

Adaptive Compression-based Lifelong Learning

Shivangi Srivastava¹

shivangi.srivastava@wur.nl

Maxim Berman²

maxim.berman@kuleuven.be

Matthew B. Blaschko²

matthew.blaschko@esat.kuleuven.be

Devis Tuia¹

devis.tuia@wur.nl

¹ Geo-Information Science and Remote Sensing, Wageningen University the Netherlands

² Center for Processing Speech & Images, KU Leuven Belgium

Abstract

The problem of a deep learning model losing performance on a previously learned task when fine-tuned to a new one is a phenomenon known as *Catastrophic forgetting*. There are two major ways to mitigate this problem: either preserving activations of the initial network during training with a new task; or restricting the new network activations to remain close to the initial ones. The latter approach falls under the denomination of *lifelong learning*, where the model is updated in a way that it performs well on both old and new tasks, without having access to the old task's training samples anymore.

Recently, approaches like pruning networks for freeing network capacity during sequential learning of tasks have been gaining in popularity. Such approaches allow learning small networks while making redundant parameters available for the next tasks. The common problem encountered with these approaches is that the pruning percentage is hard-coded, irrespective of the number of samples, of the complexity of the learning task and of the number of classes in the dataset. We propose a method based on Bayesian optimization to perform adaptive compression/pruning of the network and show its effectiveness in lifelong learning. Our method learns to perform heavy pruning for small and/or simple datasets while using milder compression rates for large and/or complex data. Experiments on classification and semantic segmentation demonstrate the applicability of learning network compression, where we are able to effectively preserve performances along sequences of tasks of varying complexity.

1 Introduction

Humans are very good at learning tasks in a sequence [1], including the case when observations from the previous tasks are not accessible anymore. On the contrary, artificial intelligence-based learning models, such as Convolutional Neural Networks (CNNs), struggle in that situation: when confronted with the new task, CNNs tend to migrate towards it and to forget the representation that helped to solve the original task. This problem is generally known as *catastrophic forgetting* [2, 16, 22, 23, 26]. After some initial empirical attempts to

understand the phenomenon [10, 80], methods dealing explicitly with the problem have been proposed in the literature under the name of *lifelong learning* [19, 25]: those methods usually either preserve activations of the initial network when training for new task, or constrain the new network’s activations to remain close to the initial ones. The promise of lifelong learning is to provide methods that are able to perform well on both tasks, even after having learned them in a sequence and without access to the labels of the former task while learning for the next one.

In parallel, approaches known as *pruning networks* for freeing network capacity during sequential learning of tasks have been gaining in popularity [20]. The weights in the network associated with each task are pruned until they occupy a fraction of the global network capacity; these pruned weights then remains frozen while learning the subsequent tasks. By doing so, one can provide capacity for learning the new tasks without having to significantly increase the model size. Moreover, such a strategy also allows reusing redundant parameters for the next tasks, while restricting the growth of the model only to a new classifier layer per new task being considered.

Pruning networks primarily rely on one parameter, the pruning percentage, which balances the compression gain and accuracy decrease. Even in the most recent models, this percentage is generally treated as a hyperparameter, and hence hard-coded. However, we argue that hard-coding the percentage is suboptimal, due to multiple reasons. First, the compression rate needs to be related to the size and complexity of the task at hand: while it makes sense to prune heavily, *i.e.* heavier network compression rates, for a small and/or simple dataset, it is more advisable to perform lower compression for larger or more complex datasets (like ImageNet). Second, the order in which the tasks are coming is also of importance: one cannot know in advance when the more complex task will come, so the ability to save as much capacity as possible is a desirable property for a sequential learning algorithm.

In this paper, we tackle the problem of learning compression of neural networks for sequential learning. Using a Bayesian optimization approach, we learn the optimal compression rate to be applied, which is optimal in the sense that it will perform compression up to an acceptable loss in performance in the previous task. By doing so, the method *guarantees* to avoid catastrophic forgetting, while saving as much network capacity as possible for the next task(s). Additionally, and since the weights on the previous tasks are never modified, there is no need to actively train for the preservation of the accuracy on the previous tasks. We showcase the interest of learning compressing CNNs both in image classification and segmentation: in the first case, we show how a learned compression rate can save capacity to learn a complex new task like ImageNet, while in the second we showcase the advantage of our proposed method in a three-tasks satellite image segmentation problem.

2 Related Works

The most common way to learn a new task from a model trained on another is to fine-tune it [6, 9]. Fine-tuning works generally very well for the new task, but at the price of a drop in accuracy for the former, since the weights are modified and tuned for the new task. A first possible solution is to keep a copy of the original model trained on the original task, but this leads to heavy memory requirements with an increase in the number of tasks. Another solution would be to perform multi-task learning [3], but this strategy relies on labeled data for all tasks to be available during training, which is typically not possible in sequential learning.

The issue of accessing the data of previous tasks is mitigated to a large extent in the ‘Learning without Forgetting’ (LwF) framework [18, 19]. LwF combines fine-tuning and distillation networks [13], where a knowledge distillation loss [13] tries to preserve the output of the former classifier on data from the new task. However, LwF uses several losses, whose number (and balancing weights involved) scales linearly with the number of tasks. The authors in [15, 16] propose approaches where the distance between parameters of the models trained on the old and new tasks is regulated via ℓ_2^2 losses. As for LwF, the number of parameters increases with the number of tasks. In [25], the authors use autoencoders in addition to LwF. This approach has an overhead of a linearly increasing number of autoencoders and task-specific classifiers, several hyperparameters and also a distillation loss between the single-task and the multitask model, making its training complex.

An alternative direction to the above is the idea of removing redundant parameters by neural network compression [20]. The authors report good results but only use a fixed pruning percentage for all the tasks, irrespective of the complexity of the data involved. Other works have used masks on networks’ weights, either using attention [29] or by learning binary weights masks end-to-end, in order to use only the weights useful for the new task [21].

Fine-tuning or lifelong learning approaches lead to an automatically learned balance between the network capacity dedicated to the old task and the network capacity dedicated to the new task through the learning objective but do not guarantee the preservation of the performance of the network on older tasks. On the contrary, compression-based approaches allow a stronger guarantee no forgetting through the preservation of the former task weights in the network but require a manual and arbitrary adjustment of the network capacity dedicated to older tasks vs. the new task. Our proposed framework represents the best of both worlds; given a sequence of tasks, it learns optimal compression rates for each task and avoids drops in accuracy by allocating parts of its tunable parameters to the different tasks in advance. The amount of allocated memory depends on how much compression is applied, and this rate is learned from the data itself through Bayesian optimization.

3 Adaptive compression-based lifelong learning (AcLL)

The compression-based lifelong learning approach of [20] prescribes a fixed pruning rate in order to compress a neural network. A model trained for a particular task after pruning frees up parameters that can be used to learn other tasks. The set of weights that are set to zero are stored as a bit mask. In training the next task, the weights that had been previously set to zero are optimized to maximize performance on the new task, disregarding their effect on the previous task, while the weights that have been retained from the first task are fixed in perpetuity. At test time, the bit mask is applied when evaluating samples from the first task to ensure that the performance of the network is unaffected by the weight changes coming from subsequent tasks. This way, catastrophic forgetting is avoided, and performance on earlier tasks is never degraded by training subsequent tasks. However, this comes at a cost: the number of weights that can be modified to train subsequent tasks is reduced. [20] propose a fixed pruning weight (either 50% or 75% of remaining weights), meaning after the first task 50% of weights remain for the second task, after which 25% remains for the 3rd task, 12.5% for the 4th task, and so on. Assuming the n th task requires a minimum fixed number of tunable parameters to achieve reasonable accuracy, the original network would need to have a size exponential in n with this fixed weighting scheme. We address this by not setting a compression rate *a priori*, but *adaptively*.

Compression algorithms are typically parametrized, *e.g.* through the rank in a low-dimensional matrix factorization, a threshold, or a fraction of weights to remove in sparsifying a network, with each parameter setting achieving a different amount of compression. As a result, there is a trade-off between the amount of network compression and the accuracy of the resulting compressed network. Our intuition is that the amount of compression at any stage of a lifelong learning algorithm should be determined by a performance target on a given task. Then, the amount of compression can be maximized subject to this performance target. Let f be an unmodified neural network, θ a vector of compression parameters, and f_θ the resulting compressed network. Also, let $\mathcal{R}(f)$ be the risk of a function f . $\mathcal{R}(f)$ is the 0-1 loss and the models are optimized on a validation set from the current task. We then consider the following optimization problem:

$$\min_{\theta \in \mathbb{R}^d} \text{size}(f_\theta) \quad (1)$$

$$\text{s.t. } \mathcal{R}(f_\theta) \leq \mathcal{R}(f) + \varepsilon \quad (2)$$

where ε is typically greater than zero and indicates the amount of loss over an uncompressed network that will be tolerated in order to reserve network capacity for future tasks.¹ The tolerance ε is a user-defined parameter that trades off the acceptable loss in performance on the current task versus higher compression rates (and the corresponding available model capacity for new tasks). In previous work with hand-selected parameters [20], the reduction in accuracy is of the order of 1-2%.

The optimization of Eq. (1) is not immediately evident, as the size of the function is not differentiable, and the constraint is over a complicated (non-differentiable in the case of *e.g.* a 0-1 loss) risk functional. In the following, we formulate this problem using a Lagrangian-based optimization strategy to transform it into a series of unconstrained optimization problems. Subsequently, we solve these unconstrained problems using Bayesian optimization.

The Lagrangian of our constrained optimization is

$$\mathcal{L}(\theta, \lambda) := \text{size}(f_\theta) + \lambda (\mathcal{R}(f_\theta) - (\mathcal{R}(f) + \varepsilon)), \quad (3)$$

which indicates that for varying $\lambda \geq 0$, each optimization over θ will be of the form

$$\arg \min_{\theta} \mathcal{L}(\theta, \lambda) = \arg \min_{\theta} \text{size}(f_\theta) + \lambda \mathcal{R}(f_\theta). \quad (4)$$

For fixed λ , we call an off-the-shelf Bayesian optimization routine [24]; indeed, Bayesian optimization can be considered to be at the state of the art for optimization of black-box, non-differentiable functions [9, 21].

We wish to determine the optimal λ suited for a target accuracy tolerance ε in the original problem (1). In general, one could solve the unconstrained problem (4) many times for different values of λ . This is however inefficient, in particular because of the time spent for the multiple evaluations of $\mathcal{R}(f_\theta)$ needed for the bayesian optimization of problem (4). We use two properties in order to reduce the optimization time:

- **Concavity.** As Eq. (4) is concave in λ [9, Sec. 5.1.2], the search for the optimal λ for a given ε can be made more efficient by using line search strategies: this can be

¹We have observed that, particularly for small amounts of data, a degree of compression can provide a regularizing effect and it is possible to achieve lower risk from a compressed network than an original uncompressed network.

done e.g. using a cutting plane approach – where each evaluation of the dual (4) gives a subgradient direction. In our experiments, we found that performing a binary search over λ , which is equivalent to taking the sign of that subgradient direction, already leads to an acceptable convergence of λ .

- **Efficient caching strategy.** We cache each evaluation of $\mathcal{R}(f_\theta)$ and reuse these evaluations among different Bayesian optimization runs, for different values of λ . Indeed, given observations at iteration $t - 1$ of the Lagrangian optimization used to model $\text{size}(f_\theta) + \lambda_{t-1}\mathcal{R}(f_\theta)$, it is straightforward to simply re-weight the saved values by a different factor λ_t to initialize the Gaussian process model for the next round of Bayesian optimization. This seeding of the Bayesian optimization process leads to a faster convergence and reduces the number of new evaluations needed.

In practice, the cost of optimizing the constrained form with this caching scheme is a very small multiple of the cost of a single unconstrained optimization. In our experiments using pruning-based compression, we have observed an overall increase in the cost of compression by a factor of approximately 6 to 8 using this Lagrangian-based optimization scheme with caching vs. a fixed compression ratio.

4 Experiments

In this section, we present the results on two challenging settings: sequential learning of models for classification with increasing complexity (Section 4.1) and sequential learning of models for semantic segmentation of satellite images (Section 4.2).

4.1 Classification

In this section, we performed experiments to verify if adaptive compression can lead to better classification performances on a complex task. We experimentally show that by applying our adaptive compression method we perform stronger compression to a model trained with a relatively simple task. This helps to improve the performance of the model for a subsequent more complex task that will have more free parameters to train. It is worth mentioning that after we train the model for the second more complex task the model is still able to perform prediction in the first task with accuracy within tolerated limits. Our motivation comes from [20], where the authors first trained a model on ImageNet [28], pruned 50% of the model weights, but then applied lifelong learning to a smaller CUBS Birds dataset [52], for which we argue not a large capacity is necessary, so *a priori* small compression rates are acceptable.

Data and setup. As a first task, we trained a model on the CUBS Birds dataset and then switched to ImageNet as a second task. Details on the number of images are provided in Table 1. We used ResNet50 [12] as base model, initialized with the pretrained weights of Places365 [63]. We then trained on CUBS for 40 epochs, with a learning rate of 0.01, divided by a factor 10 every 20 epochs. As explained in [20], when pruning the model some parameters of the model are set to zero to make them available for learning subsequent tasks. After pruning, the parameters that remain for the original model need to be finetuned [20].

Therefore, postprune finetuning for CUBS was pursued for 10 additional epochs with a learning rate of 0.01. Next, the model was trained on the second task (ImageNet) for 20 epochs, with a learning rate of 0.001, also divided by a factor 10 every ten epochs.

Task	# classes	# images	
		training	test
CUBS	200	5,994	5,794
ImageNet	1,000	1,281,144	50,000

Table 1: Datasets used for the classification experiments.

Results. Results are reported in Table 2. We can see that a compression rate of 50% leads to an accuracy of 64.14% over the ImageNet test set. Using our proposed adaptive approach, we found that the same ResNet50 network, originally trained with CUBS, could be pruned to a much higher rate, while still keeping the drop of accuracy on CUBS dataset within the 2% range, therefore saving more capacity for the second task and leading to higher accuracies on ImageNet: 66.98%. This implies that setting a hard pruning parameter is not optimal in the case of lifelong learning with tasks of different complexity and that learning such rates can make the difference in saving capacity for future tasks.

This allows us to smartly utilize the available parameters in the model according to the need of the task at hand and still perform within an acceptable loss in performance in the first task, contrarily to classical fine-tuning, where we observed a performance loss of 76.48% on CUBS (from 77% of the original model to 0.52% after finetuning).

Lifelong learning strategy	CNN compression rate (%)	Accuracy	
		CUBS (%)	ImageNet (%)
None	0	77.0	-
Finetuning ImageNet from CUBS	0	0.52	67.27
PackNet [20]	50	76.72	64.14
AcLL (us)	86	75.18	66.98

Table 2: Lifelong learning results in the classification setting where a model learned on CUBS is re-used to learn ImageNet as a second task.

4.2 Semantic segmentation of satellite images

In this experiment, we aim at learning a sequence of three models dedicated to three different tasks of semantic segmentation of satellite imagery: detecting roads, detecting buildings and mapping coarse landcover types. We investigate if allocating network capacity according to task complexity, for more than two tasks, has an impact on the overall performance.

Data and setup. We used the training portion of the DeepGlobe 2018 dataset [8], which is composed of the disclosed labeled images of the DeepGlobe challenge (deepglobe.org), as the validation and test sets were unavailable during the course of the challenge. We then divided the data into our own training, validation, and test subsets. We considered three semantic segmentation tasks: ‘Landcover’ (multi-class), ‘Roads’ and ‘Buildings’, the latter

two being binary class problems (e.g. road vs. background in the case of ‘Roads’). In each task, the expected outcome is a map per image, where every pixel is classified in one of the classes or background. The number of images available per task is provided in Table 3.

Task	# classes	# images			Size (pixels)	Resolution (cm)
		training	val	test		
Roads	2	3,984	1,121	1,121	$1,024 \times 1,024$	50
Landcover	7	562	120	121	$2,448 \times 2,448$	50
Buildings	2	3,207	687	688	650×650	31

Table 3: Datasets used for the segmentation experiments. Examples of images can be seen in Figures 1 and 2. Each image corresponds to a full semantic segmentation map with $r \times c$ pixels to be classified.

As base semantic segmentation model, we used ERFNet [27], and evaluated two task sequences: ‘Landcover’, ‘Road’, ‘Buildings’ ($1 : \mathbb{L} \rightarrow \mathbb{R} \rightarrow \mathbb{B}$) and ‘Road’, ‘Landcover’ and then ‘Buildings’ ($2 : \mathbb{R} \rightarrow \mathbb{L} \rightarrow \mathbb{B}$), respectively. We compared our proposed AcLL against four baselines: finetuning one model after the other, learning without forgetting (LwF [19]), an autoencoder-based LwF (AE [25]) and fixed-rate compression (PackNet [24]).

For all models, we used Adam optimizer [14] with weight decay of 0.0001. The models were trained for $T = 100$ epochs with an initial learning rate of 0.0005, which was then decreased by a factor of $(1 - \frac{t}{T})^{0.9}$ at each epoch t [24]. All the images were resized to 512×512 pixels before data augmentation (random horizontal flips and translation of up to two pixels in the horizontal and vertical directions). We used class weight inversely proportional to the number of pixels per class.

The distillation loss weight was set to 1 and the weight of the autoencoder-based loss component to 0.01 as in [25]. Since the task is semantic segmentation, we used a convolutional autoencoder for AE to output a grid of predictions from the bottleneck of ERFNET.

For AcLL, the multitask scheduling was as follows: ERFNet was first trained for 100 epochs with the first task, and then pruned. After pruning, the model was further fine-tuned on the first task for 30 epochs. The same scheduling was adopted for the second task, while for the third (the last) task only training with 100 epochs were performed. For each intermediate task, the drop in accuracy (within 2.0% from $\mathcal{R}(f_\theta)$, see Eq. (1)) was checked on the validation data. The accuracies reported in Tables 4 and 5 are intersection over Union (IoU) scores, evaluated on the test sets of each task.

Results. Table 4 presents the results for the first task sequence $1 : \mathbb{L} \rightarrow \mathbb{R} \rightarrow \mathbb{B}$, while Table 5 focuses on the second task sequence $2 : \mathbb{R} \rightarrow \mathbb{L} \rightarrow \mathbb{B}$. The accuracies for the three individual models trained with just one task are found in the first three rows of both Tables 4 and 5. We cannot make a direct comparison to the performance reported in the official competition as we did not have access to the official validation and test sets, but we note that the accuracy achieved by our model on our test set is comparable or exceeds that of the accuracies reported by the respective leaderboard winners of the challenge: Landcover 52.24% mIoU [8]; Roads 64.12% IoU [54]; and Buildings 74.67 F1-score [10]. This is a good indication that we are analyzing the lifelong learning framework on a strong baseline.

In both the Tables 4 and 5, the baseline Fine-tune obtains good results in the last task, *i.e.* the detection of buildings. However, its performance on the other two tasks is very poor. The baselines LwF and AE obtain the best results on the last task. However, they show

Lifelong learning strategy	CNN compression rate (% $T_{1,2}$), (% $T_{2,3}$)	Accuracy (%)			3 tasks average accuracy
		Task 1 Landcover	Task 2 Roads	Task 3 Buildings	
None (baselines)	0	48.20	-	-	
	0	-	71.03	-	
	0	-	-	80.10	
Fine-tune {T1, T2} → T3	0	3.15	47.95	79.26	43.45
LwF [19]	0	26.52	62.53	81.30	56.78
AE [25]	0	25.77	64.59	81.36	57.24
PackNet [20]	(50.0), (50.0)	49.36	67.67	75.24	64.09
	(75.0), (75.0)	<u>47.47</u>	<u>68.81</u>	78.85	<u>65.04</u>
AcLL (us)	(84.375), (72.0)	47.30	68.92	79.14	65.12

Table 4: Sequential learning of tasks: Landcover → Roads → Buildings (best result in bold, second best underlined).

Lifelong learning strategy	CNN compression rate (% $T_{1,2}$), (% $T_{2,3}$)	Accuracy (%)			3 tasks average accuracy
		Task 1 Roads	Task 2 Landcover	Task 3 Buildings	
None (baselines)	0	71.03	-	-	
	0	-	48.20	-	
	0	-	-	80.10	
Fine-tune {T1, T2} → T3	0	47.95	2.25	79.75	43.31
LwF [19]	0	62.71	22.70	81.23	55.54
AE [25]	0	62.73	30.92	<u>80.52</u>	58.05
PackNet [20]	(50.0), (50.0)	70.75	53.72	70.48	64.98
	(75.0), (75.0)	<u>70.22</u>	<u>48.80</u>	74.61	64.54
AcLL (us)	(86.25), (92.0)	69.08	48.13	77.53	<u>64.91</u>

Table 5: Sequential learning of tasks: Roads → Landcover → Buildings (best result in bold, second best underlined).

heavily degraded performances on the first task (a drop of more than 20%, see Table 4) and, to a lesser extent, on the second task (drop by 8% Table 5). It is evident that these lifelong learning baselines fail in remembering the previous tasks compared with network compression approaches (PackNet and our AcLL), which always outperform the competing methods by a large margin, leading to the best average score over the three tasks (last column of both the Tables). Moreover, our proposed AcLL achieves the best or second-best accuracy in both task orderings (Tables 4 and 5) and also allows for the compression of the network according to the task’s complexity at hand. Such optimal compression is way far greater than 50%, especially since the tasks are not so complex and an efficient network can be obtained with higher compression rates. This approach allows for freeing more redundant parameters for future, unseen tasks if the current task is small or less complex. The benefits of AcLL with respect to PackNet can be seen in the last task, where the additional freeing of

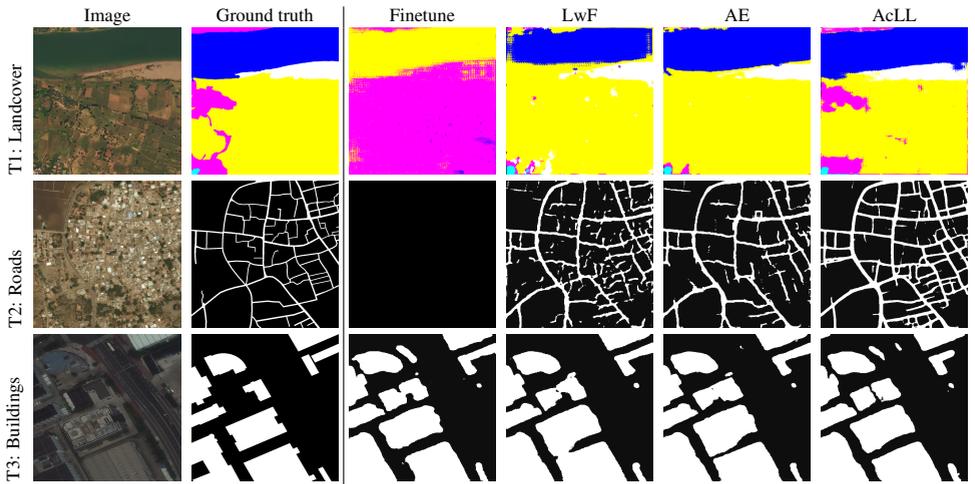


Figure 1: Comparison of the different lifelong learning strategies in the DeepGlobe data for the sequence: Landcover \rightarrow Roads \rightarrow Buildings (for AcLL: 84.375%, 72%). The legend for the Landcover task is: **forest**, **water**, **urban**, **rangeland**, **agriculture**; white shows barren land and black denotes background/unknown.

parameters provides more capacity, and therefore a more accurate CNN for the third task. In Table 5 PackNet with 50% compression ratio achieved a marginally higher average accuracy over the three tasks, it did so at the cost of a more than 7% reduction in accuracy on the final task and after exhausting 75% of its available parameters after two tasks, while the adaptive method had only used less than 21% of the available weights. We expect the benefits to become more and more evident with increase in the number of tasks.

Inspecting the segmentation maps for the two tasks sequences (Figures 1 and 2, respectively), we observe that our adaptive pruning AcLL leads to overall more accurate maps than the three competing baselines, which provide accurate maps mostly for the last task. The segmentation maps for the three tasks (‘Landcover’, ‘Roads’ and ‘Buildings’) are closer to their respective ground truths (Column 2 of both figures). By looking at the maps, it becomes evident that the three competing methods struggle to remember the correct decision function for the first task and only partially perform adequately on the second. The proposed AcLL provides plausible maps for all tasks, even if it sometimes hallucinates linear structures (though still topologically plausible) for the road task in the $2 : R \rightarrow L \rightarrow B$ sequence (see second row, column AcLL in Fig 2).

5 Conclusion

In this work, we propose a method for lifelong learning based on adaptive compression. Different from a recent compression-based method, called PackNet [20], that uses a pre-defined compression rate, we perform adaptive compression that considers the complexity of the task at hand while maintaining guarantees on accuracies of the compressed network on previous tasks. Thus, if a model was trained for a relatively simple task it can be strongly compressed in a way that more free parameters are available to train other subsequent tasks. Our experimental results show the advantage of our AcLL method over four baseline methods: standard

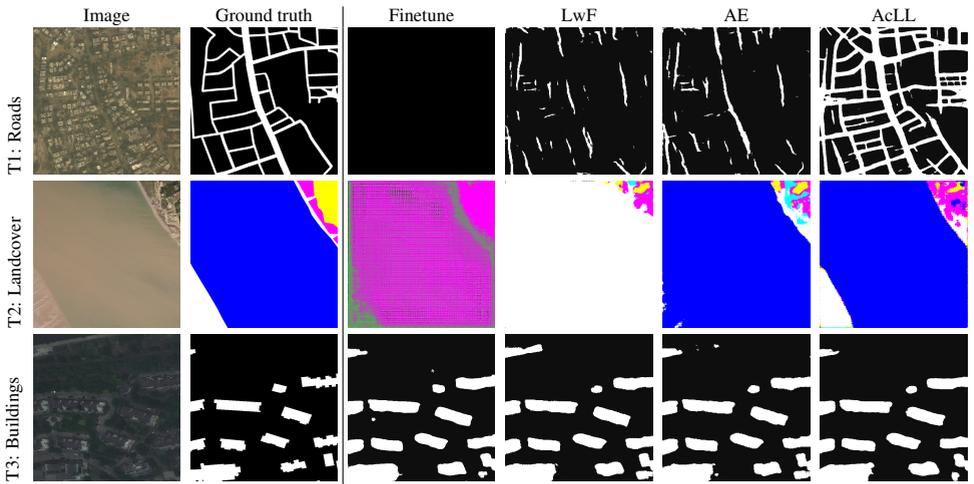


Figure 2: Comparison of the different lifelong learning strategies in the DeepGlobe data for the sequence: Roads \rightarrow Landcover \rightarrow Buildings (for AcLL: 86.25%, 92%). The legend for the Landcover task is: **forest**, **water**, **urban**, **rangeland**, **agriculture**; white shows barren land and black denotes background/unknown.

finetuning, Learning without Forgetting (LwF), Encoder-based life long learning (AE), and PackNet.

Acknowledgement

S.S. is thankful to Gabriel Peyré (“ERC project Noria”) and Marco Cuturi (“Chaire d’excellence de l’IDEX Paris Saclay”) for their support by providing access to computing facilities. M.B. and M.B.B. acknowledge support from FWO (grant G0A2716N), an Amazon Research Award, an NVIDIA GPU grant, and the Facebook AI Research Partnership. This research was carried out while S.S. was a visiting researcher at KU Leuven.

References

- [1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [2] Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *CoRR*, abs/1012.2599, 2010.
- [3] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- [4] Joseph Cichon and Wen-Biao Gan. Branch-specific dendritic ca 2+ spikes cause persistent synaptic plasticity. *Nature*, 520(7546):180, 2015.
- [5] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to

- parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–17209. IEEE, 2018.
- [6] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.
- [7] Peter I. Frazier. A tutorial on Bayesian optimization. *CoRR*, abs/1807.02811, 2018.
- [8] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.
- [9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [10] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.
- [11] Ryuhei Hamaguchi and Shuhei Hikosaka. Building detection from satellite imagery using ensemble of size-specific detectors. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 223–2234. IEEE, 2018.
- [12] 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.
- [13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NIPS Workshop*, 2014.
- [14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International conference on learning representations*, 2015.
- [15] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [16] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016.
- [17] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. In *Advances in Neural Information Processing Systems*, pages 4652–4662, 2017.
- [18] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pages 614–629. Springer, 2016.
- [19] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

- [20] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.
- [21] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–82, 2018.
- [22] 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.
- [23] 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.
- [24] Fernando Nogueira. Bayesian optimization github repository, 2018. URL <https://github.com/fmfn/BayesianOptimization>.
- [25] Amal Rannen, Rahaf Aljundi, Matthew B Blaschko, and Tinne Tuytelaars. Encoder based lifelong learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1320–1328, 2017.
- [26] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.
- [27] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2018.
- [28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [29] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4548–4557. PMLR, 2018.
- [30] Rupesh K Srivastava, Jonathan Masci, Sohrob Kazerounian, Faustino Gomez, and Jürgen Schmidhuber. Compete to compute. In *Advances in neural information processing systems*, pages 2310–2318, 2013.
- [31] Chao Tian, Cong Li, and Jianping Shi. Dense fusion classmate network for land cover classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 262–2624. IEEE, 2018.
- [32] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

-
- [33] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2018.
- [34] Lichen Zhou, Chuang Zhang, and Ming Wu. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 182–186. IEEE, 2018.