Pleydell-Bouverie, David Dupret, and Jozsef Csicsvari. Play it again: reactivation of waking experience and memory. *Trends in Neurosciences*, 33(5):220–229, 2010. 4

[36] German Ignacio Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *CoRR*, abs/1802.07569, 2018. 2

[37] Lorenzo Pellegrini, Gabriele Graffieti, Vincenzo Lomonaco, and Davide Maltoni. Latent replay for real-time continual learning. *CoRR*, abs/1912.01100, 2019. 4

[38] Quang Pham, Chenghao Liu, and Steven C. H. Hoi. Dualnet: Continual learning, fast and slow. *CoRR*, abs/2110.00175, 2021. 4

[39] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *CoRR*, abs/1611.07725, 2016. 2

[40] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *CoRR*, abs/1810.11910, 2018. 2

[41] Mohammad Rostami. Lifelong domain adaptation via consolidated internal distribution. *Advances in Neural Information Processing Systems*, 34:11172–11183, 2021. 7

[42] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 6

[43] Mark Schutera, Frank M. Hafner, Jochen Abhau, Veit Hagenmeyer, Ralf Mikut, and Markus Reischl. Cuepervision: self-supervised learning for continuous domain adaptation without catastrophic forgetting. *Image and Vision Computing*, 106:104079, 2021. 2

[44] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 2

[45] Shixiang Tang, Peng Su, Dapeng Chen, and Wanli Ouyang. Gradient regularized contrastive learning for continual domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2665–2673, 2021. 2, 7

[46] Mamatha Thota, Stefanos Kollias, Mark Swainson, and Georgios Leontidis. Multi-source domain adaptation for quality control in retail food packaging. *Computers in Industry*, 123:103293, 2020. 1, 2

[47] Mamatha Thota and Georgios Leontidis. Contrastive domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2209–2218, June 2021. 1, 2

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 2

[49] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[50] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. *CoRR*, abs/1706.07522, 2017. 6

[51] Matthew A Wilson and Bruce L McNaughton. Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679, 1994. 4

[52] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 7

[53] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2, 7