Continual Learning via Variational Bayesian Dropout

Lam Tran

November 29, 2024

Abstract

As an effective and simple approach for Continual Learning (CL), regularization-based mitigate forgetting by constraining updates of model parameters based on their importance to previous tasks. The more important a parameter is to an old task, the less it should be updated. However, prior methods either require significant memory to store parameter importance or struggle to balance performance across old and new tasks when faced with a long sequence of tasks. To address memory limitations, instead of relying on parameter importance, I opted to use neuron importance to penalize update of parameter. To better reduce forgetting over a lengthy task sequence, I sought an effective network compression method to efficiently utilize the model for more tasks. Therefore, inspired by the ability to structural sparsify networks of Variational Bayesian Dropout (VBD), I leveraged the drop-rates of individual neurons to quantify their importance. Furthermore, with a nature of favoring flat minima of Dropout, VBD can learn such minima for each task, hence facilitating the possibility to find a region that is good for all tasks. Empirically, my method achieved significantly higher average accuracy than most of regularization-based baselines[1] across several network sizes and CL benchmarks.

Introduction

Continual Learning (CL) aims to enable neural networks to continually learn from a sequence of tasks with different data distributions, without forgetting old tasks knowledge. In my thesis, I addressed a practical CL scenario that prohibits access to old data to replay when learning a new task, e.g. data privacy settings, and limits the use of a large generator to generate old data or an architecture-dynamic network that can expand for incoming tasks, e.g. edge-device applications. Hence, I focused on regularization-based approach which can combat forgetting without requiring access to old data or model expansion. These methods typically add an additional regularization term to penalize updates of parameters critical for the knowledge of old tasks. Some examples are Elastic Weight Consolidation (EWC) [1] and Synaptic Intelligence (SI) [2]. However, both require storing old parameters along with their importance values, doubling additional memory usage. Other works [3, 4, 5] consider importance of neurons or nodes so that memory to store importance is reduced (store connections versus store nodes). Among them, Adaptive Group Sparsity based Continual Learning (AGS-CL) [5] and Uncertainty-based Continual Learning (UCL) [4] still need to store old model parameters. Hard Attention to the Task (HAT) [3], although requires the least memory as it only saves node importance, performs poorly on long task sequences because its network sparsification is inefficient, leaving insufficient model capacity to accommodate later tasks.

Contributions. My thesis aimed to address the limitations of prior works by leveraging the strengths of Variational Bayesian Dropout (VBD). Specifically, I utilized its ability to combat overfitting and compress networks to propose a novel and effective method for quantifying node importance based on learnable dropout rates, which has the following benefits:

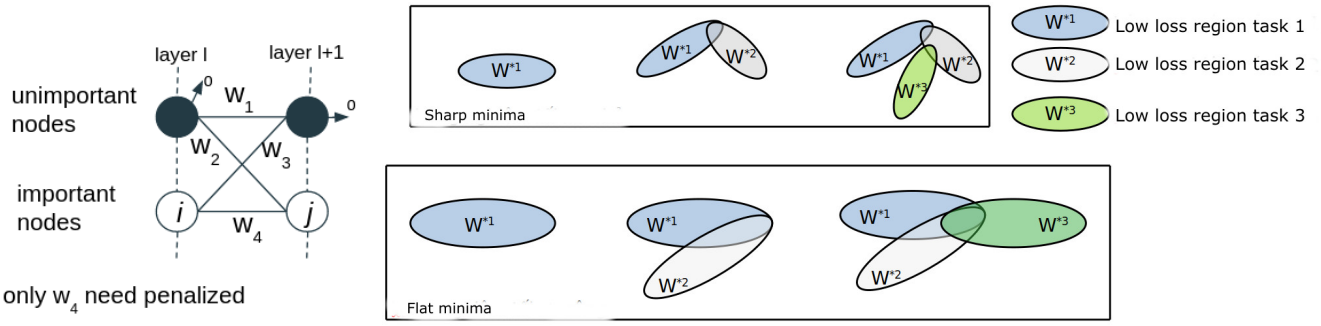
- The additional storage cost is only the node importance.
- The model can learn a long sequence of tasks due to the greate compression ability of VBD.
- The model has an 'implicit effect' of retaining old knowledge because VBD tends to find flat minima for each task, making the finding of a common low loss region for them is more feasible.

Method

Notations: Suppose we have T tasks with $D_t = \{x_i, y_i\}_{i=1}^{N_t}$ is the dataset of task t. Denote θ as model parameters. For simplicity, we consider a fully connected $W \in R^{K_{in} * K_{out}}$ between layer i^{th} layer X_i and j^{th}layer X_j.

Variational Bayesian Dropout: VBD element-wise multiplies X_i with a random noise having the same dimension with X_i, then it performs the matrix multiplication between the 'dropped' input and the weight matrix. Formally,

[1] at the time of writing this thesis, June 2021.



(a) Visualization for gradient update of weights connecting two important nodes.

(b) Flat minima for each task have a higher tendency to find a common low-loss region for all tasks.

Figure 1: Visualization for VBD-CL's gradient update and flat-minima property.

$X_j = a((X_i \odot \epsilon)W)$, where $a(.)$ is an activation function, $\epsilon \sim p(\epsilon|\gamma)$ is a random noise following a certain distribution parameterized by γ. VBD treats ϵ as learnable parameters that can be inferred from D, and further imposes a hierarchical prior on γ. Specifically, VBD assumes a fully-factorized zero-mean gaussian for the prior distribution of ϵ, $p(\mathcal{E}|\gamma) = \prod_{i=1}^{K_{in}} \mathcal{N}(0, \gamma_i)$, and a uniform in range $[a, b]$ for the prior of γ: $p(\gamma) = \prod_{i=1}^{K_{in}} \mathcal{U}(\gamma_i|a, b)$. Now to approximate the posterior distribution $p(\epsilon, \gamma|D)$, a proper choice of gaussian distribution for the variational posterior $q(\epsilon|\mu, \sigma^2)$ and delta distribution for the variational posterior of $q(\gamma|a, b)$ gives us the Evidence Lower Bound (ELBO) of the marginal likelihood: $\max_{W,\mu,\sigma} \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathcal{E})}[log\, p(y_i|x_i, W)] - 0.5 \sum_{i=1}^{K_{in}} log\left(1 + \frac{\mu_i^2}{\sigma_i^2}\right)$. To compress neurons, VBD relies on signal-to-noise ratio (SNR), $SNR_i = \frac{|\mu_i|}{\sigma_i}$, which removes a neuron by setting its output and input signals to 0 if its SNR is lower than a threshold.

Continual Learning via Variational Bayesian Dropout (VBD-CL): SNR is used to quantify neuron importance for each task, and only weights connecting two important neurons need constrained during new task updates. Indeed, since output and input signals of unimportant, i.e. pruned, neuron are set to 0, its corresponding weights can freely change without affecting output of the whole layer. This is visualized in Figure 1a. Note that, to constraint for needed weights, the more important neuron between the two is considered. Furthermore, to accumulate knowledge of all learnt tasks, importance of a neuron is taken according to the task having the biggest importance, i.e. the task needs that neuron the most. Formally, the gradient update when learning a new task t is:

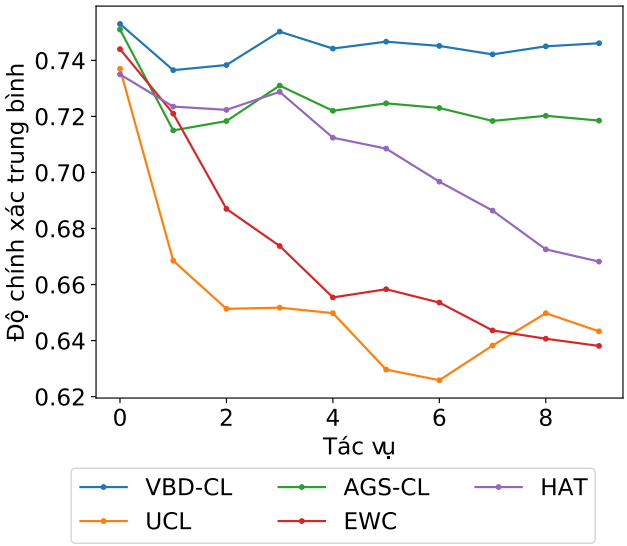
$$
\begin{aligned}
W_{ij}^{'} &= W_{ij} - \eta \Omega_{ij} \nabla ELBO, \eta \text{ is learning rate,} \\
\Omega_{ij} &= \Phi(s * [min(SNR_i^{\leq t}, SNR_j^{\leq t}) - thresh]), \\
SNR_i^{\leq t} &= min(SNR_i^{\leq t-1}, SNR_i^t), \\
\Phi(x) &= \frac{1}{1 + e^{-x}} \text{ (sigmoid function).}
\end{aligned}
$$

Here I introduced two hyper-parameters *thresh* to set the threshold for neuron pruning, and s to adjust how 'strictly' we want to preserve old knowledge, hence at the expense of losing capacity for new knowledge. By setting s at a proper value, we can have a good balance for the trade of between stability and plasticity.

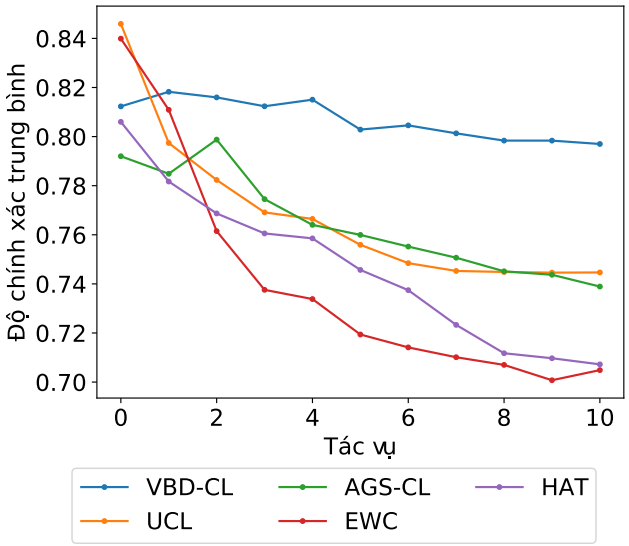
Analysis: Besides the benefit of requiring less additional memory than previous baselines, i.e. storing $O(K_{in} + K_{out})$ instead of $O(K_{in} * K_{out})$, VBD-CL possesses an implicit effect beneficial for reducing forgetting which is driven by VBD's ability to encourage the finding of flat minima. In specific, by Taylor's expansion, we can approximate the amount of forgetting of task 1 (w_1^*) after learning task 2 (w_2^*): $F_1 = L_1(w_2^*) - L_1(w_1^*) \leq \frac{1}{2}\lambda_1^{max}||w_2^* - w_1^*||^2$, where λ_1^{max} is the largest eigenvalue of the Hessian matrix $\nabla^2 L_1(w_1^*)$, which measure the flatness of a minimum. This shows that flat minima tend to reduce forgetting better than sharp one. Additionally, [6] shows that dropout implicitly minimizes curvature of Hessian matrix, i.e. increase flatness. Therefore, VBD-CL inherits this property, and thus, can implicitly lead to less forgetting, which is demonstrated in Figure 1b.

Experimental results

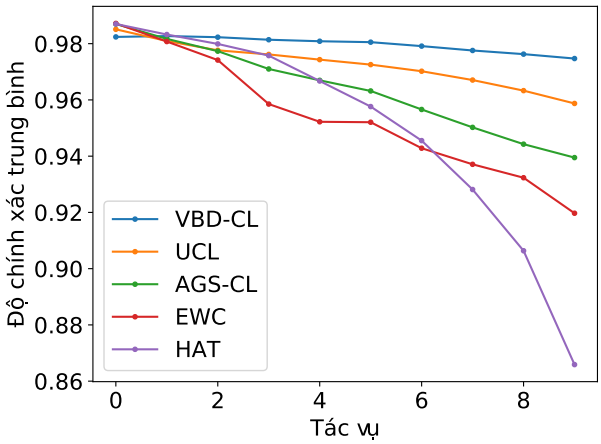
Here I provide a few results from my thesis which show the superior performance of VBD-CL across many CL benchmarks. Notably, VBD-CL maintains a significantly higher average accuracy than baselines even at then end of sequence. Source code is available at https://github.com/tunglamlqddb/VBD-CL.



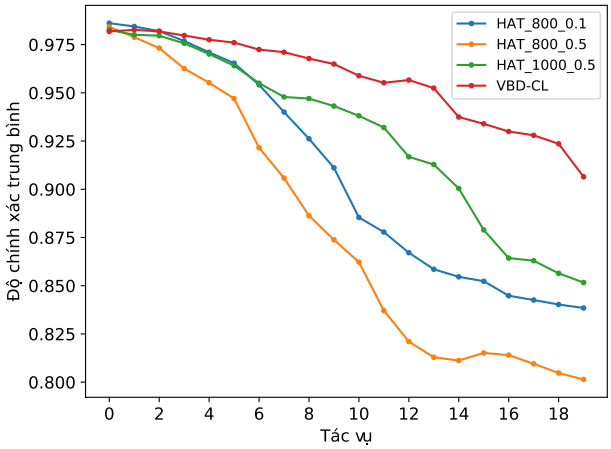
(a) Split-Cifar100 10 tasks.



(b) Cifar10-Split-cifar100 11 tasks.



(c) Permuted Mnist 10 tasks.



(d) Permuted Mnist 20 tasks.

Figure 2: The average accuracy (y-axis) along tasks (x-axis) of VBD-CL and other regularization-based methods.

References

[1] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016.

[2] 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, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995. PMLR, 2017.

[3] Joan Serrà, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4555–4564. PMLR, 2018.

[4] Hongjoon Ahn, Sungmin Cha, Donggyu Lee, and Taesup Moon. Uncertainty-based continual learning with adaptive regularization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4394–4404, 2019.

[5] Sangwon Jung, Hongjoon Ahn, Sungmin Cha, and Taesup Moon. Continual learning with node-importance based adaptive group sparse regularization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[6] Colin Wei, Sham M. Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10181–10192. PMLR, 2020.