





# Transfer Learning via Test-Time Neural Networks Aggregation

Bruno Casella<sup>1,2</sup><sup>a</sup>, Alessio Barbaro Chisari<sup>3,4</sup><sup>b</sup>, Sebastiano Battiato<sup>4</sup><sup>c</sup> and Mario Valerio Giuffrida<sup>5</sup><sup>d</sup>

<sup>1</sup>*Department of Computer Science, University of Torino, Torino, Italy*

<sup>2</sup>*Department of Economics and Business, University of Catania, Catania, Italy*

<sup>3</sup>*Department of Civil Engineering and Architecture, University of Catania, Catania, Italy*

<sup>4</sup>*Department of Mathematics and Computer Science, University of Catania, Catania, Italy*

<sup>5</sup>*School of Computing, Edinburgh Napier University, Edinburgh, UK*

*casella@di.unito.it, alessio.chisari@phd.unict.it, battiato@dmi.unict.it, v.giuffrida@napier.ac.uk*

Keywords: parameter aggregation, transfer learning, selective forgetting.

Abstract: It has been demonstrated that deep neural networks outperform traditional machine learning. However, deep networks lack generalisability, that is, they will not perform as good as in a new (testing) set drawn from a different distribution due to the domain shift. In order to tackle this known issue, several transfer learning approaches have been proposed, where the knowledge of a trained model is transferred into another to improve performance with different data. However, most of these approaches require additional training steps, or they suffer from catastrophic forgetting that occurs when a trained model has overwritten previously learnt knowledge. We address both problems with a novel transfer learning approach that uses network aggregation. We train dataset-specific networks together with an aggregation network in a unified framework. The loss function includes two main components: a task-specific loss (such as cross-entropy) and an aggregation loss. The proposed aggregation loss allows our model to learn how trained deep network parameters can be aggregated with an aggregation operator. We demonstrate that the proposed approach learns model aggregation at test time without any further training step, reducing the burden of transfer learning to a simple arithmetical operation. The proposed approach achieves comparable performance w.r.t. the baseline. Besides, if the aggregation operator has an inverse, we will show that our model also inherently allows for selective forgetting, i.e., the aggregated model can forget one of the datasets it was trained on, retaining information on the others.

## 1 INTRODUCTION


Deep Learning (DL) has demonstrated superior performance than traditional ML methods in a variety of tasks. This is due to being able to extract discriminative features from the data for the task at hand via end-to-end training. Such discriminative features are suitable for the dataset the network was trained on. However, a deep network will not perform as good as in a different dataset due to the *domain shift* (or dataset bias) (Zhao et al., 2020).


A way to address the domain shift is via Transfer Learning (TL), where the information learnt by a trained network is (re)used in another context.


Several approaches to transfer learning have been proposed in literature, such as sample reweighting (Schölkopf et al., 2007), feature distributions minimisation (Tzeng et al., 2017; Litrico et al., 2021), distillation (Hinton et al., 2015), and so on (for a recent survey on TL, please read (Zhuang et al., 2021)).


However, transfer learning techniques may be affected by catastrophic forgetting (Goodfellow et al., 2013), where a network forgets the information learnt from a previous task when transferred to a new one. Furthermore, generally, transfer learning requires further training steps to accommodate for new data, even though the learnt task remains unchanged.

The benefit of transfer learning has been demonstrated extensively in the last years (Weiss et al., 2016), even in distributed training scenario (Chen et al., 2020). In this context, a central model is trained on several datasets that have never directly seen, as they are located in different machines (feder-

<sup>a</sup> <https://orcid.org/0000-0002-9513-6087>

<sup>b</sup> <https://orcid.org/0000-0002-7831-382X>

<sup>c</sup> <https://orcid.org/0000-0001-6127-2470>

<sup>d</sup> <https://orcid.org/0000-0002-5232-677X>

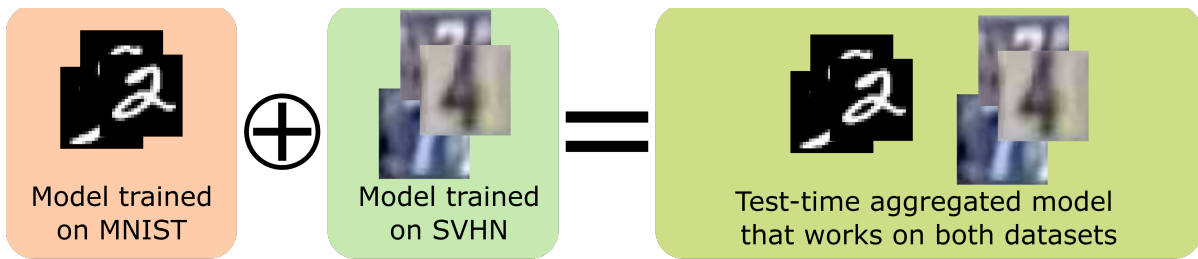


Figure 1: Pictorial representation of the proposed method that performs test-time neural network aggregation.

ated learning). However, this training paradigm raised another question: what if one (or more) datasets used to train the centrally trained model needs to be removed? Machine unlearning (Golatkar et al., 2021) is studied for several reasons, especially when sensible data are used (e.g., medical imaging). However, it is generally hard to selectively *scrub* the parameters of a model such that it cannot perform well on a portion of the dataset, whilst it retains comparable performance as before on the rest of the dataset.

In this paper, we propose a new proof-of-concept technique to TL that inherently allows for selective forgetting by aggregating the network parameters *without any further training*. This can be applied to different datasets, assuming they all share the same task. Our approach is represented in Figure 1 and works as follows: we train a VGG-like (Simonyan and Zisserman, 2015) deep neural network for each dataset – we will refer to these networks as  $N_i$ , for  $i = 1, \dots, n$ , with  $n$  being the number of datasets. In addition, a VGG-like network – named  $N^*$  – is also trained taking all the datasets as inputs. All the networks are trained end-to-end with a *aggregation* regulariser, ensuring that the weights learnt by  $N^*$  are obtained as an aggregation for all the other networks  $N_i$ . This training paradigm will ensure that the networks  $N_i$  also learn how to be aggregated. Furthermore, requiring that the aggregation function is invertible, our model inherently allows for selective forgetting. In our experiments, we set  $n = 2$  datasets, and we used the sum of weights as network aggregation function (which can easily be inverted with subtraction), which is applied to only the parameters of the feature extractors. All the networks trained within this end-to-end framework (including  $N^*$ ) share the same classifier. Experimental results show that test-time network aggregation is possible, outperforming the baseline.

The key contributions of our approach can be summarised as follows:

1. we propose the *aggregation regulariser* during training;
2. network aggregation is achieved at test time (no further training is required);

3. our transfer learning technique does not suffer from catastrophic forgetting;
4. our approach can also be used for selective forgetting (assuming networks are aggregated via an invertible function).

The rest of the paper is organised as follows. In Section 2, we discuss the recent related works. Section 3 outlines our proposed approach. In Section 4, experimental results are shown and discussed. Finally, Section 5 concludes the paper.

## 2 RELATED WORK

The aggregation of network parameters is a form of transfer learning. Typically, TL generally addresses a better initial and steeper growth performance (Tommasi et al., 2010) by reuse of the convolutional filter parameters of CNNs. For example, fine-tuning is the simplest way to achieve transfer learning: a model, pre-trained on a dataset, e.g. ImageNet (Deng et al., 2009), is used as starting point for other datasets and tasks (Reyes et al., 2015). Although intuitive and easy to do, fine-tuning typically underperforms wrt other transfer learning approaches (Shu et al., 2021; Han et al., 2021). More sophisticated methods have been proposed (Oquab et al., 2014), but several of them suffer from *negative transfer* (Rosenstein et al., 2005; Pan and Yang, 2010; Torrey and Shavlik, 2010; Wang et al., 2018; Zhuang et al., 2021): the process of transferring knowledge is harmful because the knowledge is not transferable across all the domains (in particular when the source and target datasets are not related).

Another issue affecting transfer learning approaches is *catastrophic forgetting*, where new knowledge permanently replaces information learnt from previous tasks (Goodfellow et al., 2013). In fact, several approaches to TL, such as *Batch Spectral Shrinkage* (Chen et al., 2019), attempts to solve such an issue. However, these approaches still rely on a training procedure to adapt to a new dataset (or task). However, we asked ourselves the following question:

is it possible achieving TL without catastrophic forgetting at test time? We achieve that by aggregating the weights of trained networks together.

The idea of aggregating the parameters of deep neural networks is not new in the literature. A framework that aggregates knowledge from multiple models is the *Transfer-Incremental Mode Matching* (T-IMM) (Geyer et al., 2019), which enables for adaptive merging of models. It is a re-interpretation of IMM (Lee et al., 2017), a work in the context of lifelong learning aiming at the sequential aggregation of models retaining good performance on all the prior tasks, rather than on transfer learning. T-IMM belongs to the field of incremental learning, a subtly different area concerning lifelong learning, in which the parameters of the  $i$ -th model are used as initialisation for model  $i + 1$ . More recently, Zoo-Tuning was proposed to adaptively aggregate multiple trained models (Shu et al., 2021). To achieve network aggregation, the authors proposed the *AdaAgg* layer. However, this approach assumes that models are already pre-trained before being aggregated (involving a two-step learning). In our work, models are randomly initialised and then trained once end-to-end and simultaneously.

Lifelong (or continual) learning describes the scenario in which new tasks arrive sequentially and should be incorporated into the current model, retaining previous knowledge (Parisi et al., 2019). Approaches to lifelong learning are mainly aimed to mitigate catastrophic forgetting (Rao et al., 2019; Ramapuram et al., 2020; Ye and Bors, 2020). According to Parisi *et al.* (2019), there are three main approaches to lifelong learning: (i) retraining with regularisation; (ii) network expansion; (iii) selective network retraining and expansion. In the first case, neural networks are retrained with constraints to prevent forgetting. Network expansion approaches perform architectural changes (e.g., adding neurons) to the network to add novel information. The last approaches update only a subset of neurons and allow expansion (if necessary). Our proposed method loosely follows the paradigm of regularisation approaches with an important difference: no retraining of the architecture is performed neither transfer learning nor selective forgetting.

Our approach to network aggregation inherently allows network decomposition for selective forgetting. Recently, several related works have focused on machine unlearning (Golatkhar et al., 2020; Golatkhar et al., 2021). Overall, these approaches assume that the portion of the dataset that the model should unlearn is given to a *scrub* function that aims to remove the information learnt from the dataset to be forgotten, impacting (although minimally) the performance of the scrubbed model on the rest of the dataset. Our

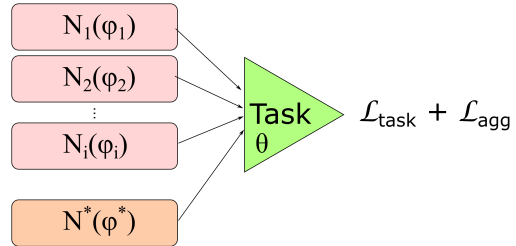


Figure 2: Graphical representation of the proposed method. Each network  $N_i$  is parametrised by a set of weights  $\phi_i$ . The aggregated network  $N^*$  is parametrised by  $\phi^*$ . There is no weight sharing between these networks. However, all the networks share the same task network (i.e., a classifier). The total objective function is given as a combination two loss functions: (i) task loss (i.e., cross-entropy); (ii) aggregation loss (see Section 3.3).

approach is different: we do not require data to be provided for selective forgetting. Instead, the aggregated model trained on two (or more) datasets can be changed by applying the inverse of the aggregation function (in our case, a simple subtraction).

### 3 PROPOSED METHOD

Figure 2 displays the proposed approach: the general idea is to aggregate the weights of two different neural networks trained on two different datasets (sharing the same underlying task). The ultimate goal is to obtain  $n$  individual networks  $N_i$  such that their composition  $N_1 \oplus N_2 \oplus \dots \oplus N_i \approx N^*$  (note that the operator  $\oplus$  refers to a generic network aggregation operator, which details will be provided in Section 3.3). Below, we will refer to an individual network  $N_i$  as *dataset-specific network*, whereas  $N^*$  will be referred to as *aggregated network*. As anticipated in Section 1, all these networks used as feature extractors share the same task network.

#### 3.1 Task Network

As shown in Figure 2, the task network is shared across the aggregated and dataset-specific networks. This network is parametrised by the set of weights  $\theta$ : the output provided by all of the  $N_i$  and  $N^*$  is used as input of the task network, and its output is the prediction (i.e., softmax activation in case of classification).

The task network is trained with a task-specific loss function as  $\mathcal{L}_T(z, y; w)$ , that takes the training data  $z$  (in form of representation) and target variables  $y$  as inputs, and it is parametrised by a set of weights  $w$  (that includes  $\theta$  and the parameters of the feature ex-

tractors). We used cross-entropy loss for  $\mathcal{L}_T$  in this paper. For other tasks (i.e., regression), a different loss function may be used (e.g., mean squared error).

### 3.2 Dataset-Specific Network

Each dataset-specific network  $N_i$  acts as a feature extractor for the dataset it is trained on. We opted to use a VGG-like network (Simonyan and Zisserman, 2015) in our experiments. In particular, following the same architectural tweaks as others (Loh et al., 2021), we used a VGG-16 network with Group Normalisation (Wu and He, 2018).<sup>1</sup>

Each network  $N_i$  is parametrised by the set of weights  $\phi_i$  that is trained via standard supervised learning. In fact, each  $N_i$  is trained with a different label dataset; all of those datasets maintain the same underlying task. This means that each dataset  $\mathcal{D}_i$  contains a set of input data  $\mathcal{X}_i$ , such that  $x^{(i)} \in \mathcal{X}_i$ , and a set of target values  $\mathcal{Y}_i$ , such that  $y^{(i)} \in A$  (the set  $A$  is a generic set defined by the task, i.e. if the task is classification,  $A$  will contain all the possible classes).

For each of these networks, a specific loss function is used during training:

$$\mathcal{L}_{T_i}(x^{(i)}, y^{(i)}; \phi_i \cup \theta) = \mathcal{L}_T(N_i(x^{(i)}), y^{(i)}; \phi_i \cup \theta). \quad (1)$$

### 3.3 Aggregated Network

The aggregated network is similar to the dataset-specific networks: it shares the same architecture but not the weights. In fact, this network is parametrised by the set of weights  $\Phi^*$ .

The aggregated network is trained such that its weights can be expressed as a sum of the dataset-specific networks. To achieve this, we proposed the *aggregation* regulariser. Each network  $N_i$  is made of  $L_i$  layers, and each layer  $\ell$  is parametrised by some weights  $W_i^\ell \in \phi_i$ .<sup>2</sup> Similarly, the aggregated network  $N^*$  includes several layers  $L$ , each of those is parametrised by  $W^\ell$ . However, it is important to emphasise that all the feature extractors share the same architecture, that is,  $L = L_1 = L_2 = \dots = L_n$ .

During training, we want that:

$$W^\ell = W_1^\ell \oplus W_2^\ell \oplus \dots \oplus W_n^\ell, \quad (2)$$

<sup>1</sup>We also tried either Batch Normalisation (Ioffe and Szegedy, 2015) or no normalisation with no success.

<sup>2</sup>Some types of layers, for example, convolutional layers, can be parametrised by multiple weights, such as kernels and bias. For the sake of clarity, we incorporate these weights within  $W_i^\ell$ .

i.e., the weights at layer  $\ell$  in the aggregated network should be equal to the aggregation of the corresponding layer weights in the dataset-specific networks. We reformulate this constrain as a regulariser during training. Assuming that  $\ominus$  is the inverse operator of  $\oplus$ , the aggregation regulariser is then expressed as:

$$\mathcal{L}_{agg}(\Phi) = \sum_{\ell=1}^L W^\ell \ominus [W_1^\ell \oplus W_2^\ell \oplus \dots \oplus W_n^\ell], \quad (3)$$

where  $\Phi = \phi^* \cup (\bigcup_{i=1}^n \phi_i)$  is the set of all the weights in all the feature extractors.<sup>3</sup> Although the networks learn a non-linear mapping w.r.t the task, the aggregation regulariser in Equation (3) learns the weights  $\Phi$  such that the network aggregation can be performed with a linear operation (assuming that  $\oplus$  is linear).

The aggregated network takes all the input data that are used for each dataset-specific network  $\mathcal{D} = \bigcup_{i=1}^n \mathcal{D}_i$  and it is trained in a supervised manner w.r.t. the task  $\mathcal{L}_T$  as follows:

$$\mathcal{L}_{T^*}(x, y; \phi^* \cup \theta) = \mathcal{L}_T(N^*(x), y, \phi^* \cup \theta), \quad (4)$$

where  $(x, y) \in \mathcal{D}$ , i.e. inputs and labels are taken from all the datasets used to train the dataset-specific networks.

### 3.4 Objective Function

As shown in Figure 2, the objective functions used to train our model is the following:

$$J(x, y; \Theta) = \mathcal{L}_{task} + \mathcal{L}_{agg}, \quad (5)$$

where  $\Theta = \Phi \cup \theta$  is the set of all the parameters in the network. the loss function  $\mathcal{L}_{task}$  is given as the sum of all the task-specific loss functions expressed in Equation (1) and Equation (4):

$$\mathcal{L}_{task}(x, y; \Theta) = \mathcal{L}_{T^*}(x, y; \phi^* \cup \theta) + \sum_{i=1}^n \mathcal{L}_{T_i}(x^{(i)}, y^{(i)}; \phi_i \cup \theta).$$

After training, there is no guarantee that the regulariser in Equation (3) ensures that Equation (2) is satisfied. However, the optimisation of Equation (5) will make sure that  $W^\ell \approx W_1^\ell \oplus W_2^\ell \oplus \dots \oplus W_n^\ell$ . Hence, we can retrieve  $N^*$  by aggregating all the weights trained for each dataset-specific network  $N_i$  – we will refer to this model as  $\hat{N}^*$ , such that  $\hat{N}^* \approx N^*$ .

<sup>3</sup>Similarly as in (Geyer et al., 2019), we only aggregate weights of convolutional layers.

We can selectively forget one of the datasets from  $\hat{N}^*$  by applying a simple arithmetic operation. Assuming that we wanted to remove the  $k$ -th dataset, we can perform the operation  $\hat{N}^* \ominus N_k$ , without any further training (or adaptation) steps.

### 3.5 Implementation Details

We set the number of datasets (and thus the number of dataset-specific networks) to  $n = 2$ . This allowed us to demonstrate whether our approach works and to set a baseline. We set the number of groups for Group Normalisation to 32. We used as aggregation operator the sum for the following reasons: (i) it has an inverse – the subtraction; (ii) it is differentiable. We set as task-specific loss function the cross-entropy loss as we train the whole network for a classification task. Stochastic Gradient Descent (SGD) was used as optimiser for training with a learning rate  $\eta = 0.01$ . The baseline was trained for 20 epochs, while training of our proposed method lasted 200 epochs. We implemented our approach in PyTorch (Paszke et al., 2019) on Google Colaboratory.

## 4 EXPERIMENTAL RESULTS

**Dataset:** we used MNIST (LeCun et al., 2010) as  $\mathcal{D}_1$  and SVHN format 2 (Netzer et al., 2011) as  $\mathcal{D}_2$ . MNIST contains 60,000 binary images of size  $28 \times 28$  for training and 10,000 images for testing. SVHN contains 73,257 colour images of size  $32 \times 32$  for training and 26,032 for testing. We chose these two datasets for the following reasons: (i) they are designed for the same classification (10-class) task; (ii) data are drawn from different distributions.

**Preprocessing:** In order to use the same architecture for both datasets, MNIST images were rescaled to  $32 \times 32$ . We converted the SVHN images in grayscale. As for data augmentation, we performed random horizontal flips with a probability of 50%.

**Baseline:** We compared our method with a standard VGG-16 network with Batch Normalisation (Ioffe and Szegedy, 2015). We ran the following baseline experimentation:

1. trained it only on MNIST – following the notation adopted in this paper, we called this trained network  $N_1$ ;
2. trained it only on SVHN – named  $N_2$ ;
3. trained it on both ( $N^*$ );
4. The weights trained on  $N_1$  and  $N_2$  were taken to perform  $N_1 \oplus N_2$  at test time.

Experimental results are shown in Table 1.

	Trained on		Tested on		
	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_1 \cup \mathcal{D}_2$
<i>Baseline</i>					
$N_1$	✓	–	99.04%	8.67%	33.75%
$N_2$	–	✓	57.63%	90.29%	80.78%
$N^*$	✓	✓	98.73%	90.58%	92.85%
$N_1 \oplus N_2$	✓	✓	9.75%	7.59%	8.19%
<i>Proposed Method</i>					
$N_1$	✓	–	98.90%	41.45%	57.35%
$N_2$	–	✓	45.30%	92.41%	79.31%
$N^*$	✓	✓	98.40%	86.47%	89.77%
$N_1 \oplus N_2$	✓	✓	96.41%	68.03%	75.88%

Table 1: Testing performance of the proposed method compared to the baseline performance.  $\mathcal{D}_1$  indicates MNIST;  $\mathcal{D}_2$  indicates SVHN. The models obtained via aggregation (i.e.,  $N_1 \oplus N_2$ ) are obtained at test time by aggregating the weights of the networks.

### 4.1 Discussion

Our purpose is to demonstrate that the performance of our aggregated network  $N_1 \oplus N_2$  is better than the baseline. Overall, our method achieves comparable performance with the baseline for individual tasks (i.e.,  $N_1$  and  $N_2$ ). However, there is a slight loss in performance in  $N^*$ , that is, the network trained on both MNIST and SVHN, with our method. The baseline achieves approx. 92% accuracy, whereas  $N^*$  trained with our method achieves approx. 89%.

Although this minor performance reduction, our method achieves high performance with test-time weight aggregation. After  $N_1$  and  $N_2$  are trained with the baseline and our method, weights are aggregated by applying the  $\oplus$  operator. Table 1 clearly shows that our training procedure outperforms the baseline (8% vs 75% testing accuracy). This demonstrates that a traditional training of two networks cannot be aggregated, leading to catastrophic forgetting on both datasets. Our training approach with Equation (3) enables the networks to explicitly learn an aggregation operation that can be reproduced at test time.

Ideally, the performance of  $N_1 \oplus N_2$  should be as close as possible to  $N^*$ . As shown in the last two lines of Table 1, there is an approximate loss of 14% accuracy. We hypothesise several reasons for this gap in accuracy: (i) our method may require more training time; (ii) Group Normalisation may be having an impact at test time (as specified in Section 3.3, we only aggregate the weights of convolutional layers); (iii) use of Weight Standardisation (WS) can improve performance (Qiao et al., 2019; Loh et al., 2021).

In relation to training time, we plot the training and validation accuracies and losses of our method in Figure 3. Overall, it can be noted that 50 epochs

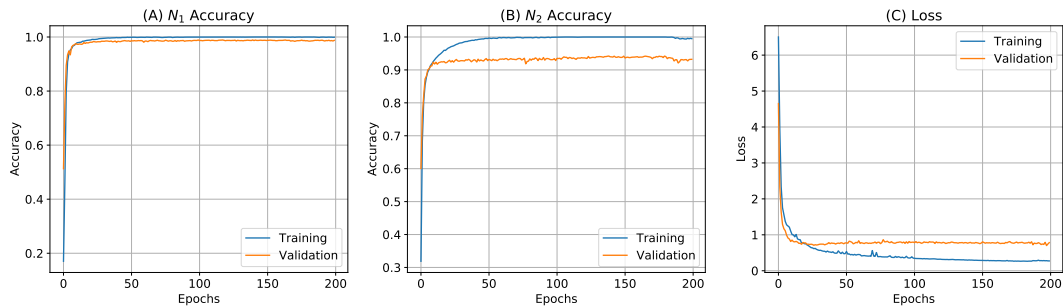


Figure 3: Training and validation accuracies and losses of our proposed method. (A)  $N_1$  training and validation accuracies; (B)  $N_2$  training and validation accuracies; (C) Total training and validation loss.

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_1 \cup \mathcal{D}_2$
<i>Commutativity</i>			
$N_1 \oplus N_2$	96.41%	68.03%	75.88%
$N_2 \oplus N_1$	96.41%	68.03%	75.88%
<i>Selective forgetting</i>			
$N_1$	98.90%	41.45%	57.35%
$(N_1 \oplus N_2) \ominus N_2$	98.55%	47.90%	61.92%
$(N_2 \oplus N_1) \ominus N_2$	98.55%	47.90%	61.92%
$N_2$	45.30%	92.41%	79.31%
$(N_1 \oplus N_2) \ominus N_1$	34.67%	90.88%	75.28%
$(N_2 \oplus N_1) \ominus N_1$	34.67%	90.88%	75.28%

Table 2: Commutativity and Selective forgetting testing results. The training is performed in both  $\mathcal{D}_1$  and  $\mathcal{D}_2$ . The two datasets are the same as in Table 1. Highlighted rows are copied from Table 1 to ease comparison.

should be enough to accommodate for both datasets. However, we found experimentally that more training time results in higher performances in our aggregated model. We hypothesise that the optimisation of Equation (3) could require more time to learn better aggregable models. In relation to Weight Standardisation, we posit that an increase in performance may be achievable if the aggregation function is changed to the mean instead of the sum. Otherwise, the aggregated weights will no longer be zero-centred.

## 4.2 Commutativity

Here, we want to demonstrate whether our method is commutative: does the performance of  $N_1 \oplus N_2$  match the performance of  $N_2 \oplus N_1$ ? Theoretically, commutativity should be strictly related to the  $\oplus$  operator. However, this is not exactly guaranteed in our framework because we only aggregate the convolution weights in the networks  $N_i$  (see Section 3.3). Group Normalisation layers also include learnable parameters that are not included during network aggregation. To confirm whether our method is commu-

tative, we also performed the operation  $N_2 \oplus N_1$  and results are reported in Table 2. It can be seen that the performances in both scenarios are the same. Hence, we can conclude that our approach is commutative.

## 4.3 Selective Forgetting

For the same reasons as in Section 4.2, we also experimentally show whether selective forgetting is possible with our method. Because of Group Normalisation layers,  $(N_1 \oplus N_2) \ominus N_1 \approx N_2$ , i.e., by removing the contribution of  $N_1$ , we do not exactly obtain  $N_2$  (and vice versa). Therefore, we asked the following question: does the network obtained by  $(N_1 \oplus N_2) \ominus N_1$  perform as good as  $N_2$ ? Table 2 shows the experimental results of selective forgetting.

**Forgetting SVHN:** We removed the weights of  $N_2$  from the aggregated networks (we considered  $N_1 \oplus N_2$  and  $N_2 \oplus N_1$  as aggregated networks), and we provided the SVHN testing set to this new network. The testing accuracy is 47.90%, against the 41.45% of  $N_1$ . Therefore, the resulting network does forget about SVHN, although not completely (there is approx +6% increase of performance).

**Forgetting MNIST:** A similar experiment was performed by removing the weights of  $N_1$  from the aggregated networks. Differently than before, the testing accuracy of the resulting network is 34.67%, compared with 45.30%. This experimentally demonstrates that our method has completely forgotten the information learnt from the MNIST dataset.

**Retained information:** In the two previous experiments, it was shown that the resulting network had forgotten information from either of the two datasets. However, we must check whether the network can still perform well in the other dataset. Overall, the performance on MNIST dataset is very similar (from 98.90% to 98.55%), whereas in the case of SVHN there is approx 2% loss of performance (from 92.41% to 90.88%) – although the overall testing error is

above 90%. This also demonstrates that the proposed method retains information from both tasks with a loss in performance up to 2%.

## 5 CONCLUSIONS

In this paper, we proposed a novel and simple proof-of-concept transfer learning approach that inherently allows for selective forgetting. Our training method enables for network aggregation at test time, i.e. the weights of two networks (trained on two different datasets) are aggregated together, such that the resulting network can work on both datasets without any further training/adaptation step.

We achieve that by introducing an *aggregation regulariser*, that enables the networks to also learn the aggregation operation in an end-to-end training framework. We used the sum as aggregation operator, as it is invertible and differentiable. VGG-like architectures were used as feature extractors, using Group Normalisation in lieu of Batch Normalisation.

Our experimental results demonstrated that the proposed approach allows for test-time transfer learning without any further training steps. Furthermore, we showed that our training procedure is commutative: the aggregated network  $N_1 \oplus N_2$  obtains the same performance of  $N_2 \oplus N_1$ . Moreover, we demonstrated that our method allows for selective forgetting (at the cost of up to 2% testing performance).

The proposed method has some limitations: (i) it requires that all the networks involved in the training share the same architecture; (ii) the selective forgetting does not allow to forget a subset of the dataset; (iii) we evaluated it on just two benchmark datasets (although the proposed framework can easily accommodate for multiple datasets). As future work, we will generalise our approach exploring the training with  $N_i$  deep neural networks, for  $i = 1, \dots, n$ , with  $n$  being the number of datasets, in a federated learning scenario.

## ACKNOWLEDGEMENTS

This work was funded by the Edinburgh Napier University internally funded project “Li.Ne.Co.”

## REFERENCES

Chen, X., Wang, S., Fu, B., Long, M., and Wang, J. (2019). Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-

Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 1906–1916. Curran Associates, Inc.

Chen, Y., Qin, X., Wang, J., Yu, C., and Gao, W. (2020). Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35(4):83–93.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Geyer, R., Corinzia, L., and Wegmayr, V. (2019). Transfer learning by adaptive merging of multiple models. In Cardoso, M. J., Feragen, A., Glocker, B., Konukoglu, E., Oguz, I., Unal, G., and Vercauteren, T., editors, *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 185–196. PMLR.

Golatkar, A., Achille, A., Ravichandran, A., Polito, M., and Soatto, S. (2021). Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 792–801.

Golatkar, A., Achille, A., and Soatto, S. (2020). Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. (2013). An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Han, X., Huang, Z., An, B., and Bai, J. (2021). Adaptive transfer learning on graph neural networks.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR.

LeCun, Y., Cortes, C., and Burges, C. (2010). Mnist handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2.

Lee, S.-W., Kim, J.-H., Jun, J., Ha, J.-W., and Zhang, B.-T. (2017). Overcoming catastrophic forgetting by incremental moment matching. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.

Litrico, M., Battiato, S., Tsafaris, S. A., and Giuffrida, M. V. (2021). Semi-supervised domain adaptation for holistic counting under label gap. *Journal of Imaging*, 7(10).

Loh, A., Karthikesalingam, A., Mustafa, B., Freyberg, J., Houlsby, N., MacWilliams, P., Natarajan, V., Wilson, M., McKinney, S. M., Sieniek, M., Winkens, J., Liu, Y., Bui, P., Prabhakara, S., and Telang, U. (2021). Supervised transfer learning at scale for medical imaging.

- Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.
- Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, page pages 1717–1724.
- Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22.
- Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Qiao, S., Wang, H., Liu, C., Shen, W., and Yuille, A. (2019). Micro-batch training with batch-channel normalization and weight standardization.
- Ramapuram, J., Gregorova, M., and Kalousis, A. (2020). Lifelong generative modeling. *Neurocomputing*, 404:381–400.
- Rao, D., Visin, F., Rusu, A. A., Teh, Y. W., Pascanu, R., and Hadsell, R. (2019). Continual unsupervised representation learning. *arXiv preprint arXiv:1910.14481*.
- Reyes, A. K., Caicedo, J. C., and Camargo, J. E. (2015). Fine-tuning deep convolutional networks for plant recognition. *CLEF (Working Notes)*, 1391:467–475.
- Rosenstein, M. T., Marx, Z., Kaelbling, L. P., and Dietterich, T. G. (2005). To transfer or not to transfer. In *In NIPS'05 Workshop, Inductive Transfer: 10 Years Later*.
- Schölkopf, B., Platt, J., and Hofmann, T. (2007). *Correcting Sample Selection Bias by Unlabeled Data*, pages 601–608.
- Shu, Y., Kou, Z., Cao, Z., Wang, J., and Long, M. (2021). Zoo-tuning: Adaptive transfer from a zoo of models. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9626–9637. PMLR.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ICLR 2015*.
- Tommasi, T., Orabona, F., and Caputo, B. (2010). Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *roceedings of IEEE Computer Vision and Pattern Recognition Conference*.
- Torrey, L. and Shavlik, J. (2010). Transfer learning. *Handbook of Research on Machine Learning Applications and Trends*.
- Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wang, Z., Dai, Z., Póczos, B., and Carbonell, J. (2018). Characterizing and avoiding negative transfer.
- Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1):1–40.
- Wu, Y. and He, K. (2018). Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.
- Ye, F. and Bors, A. G. (2020). Learning latent representations across multiple data domains using lifelong vae-gan. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, pages 777–795, Cham. Springer International Publishing.
- Zhao, S., Yue, X., Zhang, S., Li, B., Zhao, H., Wu, B., Krishna, R., Gonzalez, J. E., Sangiovanni-Vincentelli, A. L., Seshia, S. A., et al. (2020). A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*.
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.