Bachelor's Thesis

Survey on Regularization Methods in Continual Learning

Jörg Schantz

Ludwig-Maximilian Universität München

April 10, 2025

Supervised by Dr. Julian Rodemann

Outline

Introduction

Introduction

What is Continual Learning (CL)?

- Training a model with sequentially arriving data [9]

Introduction

What is Continual Learning (CL)?

- Training a model with sequentially arriving data [9]

Why is CL important?

- Cost-effectiveness of training large models and physical limitations, such as memory [2]

Introduction

What is Continual Learning (CL)?

- Training a model with sequentially arriving data [9]

Why is CL important?

- Cost-effectiveness of training large models and physical limitations, such as memory [2]

What's the catch?

- Preserving old information about the model without inhibiting new learning and v.v. [9]

Introduction

How is catastrophic forgetting handled via Regularization?

- (In-)direct parameter penalties: restrict movement in Output- or Parameter-space

- Regularization requires some degree of task similarity to be successful

CL Environment

CL Environment

Which type of model/agent/learner is used?

- Typically neural networks

CL Environment

Which type of model/agent/learner is used?

- Typically neural networks

What is a continual learner? [9]

- Models the joint probability distribution over all tasks
- Each sample is assumed to be conditionally independent
- Can only access current sample, all else are unavailable

CL Environment

Which type of model/agent/learner is used?

- Typically neural networks

What is a continual learner? [9]

- Models the joint probability distribution over all tasks
- Each sample is assumed to be conditionally independent
- Can only access current sample, all else are unavailable

Can we quantify forgetting?

- Upper bound: $F_t \leq 0.5 * \lambda_t^{max} \|\Delta W\|^2$ [7]

Explored Methods

Explored Methods

Elastic Weight Consolidation (EWC)[4]

- Direct parameter penalty

- Bayesian View of Parameters

- Approximate the parameters old posterior with a normal distribution

- Mean is estimated parameters and variance is inverted diagonal Fisher Information Matrix

- $\text{pen}_{EWC}(w) = \frac{\lambda}{2}(w - \hat{w}^{(t-1)})^\top F(w - \hat{w}^{(t-1)})$

Explored Methods

Adaptive Group Sparsity based Continual Learning (AGS-CL) [3]

- Direct parameter penalty

- Uses Grouped-LASSO penalty

- Determines Importance based on node activation

- Node Importance decides if Grouped-LASSO is centered around 0 or old weights

Explored Methods

Functional Regularization for Continual Learning (FRCL)[8]

- Indirect parameter penalty

- Uses Gaussian Process to approximate posterior of the old labels

- Stores inducing points and distribution parameters for all old tasks

- Penalizes deviations from old posteriors via KL-Divergence

Explored Methods

Other Methods

- Continual Ridge Regression [5]
- Generalized l2-regression [13]
- Synaptic Intelligence [12]
- Memory Aware Synapses [1]
- Dynamically Expandable Network [11]
- Learning without Forgetting [6]
- Deep Retrieval and Imagination [10]

Results

Results

Weighted Ridge Penalties

- Make use of importance measures (IM) to penalize shifts from previous parameters
- All presented methods use approximations of the FIM or Hessian of the loss as IM
- Ensures stability but hinders plasticity

Results

Ridge-like & other quadratic Penalties

- Make use of importance measures (IM) to penalize shifts from previous parameters
- All presented methods use approximations of the FIM or Hessian of the loss as IM
- Ensures stability but hinders plasticity

LASSO inspired Penalties

- Also use IM to identify important parameters
- Depending on IM they impose (Grouped-)LASSO penalties on nodes
- Slows down learning decline while maintaining stability

Results

Output-based Penalties

- Simulate/ store old data
- penalize movement in the output-space
- mimic learning the "true" model
- penalize shifts in posterior of y

Conclusion

Conclusion

What have I learned?

- Regularization can mitigate forgetting

- Too much stability hinders "life-long" learning and task variety

- To keep learning, models need structural updates or focus on similar tasks

- Mostly differ from the "true" loss by an approximation error

Conclusion

What have I learned?

- Regularization can mitigate forgetting

- Too much stability hinders "life-long" learning and task variety

- To keep learning, models need structural updates or focus on similar tasks

- Mostly differ from the "true" loss by an approximation error

Open questions:

- What are potential challenges in hyper-parameter estimation?

- Task similarity: a clear definition and can it exploited in regularization?

Appendix

Individual Penalties I

$$\text{pen}_{CRR}(w) = \lambda \|w - \hat{w}^{(1)}\|_2^2 \tag{1}$$

$$\text{pen}_{l2}(w) = \lambda (w - \hat{w}^{(t)})^\top A (w - \hat{w}^{(t)}) \tag{2}$$

$$\begin{aligned}\text{pen}_{AGSCL}(W) =& \mu \sum_{j,k \le l,\nu} \text{id}(\Omega_{j,k}^{(t-1)} = 0) \|W_{j,k}\|_2 \\ &+ \lambda \sum_{j,k \le l,\nu} \text{id}(\Omega_{j,k}^{(t-1)} > 0) \|W_{j,k} - \hat{W}_{j,k}^{(t-1)}\|_2\end{aligned} \tag{3}$$

$$\text{pen}_{LwF}(W) = \lambda \sum_{i=1}^{\#classes} -y_{o,i}^{(t)} \log \hat{y}_{o,i}^{(t)} \tag{4}$$

Individual Penalties II

$$\hat{W}^{(t)} = \arg\min_{W} L(W, D^{(t)} \cup M) + \beta L(W, M) + \frac{\alpha}{n^{(M)}} \sum_{i}^{n^{(M)}} \| f(W, x_i^{(M)}) - f(W^{(t-1)}, x_i^{(M)}) \|_2^2 \tag{5}$$

$$\text{pen}_{FRCL}(\theta, q(w^{(t)})) = -\sum_{j=1}^{t-1} \text{KL}(q(\tilde{y}^{(j)}) \| p_\theta(\tilde{y}^{(j)})) \tag{6}$$

Sources I

[1] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. Memory aware synapses: Learning what (not) to forget, 2018. URL `https://arxiv.org/abs/1711.09601`.

[2] f. given i=K, given=Katharina. The extreme cost of training ai models. URL `https://www.forbes.com/sites/katharinabuchholz/2024/08/23/the-extreme-cost-of-training-ai-models/`.

[3] S. Jung, H. Ahn, S. Cha, and T. Moon. Continual learning with node-importance based adaptive group sparse regularization, 2021. URL `https://arxiv.org/abs/2003.13726`.

[4] J. Kirkpatrick, R. Pascanu, N. Rabinowitza, J. Veness, A. A. R. Guillaume Desjardins, K. Milan, J. Quan, T. Ramalho, D. H. Agnieszk Grabska-Barwinska, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *arXiv:1612.00796v2*, 2017.

Sources II

[5] H. Li, J. Wu, and V. Braverman. Fixed design analysis of regularization-based continual learning, 2024. URL https://arxiv.org/abs/2303.10263.

[6] Z. Li and D. Hoiem. Learning without forgetting, 2017. URL https://arxiv.org/abs/1606.09282.

[7] S. I. Mirzadeh, M. Farajtabar, R. Pascanu, and H. Ghasemzadeh. Understanding the role of training regimes in continual learning, 2020. URL https://arxiv.org/abs/2006.06958.

[8] M. K. Titsias, J. Schwarz, A. G. de G. Matthews, R. Pascanu, and Y. W. Teh. Functional regularisation for continual learning with gaussian processes, 2020. URL https://arxiv.org/abs/1901.11356.

[9] L. Wang, X. Zhang, H. Su, J. Zhu, Fellow, and IEEE. A comprehensive survey of continual learning: Theory and method and application, 2024. URL https://arxiv.org/abs/2302.00487.

Sources III

[10] Z. Wang, L. Liu, Y. Duan, and D. Tao. Continual learning through retrieval and imagination. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8594–8602, Jun. 2022. doi: 10.1609/aaai.v36i8.20837. URL https://ojs.aaai.org/index.php/AAAI/article/view/20837.

[11] J. Yoon, E. Yang, J. Lee, and S. J. Hwang. Lifelong learning with dynamically expandable networks, 2018. URL https://arxiv.org/abs/1708.01547.

[12] F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence, 2017. URL https://arxiv.org/abs/1703.04200.

[13] X. Zhao, H. Wang, W. Huang, and W. Lin. A statistical theory of regularization-based continual learning, 2024. URL https://arxiv.org/abs/2406.06213.