1st International Electronic Conference on Applied Sciences

A Novel Layer Sharing-based Incremental Learning via Bayesian Optimization

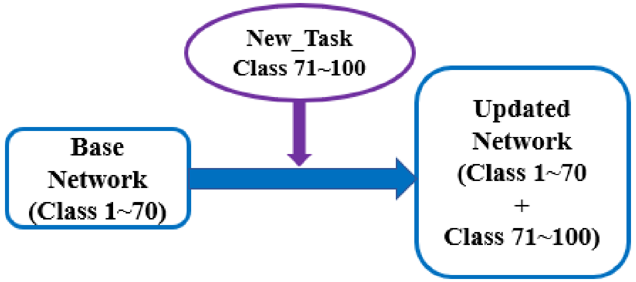
Bomi Kim, Taehyeon Kim and Yoonsik Choe

Department of Electrical and Electronic Engineering,
Yonsei University
bbboming@yonsei.ac.kr

Introduction

- **Incremental learning**
  - One of the significant challenges in neural network-based computer vision algorithms is learning new tasks incrementally, like the cognitive process of human learning
  - Human learns for lifetime acquiring new skills. However Deep Neural Networks and CNNs are designed to learn multiple tasks only if the data is presented all at once.

Task	Dataset
Base_Task	Class 1~70
New_Task	Class 71~100



New_Task
Class 71~100

Base
Network
(Class 1~70)

Updated
Network
(Class 1~70
+
Class 71~100)

※ **Incremental learning model**
the network should grow its capacity to accommodate classes of new task.

Introduction

- **3 Conditions for successful incremental learning algorithm**

1) The subsequent data from new tasks should be trainable and be accommodated incrementally without forgetting any knowledge in old tasks, i.e., it should not suffer from catastrophic forgetting.
2) The overhead of incremental training should be minimal.
3) The previously seen data of old task should not be accessible when it is training incrementally.

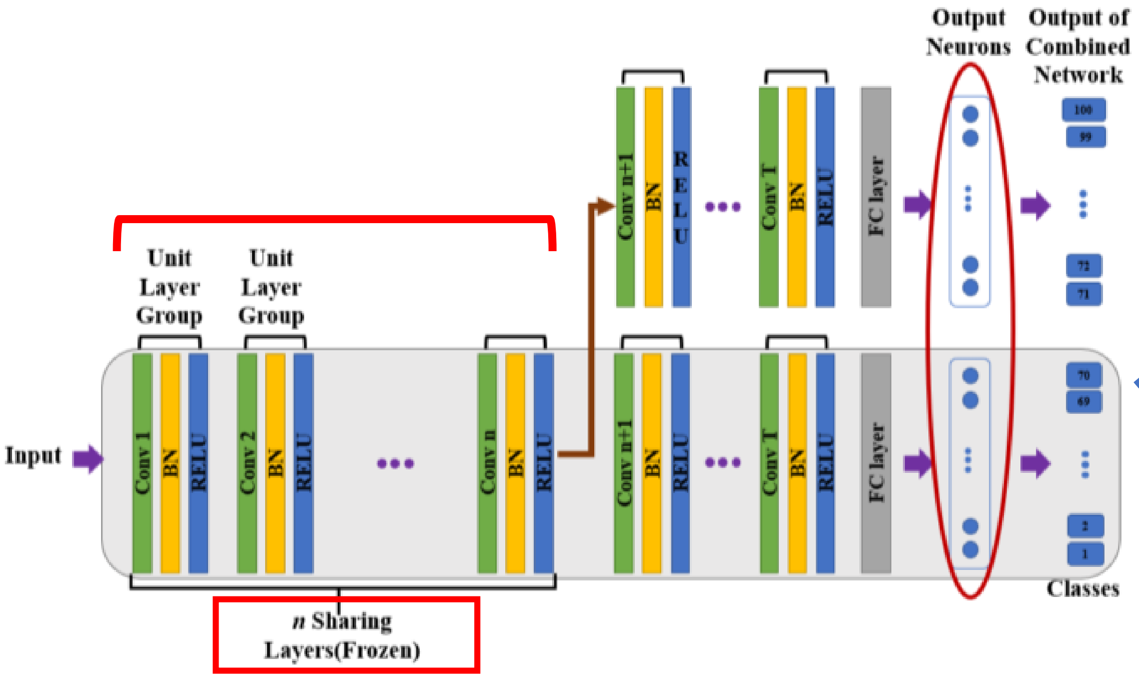
Preliminaries

- A deep convolutional neural network (DCNN) consists of multiple convolutional layers to extract hierarchical visual features.
- The earlier layers in DCNN extract the most basic part of an image, while the later layers extract much more detailed and sophisticated structures.

Erhan, Dumitru, et al. Visualizing higher-layer features of a deep network. University ofMontreal 1341.3 2009: 1.

Preliminaries

- **Partial Layer sharing algorithm for incremental learning**



Base Network

Fig. 1. The incremental network structure called *'clone and branch'*. The shaded gray rectangle is the base network. Each time we update the network for the new task, some initial parts of convolutional layers are shared and frozen, so only the branch part of layers is retrained with the new task data, while the last layers of the base network remain disconnected.

- Incremental learning based on layer sharing technique leverages general knowledge from previously learned tasks to learn subsequent new tasks by sharing initial convolutional layers of base networks especially in a similar domain of input used in new task

Preliminaries

- ***Clone and branch*** **Technique**
  - In *'clone and branch'* technique, there are two training methodologies.

1) ***Empirical searching method:***
   To select sharing layer number, they generate an 'accuracy vs sharing' trade-off curve in a brute force manner.
   : A large overhead for training all possible cases.

2) Using similarity score:
Few random samples of each class in a new task are passed through the pre-trained base network, and the number of repeating classes is regarded as a similarity score.
: The similarity score can not be robust on randomly few sampled data and it essentially has approximation errors.

Syed Shakib Sarwar, Aayush Ankit, and Kaushik Roy.. Incremental Learning in deep convolutional neural networks using partial network sharing., *IEEE Access 8* **2019**: 4615-4628.

Proposed Algorithm

1) Combined Classification Accuracy

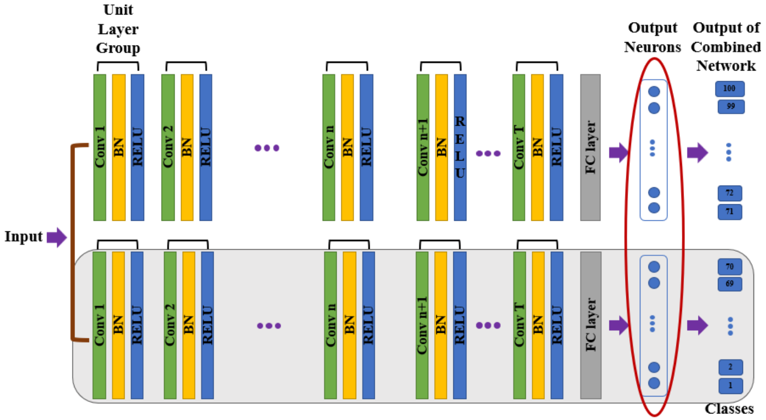
$$L_{Acc}(n) = \frac{1}{N} \sum_{x_i \in D_N} n(x_i, d_i), \quad n(x_i, d_i) = \begin{cases} 1 & \text{if } F_{base,new}(x_i) = d_i. \\ 0 & \text{otherwise.} \end{cases}$$

- where N denotes the total number of data for testing combined classification, n_i is the i th data for testing, d_i is the label of x_i, and $n(x_i, d_i)$ denotes accuracy on x_i , respectively. Therefore, if $F_{base,new}(x_i) = d_i$, which means the output of the combined network has the same value with the ground truth on x_i. i.e. di, it provides 1 or otherwise 0.

Proposed Algorithm

2) Target combined classification accuracy

$$L_{Target} = L_{Acc}(0) - T_{Deg}$$



※ incremental network structure without any network sharing

- where $L_{Acc}(0)$ is the baseline and T_{Deg} is the threshold accuracy degradation value. Then, L_{Target} is the target combined classification accuracy.
- The reason of utilizing $L_{Acc}(0)$ to define L_{Target} is that $L_{Acc}(0)$ is the upper-bound value of accuracy, where every layer of the network is updated for new tasks <u>without any network sharing.</u>

Proposed Algorithm

3) Proposed objective function

$$n^{*} = \arg\min_{n} L(n), \quad L(n) = ||L_{Acc}(n) - L_{Target}||_{1}.$$

- The objective function $L(n)$ with sharing some of the initial convolutional layers n is the linear-combination between $L_{Acc}(n)$ and L_{Target}. The $n*$ is the global optimal configurations for the target combined classification accuracy degradation in the incremental learning modeling, and it minimizes the objective function.

Proposed Algorithm

4) Global optimal layer selection via BayesOpt

- **Bayesian Optimization (BayesOpt)**
  - BayesOpt is designed for black-box derivative-free global optimization

    ① It does not require the structural information of objective function (black-box)

    ② It does not used the derivatives of objective function (derivative-free)

    ③ It finds the global optimum by calculating the uncertainty of the objective function at unobserved points (global optimization)

Frazier, Peter I. "A tutorial on bayesian optimization." *arXiv preprint arXiv:1807.02811* (2018).

Proposed Algorithm

4) Global optimal layer selection via BayesOpt

- **Bayesian Optimization (BayesOpt)**
  - BayesOpt is a class of machine learning based optimization methods.
  - **BayesOpt** consists of two major components:

    ① A **Bayesian statistical model** for modeling the objective function,

    ✓ Bayesian statistical model provides *quantified uncertainty of objective function values at an unobserved data.*

    ② An **acquisition function** for deciding the next sampling points.

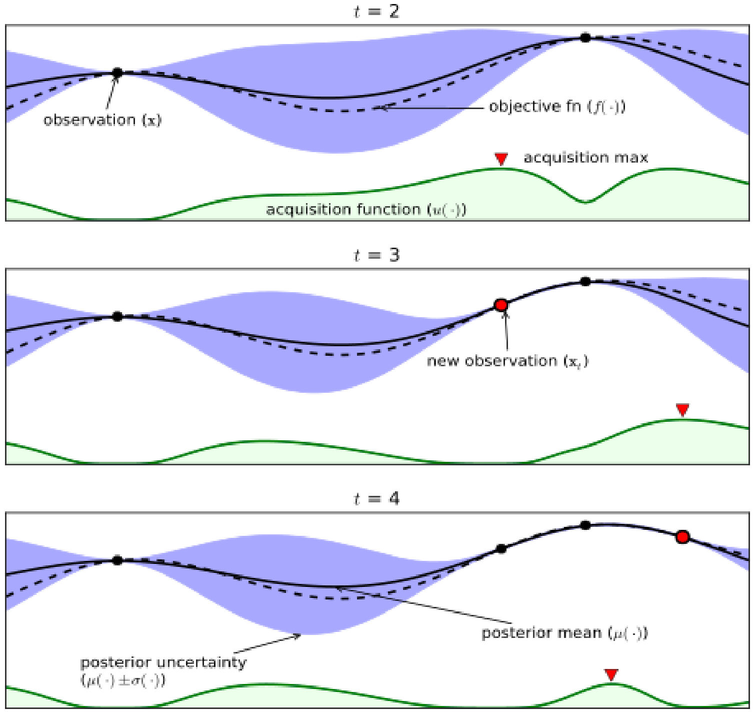
    ✓ The acquisition function measures *the predictive enhancement at an unobserved data, to determine the next sampling point.*

Frazier, Peter I. "A tutorial on bayesian optimization." *arXiv preprint arXiv:1807.02811* (2018).

Proposed Algorithm

4) Global optimal layer selection via BayesOpt

- **Bayesian Optimization (BayesOpt)**



- **Black dotted line:**

  actual objective function

- **Black solid line:**

  estimated mean function

- **Shades of blue:**

  estimated standard deviation

- **Black points & Red points:**

  observed data

- **Red triangle:**

  next sampling point

- **Green solid line:**

  acquisition function

Proposed Algorithm

4) Global optimal layer selection via BayesOpt

The proposed method builds a statistical model for quantifying uncertainty using GP regression:

Prior Distribution: $\mathrm{L}(n_{1:k}) \sim Normal(\mu_0(n_{1:k}), \Sigma_0(n_{1:k},\ n_{1:k})),$

Conditional distribution: $L(n)|L(n_{1:k}) \sim Normal\ \left(\mu_k(n), \sigma_k^2(n)\right),$

$$\mu_k(n) = \Sigma_0(n, n_{1:k})\Sigma_0(n, n_{1:k})^{-1}\left(L(n_{1:k}) - \mu_0(n_{1:k})\right) + \mu_0(\hat{n}),$$

$$\sigma_k^2(n) = \Sigma_0(n, n) - \Sigma_0(n, n_{1:k})\Sigma_0(n_{1:k}, n_{1:k})^{-1}\Sigma_0(n_{1:k}, n)\ .$$

Proposed Algorithm

4) Global optimal layer selection via BayesOpt

The proposed algorithm uses an expected improvement (EI) acquisition function to decide the next observation points.

Expected Improvement: $EI_k(n) \coloneqq E_k\left[\min\left(L(n) - L(n_k^*)\right), 0\right],$

$$n_{n+1} = argminEI_k(n).$$

Experiment Results – Implementation details

※ Network

Network
ResNet 50: 53 Convolution, 53 Batch Normalization, 49 ReLU, 1 averaging pooling 1 FC layer

※ Dataset

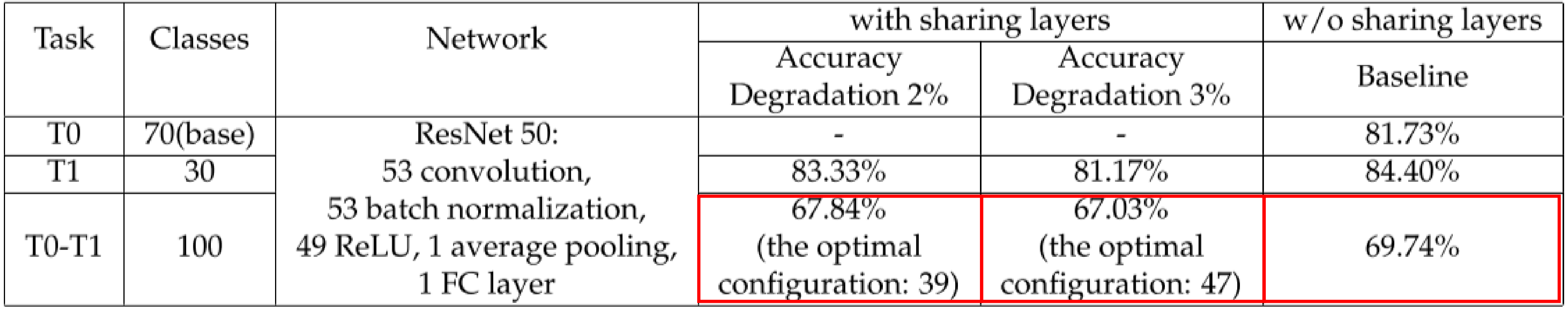
Dataset	Case 1 (accuracy degradation 2% or 3%)		Case 2 (Comparison)
CIFAR-100 **100 (classes)**	Task 0 (Base)	70	60
	Task 1 (incremental)	30	30
	Task 2 (incremental)	-	10

He, Kaiming, et al. Deep residual learning for image recognition. Proceedings ofthe IEEE conference on computer vision and pattern recognition. 2016.

Krizhevsky, Alex, and Geoffrey Hinton. Learning multiple layers of features from tiny images.Citeseer 2009: 7.

Experiment Results

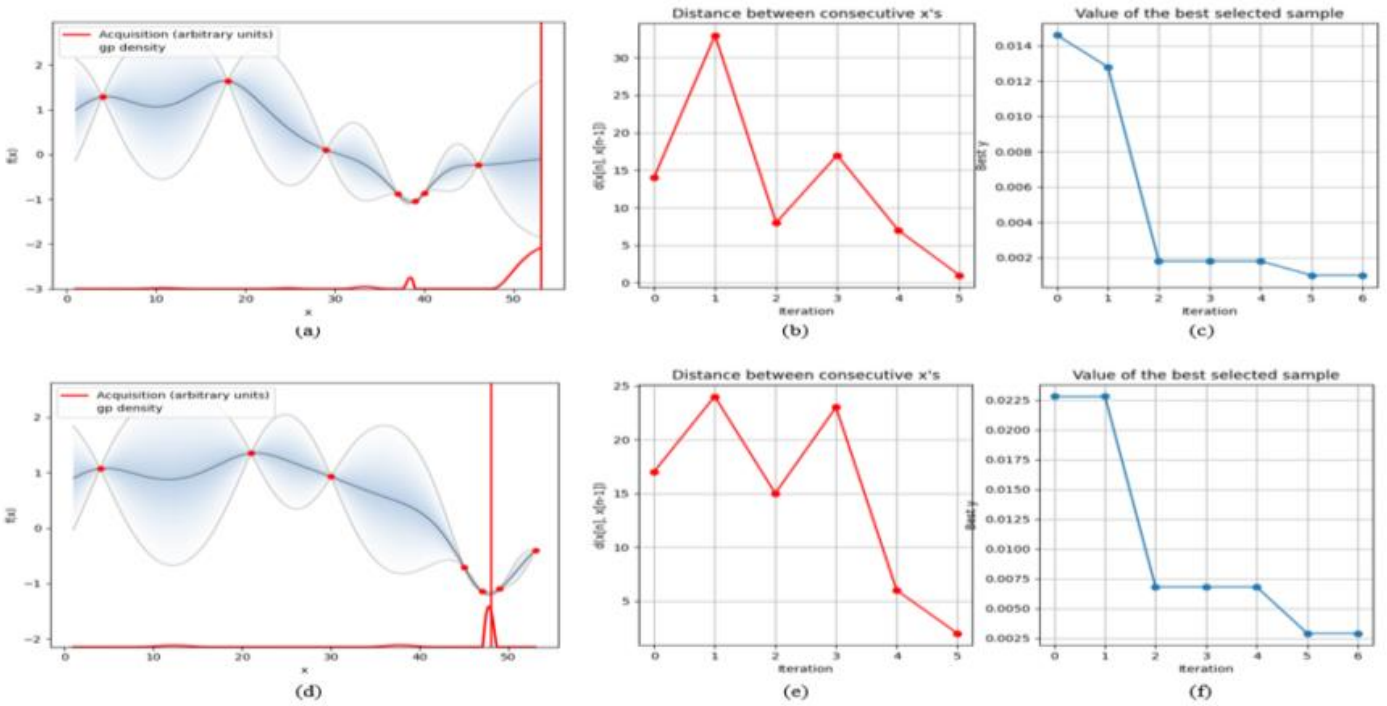
Table 1. Experimental results on the proposed algorithm



Task	Classes	Network	with sharing layers		w/o sharing layers
			Accuracy Degradation 2%	Accuracy Degradation 3%	Baseline
T0	70(base)	ResNet 50: 53 convolution, 53 batch normalization, 49 ReLU, 1 average pooling, 1 FC layer	-	-	81.73%
T1	30		83.33%	81.17%	84.40%
T0-T1	100		67.84% (the optimal configuration: 39)	67.03% (the optimal configuration: 47)	69.74%

The proposed method produces the global optimal sharing layer number in only 6 iterations without searching for all possible layer cases.

Experiment Results



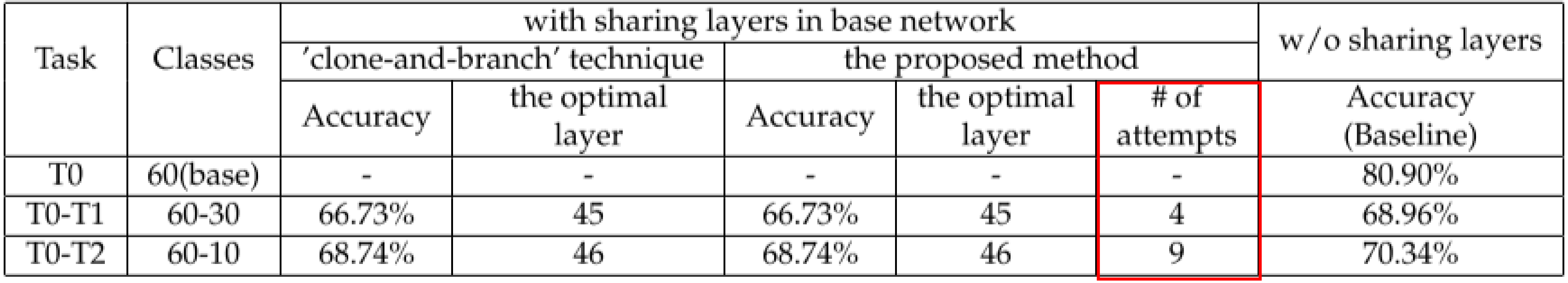
(a) (b) (c)

(d) (e) (f)

Figure 1. A visualization results of the optimal sharing layer configuration through BayesOpt. **(a),(d)** The result of the selected optimal number of sharing layers. **(b),(e)** The L2 distance of the consecutive observed points. **(c),(f)** The value of the calculated best-selected sample for every iteration.

Experiment Results

Table 2. Comparison of experimental results for *’clone and branch’*



Task	Classes	with sharing layers in base network					w/o sharing layers
		’clone-and-branch’ technique		the proposed method			
		Accuracy	the optimal layer	Accuracy	the optimal layer	# of attempts	Accuracy (Baseline)
T0	60(base)	-	-	-	-	-	80.90%
T0-T1	60-30	66.73%	45	66.73%	45	4	68.96%
T0-T2	60-10	68.74%	46	68.74%	46	9	70.34%

Syed Shakib Sarwar, Aayush Ankit, and Kaushik Roy.. Incremental Learning in deep convolutional neural networks using partial network sharing., *IEEE Access 8* **2019**: 4615-4628.

Conclusions

- The proposed methodology can adeptly find the number of sharing layers according to a given condition of accuracy degradation by adjusting the threshold accuracy parameter.
- The experimental results demonstrate that our method finds the precise sharing capacity of a base network for subsequent new tasks and converges in a few iterations.
- We solve the discrete combinatorial optimization problems for incremental learning by BayesOpt, which ensures global convergence.

Thank you!

A Novel Layer Sharing-based Incremental Learning via Bayesian Optimization

Bomi Kim, Taehyeon Kim and Yoonsik Choe

Department of Electrical and Electronic Engineering,

Yonsei University

bbboming@yonsei.ac.kr