Continual learning: an approach via feature maps extrapolation

Edoardo De Rose[1,*,†]

[1]Department of Mathematics and Computer Science, University of Calabria, Italy

Abstract

The semantic image segmentation task consists of classifying each pixel of an image into an instance, where each instance corresponds to a class. This task is a part of the concept of scene understanding or better explaining the global context of an image. In the medical image analysis domain, image segmentation can be used for image-guided interventions, morphology characterization, and diagnostics. The state-of-the-art methods for this task use deep convolutional neural networks, which can learn from data how to perform the segmentation. The recent advances in deep learning allow training networks on small datasets, which is a critical issue for bio-medical images. Moreover, these methods are designed for a static learning scenario, where the data distribution does not change over time. This is not realistic for the dynamic medical imaging environment, where new tasks and data may appear continuously. Therefore, we propose a new continual learning approach, which aims to enable neural networks to learn new tasks sequentially without forgetting the previous ones. The main challenge in this learning scenario is to prevent catastrophic forgetting, which occurs when the network overwrites the knowledge of previous tasks with the knowledge of the current task. The goal is to develop a method that can overcome this challenge and allow models to learn continuously and accumulate knowledge from different tasks.

Keywords

Medical imaging, Deep Learning, Semantic segmentation, Continual learning

1. Introduction

Medical imaging techniques are used to create visual representations of the internal anatomy of the human body for clinical analysis and medical intervention. Medical imaging techniques include X-rays, magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and ultrasound imaging. Extrapolating information from these images requires, for example, accurate semantic segmentation of the anatomical parts. This is a key process where a given input signal is divided into constituent regions or partitions [1]. The pixels of a partition should share some local properties such as intensities, continuity and regularity of the signal, variance, texture information, and others. With the advancement of technology and the availability of data, Artificial Intelligence (AI) has become an essential tool for medical professionals to improve patient outcomes, optimize treatment plans, and facilitate the diagnosis of various medical conditions. Recently, Deep Learning (DL) techniques

AIxIA'23: 22nd International Conference of the Italian Association for Artificial Intelligence, November 06–09, 2023, Rome, Italy

*Corresponding author.

†Doctoral project supervised by Francesco Calimeri (University of Calabria) and Pierangela Bruno (University of Calabria).

edoardo.derose@unical.it (E. D. Rose)

gained significant attention in handling a lot of computer vision problems [2]. Specifically, the Convolutional Neural Networks (CNN), the Fully CNNs (FCNs) and Transformers-based [3] architectures achieved significant success and rapidly became state-of-the-art methodologies in medical image segmentation and classification [4] due to their ability to automatically learn relevant features from images and provide accurate results. However, this remarkable success is achieved in a static learning paradigm where the model is trained using large training data of a specific task and deployed for testing on data with a similar distribution to the training data. This paradigm contradicts the real dynamic world medical environment which changes very rapidly. Standard retraining of the neural network model on new data leads to significant performance degradation on previously learned knowledge, a phenomenon known as catastrophic forgetting [5]. Continual learning (CL) approaches come to address this dynamic learning paradigm. It aims at building neural network models capable of learning sequential tasks while accumulating and maintaining the knowledge from previous tasks without forgetting. In general, the main components of a continual learning problem are:

- a sequence of tasks $1; 2; ...; t; ...; T$ where T is the total number of tasks;
- each task t refer to its own dataset D_t;
- the neural network model faces tasks one by one;
- the capacity of the model should be utilized to learn the sequence of the tasks without forgetting any of them;
- all samples from the current task are observed before switching to the next task;
- the data across the tasks is not assumed to be identically and independently distributed.

In this paper, we propose a new method to solve continual learning problems, in the field of semantic segmentation, that is based on two main intuitions:

- selecting the model weights that are relevant for each task during the training phase;
- updating the weights that only learn new features during the backpropagation without forgetting the previous ones.

Our method is motivated by the observation that different tasks may have different data distributions, but they may also share common features that are learned by the model. We tested our method on a case study of computed tomography (CT). It is a 3D x-ray imaging technique that generates 3D digital gray-scale images of the organs' internal structures. These images can be semantically segmented to identify specific morphological components, such as tissues, vessels, or tumors. Continual learning methods are useful for CT image segmentation because they allow the system to segment different organs or diseases as they become available or relevant, without losing the ability to segment the ones that were learned before. This can lead to more efficient and general segmentation of CT images of various organs and diseases.

1.1. Related Works

Several continual learning methods have been proposed to tackle catastrophic forgetting. Following De Lange et Al [6] continual learning algorithms can be divided into three general groups. The first group consists of replay-based methods that build and store a memory of the

knowledge learned from old tasks. iCaRL [7] learns in a class-incremental way by having a fixed memory that stores samples that are close to the center of each class while ER-Reservoir [8] uses a Reservoir sampling method as its selection strategy. The methods in the second group use explicit regularization techniques to supervise the learning algorithm such that the network parameters are consistent during the learning process. As a notable work, Elastic weight consolidation (EWC) [9], uses the Fisher information matrix as a proxy for weights' importance and guides the gradient updates. Some other regularization-based methods have utilized gradient information to protect previous knowledge, like Orthogonal Gradient Descent (OGD) in [10] uses the projection of the prediction gradients from new tasks on the subspace of previous tasks' gradients to maintain the learned knowledge. Finally, in parameter isolation methods, in addition to potentially a shared part, different subsets of the model parameters are dedicated to each task [11, 12, 13]. This approach can be viewed as a flexible gating mechanism, which enhances stability and controls plasticity by activating different gates for each task.

2. Experimental design and tasks

We evaluated and validated our continual learning method on a dataset of CT images, but it can be generalized to any dataset. The tasks involve the semantic segmentation of CT images of insects, with an emphasis on their internal organs, such as testicles and glands. We treated each organ segmentation as a different task; to deal with the continual learning scenarios, the model observes all samples from the current task before moving to the next task, and we assumed that the data from the previous tasks is inaccessible after learning them. As baseline models to perform semantic segmentation we used two state-of-the-art deep learning models: SegNet [14] and U-Net [2]. U-Net and SegNet are both convolutional networks for biomedical image segmentation that use an encoder-decoder structure. U-Net uses skip connections to transfer features from the encoder to the decoder, while SegNet uses non-linear upsampling with pooling indices to do the same. In general, U-Net also has more feature channels in the upsampling path than SegNet.
For the specific task and dataset, we fine-tuned and optimized the models. The dataset consisted of 30 insect samples, differentiated by age and type, and approximately $4000 \times 2048 \times 2048$ images reconstructed for each sample. We measured the models' performance on various metrics, such as intersection over union. We contrasted the models' results on training all organs jointly or training them on each organ individually. We employed these results to compare and assess our proposed continual learning method and how much knowledge is retained by the models with this strategy.

3. Continual learning strategy

Once the continual learning problem is defined, we performed training and optimization of parameters and weights of the model for the first task T_1. Next, we designed the training strategy for the next tasks T_2, without the model forgetting the previous task. The first step in devising the continual learning training strategy was to examine the similarities and differences among the features learned by the models for the old and new tasks. We aimed to identify which parts

of the models could be shared and which parts required fine-tuning for next tasks. To do this we used the cosine similarity between tensors:

$$S_{l,i} = \frac{F_{l,i}^{n} \cdot F_{l,i}^{m}}{||F_{l,i}^{n}||||F_{l,i}^{m}||} \tag{1}$$

where n, m indicates the different tasks, l indicates the layer, and i the images features. By comparing these for different tasks, we identified which ones were shared and which ones were specific. This allowed us to optimize our models and reduce the computational cost of continual learning training. Once the model parameters, that need to be changed to learn new tasks, were determined we defined a strategy to update the weights preventing catastrophic forgetting. In the high-dimensional parameter space of the model just trained, there could be update directions causing large changes in the predictions from $x \in T_1$, while there also exist updates that minimally affect such predictions. In particular, moving locally along the j-th direction of the gradient of the model $\pm\nabla f_j(x; w)$, where w are the weights, leads to the biggest change in model prediction $f_j(x; w)$. Inspired by M. Farajtabar et Al [15] we updated the weights moving orthogonal to $\nabla f_j(x; w)$. This leads to the least change (or no change, locally) to the prediction of $x \in T_1$ and a learning direction for the prediction $y \in T_2$. As mentioned above, although tasks 1 and 2 are different and have different distributions, they may have some similar features that could interfere with the finding of orthogonal directions. From this intuition, we deduced that before calculating the possible orthogonal directions, the similar features between the tasks should be eliminated on the new task, because they were already part of the model's knowledge. This is done using the Grad-CAM [16] approach, which highlights the important weights for the model's old predictions. Finally to perform the entire steps just described, in particular the orthogonal gradient descent, the only things that need to be stored from the previous task are the old gradient. These are necessary to calculate the new orthogonal directions for the new task.

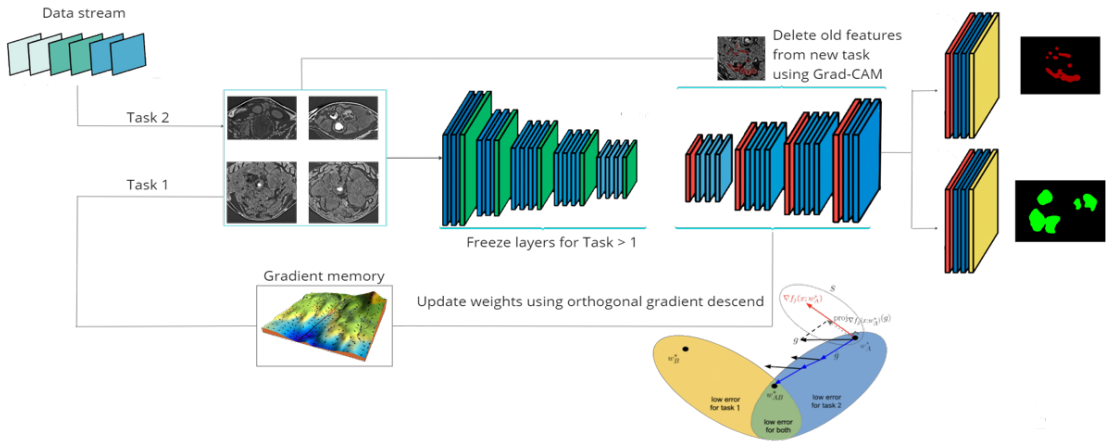

Figure 1: Continual learning workflow proposed in this paper.

The continual learning strategy just proposed is shown in figure 1 and can be resumed in these steps:

1. The model is trained for the current task and their gradients are stored;
2. Layers that need to be updated for the next task are identified using the cosine similarity;
3. The weights of the layers that were common or "similar" to both tasks are frozen;
4. The features in the new task similar to the old task are deleted using Grad-CAM;
5. The model is trained on the new task using the orthogonal gradient direction.

4. Conclusions

We proposed a new method for continual learning of image segmentation, inspired by M. Farajtabar et Al [15]. Our method uses feature map extrapolation, orthogonal gradient descent, and similarity measures to find out which layers and weights need to change for each new task. We freeze the weights of the shared layers and only use the old gradients to avoid forgetting the previous task and learning the new task. Our method is also faster and cheaper than other methods because we optimize the parameters for each task and we only save the old gradients instead of reply-based methods.

Acknowledgments

I am thankful to my supervisors, Francesco Calimeri and Pierangela Bruno, for introducing and guiding me to this exciting and challenging research field. I also appreciate the University of Calabria for offering this engaging doctoral program and for the high-quality education that I received during my studies.

References

[1] L. Carvalho, A. C. Sobieranski, A. von Wangenheim, 3d segmentation algorithms for computerized tomographic imaging: a systematic literature review, Journal of digital imaging 31 (2018) 799–850.

[2] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[3] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, M. Shah, Transformers in vision: A survey, ACM computing surveys (CSUR) 54 (2022) 1–41.

[4] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical image analysis 42 (2017) 60–88.

[5] M. McCloskey, N. J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, in: Psychology of learning and motivation, volume 24, Elsevier, 1989, pp. 109–165.

[6] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, T. Tuytelaars, A continual learning survey: Defying forgetting in classification tasks, IEEE transactions on pattern analysis and machine intelligence 44 (2021) 3366–3385.

[7] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, C. H. Lampert, icarl: Incremental classifier and representation learning, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 2001–2010.

[8] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. Torr, M. Ranzato, On tiny episodic memories in continual learning, arXiv preprint arXiv:1902.10486 (2019).

[9] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, Proceedings of the national academy of sciences 114 (2017) 3521–3526.

[10] M. Farajtabar, N. Azizan, A. Mott, A. Li, Orthogonal gradient descent for continual learning, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 3762–3773.

[11] J. Yoon, E. Yang, J. Lee, S. J. Hwang, Lifelong learning with dynamically expandable networks, arXiv preprint arXiv:1708.01547 (2017).

[12] G. Jerfel, E. Grant, T. Griffiths, K. A. Heller, Reconciling meta-learning and continual learning with online mixtures of tasks, Advances in neural information processing systems 32 (2019).

[13] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, R. Hadsell, Progressive neural networks, arXiv preprint arXiv:1606.04671 (2016).

[14] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE transactions on pattern analysis and machine intelligence 39 (2017) 2481–2495.

[15] M. Farajtabar, N. Azizan, A. Mott, A. Li, Orthogonal gradient descent for continual learning, in: S. Chiappa, R. Calandra (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 3762–3773. URL: https://proceedings.mlr.press/v108/farajtabar20a.html.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.