Continual Classification Learning Using Generative Models

Frantzeska Lavda[*]
Frantzeska.Lavda@etu.unige.ch

Jason Ramapuram[*]
Jason.Ramapuram@etu.unige.ch

Magda Gregorova[*]
Magda.Gregorova@hesge.ch

Alexandros Kalousis[*]
Alexandros.Kalousis@hesge.ch

Abstract

Continual learning is the ability to sequentially learn over time by accommodating knowledge while retaining previously learned experiences. Neural networks can learn multiple tasks when trained on them jointly, but cannot maintain performance on previously learned tasks when tasks are presented one at a time. This problem is called catastrophic forgetting. In this work, we propose a classification model that learns continuously from sequentially observed tasks, while preventing catastrophic forgetting. We build on the lifelong generative capabilities of [10] and extend it to the classification setting by deriving a new variational bound on the joint log-likelihood, $\log p(x, y)$.

1 Introduction

Continual learning tries to mimic the ability of humans to retain or accumulate previous knowledge and use it to solve future problems with possible adaptations. In this paper we propose a new method for continual learning in the classification setting. Our model combines the encoder and decoder of a variational autoencoder (VAE) [6] with a classifier. To do this we derive a new variational bound on the joint log-likelihood $\log p(x, y)$.

To enable the continual discriminative learning we build on the work of Ramapuram et al. [10] on lifelong generative modelling. The model has a student-teacher architecture (similar to that in distillation methods, [4], [2]), where the teacher contains a summary of all past distributions and is able to generate data from the previous tasks once we no longer have access to the original data. Every time a new task arrives, a student is trained on the new data together with the data generated by the teacher from the old tasks. The proposed method thus does not need to store the previous models (it only stores their summary within the teacher model) nor data from the previous tasks (it can generate them using the teacher model).

1.1 Related work

Several approaches have been proposed to solve catastrophic forgetting over the last few years. We can roughly distinguish 2 streams of work: a) methods that rely on a dynamic architecture that evolves as they see new tasks b)methods with regularization approaches that constrain the models learned in new tasks so that the network avoids modifying the important parameters of the previous tasks. In dynamic architectures parameters of the models learned on the old tasks are passed over to the new tasks while the past models for each task are preserved ([11], [1]). In contrast, our method does not need to keep the past models. Regularization approaches ([7], [13]) impose constrains to the objective function to minimize changes in parameters important for previous tasks. However, these methods need to store the parameters of the previous tasks, something that is not required in our proposed method.

In Variational Continual Learning, [9], the authors propose a method which is applicable to discriminative and generative models but not both at the same time while our method is. While VCL shows

[*]University of Geneva & Geneva School of Business Administration, HES-SO

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

rather impressive results, it achieves those relying on the reuse of some of the previous data through the use of *core-sets* and by maintaining task-specific parameters, called *head networks*. It therefore relaxes the continual learning paradigms of no access to past data and no storage of past task-specific models; paradigms that our method fully takes on board.

2 Model

In the continual classification setting, we deal with data that come sequentially in pairs $(\mathbf{X}, \mathbf{Y}) = \{(x_1, y_1), ..., (x_n, y_n)\}$. For each task j the network receives a new data set $\{(x_j, y_j)\}$ and does not have access to any of the data sets of the previously seen tasks.

To perform the classification, we use a latent variable model as shown in Fig.1. In this model, each observation x has a corresponding latent variable z, that is used to generate the correct label class y. The joint distribution of the latent variable model that we consider factorizes as $p(x, y, z) = p(x|z)p(y|z)p(z)$ where (x, y) are labeled data pairs and z are the latent variables. The data variables x, y are assumed to be conditionally independent given the latent variables z, $((x \perp y)|z)$, such that $p(x, y|z) = p(x|z)p(y|z)$.

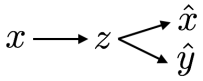


Figure 1: Graphical model

Following the classical VAE approach we will use variational inference to approximate the intractable posterior $p(z|x, y)$. Instead of the natural $q(z|x, y)$ we use $q_\phi(z|x)$ to approximate the true posterior $p(z|x, y)$ since in the test phase of the classification y is not available. To measure the similarity between the true posterior $p(z|x, y)$ and its approximation $q(z|x)$ we minimize the Kullback-Leibler divergence between the approximate posterior and the true posterior.

$$D_{KL}\left(q_\phi(z|x)\|p(z|x,y)\right) = -E_{q_\phi(z|x)}\left[\log p(x,y,z) - \log q_\phi(z|x)\right] + \log p(x,y) \tag{1}$$

The term $\log p(x, y)$ in Eq.1 is a constant. This means that in order to minimize the KL-divergence we minimize $-E_{q(z|x)}\left[\log p(x, y, z) - \log q_\phi(z|x)\ \right] = -L(x, y)$ which is equivalent to maximizing $L(x, y)$.

$$L(x,y) = E_{q_\phi(z|x)}\left[\log p(x,y,z) - \log q_\phi(z|x)\ \right] \tag{2}$$

Rearranging Eq.1 as:

$$\begin{aligned}\log p(x,y) &= E_{q_\phi(z|x)}\left[\log p(x,y,z) - \log q_\phi(z|x)\ \right] + D_{KL}\left(q_\phi(z|x)\|p(z|x,y)\right)\\ &= L(x,y) + D_{KL}\left(q_\phi(z|x)\|p(z|x,y)\right)\end{aligned} \tag{3}$$

we can see that the $L(x, y)$ is a lower bound of the *joint* log-likelihood, $\log p(x, y)$: a new variational bound for the joint generative and discriminative VAE learning.

To gain better intuition into our newly derived variational bound, we show the relation to the classical ELBO (variational bound on the marginal likelihood $p(x)$) used in VAEs. Rearranging the terms in Eq.3, under the conditional independence assumption $p(x, y|z) = p(x|z)p(y|z)$ and using the fact that the KL-divergence is always positive, we arrive at:

$$\begin{aligned}\log p(x,y) \geq L(x,y) &= E_{q_\phi(z|x)}\left[\log p(x,y,z) - \log q_\phi(z|x)\right]\\ &= \underbrace{E_{q_\phi(z|x)}\left[\log p(x|z)\right] - D_{KL}\left(q_\phi(z|x)\|p(z)\right)}_{\text{ELBO}} + \underbrace{E_{q_\phi(z|x)}\left[\log p(y|z)\right]}_{\text{classification loss}}\end{aligned} \tag{4}$$

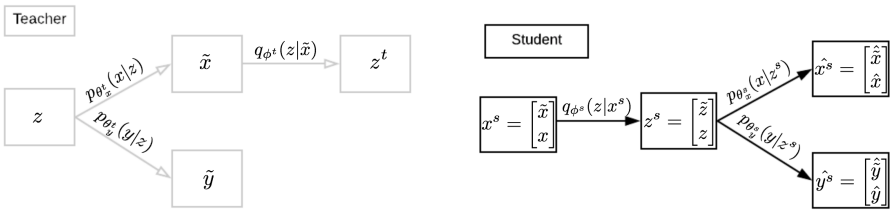
The first term $E_{q(z|x)}\left[\log p(x|z)\right] - D_{KL}(q(z|x)\|p(z))$, as in the standard VAE is the variational bound on the marginal likelihood $p(x)$ (ELBO). The second term $E_{q(z|x)}\left[\log p(y|z)\right]$ is the expectation of the conditional log-likelihood of the labels y on the latent variable z, the classification loss. This term allows our variational bound to be used in classification settings. This means that we solved the two problems of producing the labels y, and generating input data x jointly, resulting in a common latent variable z which is good for classification and reconstruction at the same time.

Furthermore, it is easy to show that under our conditional independence assumption $p(z|x, y)/p(z|x) = p(y|z)/p(y|x)$. Assuming that z summarizes x well for the classification of y ($p(y|z) \approx p(y|x)$) both of the ratios are close to 1. Replacing the intractable posterior $p(z|x)$ by the approximation $q_\phi(z|x)$ results in $p(z|x, y)/q_\phi(z|x) \approx 1$ which is what the minimization of the

KL-divergence in Eq.(1) tries to achieve. This therefore provides and alternative argument for the validity of our approach described above.

Our goal in this paper is to correctly classify data from different tasks that arrive continuously, requiring us to handle the catastrophic forgetting problem. For this we use the lifelong generative ability of [10] and extend their VAE based generative model to include a classifier that remembers all the classification tasks it has seen before. The method uses a dual architecture based on a student-teacher model. The main goal of the student model is to classify the input data. The teacher model's role is to preserve the memory of the previously learned tasks and to pass this knowledge onto the student.

Both the teacher and the student consist of an encoder $q_{\phi^m}(z|x)$, a decoder $p_{\theta_x^m}(x|z)$ and newly a classifier $p_{\theta_y^m}(y|z)$ following the graphical model in Fig.1. In the above notation $m \equiv t, s$ represents the teacher and student model respectively. The teacher model remembers the old tasks and generates data from them $\{(\tilde{x}, \tilde{y})\}$ for the student to use in learning once the old data are no longer available. The student model learns to generate and classify over the new labeled data pairs $\{(x, y)\}$ and the old-task data genrerted by the teacher $\{(\tilde{x}, \tilde{y})\}$. Every time a new task is initiated, the student passes the latest parameters to the teacher and starts learning over data from the new task augmented with data generated by the teacher from all the previous tasks. In this way the acquired information of the previous tasks is preserved and the proposed model learns to classify correctly even over data distributions seen in previous tasks. The proposed architecture does not need to store the task-specific models for the previous data distributions nor the previous data themselves.



(a) (b)

Figure 2: The architecture of the learning procedure. Fig.2a The teacher model generates input-output pairs from the previously seen tasks and passes them onto the student. Moreover the teacher evaluates the posterior $q(z|\tilde{x})$ Fig.2b The student model learns to classify and generate new data augmented by data from the teacher.

The student optimizes the variational bound of the joint log-likelihood $\log p(x, y)$ Eq.(4) instead of the marginal log-likelihood $\log p(x)$ over which the classical VAE operates. As a result our model is able to both generate the input data x and learn the labels y at the same time. We should note that previous approaches to classification with VAEs ([5]) do so by adding an ad-hoc manner to the VAE optimization function terms that relate to classification performance. Here we naturally extend the VAE setting to classification.

Following [10] we add an additional term ($D_{KL}[q_{\phi^s}(z^s|\tilde{x})\|q_{\phi^t}(z^t|\tilde{x})$) to our objective to preserve the posterior representation of all previous tasks to speed up the training and a negative information gain regularizer, $L_I(z, \tilde{x})$, between the latent representation z and the generated data $\tilde{x}$ from the teacher. The final loss that we optimize is given by Eq.5.

$$E_{q_{\phi^s}(z|x^s)}\left[\log p_{\theta_x^s}(x|z^s) + \log p_{\theta_y^s}(y|z^s)\right] - D_{KL}(q(z|x)\|p(z)) - D_{KL}[q_{\phi^s}(z^s|\tilde{x})\|q_{\phi^t}(z^t|\tilde{x})] - L_I(z, \tilde{x}) \tag{5}$$

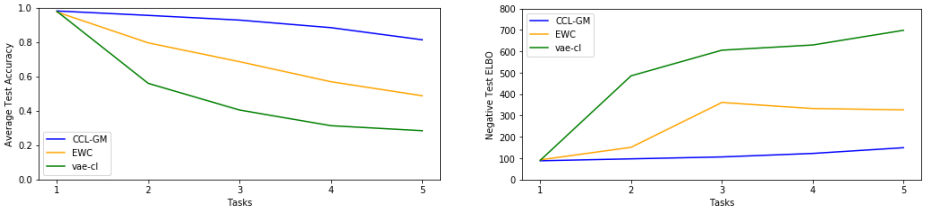
3 Experiments

In this section we present preliminary results achieved with the proposed model. We investigate the problem of whether our model is able to learn a set of different tasks that are coming in sequence without forgetting the previously trained tasks.

We evaluated our approach for continual learning on permuted MNIST [8], [3]. Each task is a 10-way classification (0-9 digits) over images with the pixels shuffled by a random fixed permutation. We train on a sequence of 5 tasks (original MNIST and 4 random permutations). After the training of each task we allow no further training or access to that task's data set. For training we process the data in mini batches of 256 (random data shuffling) and use early stopping on the classification accuracy.

We use two baseline models for comparisons. The first is a standard VAE augmented by our classifier (vae-cl) using our variational bound but without the teacher-student architecture. In the second, we adapt the elastic weight consolidation (EWC) regularisation approach of [7] to our setting. We use the teacher here to keep the summary of all the previous distributions[2] and employ the EWC-like regularisation $\sum_i F_i(\psi_i^s - \psi_i^t)^2$ over the parameters of the teacher and student ($\psi^m = [\theta^m, \phi^m]$) models where $F_i = \text{diag}E[(\partial L_\psi(x, y)/\partial\psi)^2]$.

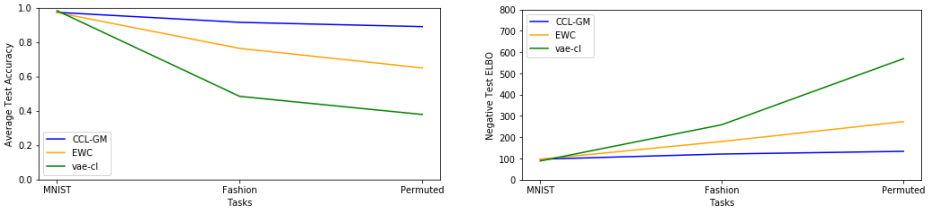
We measure the performance by the ability of the network to solve all tasks seen up till the current point. For all tested methods we performed a random hyper-parameter sweep over convolutional and dense network architectures. We present the results of the best obtained models[3] in Fig. 3. For the naive vae-cl method the performance drops dramatically already when the training regime switches from the MNIST to the first permuted task. For the EWC method the performance after the first task degrades less severely, but it still forgets the previous tasks. Our model, continual classification learning using generative models (CCL-GM) retains high average classification accuracy Fig. 3a and low average reconstruction ELBO 3b. This shows that our model is able to learn continuously and concurrently for both classification and generation.



(a) Average test classification accuracy (b) Average test negative reconstruction ELBO

Figure 3: Average performance over all learned tasks from the permuted MNIST data set as a function of the number of tasks. Our approach, CCL-GM maintains high accuracy and low negative ELBO as the number of tasks increases. Vanilia VAE our classifier performs far worse. EWC degrades less severely, but still forgets the previous tasks

To support our initial results from the above experiments, we conducted a second set of experiments on a sequence of three different tasks: MNIST, FashionMNIST [12] and one MNIST permutation. The results presented in Fig. 4 show that our method outperforms the baselines and confirm our preliminary conclusions that our new model CCL-GM has the ability to mitigate catastrophic forgetting in joint generative and discriminative problems.



(a) Average test classification accuracy (b) Average test negative reconstruction ELBO

Figure 4: Average performance over all learned tasks.

4 Conclusion

In this work we propose a method to address continual learning in the classification setting. We use a generative model to generate input-output pairs from the previously learned tasks and use these to augment the data of the current tasks for further training. In this way our classification model overcomes catastrophic forgetting. Our model does not reuse data nor previous task-specific models and it continuously learns to concurrently classify and reconstruct data over a number of different tasks.

[2] In our EWC baseline the teacher is not used to generate data for the student

[3] Convolutional for ours and vae-cl, dense for EWC

References

[1] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

[2] Tommaso Furlanello, Jiaping Zhao, Andrew M Saxe, Laurent Itti, and Bosco S Tjan. Active long term memory networks. *arXiv preprint arXiv:1606.02355*, 2016.

[3] 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.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[5] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[7] 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. *arXiv preprint arXiv:1612.00796*, 2016.

[8] Yann LeCun, Corinna Cortes, and CJC Burges. The mnist dataset of handwritten digits. 1998.

[9] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. In *International Conference on Learning Representations, ICLR*, 2018.

[10] Jason Ramapuram, Magda Gregorova, and Alexandros Kalousis. Lifelong generative modeling. *arXiv preprint arXiv:1705.09847*, 2017.

[11] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[12] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[13] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.