Continual/Incremental/Lifelong Learning

Lei LIU

25/08/2021

Outline

Background

- Static models require restarting the training process each time new data becomes available.
- It is intractable due to storage constraints or privacy issues in real world.
- Human can learn new knowledge without catastrophic forgetting by rehearsal.
- How to gradually extend acquired knowledge from an infinite stream of data?

Parameter-Level Algorithms

These two papers tried to reduce the changes of the model parameters, which is more related to the old tasks.

- Ebrahimi, Sayna, et al. "Uncertainty-guided Continual Learning with Bayesian Neural Networks." International Conference on Learning Representations. 2019.
- Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." Proceedings of the national academy of sciences 114.13 (2017): 3521-3526.

Preliminaries

- $\mathcal{D} = (x, y)$ is a training set.
- $x \in \mathbb{R}^{\boldsymbol{n}}$ is a set of observed variables.
- $P(y|x, w)$ is a probabilistic model.
- w is a set of latent variables as model parameters.
- BNN aims to model the following distribution for prediction:

$$\begin{aligned} P(Y^*|X^*, \mathcal{D}) &= \int P(Y^*|X^*, \mathcal{W})P(\mathcal{W}|\mathcal{D})d\mathcal{W} \\ \text{where } P(\mathcal{W}|\mathcal{D}) &= \frac{P(\mathcal{W})P(\mathcal{D}|\mathcal{W})}{P(\mathcal{D})} \end{aligned} \tag{1}$$

- $P(\mathcal{W}|\mathcal{D})$ is intractable.

Variational Bayes-by-backprop

- Use a latent variable θ to control the distribution of $q(w|\theta)$, which can be approximated to $p(w|\mathcal{D})$

$$\begin{aligned} \theta^* &= \arg\min_\theta D_{\mathrm{KL}}[q(\mathrm{w} \mid \theta) \| P(\mathrm{w} \mid \mathcal{D})] \\ &= \arg\min_\theta \int q(\mathrm{w} \mid \theta) \log \frac{q(\mathrm{w} \mid \theta)}{P(\mathrm{w}) P(\mathcal{D} \mid \mathrm{w})} d\mathrm{w} \\ &= \arg\min_\theta D_{\mathrm{KL}}[q(\mathrm{w} \mid \theta) \| P(\mathrm{w})] - \mathbb{E}_{q(\mathrm{w}|\theta)}[\log P(\mathcal{D} \mid \mathrm{w})] \end{aligned} \tag{2}$$

- To solve the expectation value, use Monte Carlo Method:

$$\mathcal{L} = \sum_i \log q\left(w_i \mid \theta_i\right) - \sum_i \log P\left(w_i\right) - \sum_j \log P\left(y_j \mid w, x_j\right) \tag{3}$$

Learning from Uncertainty

- When learning on the new data, preserve the important parameters.
- By scaling the learning rate of each parameter:

$$\alpha \leftarrow \alpha/\Omega \tag{4}$$

  where Ω is the importance of the parameter.
- $\Omega = 1/\sigma$ yields the highest accuracy, where σ is the standard deviation.

Class incremental task

- Dataset: 2/5-Split MNIST

Method	μ	ρ	Importance Ω	BWT (%)	ACC (%)
UCB	x	-	$1/\sigma$	0.00	99.2
UCB	-	x	$1/\sigma$	-0.04	98.7
UCB	x	x	$1/\sigma$	-0.02	98.0
UCB	x	-	$\|\mu\|/\sigma$	-0.03	98.4
UCB	-	x	$\|\mu\|/\sigma$	-0.52	98.7
UCB	x	x	$\|\mu\|/\sigma$	-0.32	98.8

Figure 1: Variants of learning rate regularization and importance measurement on 2-Split MNIST.

Class incremental task

Method	BWT	ACC
VCL-Vadam†	-	99.17
VCL-GNG†	-	96.50
VCL	-0.56	98.20
IMM	-11.20	88.54
EWC	-4.20	95.78
HAT	0.00	99.59
ORD-FT	-9.18	90.60
ORD-FE	0.00	98.54
BBB-FT	-6.45	93.42
BBB-FE	0.00	98.76
UCB-P (Ours)	-0.72	99.32
UCB (Ours)	0.00	**99.63**

Figure 2: 5-Split MNIST, 5 tasks.

Elastic weight consolidation (EWC)

- $\theta^*_{A,i}$ is the i-th parameter learned from the task A.

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i \left(\theta_i - \theta^*_{A,i}\right)^2 \tag{5}$$

where F_i is the fisher information.

Fisher Information

- $X_i, \cdot, X_n \sim f(X;\theta)$
- The likelihood function is

$$L(\mathrm{X};\theta) = \prod_{i=1}^{n} f\left(X_i;\theta\right) \tag{6}$$

- Maximum Likelihood Estimate(MLE):

$$S(\mathrm{X};\theta) = \sum_{i=1}^{n} \frac{\partial \log f\left(X_i;\theta\right)}{\partial\theta} \tag{7}$$

- Standard Definition of fisher information:

$$I(\theta) = E\left[S(X;\theta)^2\right] - E\left[S(X;\theta)\right]^2 = Var[S(X;\theta)]. \tag{8}$$

Fisher Information

Two variant definitions:

- $E[S(X;\theta)] = 0$
- $I(\theta) = E\left[S(X;\theta)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\log L(\boldsymbol{X};\theta)\right]$ (second moment and second order derivation)
- Fisher information is increasing along with the data scale.
- Second order derivation can reflect the convexity of the likelihood function. Thus fisher information can measure the estimation ability.

Fisher information

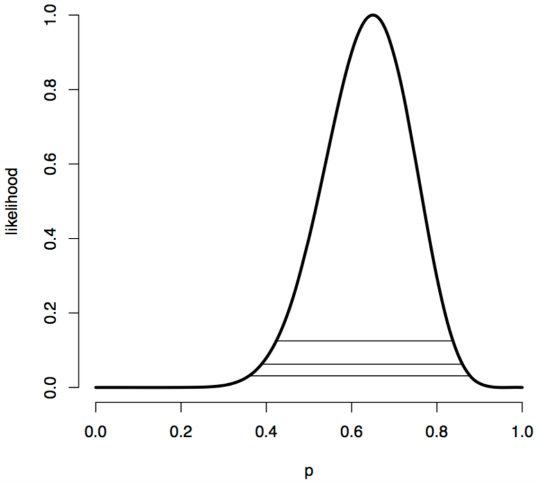

Figure 3: Second order derivation.

Experiments

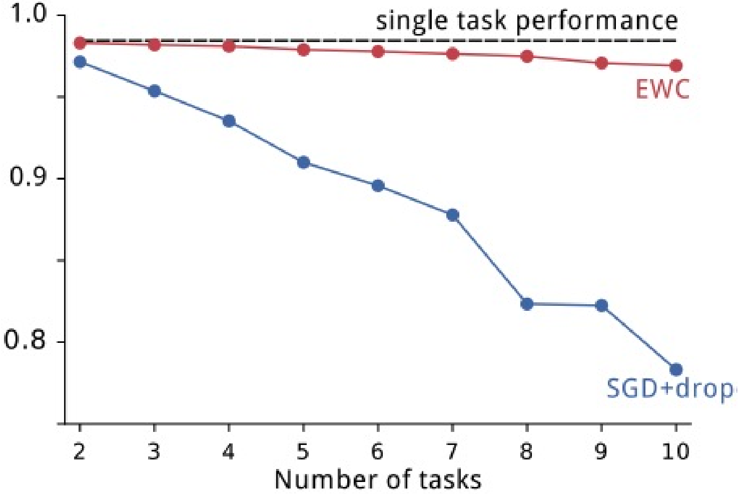

Figure 4: Performance on MNIST.

Summary

- Relationship between parameters and tasks.
- Rehearsal old data of the old task.