Towards Exemplar-Free Continual Learning in Vision Transformers: an Account of Attention, Functional and Weight Regularization

Francesco Pelosin*
Ca’ Foscari University
Venice, Italy
francesco.pelosin@unive.it

Saurav Jha*
University of New South Wales
Sydney, Australia
saurav.jha@unsw.edu.au

Andrea Torsello
Ca’ Foscari University
Venice, Italy
torsello@unive.it

Bogdan Raducanu
Computer Vision Center
Barcelona, Spain
bogdan@cvc.uab.es

Joost van de Weijer
Computer Vision Center
Barcelona, Spain
joost@cvc.uab.es

Abstract

In this paper, we investigate the continual learning of Vision Transformers (ViTs) for the challenging exemplar-free scenario, with special focus on how to efficiently distill the knowledge of its crucial self-attention mechanism (SAM). Our work takes an initial step towards a surgical investigation of SAM for designing coherent continual learning methods in ViTs. We first carry out an evaluation of established continual learning regularization techniques. We then examine the effect of regularization when applied to two key enablers of SAM: (a) the contextualized embedding layers, for their ability to capture well-scaled representations with respect to the values, and (b) the prescaled attention maps, for carrying value-independent global contextual information. We depict the perks of each distilling strategy on two image recognition benchmarks (CIFAR100 and ImageNet-32) – while (a) leads to a better overall accuracy, (b) helps enhance the rigidity by maintaining competitive performances. Furthermore, we identify the limitation imposed by the symmetric nature of regularization losses. To alleviate this, we propose an asymmetric variant and apply it to the pooled output distillation (POD) loss adapted for ViTs. Our experiments confirm that introducing asymmetry to POD boosts its plasticity while retaining stability across (a) and (b). Moreover, we acknowledge low forgetting measures for all the compared methods, indicating that ViTs might be naturally inclined continual learners.¹

¹Equal Contribution

¹Code will be made available at https://github.com/srvCodes/continual_learning_with_vit.

1. Introduction

Transformers have shown excellent results for a wide range of language tasks [6, 40] over the course of the last couple of years. Influenced by their initial results, Dosovitskiy et al. [13] proposed Vision Transformers (ViTs) as the first firm yet competitive application of transformers within the computer vision community.² ViTs’ applications have since spanned a range of vision tasks, including, but not limited to image classification [44], object recognition [25], and image segmentation [46]. The single most essential element of their architecture remains the self-attention mechanism (SAM) that allows the learning of long-range interdependence between the elements of a sequence (or patches of an image). Another feature vital to their performance is the way they are pretrained in an often unsupervised or self-supervised manner over a large amount of data. This is then followed by the finetuning stage where they are adapted to a downstream task [12].

For ViTs to be able to operate in real-world scenarios, they must exploit streaming data, i.e., sequential availability of training data for each task.³ Storage limitations or privacy constraints further imply the restrictions on the storage of data from previous tasks. Task-incremental continual learning (CL) seeks to find solutions to such constraints by alleviating the event of catastrophic forgetting - a phenomena where the network has a dramatic drop in performance on data from previous tasks. Several solutions have been proposed to address forgetting, including regularization [1, 21, 39, 53], data replay [2, 8, 27] and parameter isolation [3, 20, 31, 41]. Nowadays most works on CL study recurrent [9, 43] and convolutional neural networks

²By firmness, we refer to the non-reliance on convolutional operations.

³A task may encompass training data of one or more classes.
However, little has been done to investigate different CL settings in the domain of ViTs. We, therefore, mark the first step for the domain by considering the further restrictive setting of exemplar-free CL with a zero overhead of storing any data from previous tasks. We consider this restriction for its real-world aptness to scenarios involving privacy regulations and/or data security considerations.

Given that regularization-based methods form one of the main techniques for exemplar-free CL, we consider an in-depth analysis of these for ViTs. Regularization-based techniques are mainly organized along two branches: weight regularization methods (such as EWC [19], SI [53], MAS [1]) and functional regularization methods (such as LwF [23], PODNET [14]). As discussed above, the architectural novelty of transformers lies in the SAM building a representation of a sequence by exhaustively learning relations among query-key pairs of its elements [45]. We show that for ViTs (and subsequently, all other architectures leveraging SAM), this property allows for a third form of regularization, which we coin as Attention Regularization (see Figure 1). We ground our idea in the hypothesis that when learning new tasks, the attention of the new model should still remain in the neighborhood of the attention of the previous model. As another contribution, we question the temporal symmetry currently applied to regularization losses; referring to the fact that they penalize the forgetting of previous knowledge and the acquiring of new knowledge equally (see Figure 2). With the aim of countering forgetting while mitigating the loss of plasticity, we then propose an asymmetric regularization loss that penalizes the loss of previous knowledge but not the acquiring of new knowledge. We index the major contributions of our work below:

- We are the first to investigate continual learning in vision transformers in the more challenging exemplar-free setting. We perform a full analysis of regularization techniques to counter catastrophic forgetting.
- Given the distinct role of self-attention in modeling short and long-range dependencies [49], we propose distilling the attention-level matrices of ViTs. Our findings show that such distillation offers accuracy scores on par with that of their more common functional counterpart while offering superior plasticity and forgetting. Motivated by the work of Douillard et al. [14], we pool spatiality-induced attention distillation across our network layers.
- We propose an asymmetric variant of functional and attention regularization which prevents forgetting while maintaining higher plasticity. Through our extensive experiments, we show that the proposed asymmetric loss surpasses its symmetric variant across a range of task incremental settings.

2. Related Works

Continual learning has been gaining contributions from the deep learning research community during the last few years [11, 32]. In the following, we list the most prominent ones:

- Weight-based: these methods operate in the parameter space of the model through gradient updates. Elastic Weight Consolidation (EWC) [19] and Synaptic Intelligence (SI) [54] are two widely used methods in this family with the former being probably, the most well-known. EWC uses Fisher information [36] to identify the parameters important to individual tasks and penalizes their updates to preserve knowledge from older tasks. SI makes the neurons accumulate and exploit old task-specific knowledge to contrast forgetting.
- Functional-based: these methods rely upon trading the plasticity for stability by training either the current (new) model on older data or vice-versa. Learning Without Forgetting (LwF) [23] remains among the most widely known approaches in this family. It employs Knowledge Distillation [17] upon the logits of the network.
- Parameter Isolation-based: also known as architectural approaches, these methods tackle CF through a dynamic expansion of the network’s parameters as the number of tasks grow. Among the well known methods in this family remain Progressive Neural Network
(PNN) [42] followed by Dynamically Expandable Network (DEN) [50] and Reinforced Continual Learning (RCL) [48].

The majority of the aforementioned works target CL in CNNs mainly due to their inductive bias allowing them to solve almost all problems that involve visual data. This can also be seen in several reviews [4, 5, 11, 29, 30, 35] reporting few approaches that consider architectures besides CNNs, despite the attempts to investigate CL in RNNs [9, 43].

Only recently have some works analyzed catastrophic forgetting in transformers. Among the earliest to do so remains that of Li et al. [22] proposing the continual learning with transformers (COLT) framework for object detection in autonomous driving scenarios. Using the Swin Transformer [26] as the backbone for a CascadeRCNN detector, the authors show that the extracted features generalize better to unseen domains hence achieving lesser forgetting rates compared to ResNet50 and ResNet101 [16] backbones. In case of ViTs, Yu et al. [52] show that their vanilla counterparts are more prone to forgetting when trained from scratch. Alongside heavy augmentations, they employ a set of techniques to mitigate forgetting: (a) knowledge distillation, (b) balanced re-training of the head on exemplars (inspired by LUCIR [18]), and (c) prepending a convolutional stem to improve low-level feature extraction of ViTs.

In their work studying the impact of model architectures in CL, Mirzadeh et al. [33] also experiment with ViTs in brief (with the rest of the work focusing mainly on CNNs). While they vary the number of attention heads of ViTs to show that this has little effect on the accuracy and forgetting scores, they further conclude that ViTs do offer more robustness to forgetting arising from distributional shifts when compared with their CNN-based counterparts with an equivalent number of parameters. The conclusion remains in line with previous works [37]. Finally, Douillard et al. [15] attempt to overcome forgetting in ViTs through a parameter-isolation approach which dynamically expands the tokens processed by the last layer. For each task, they learn a new task-specific token per head. They then couple such approach with the usage of exemplars and knowledge distillation on backbone features. It is worth noting that these works rely either on pretrained feature extractors [22] or rehearsal [15, 52] to defy forgetting. Thus the challenging scenario of exemplar-free CL in ViTs remains unmarked.

3. Methodology

We start by shortly describing the two main existing regularization techniques for continual learning. We then propose attention regularization as an alternative approach tailored for ViTs. Lastly, we put forward an adaptation for functional and attention regularization which is designed to elevate plasticity while retaining stability levels.

3.1. Functional and Weight Regularization

**Functional Regularization:** We include LwF [23] in this component since it constitutes one of the most prominent, and perhaps the most widely used regularization method acting on data. The appealing property of LwF lies in the fact it is exemplar-free, i.e., it uses only the data of the current task and maintains only the model at task \( t-1 \) to exploit Knowledge Distillation [17]. Formally, LwF is defined as:

\[
L_{LwF}(\theta) = \xi_o L_{KD} \left( Y_o, \hat{Y}_o \right)
\]  

(1)

where \( L_{KD} \) is the knowledge distillation loss incorporated to impose stability on the outputs, \( Y_o \) the predictions on the current task data from the old model and \( \hat{Y}_o \) the ground truth of such data. \( \xi_o \) remains the temperature annealing factor for softmax logits. To ensure better plasticity, \( L_{KD} \) is often combined with the standard cross entropy loss \( L_{CE} \) calculated upon the new task examples.

Some works have investigated the usage of functional regularization at intermediate layers of the network [14, 24, 51]. This can be generalized to ViTs by applying a regularization on the contextualized embedded outputs (see purple block in Figure 1). In this case, the regularization loss could be defined in terms of functional distillation (FD):

\[
L_{FD} = \left\| \sum_{w=1}^{W} \sum_{h=1}^{H} z_{l,w,h}^{t-1} - \sum_{w=1}^{W} \sum_{h=1}^{H} z_{l,w,h}^{t} \right\|_2
\]  

(2)

where \( \left\| \cdot \right\| \) denotes the L2-norm, \( z_{l}^{(t-1)} \) refers to the \( l \)-th layer output of the model trained at task \( t-1 \), and \( W \) and \( H \) are the width and height of the SAM outputs, respectively.

**Weight Regularization:** These methods encourage the network to adapt to the current task data mainly by using those parameters of the network that are not considered important for previous tasks. As representative method we select EWC [19]. EWC exploits second-order information to estimate the importance of parameters for the current task. The importance is approximated by the diagonal of the Fisher Information Matrix \( F \):

\[
L_{EWC}(\theta) = L_X(\theta) + \sum_j \frac{\lambda}{2} F_j (\theta_j - \theta_{Y,j})^2
\]  

(3)

where \( L_X(\theta) \) is the loss for task \( X \), \( \lambda \) the regularization strength, and \( \theta_{Y,j} \) the optimal value of \( j^{th} \) parameter after having learned task \( Y \).

3.2. Attention Regularization

**Self-Attention Mechanism:** The self-attention mechanism (SAM) [45] forms the core of Transformer-based models and can be defined as:
where $Q$, $K$, and $V$ are respectively the projections of the Query, Key, and Values of the $\mathbb{R}^{d_e}$ input embeddings while $z$ constitutes the new contextualized embeddings. Our novel attention-based regularization intervenes prior to the computation of the softmax operation of the standard self-attention mechanism as illustrated in Figure 1.

In particular, given a ViT model at incremental step $t$ and an SAM head $k$ of layer $l$, we define the prescaled attention matrix $A^t_{k l}$ prior to the softmax operation as:

$$A^t_{k l} = \frac{QK^T}{\sqrt{d_e}}$$ (5)

The attention matrix for time step $(t - 1)$ can be similarly computed as $A^{t-1}_{k l}$. We employ this predecessor in the calculation of knowledge distillation in what follows.

**Pooled Attention Distillation:** Functional approaches leverage network’s submodules typically to apply knowledge distillation [17]. When the regularization takes place in intermediate layers, the model can experience excessive stability, therefore loosing in plasticity abilities [14, 24, 51]. Amongst these methods, PODNet [14] clearly identifies the problem of excessive stability. We devise a regularization approach which instead of regularizing functional submodules targets attention maps, the core mechanisms of SAMs.

More formally, given the attention maps at steps $t$ and $(t-1)$, we define $L_{\text{PAD}}\left(A^{t-1}_{k l}, A^t_{k l}\right)$ [14] to be:

$$L_{\text{PAD-width}}\left(A^{t-1}_{k l}, A^t_{k l}\right) + L_{\text{PAD-height}}\left(A^{t-1}_{k l}, A^t_{k l}\right)$$ (6)

where $L_{\text{PAD-width}}\left(A^{t-1}_{k l}, A^t_{k l}\right) = \sum_{h=1}^{H} D_W \left(A^{t-1}_{k l}, A^t_{k l}\right)$,

$$L_{\text{PAD-height}}\left(A^{t-1}_{k l}, A^t_{k l}\right) = \sum_{w=1}^{W} D_H \left(A^{t-1}_{k l}, A^t_{k l}\right)$$

$$D_X \left(A^{t-1}_{k l}, A^t_{k l}\right) = \left\| \sum_{x=1}^{X} A^{t-1}_{x w, h} - \sum_{x=1}^{X} A^t_{x w, h} \right\|^2$$ (8)

where, $W$ and $H$ indicate the width and height of the attention maps, and $D_X(a, b)$ is the sum total of the distance measure between maps $a$ and $b$ along the arbitrary dimension $X$. As shown in equation 8, the standard $L_{\text{PAD}}$ uses the difference operator as the choice for $D$. We point out the limitation of such symmetric $D$ and introduce, in the next section, the notion of asymmetry into our distance measure.

As previously mentioned, Douillard el al. [14] propose the pooled outputs distillation PODNet loss which leverages the symmetric Euclidean distance between the L2-normalized outputs of the convolutional layers of models at $t$ and $(t-1)$ after pooling them along specific dimension(s). They achieve the best results upon combining the pooling along the spatial width and height axes which they term as the POD-spatial loss. Given the generic correspondence among the various pooling variants in their paper, our work is particularly influenced by POD-spatial as we pool attention maps of ViTs along two dimensions. In fact, throughout the experiments, we analyze this formulation when applied to the contextualized embeddings $z$ resulting from a SAM operation. We would like to highlight that PAD differs from PODNet in two important factors: it is applied to the attention and not directly on the layer outputs, and its marginalization is not on the spatial dimensions due to the fact that $A$ no longer encodes spatiality.

![Figure 2](image-url)  
Visual illustration of the asymmetric loss with two generated attention maps (a) and (b) while training on task 2. In scenario (a), when previous knowledge is lost, both the symmetric and asymmetric regularization work correctly. However, in scenario (b), acquiring new knowledge is penalized by the symmetric loss but not by the asymmetric loss.
3.3. Asymmetric Regularization

The proposed attention regularization prevents forgetting of previous tasks by ensuring that the old attention maps are retained while the model learns to attend to new regions over tasks. However, the symmetric nature of $D_X$ (with respect to the two attention maps) means that any differences between the older and the newly learned attention maps lead to increased loss values (see Equation 7). We agree that penalizing a loss in attention with respect to previous knowledge is crucial in addressing forgetting. However, also penalizing a gain in attention for newly learned knowledge is undesirable and may actually hurt the performance over subsequently learned tasks. In other words, punishing additional attention can be counterproductive. As a result, we classify our final PAD attention loss as asymmetric.

Based on our choice for $D_X$ from equations 8 and 9, we classify our final PAD attention loss as asymmetric $\mathcal{L}_{\text{PAD-asym-att}}$ or symmetric $\mathcal{L}_{\text{PAD-sym-att}}$, respectively. Each of these losses is computed separately for each of the SAM head $k$ and model layer $lL$. The final asymmetric variant for attention regularization can be stated as:

$$\mathcal{L}_{\text{PAD-asym-att}}(A_{^t_k}^{l-1}, A_{^t_k}^l) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{\text{PAD}}(A_{^t_k}^{l-1}, A_{^t_k}^l)$$

(10)

where, $K$ and $L$ denote the total number of heads per layer and the total number of layers in the model, respectively. We now show that $\mathcal{L}_{\text{PAD-asym-att}}$ can accommodate functional distillation (FD) when applied to the contextualized global embeddings. To do so, we first replace the attention maps $A_{^t_k}$ in Equation 6 by the contextualized vectors $z_{^t_k}$ from Equation 4. We denote the resulting functional counterpart of $\mathcal{L}_{PAD}$ by $\mathcal{L}_{FD}$. The final asymmetric variant for functional regularization can therefore be defined as:

$$\mathcal{L}_{\text{FD-asym-func}}(z_{^t_k}^{l-1}, z_{^t_k}^l) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{\text{FD}}(z_{^t_k}^{l-1}, z_{^t_k}^l)$$

(11)

Overall loss: We augment the PAD losses from equations 10 and 11 with knowledge distillation loss $\mathcal{L}_{kdw}$ and standard cross entropy loss $\mathcal{L}_{CE}$. The overall loss term thus takes the form:

$$\mathcal{L} = \mu \mathcal{L}_{\text{asym}} + \lambda \mathcal{L}_{\text{LwF}} + \mathcal{L}_{\text{CE}}$$

(12)

where $\mu, \lambda \in [0, 1]$ are two hyperparameters regulating the respective contributions. Note that when $\mu = 0$, $\mathcal{L}$ degenerates to baseline finetuning for $\lambda = 0$ and to LwF for $\lambda = 1$.

Stability-Plasticity Curves: Several measures have been proposed in the CL literature to assess the performance of an incremental learner. Besides the standard incremental accuracy, Lopez-Paz et al. [27] introduce the notion of Backward Transfer (BWT) and Forward Transfer (FWT). BWT measures the ability of a system to propagate knowledge to past tasks, while FWT assesses the ability to generalize to future tasks. The CL community, however, still lacks consensus on a specific definition of the stability-plasticity dilemma. An elemental formulation for such quantification is thus desirable for allowing us to better grasp the trading capabilities of an incremental learner between acquiring new knowledge and discarding previous concepts. To this end, we introduce stability-plasticity curves computed using task accuracy matrices.

A task accuracy matrix $M$ for an incremental learning setting composed of $T$ tasks is defined to be a $[0, 1]^{T \times T}$ matrix, whose entries are the accuracies computed at each incremental step.\(^4\) For instance, $M_{i,j}$ would constitute the test accuracy of task $j$ when the system is learning task $i$. Subsequently, the diagonal entries $M_{i,i}$ give us the accuracies at the respective current tasks while the entries below the diagonal, i.e., $j < i$, give the performance of the model on past tasks. A visual depiction can be seen in Figure 3.

\(^4\)This calls for $M$ to be lower trapezoidal.
We define the stability to be the performance on the first experienced task at any given time and plasticity to be the ability of the model to adapt to the current task. Namely, these constitute the first column $M_{i,0}$ and the diagonal of the matrix $\text{diag}(M)$. We employ the curves derived from these definitions to better dissect the stability-plasticity dilemma of the methods analyzed in our work.

4. Experiments

In this section, we compare regularization-based methods for exemplar-free continual learning. We evaluate the newly proposed attention regularization and compare it with the existing functional ($LwF$) and weight ($EWC$) regularization methods. We then ablate the usefulness of the newly proposed asymmetric loss as well as the importance of pooling prior to applying the regularization.

4.1. Experimental Setup

Setting: For our experiments, we adopt the variation of ViTs introduced by Xiao et al. [47]. Here, the standard linear embedder of a ViT model is replaced by a smaller convolutional stem which helps build more resilient low-level features. Convolutional stems have previously been shown to improve performance and convergence speed in incremental learning settings [52]. We therefore define our architecture to be a lightweight variation of a ViT-Base by setting $L = 12$ layers, $K = 12$ heads per layer and a $d_e = 192$ embedding size. The choice of a small embedding size has been made to speed up the training procedure and unlock the ability to handle larger batch sizes (1024 for our work).

We analyze our task-incremental setting on two widely used image recognition datasets - namely CIFAR100 and ImageNet-32 with 100, and 300 classes each. Both datasets host $32 \times 32$ images. On CIFAR100, we consider a split of 10 tasks (denoted as CIFAR100/10 setting) where each incremental task is composed of 10 disjoint set of classes. On ImageNet-32, we split 6 tasks with 50 disjoint set of classes each (denoted as ImageNet/6 setting).

Our total training epochs remain 200 (per task) for CIFAR100 and 100 for ImageNet32 with an initial learning rate of 0.01 and patience set of 20 epochs. We report our scores averaged over 3 random runs. We apply a constant padding of size 4 across all our datasets. The train images are augmented using random crops of sizes $32 \times 32$ and random horizontal flips with a flipping probability of 50%. For test images, we only apply center crops of sizes $32 \times 32$.

We compare the attention and functional symmetric and asymmetric versions of $L_{\text{sym}}$. Our basic functional and weight regularization approaches include $LwF$ [23] and $EWC$ [19], respectively. For all our experiments relying on PAD losses, we performed a hyperparameter search (using equation 12) for $\mu$ and $\lambda$ by varying each in the range $[0.5, 1.0]$ and found $\mu = \lambda = 1.0$ to perform reasonably well. We thus stick to these values unless otherwise specified. For the sake of readability, the functional and attentional variants of $L_{\text{sym}}$ are indicated as $\text{FUNC}(\text{sym})$ and $\text{ATT}(\text{sym})$, respectively. Similarly, those corresponding to $L_{\text{asym}}$ are denoted by $\text{FUNC}(\text{asym})$ and $\text{ATT}(\text{asym})$. Note that all these four losses specify instantiations of equation 12. The functional approaches regularize the contextualized embeddings while the attention approaches leverage prescaled attention maps (see Figure 1).

4.2. Results

We report accuracy as well as forgetting [7] scores in task aware (taw) setting. We further report taw plasticity-stability curves (based on Figure 3) to provide insights upon how well the different models trade the dilemma.

Accuracy and Forgetting: As seen in Figure 4, all asymmetric approaches show better performances with respect to their symmetric counterparts on CIFAR100/10 with $\text{ATT}(\text{asym})$ offering the best accuracy of 57.3% on the last task. The trend continues for ImageNet/6 with an exception of asymmetric functional approach with an accuracy of

\[\text{Accuracy} = \frac{\text{Number of correctly classified images}}{\text{Total number of images}} \times 100\%\]

\[\text{Forgetting} = \frac{\text{Number of incorrectly classified images}}{\text{Total number of images}} \times 100\%\]

\[\text{Stability} = \frac{\text{Performance on the first experienced task}}{\text{Final performance}} \times 100\%\]

\[\text{Plasticity} = \frac{\text{Final performance}}{\text{Performance on the first experienced task}} \times 100\%\]

\[\text{Dilemma} = \text{Stability} - \text{Plasticity}\]

5 Refer to Appendix A for experiments on additional settings.

6 The corresponding task agnostic scores can be found in Figure 8, Appendix A.
Figure 5. Mean and standard deviation of task-aware plasticity-stability scores for CIFAR100/10 and ImageNet/6 settings (over 3 random runs). Asymmetric approaches are more plastic compared to their symmetric counterparts while retaining competitive stability.

27.55% falling behind its symmetric counterpart by 0.44%.

In general, the asymmetric and symmetric losses lead to improved accuracy scores with respect to other methods. Moreover, we observe that all the methods depict good forgetting resilience with their forgetting scores running at \( \approx 0.01\% \) except for EWC. This suggests us that vision transformers are better incremental learners but require more training and tuning efforts to achieve reasonable accuracies. This remark remains in accordance with prior studies [33,37]. In the particular case of EWC, we observe poor compatibility in terms of accuracy as well as forgetting – with the scores falling behind finetuning at times. We suspect that the method might be less suited for ViTs due to its reliance on exhaustive fisher information estimation.

Plasticity-stability trade off: We compare the dilemma for various methods in Figure 5. With no distillation, fine-tuning is prone to the worst trading of plasticity for stability. Meanwhile, our asymmetric losses can be seen to be more plastic with respect to their symmetric counterparts while depicting comparable stability scores. This confirms our hypothesis regarding the nature of the asymmetry keeping it from discarding older attention while favoring the integration of new attention at the same time. Although, LwF with a last task score of 47.74% on CIFAR100/10 and 32.0% on ImageNet/6, reports the best plasticity among our approaches, it clearly lags behind the pooling-based approaches at retaining stability. On the contrary, the (a)symmetric attention losses and the symmetric functional loss perform similar with a last task stability score of \( \approx 0.23\% \) on ImageNet/6 and \( \approx 53\% \) on CIFAR100/10. EWC shows good plasticity but virtually zero stability. This trend is in line with our previous remark on the limitation of EWC in Figure 4.

### 4.3. Ablation study

Towards the end goal of evaluating the effectiveness of the best performing FD losses, we ablate the contribution of pooling on the CIFAR100/10 setting. In particular, we consider distilling the contextualized embeddings when these are: (a) pooled along both dimensions, i.e., \( \text{FUNC(asymsym)} \) Spatial (see Equation 6), and (b) not pooled at all, i.e., \( \text{FUNC(asymsym)} \) Intact. Distilling the intact embedding vectors of the latter setting obviously implies enhanced stability over its pooled counterpart. Our standard accuracy and plasticity-stability measures across tasks can therefore be deemed redundant in this setting.

As a consequence, we choose to compare the task-aware average incremental accuracy [38] and the last task accuracy across (a) and (b) while contrasting these with LwF as a strong baseline. As shown in Table 1, we find that \( \text{FUNC(asymsym)} \) Spatial consistently performs the best on both the metrics (with a gain of \( > 2\% \)) over \( \text{FUNC(asymsym)} \) Intact and \( > 1\% \) over LwF in either metric). In general, distilling the intact embeddings can be seen to be hurting the performance of the models as their accuracies drop below that of the baseline LwF. This suggests us that strong regularization effects such as that from the distillation of intact contextualized embeddings can often lead to ungenerous trading of stability for plasticity.

<table>
<thead>
<tr>
<th>CIFAR100/10 (taw)</th>
<th>( \text{FUNC(asymsym)} ) Spatial</th>
<th>( \text{FUNC(asymsym)} ) Intact</th>
<th>( \text{FUNC(asymsym)} ) Intact</th>
<th>( \text{LwF} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Incr.</td>
<td>( 56.18% )</td>
<td>( 54.43% )</td>
<td>( 53.12% )</td>
<td>( 54.81% )</td>
</tr>
<tr>
<td>Last Task</td>
<td>( 57.26% )</td>
<td>( 56.04% )</td>
<td>( 54.59% )</td>
<td>( 55.93% )</td>
</tr>
</tbody>
</table>

Table 1. Comparison of intact (no pooling), spatial (pooling along width and height), and LwF.

### 5. Conclusion

In this work, we adapted and analyzed several continual learning methods to counter forgetting in Vision Transformers mainly with the help of regularization. We then introduced a novel self-attention inspired regularization, based on the attention maps of self-attention mechanisms which we termed as Pooled Attention Distillation (PAD). Discerning its limitation at learning new attention, we devised its asymmetric version that avoids penalizing the addition of
new knowledge in the model. We validated the superior plasticity of the asymmetric loss on several benchmarks.

Besides the detailed comparison with a range of regularization approaches, i.e., functional (LwF), weight (EWC), and the proposed PAD regularization, we extended the application of PAD to the functional submodules of Vision Transformers. To this end, we investigated the functional distillation (FD) loss applied in the contextualized embeddings of ViTs. The latter exploration led us to discover that the regularization of functional submodules can help achieve the best overall performances while the regularization of their attentional counterparts endow CL models with superior stability. Finally, we remarked the low forgetting scores of vision transformers across the incremental tasks and concluded that their enhanced generalization capabilities may offer them a natural inclination towards incremental learning. By releasing our code, we hope to open the doors for future research along the direction of efficient continual learning with transformer-based architectures.

Acknowledgment

We acknowledge the support from Huawei Kirin Solution, the Spanish Government funded project PID2019-104174GB-I00/ AEI / 10.13039/501100011033, and European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement number 777720 (CytbSpeed).

References


[42] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Raz-


