Incremental learning of fetal heart anatomies using interpretable saliency maps

Arijit Patra and J. Alison Noble

Institute of Biomedical Engineering, University of Oxford, Oxford, United Kingdom
arijit.patra@eng.ox.ac.uk

Abstract. While medical image analysis has seen extensive use of deep neural networks, learning over multiple tasks is a challenge for connectionist networks due to tendencies of degradation in performance over old tasks while adapting to novel tasks. It is pertinent that adaptations to new data distributions over time are tractable with automated analysis methods as medical imaging data acquisition is typically not a static problem. So, one needs to ensure that a continual learning paradigm be ensured in machine learning methods deployed for medical imaging. To explore interpretable lifelong learning for deep neural networks in medical imaging, we introduce a perspective of understanding forgetting in neural networks used in ultrasound image analysis through the notions of attention and saliency. Concretely, we propose quantification of forgetting as a decline in the quality of class specific saliency maps after each subsequent task schedule. We also introduce a knowledge transfer from past tasks to present by a saliency guided retention of past exemplars which improve the ability to retain past knowledge while optimizing parameters for current tasks. Experiments on a clinical fetal echocardiography dataset demonstrate a state-of-the-art performance for our protocols.

Keywords: Saliency· Interpretability · Continual learning

1 Introduction

Medical image analysis pipelines have made extensive use of deep neural networks in recent years with state-of-the-art performances on several tasks. In diagnostic ultrasound, the availability of trained sonographers and capital equipment continues to be scarce. For congenital heart disease diagnosis in particular, the challenges become even more pronounced with the actual identification and processing of relevant markers in sonography scans being made difficult through the presence of speckle, enhancements and artefacts over a small region of interest. As with other applications of deep networks in medical image analysis[1], the retention of performance on already learnt information while adapting to new data distributions has been a challenge. Often, a requirement for deep networks is the availability of large labeled datasets. In medical imaging tasks however, data is often not abundant or legally available. There exist intra-patient variations, physiological differences, different acquisition methods and so on. Not all necessary data may be available initially but accumulated over time, and be used to establish overall diagnosis. Incremental learning systems are those that leverage accumulated knowledge gained over past tasks to optimize adaptation to new tasks. Such optimization may not always ensure the traversal through the parameter space in a manner suitable for old tasks. This causes degradation in the performance of old tasks while adapting to new ones, and a balance is desired between stability of old knowledge and plasticity to absorb new information.

Literature Review The loss of learnt features from prior tasks on retraining for new tasks leading to a diminished performance on old tasks is a phenomenon called 'catastrophic forgetting'[3]. Many methods have been proposed to build a lifelong learner. These are broadly classified[4] into i)architectural additions to add new parameters for new data distributions, such as Progressive Networks [11] where new parameter sets get initialised for new tasks with a hierarchical conditional structure imposed in the latter case and the memory footprint is of the order of the number of parameters added ii) memory and rehearsal based methods where some exemplars from the past are retained for replay (rehearsal[16], or are derived by generative models in pseudo-rehearsal, for replay while learning on new tasks. Examples include iCaRL [5],end-to-end lifelong learning[6] etc. In these, certain informative exemplars from the past are retained and used as representative of past knowledge on future learning sessions iii) regularization strategies, which include methods to enforce preservation of learnt logits in parts of the network like in distillation strategies in Learning without Forgetting [7], or estimate parameter importance and assign penalties on them to ensure minimal deviation from learnt values over future tasks like in Elastic Weight Consolidation [8] and Synaptic Intelligence based continual learning [9]. In medical imaging, incremental addition of new data has been sporadically addressed,notably in [2,12], despite clinical systems often acquiring data in non-deterministic phases. While there have been efforts apart from transfer learning [10] to resolve the paucity of labeled data, these have concentrated on augmentation ,multitask learning [1], and so on. In the domain of interpretability of medical images, there has been a focus on utilizing attention mechanisms to understand decisions of machine learning models, such as attention mechanisms for interpretation in ultrasound images in fetal standard plane classification[13],pancreas segmentation [14] and so on. Utilizing the notions of interpretability in a continual setting or for enabling learning in incremental sessions is yet to be studied in literature and we introduce notions of class saliency and explainability for assessing and influencing continual learning mechanisms.

Contributions Our contributions are a)a novel method to avoid catastrophic forgetting in medical image analysis b)quantifying model forgetting and incremental performance via saliency map quality evolution over multiple learning sessions c)saliency quality in individual sessions to choose informative exemplars for class-wise rehearsal over successive learning sessions.Usage of saliency map quality for evaluating incremental learning performance and saliency maps to se-

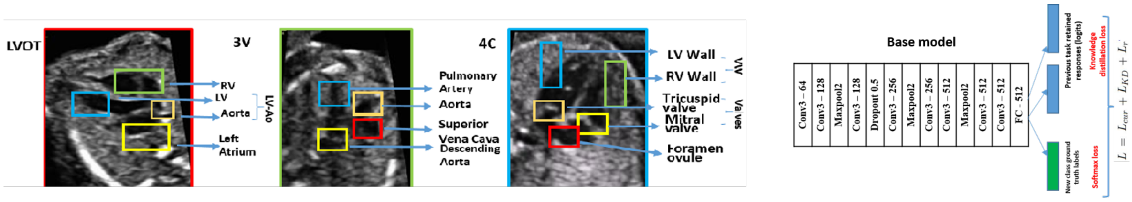

Fig. 1. (a) Fetal cardiac anatomical classes (b) Classification and knowledge distillation scheme for our model.

lect exemplars to retain for replay and distillation based knowledge regularization are new to computer vision and medical imaging to our knowledge.'Incremental learning' and 'continual learning' are used interchangeably in literature.

2 Methodology

The aim of the study is to introduce the usage of interpretability as a building block for incremental learning in clinical imaging using fetal cardiac anatomy classification as a proof-of-concept.Our classes of interest are anatomical structures apparent in standard fetal cardiac planes (four-chamber or 4C, three-vessels or 3V and left ventricular outflow tract or LVOT view). Our problem deals with a class-incremental setting for detecting fetal cardiac structures Ventricle Walls (VW),Foramen ovale (FO), valves (mitral,tricuspid) (4C view structures), left atrium (LA), right ventricles (RV), Aorta (LV and aorta are seen as a continuous cavity and labeled as LV-Ao hereafter) and right atrium (LVOT structures), Pulmonary Artery (PA), Aorta, Superior Vena Cava (SVC) (3V view). These structures are considered for study because of their relevance in assessment of congenital heart disease [17]. Out of these, structures are learnt in sets of 3, first as base categories in the initial task, followed by incremental task sessions. The remaining 3 are shown to the base-trained model in incremental stages in our class incremental learning experiments. (VW, Valves, FO) and (RV, PA, Aorta) and (LV-Ao, SVC, LA) are then the class groups introduced in successive task sessions. This simulates a setting where the algorithm needs to adapt to new data distributions in the absence of a majority of exemplars from past distributions.

We propose a dual utilization of saliency to implement this continual learning setting. First, we define novel quality metrics for class averaged attention maps that also quantify the ability of a model to learn continually.CNNs learn hierarchical features that are aggregated towards a low-dimensional representation and the inability of the model to retain knowledge is manifested in this hierarchy as well. Since attention maps point out the most relevant pixels in an input image towards the classification decision made on it by a model, a digression of focus from these pixels is indicative of degradation of past knowledge. Thus, attention map quality can be used to quantify not only the overall decline in performance over old tasks, but offer detailed insights into relative decline at the level of individual classes in the task, and also for individual instances in a class (which can be used as a measure of some examples being especially difficult to retain). This attention based analysis is motivated by the fact that there has been no standard agreement on how continual learning performance ought to be evaluated. Present measures of forgetting and knowledge transfer do not allow a granular assessment of learning processes or distinctions on the basis of difficulty of an instance, nor do they allow a scope for explainability of the continual learning process. Creating attention maps by finding class activations allows for feature level explainability of model decision on every learning cycle.

2.1 Saliency map quality.

At the conclusion of a given task of N classes, the validation set of each class is passed through the model which is then subjected to class activation mappings (CAM) to obtain attention maps using the GradCAM approach [15]. Note that the specific method to obtain attention/saliency maps here is for demonstration only and any suitable method may replace it. We consider only the maps resulting from correct predictions because the explanations are generated even otherwise but is suboptimal for further inference. Each instances map represents the understanding of the model for the decision taken on that instance. Averaged over the validation set of a class in a task, this average saliency map represents the average explanation of the model for the decisions of classification. In the absence of a ground truth, estimating a quality metric for the explanation based saliency map is non-trivial. We try to assess the extent of forgetting by tracking the difference between the activations obtained just after the class or task has been learnt and that after a few other tasks are learnt subsequently. This can be performed both for individual instances of classes and by considering classwise average saliency maps. Past literature has explored saliency map quality in terms of being able to mimic human gaze fixations or as a weak supervision for segmentation or detection tasks,in which ground truth signals were enforced, even if weakly. That apart, saliency maps were evaluated by [16] in context of their attempts at designing explainable deep models. They interpret the efficiency of saliency maps as a reasonable self-sufficient unit for positive classification of the base image. Then, the smaller the region that could give a confident classification in the map the better the saliency. Mathematically, this was expressed as a log of the ratio of the area and the class probability if that area was fed to the classifier alone. This quantity called the SSR (Smallest Sufficient Region) is expressed in [16] as $|log(a)\text{-}log(p)|$, where a is the minimum area that gives a class probability for the correct class as p. This method assumes that the concentration of informative pixels is a good indicator of attentive features, and is unsuitable for cases where features of interest are distributed spatially in a non-contiguous manner (say a fetal cardiac valve motion detection, lesions in x-rays for lung cancer classification etc.). As such, considering our dataset of fetal heart ultrasound, attentive regions may be distributed over a spatial region and the non-contiguous informative regions can be adequately quantified only through metrics designed for multiple salient region estimation and cannot be adequately expressed by SSR. We extend from the SSR concept, and instead of thinking in terms of concentration of information consider the extent of regions of informative content. To do so, we consider a grid of fixed uniform regions on the input image. Each grid region is taken as a small rectangular space whose information content is evaluated, the size of these small regions is fixed as a hyperparameter (we consider 224 x 224 image inputs, and 16 x 16 grid regions by optimizing for computational cost and accuracy). Each of these grids is evaluated by the trained model after a task for their prediction probabilities by themselves. Then for each grid region, a quantity $|log(A_g) - log(p)|$ can be used, with $log(A_g)$ being constant for fixed size grid regions, to estimate the contribution of the region to the overall saliency map. The smaller this quantity, the more informative this region is. A threshold can be imposed and all n grid regions contributing can be summed up to express the Overall Saliency Quality (OSQ)as:

$$\sum_{k=1}^{n} |\log\left(A_g^k\right) - \log\left(p\right)| \tag{1}$$

This threshold depends on the desired class probability for the correct class (this expression would be valid even for incorrect classifications, but we choose regions only with correct predictions). Thus, a quality measure can be derived for each saliency map at all stages of tasks. The expression above essentially gives an absolute measure of saliency map quality. Retention of exemplars for efficient continual learning was done by random selection, by nearest class mean and so on. These methods do not consider actual performance while making the retention decision but derive from input distributions if not selecting randomly. While in ideal cases, the class mean will reflect exemplars with high quality saliency maps (as one would trivially expect that the average class representation has the most volume of data and hence is more influential on the learning curve for these classes), it is not always true in case of classes with significant diversity and multiple sub-clusters of exemplars. In the latter, selecting exemplars by methods like herding [5], the need for retaining exemplars close to the class means is not optimal for retaining the diversity of class information, and the

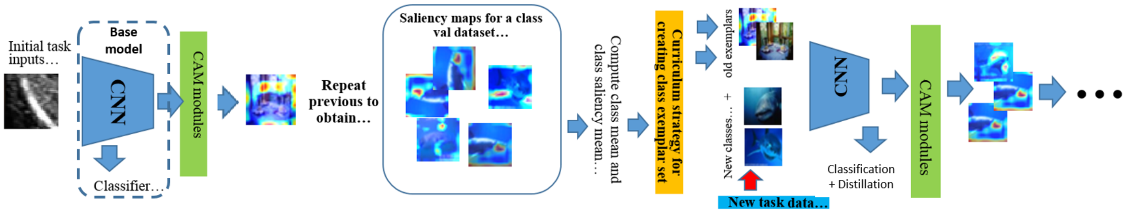
Initial task inputs…
Base model
CNN
Classifier…
CAM modules
Repeat previous to obtain…
Saliency maps for a class val dataset…
Compute class mean and class saliency mean…
Curriculum strategy for creating class exemplar set
New classes… + old exemplars
New task data…
CNN
Classification + Distillation
CAM modules

Fig. 2. Schematic of saliency curriculum based exemplar retention - After training for a session, CAM modules compute attention maps from instances for OSQ calculations and selecting retention examples. Figures for representation purposes only

diversity of informative representations needs to be accounted for, which can be informed by saliency based retention strategies.

2.2 Saliency driven continual curriculum.

We attempt to use the attention maps of past representations to help actively preserve knowledge and at the same time improve generalization to future classes. This is similar to transfer of knowledge from *Task N* to *Task (N+1)* through the attention maps of the former being used to condition the latter. After each task schedule we generate class activation maps using [9] for the validation data per class and estimate the quality measure defined in the first part. Following our need for retaining prior knowledge while moving to the next task, we consider a selective retention strategy. In order to account for fixed memory allocations per class, a fixed number of exemplars may be retained. We try to establish the most informative of instances through an optimal map quality curriculum over available instances in a class. This relies on studies of explainable representation learning that if a classification decision were established through an empirical risk minimization objective, a majority of instances for that category would have low-dimensional feature representations in a close vicinity and away from hyperplanes separating different clusters [3]. We propose two ways of preparing a saliency based exemplar retention: 1) Assuming the presence of definitely identifiable salient cues in available instances, an average class saliency map can be understood as an overall class decision explanation. A relative proximity at the pixel space of a given saliency map of a given instance to this average saliency map would indicate the suitability of such an instance for being stored as a representative exemplar for its class. This approach is termed the Average Representative Distance Selection (ARDS). 2) Another alternative is to pre-select a set of most representative exemplars from a validation dataset of the class, in terms of the class confidence probabilities, and use the normalized cumulative distance from their saliency maps to every other instances saliency map. This is termed Distributed Exemplar Distance Selection (DEDS), and is primarily useful for cases where the salient regions of interest have a non-trivial spatial variation within the same class exemplar set. ARDS is suited when strong cues are localised and class prototype saliency maps are useful. DEDS is useful in cases of a diversity of cues not similarly located on all exemplars or when a significant shift is caused by affine transformations. ARDS is computationally more efficient as a prototype-to-exemplar distance is computed in a single step. Actual choices between the two would depend on data characteristics. In both cases, the saliency map is treated as a probability distribution and similarity is assessed by KL divergence between reference saliency maps and instance maps,

$$D_{KL}[e,d] = \int (e(x) - d(x)) \log(e(x)/d(x)) dx \tag{2}$$

considering that *e(x)* and *d(x)* are the respective saliency maps being compared. As the computation over the pixel space is discretized, the integral form is replaced with a discrete summation,

$$D_{KL}[e,d] = \sum_{x=1}^{N_{pixels}} (e(x) - d(x))\log(e(x)/d(x)) \tag{3}$$

As for choosing instances for the ARDS or DEDS calculations, it would be superfluous to compute for entire training sets. We choose 100 exemplars from each class's training set (based on their confidence probabilities in final epoch) for this process. The retention examples are chosen from training sets as they are replayed in future sessions and hence validation set examples cannot be used except as saliency benchmarks, i.e, while the benchmarks for average saliency or representative exemplar sets are derived from validation sets, these can't be retained as validation data can't be used in any part of training. In both cases, 30 instances are finally chosen after a grid search over integral number of samples between 10 and 50 (with the upper bound governed by memory constraints and lower bound on performance thresholds), to be retained in memory for future rehearsal. Choosing higher numbers of exemplars was found to lead to minimal performance gains (a detailed study of these trade-offs is kept for future work).

3 Model and Objective functions

For our architecture, we implement a convolutional network with 8 convolutional layers, interspersed alternately with maxpooling and a dropout of 0.5 (Fig.1(b)). This is followed by a 512-way fully-connected layer before a softmax classification stage. The focus of the work is not to achieve possible state-of-the-art classification accuracies on the tasks and datasets studied, but to investigate catastrophic forgetting in learning incrementally. Thus, the base network used is significantly simplified to keep the order of magnitude of the number of parameters within that of other continual learning approaches reported in literature [4] and enable a fair benchmarking. For a loss function, we implement a dual-objective of minimizing a shift on learnt representations in the form of prior logits in the final model layers using a knowledge distillation approach[17], and performing a cross-entropy classification on the current task classes. A quadratic regularization is additionally imposed with a correction factor that is set as a hyperparameter by grid search. The overall objective function is:

$$L = L_{cur} + L_{KD} + L_r \tag{4}$$

where L is the total objective composed of the softmax cross entropy as the current loss, the knowledge distillation loss on past logits, and the regularization term for the previously trained parameters.

$$L_{cur}(X_n, Y_n) = -\frac{1}{|N_n|} \sum_{i=1}^{N_n} \sum_{j=1}^{J_n} y_n^{ij}.\log(p_n^{ij}) \tag{5}$$

where $N_n$is the number of examples in a batch, $J_n$is the number of classes, ${y_n}^{ij}$is the one-hot encoded label for an instance, and ${p_n}^{ij}$is the softmax prediction. For the distillation terms, the original labels not being available,${y_o}^{ij}$is computed with new and retained examples and compared to stored logits for the old examples${p_o}^{ij}$, giving a loss term:

$$L_{KD}(X_n, Y_o) = -\frac{1}{|N_n|}\sum_{i=1}^{N_n}\sum_{j=1}^{J_o} y_o^{ij'}.\log(p_o^{ij'}) \tag{6}$$

where ${y_o}^{ij'} = \frac{({y_o}^{ij})^{1/\lambda}}{\sum_j ({y_o}^{ij})^{1/\lambda}}$ and ${p_o}^{ij'} = \frac{({p_o}^{ij})^{1/\lambda}}{\sum_j ({p_o}^{ij})^{1/\lambda}}$, where the distillation temperature λ is set at 2.0 over hyperparameter search on values from 1 to 10.

The parameter regularization is imposed for already trained weights for the past classes and penalises the shift for current adaptation through a Frobenius norm over the parameters as $L_r = \mu \sum_j ||w_o - w'||^2 . \mu$ is set to 0.4 by grid search. The idea here is that while the retained exemplars from the past tasks are seen with the present data, the process of optimization should update parameters in a manner compatible with past prediction features. This ensures that parameters are adapted to present distributions without drastic shifts that adversely affect their ability to arrive at the optimal representations for previously seen examples. A distillation framework is implemented here as such a loss term in conjunction with a cross-entropy objective can enforce a regularization on representations from the past, which is not achieved by a simple parametric regularization.

Data. For fetal echocardiography data, we consider a clinically annotated dataset of 91 fetal heart videos from 12 subjects of 2-10 s with 25-76 fps. Obtaining 39556 frames of different standard planes and background, we crop out relevant anatomical structures in patches of 100 x 100 from the frames, leading to a total of 13456 instances of 4C view structures (Ventricle Walls, Valves, Foramen ovule), 7644 instances of 3V view structures (Pulmonary Artery, Superior Vena Cava, Descending Aorta) and 6480 LVOT structures (Right Atrium, Left ventricle-Aorta continuum and Right Ventricle). A rotational augmentation scheme is applied with angular rotations of 10 degrees considering the rotational symmetry of the actual acquisition process.Instances sourced from 10 subjects are used for training sets and the rest for validation.

Training. For initial training, the number of base classes (N) are taken as 3. In the (N+1)th task (N¿1) following the creation of exemplar sets of the immediate past task, the training process is started for the (N+1)th task and so on (this goes on for 3 sessions in total in our case as we deal with a total of 9 classes of sub-anatomies). Batches are created between old and new data, and to further improve performance a distillation based regularization with representative logits of the past tasks is used along with the cross-entropy loss for the present task and exemplars. The training stage essentially involves a base training with the first set of classes, carried over 50 epochs with a learning rate of 0.001. This is followed by a class activation mapping stage with the recently trained model, and an averaged saliency map calculation per class. This model is now fine-tuned for the group of classes for the next task with a joint distillation and cross-entropy loss over past logits and new labels for 50 epochs. The process continues over the remaining task sessions. CAM stages are carried out only after entire task session is completed and not in-between epochs. In those models where past exemplars are retained and rehearsed, these are interspersed with the batches during fine-tuning over the new class sets.

Baselines. For baseline comparisons, the network model described is adapted with the protocol in iCaRL [5], where the representation learning stages are followed by a storage of class-specific exemplars using a herding algorithm [5]. This is implemented in our datasets by computing average prototype representations through the penultimate fully-connected layers for the classes seen till the previous task. A multi-class adaptation of Learning without Forgetting (LwF.MC) [7] is attempted with our network and objective function without the weight regularization term, and logits for distillation are retained. For the end-to-end learning (E2EL) [8] method adapted to our network, a fixed memory version is followed to be a more accurate benchmark to our own fixed memory per class assumptions. The representative memory fine-tuning protocol in E2EL is implemented for baselines with the same training configurations as the initial training, except that the learning rate is reduced to 1/10th the initial value. A progressive distillation and retrospection scheme (PDR) [9] is implemented with replicated versions of our network serving as the teacher network for the distillation and the retrospection phases, with the exemplars generating retrospection logits for regularization while progressively learning on new data which are presented as a second set of logits which have been learned separately in another replication of the base network. In these implementations, storage of prior exemplars follows the same protocols as used in the original implementations.

4 Results and Discussion

We consider configurations of our approach in terms of the exemplar retention method used and adopt OSQ in tandem: 1)map quality with OSQ and ARDS, 2)map quality with OSQ + DEDS. Variants of our approach without storing rehearsal exemplars are also considered in terms of training a vanilla CNN with similar architecture as the base CNN used in other approaches. This is same as using a simple CNN baseline with transfer learning over task sessions. Another version CNN(TL) also functions without retention strategies but with convolutional layers frozen while further task finetuning(this would reflect in a much higher difference between initial task performance and subsequent task values).The reported performances include the average accuracies at incremental levels, with and without using the salient retention scheme in step 2, benchmarked with our adaptations of methods in iCaRL[5], LwF.MC[7], E2EL[18] and Progressive distillation (PDR)[19]. Using these benchmarks indirectly also allow comparison between different exemplar retention strategies, such as with naive and herding based methods explored in iCaRL and LwF. Also reported is the average saliency map quality in each stage. This OSQ metric is a proxy for the level of forgetting over multiple stages, and a difference between consecutive stages in the OSQ represents the decline in the models ability to seek out most salient image regions over classes. Also, the OSQ is an indicator of stagewise model interpretability since accurate model explanations rely directly on the quality of saliency maps for medical images.

The reported OSQ over learning sessions is the saliency quality averaged over all validation exemplars available for previously seen classes. We report the average value for tasks so far, since we want to look at broad trends in the overall saliency to assess the overall ability to retain knowledge. For future extensions, it is straight-forward to obtain these values at both class and instance level and only requires them to be input to the trained model and class activation maps processed before the OSQ calculation. There is a difference in accuracies for baseline methods compared to original implementations due to our using the same base network for all baselines and models for uniform assessments (e.g. iCarL originally used embeddings from 32 layer ResNet on CIFAR 100 but we use the iCaRL baseline on our data and our base network). A trend is established where the inclusion of exemplars based on a saliency driven approach is seen to have a marked improvement on mitigation of forgetting, based on the OSQ metrics introduced here, and also on the validation accuracies averaged on previously seen classes for the task sessions considered (the past accuracy % in Table 1 for a task stage refers to validation accuracy of validation sets from previously seen classes, and the present accuracy % is the validation accuracy obtained on the present validation set). The saliency quality variation roughly corresponds to the past accuracy percentages demonstrating the efficacy of using map quality as a metric for evaluation of continual learning algorithms in medical image analysis. The methods that consider the retention of exemplars from the past overall show not only a better performance with respect to past task accuracies, but also demonstrate considerably higher current learning performances. This implies that a feature importance based identification of informative examples for classes of physiological markers not only improves the ability to better rehearse on past data during future learning stages but also transfers salient representative knowledge leading to better initialization of the parameters for improving performances on the present as well.Here, the diversity of the examples that can constitute a single class type representing a physiological region requires that multiple salient features can be used to explain the final optimization decisions and a diversity of informative exemplars need to be chosen for optimal forward propagation of knowledge while learning on future increments of tasks. The proposed pipeline ensures an inherent continual explainability of the decisions and how they shift over new data arrivals.

Conclusion. In this proof-of-concept for saliency aware continual learning paradigms, we presented metrics for assessment of continual learning in terms of saliency allowing for instance and class level understanding of the basis for prediction and a shift in the learning of the same. We also utilised the saliency from the past task as a selected representative for prior tasks during subsequent learning and developed a joint curriculum for creating such sets of exemplars. Our method makes the continual learning process interpretable to a degree,

Table 1. Evolution of model performance over task sessions.Past accuracy refers to validation accuracy on past session classes (or past tasks). Map quality (MQ) is reported for present task session (leftmost column for each task session head) and for previous session classes.

Method	**Task 1**		**Task 2**				**Task 3**			
	MQ	***Task 1 acc %***	***MQ (T2)***	***MQ (T1)***	***Task 2 acc. %***	**Past acc %**	***MQ (T3)***	***MQ (T2)***	***Task 3 acc. %***	**Past acc %**
Ours(OSQ + ADRS)	0.933	0.812	0.942	0.915	0.845	0.704	0.913	0.876	0.632	0.568
Ours(OSQ + DEDS)	0.946	0.812	0.938	0.923	0.863	0.691	0.887	0.843	0.636	0.593
OSQ + std. CNN	0.871	0.811	0.840	0.631	0.778	0.592	0.852	0.802	0.661	0.455
OSQ + CNN(TL)	0.827	0.813	0.822	0.612	0.702	0.511	0.772	0.647	0.560	0.322
iCaRL	0.811	0.775	0.831	0.622	0.713	0.616	0.834	0.674	0.658	0.321
LwF.MC	0.773	0.762	0.767	0.668	0.732	0.529	0.822	0.731	0.621	0.301
E2EL	0.818	0.793	0.792	0.703	0.742	0.554	0.785	0.706	0.603	0.295
PDR	0.842	0.802	0.837	0.711	0.759	0.514	0.791	0.721	0.581	0.342

and thereby ensures that the forgetting and retention characteristics of models are explainable. Given the foundations laid here, multiple future directions are possible starting with an exploration of classwise chracteristics in terms of forgetting and retention performances and the trends of decline in map quality over intra-class variations. This is likely to be a natural follow-up of the ideas proposed here. Another immediate direction would be to study other strategies for saliency-curriculum driven exemplar retention. While we used a fixed number of exemplars for retention and rehearsal over future tasks, other approaches like a variable retention based on class difficulty are possible. Future directions can also include expanding to different tasks and datasets, using the saliency based exemplar scheme with other lifelong learning methods, using generative replay for estimating past saliency maps and images without need to retain exemplars, and so on. While we have demonstrated on approaches on distillation-based preservation, and using class activation derived saliency maps, the concept is generally applicable with any other continual learning pipeline, and can use other methods of estimating saliency maps, with different base architectures or objectives.

5 References

1. A. Patra, W. Huang, and J. A. Noble, Learning Spatio-Temporal Aggregation for Fetal Heart Analysis in Ultrasound Video', In DLMIA 2017.

2. F.Ozdemir et al,'Learn the new, keep the old: Extending pretrained models with new anatomy and images." MICCAI 2018.

3. I. Goodfellow et al., An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks, https://arxiv.org/abs/1312.6211, 2015.

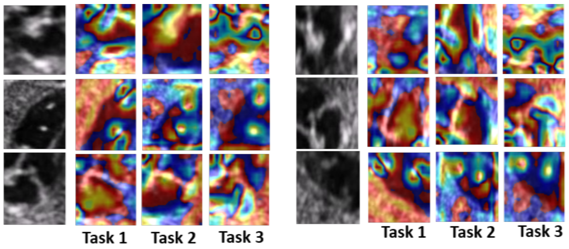

Fig. 3. Map quality variation over successive tasks sequentially, for class examples seen originally in task session 1. Saliency maps for these task 1 validation instances show a shift in attentive features over subsequent task sessions (areas in red are of higher salience) where models learn other classes. Quantification of this shift with the OSQ is treated as a proxy metric for forgetting here.

4. G. I. Parisi, R. Kemker, J. L. Part, C. Kanan and S. Wermter,'Continual Lifelong Learning with Neural Networks: A Review',https://arxiv.org/abs/1802.07569.

5. S. A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, iCaRL: Incremental classifier and representation learning, in CVPR 2017.

6. R. Aljundi, P. Chakravarty, P., and T. Tuytelaars, Expert Gate: Lifelong Learning with a Network of Experts, In ICCV 2017.

7. Z. Li and D. Hoiem, "Learning without Forgetting", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-1, 2017.

8. J. Kirkpatrick et al., Overcoming catastrophic forgetting in neural networks, in Proc. Nat. Acad. Sci. ,vol. 14, Mar 2017.

9. Zenke et al.,'Continual learning through synaptic intelligence',ICML 2017.

10. S. J. Pan, and Y. Qiang, "A survey on transfer learning", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no.10, pp. 1345-1359, 2010.

11. A. A. Rusu et al.,Progressive neural networks,arxiv:1606.04671, 2016

12. H.E.Kim et al.,'Keep and Learn: Continual Learning by Constraining the Latent Space for Knowledge Preservation in Neural Networks. In MICCAI 2018.

13. J. Schlemper et al.,'Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images' arXiv preprint arXiv:1808.08114.

14. O. Oktay et al., Attention U-Net: Learning Where to Look for the Pancreas. arXiv preprint arXiv:1804.03999.

15. R. R. Selvaraju et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In ICCV 2017.

16. Dabkowski, P., Gal, Y., Real time image saliency for black box classifiers. In Advances in Neural Information Processing Systems 2018.

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

18. Castro, F. M., Marn-Jimnez, M., Guil, N., Schmid, C., Alahari, K. (2018, September). End-to-end incremental learning. In ECCV 2018.

19. Hou, S., Pan, X., Loy, C. C., Wang, Z., Lin, D., Lifelong learning via progressive distillation and retrospection. In ECCV 2018.