CONSISTENT RELATIVE CONFIDENCE AND LABEL-FREE MODEL SELECTION FOR CONVOLUTIONAL NEURAL NETWORKS

Bin Liu*

Zhejiang Lab, Hangzhou, 311100 China

ABSTRACT
This paper is concerned with image classification based on deep convolutional neural networks (CNNs). The focus is centered around the following question: given a set of candidate CNN models, how to select the right one that has the best generalization property for the current task? Present model selection methods require access to a batch of labeled data for defining a performance metric, such as the cross-entropy loss, the classification error rate, the negative log-likelihood, and so on. In many practical cases, however, labeled data are not available in time as labeling itself is a time-consuming and expensive task. To this end, this paper presents an approach to CNN model selection using only unlabeled data. This method is developed based on a principle termed consistent relative confidence (CRC). The effectiveness and efficiency of the presented method are demonstrated by extensive experimental studies based on datasets MNIST and FashionM-NIST.

Index Terms— convolutional neural network, consistent relative confidence, label-free model selection, image classification

1. INTRODUCTION
In the last decade, CNNs have become the dominant modeling and algorithmic framework for various computer vision tasks such as image classification [1, 2]. In this paper, we are concerned with a fundamental issue related to CNNs’ practical applications, namely, given a set of candidate CNN models, how to select the one that has the best generalization properties for the present task? For ease of presentation, we only focus on the image classification task, while the idea and method described here may be easily extended to other application scenarios of CNNs.

The aforementioned model selection problem can appear in different cases. For example, in transfer learning, one may have several candidate models pre-trained from different source domains. Then, which is the one to choose for transferring knowledge to the target domain? In the context of continual or life-long learning [3], the candidate models can be obtained from previous tasks. Again, one needs to select the right one for use for the forthcoming task. Edge learning in the context of cloud computing, see e.g., [4], also calls for lightweight learning methods such as selecting one model for use at the edge node from a set of candidate models pre-trained at the server side. Broadly speaking, any supervised training or fine-tuning operation on a model can be taken as a model selection process, where the candidate models span over the whole parameter space. The optimized parameters resulted from the training or fine-tuning process define the model finally selected. In this generalized model selection problem, the number of candidate models can be huge; while here we only consider conventional model selection problems where the number of candidate models is limited.

Present approaches to CNN model selection require access to a labeled dataset, used for specifying the performance metric, such as the cross-entropy loss, the classification error rate, and the negative log-likelihood. But how to do model selection only using unlabeled data? This problem is crucial for many practical scenarios where labeled data can not be available in time. As we know, labeling itself is usually a time-consuming and expensive task. We call the above issue label-free model selection (LFMS). LFMS is in spirit related to self-supervised learning, see e.g., [5], and unsupervised clustering, see e.g., [6, 7], while they are different in the problem setting. Compared with self-supervised learning or unsupervised clustering, LFMS is characterized by an underlying assumption that a set of pre-trained models is available beforehand in its problem setting.

In this paper, we present a simple while highly effective LFMS approach in the context of CNN based image classification (Section 3). This approach is developed based on a principle termed consistent relative confidence (CRC) (Section 2). To the best of our knowledge, this principle is for the first time revealed in the literature. We verify the effectiveness and efficiency of our method by experiments in Section 4 and conclude the paper in Section 5.

2. CONSISTENT RELATIVE CONFIDENCE
Here we present the consistent relative confidence (CRC) principle through a thought experiment.

Consider a team of volunteers asked to finish an image
classification task. They all behave under the assumption of perfect rationality [8], having the same intention to make choices as accurate as possible. They make decisions according to their experiences previously collected. Consider the following question: is there a positive relationship between a volunteer’s confidence and the quality of his (her) decision? The term confidence reflects an internal judgment of the volunteer on the probability of his (her) decision being correct. There are different answers to this question. For example, [9] reports a positive relation, while [10] negates it.

What we wonder is that, if we think a pre-trained CNN model as a volunteer, is there a positive relationship between its confidence and its decision quality? We conduct experiments to find an answer to this question. The experimental results show that if a model behaves more confidently than the others in a consistent way for an image classification task, then it will be the model that gives the most accurate decisions to their experiences previously collected. Consider the phenomenon consistent relative confidence (CRC). See Section 4 for details on the experimental results.

We define the CRC principle as follows: if there is one volunteer who is more expert than the others for the current task, this volunteer’s expertise can be reflected by his (her) relative confidence that appears consistently in the decision making process. The basic idea of our model selection approach is to find out the expert (resp. the right model) among the volunteers (resp. the candidate models) using the CRC principle. Such a model selection procedure only draws information from each model’s internal judgment (i.e., the level of its confidence), while has no requirement for any labeled data. Therefore, we call it a label-free approach.

3. CRC BASED LABEL-FREE CNN MODEL SELECTION

Consider an image classification task for which we have at hand K pre-trained models $M_1, M_2, \ldots, M_K$, and N images $I_1, I_2, \ldots, I_N$ to be categorized. Denote the number of classes as $C$ and the unknown ground-truth label of $I_n$ as $l_n \in \{1,2,\ldots, C\}$, $n = 1, \ldots, N$. The last layer of the CNN architecture considered here is fixed to be a softmax layer. Given an image $I_n$ as the input of $M_k$, $n = 1, \ldots, N$, $k = 1, \ldots, K$, one forward propagation running of $M_k$ outputs a probability vector $[p_{1,c,n,k}, p_{2,c,n,k}, \ldots, p_{C,c,n,k}]$ that satisfies $0 \leq p_{c,n,k} \leq 1, c = 1, \ldots, C$ and $\sum_{c=1}^{C} p_{c,n,k} = 1$. Define the “confidence” of $M_k$ on image $I_n$ to be

$$C_{k,n} = \max\{p_{1,n,k}, p_{2,n,k}, \ldots, p_{C,n,k}\}. \quad (1)$$

Denote the mean and the standard error of $\{C_{k,1}, \ldots, C_{k,N}\}$ as $MC_k$ and $SC_k$, respectively. Define the lower confidence bound (LCB) of $M_k$ on the image set $\{I_1, I_2, \ldots, I_N\}$ as

$$LCB_k = MC_k - SC_k. \quad (2)$$

The CRC score of $M_k$ is defined to be

$$S_k = LCB_k - \max_{i \in \{1:k-1, k+1:K\}} LCB_i. \quad (3)$$

Based on the above setting, the model selection procedure is simply to find out the index $k^*$ that satisfies

$$k^* = \arg \max_{k=1:K} S_k = \arg \max_{k=1:K} LCB_k, \quad (4)$$

and then select $M_{k^*}$ for use in the present task. The maximization operation in Eqn.(4) guarantees that the model being selected is relatively more confident than the others in a consistent way. The consistency property is obtained by the usage of the LCB in Eqns.(2) and (4). Specifically, the smaller is $SC_{k^*}$, the larger is the extent to which $M_{k^*}$’s confidence is consistently greater than the others’.

The CRC score of the selected model, namely $S_{k^*}$, takes a value between 0 and 1. It can be seen as an internal judgment of the reliability (or accuracy) of the model selection result. The closer it is to 1, the more reliable (or accurate) is the model selection process.

4. EXPERIMENT

Here we present the experimental results. The purpose of the experiment is twofold. First, to check whether the CRC principle and the resulting model selection approach described in Section 3 is effective. Second, to evaluate the efficiency of our method.

4.1. Experimental setting

We use 2 datasets, MNIST and FasionMNIST, in our experiments. Each dataset contains 60,000 training samples, 10,000 testing samples, and 10 image classes. Each sample consists of an image and an associated label indicating which class this image belongs to.

For each dataset, we generate 4 additional related datasets, which are of the same size and have the same labels as the original dataset. Each generated dataset is associated with a specific image processing operation, which is performed on every image included in the original dataset. The operations involved here include:

- Operation 1: image filtering with a rotationally symmetric Gaussian lowpass filter of size 7 with standard deviation 1;
- Operation 2: image filtering with a 3-by-3 filter whose shape approximates that of the 2D Laplacian operator;
- Operation 3: adding Gaussian white noise with a mean 0 and variance 0.02 to each image;
- Operation 4: adding Gaussian white noise with a mean 0 and variance 0.1 to each image.
Fig. 1: Image samples drawn from the MNIST dataset.

Fig. 2: Counterpart images of those shown in Fig.1. The 4 sub-figures correspond to 4 different image processing operations, respectively. See the text for details on the operations.

Fig. 3: Image samples drawn from the FashionMNIST dataset.

Fig. 4: Counterpart images of those shown in Fig.3. The 4 sub-figures correspond to 4 different image processing operations, respectively. See the text for details on the operations.

See Figs.1-4 for a sample of the original images and their counterparts in the generated datasets.

Denote the original dataset MNIST by MNIST-D₀ and the newly generated dataset by MNIST-Dᵢ, i = 1, . . . , 4. MNIST-Dᵢ is associated with Operation i, i = 1, . . . , 4, described above. We obtain FashionMNIST-Dᵢ, i = 0, . . . , 4, in the same way. Then we train one CNN model based on each dataset. Only the training samples included in each dataset are used for model training. The remaining testing samples are used for performance evaluation of the involved model selection methods.

The CNN architecture is fixed to have 7 layers, namely, the input layer, the convolution layer, an average pooling layer, the flatten layer, 2 fully connected layers, and a softmax layer. In the convolution layer, 20 9-by-9 sized filters are used. The dimensions of the fully connected layers are 360 and 60, respectively. The activation function used here is ReLU. One of the most popular gradient-based optimization algorithms, Momentum, is employed for minimizing the cross-entropy loss. The learning rate is set to 0.1, the momentum term to 0.95, the minibatch size to 128, and the epoch number to 3.

4.2. Experimental results

Based on the above setting, we evaluate the performance of our approach by comparing its result with that of conventional methods that make decisions using labeled data. We consider methods that use the cross-entropy loss, the error rate, or the negative log-likelihood as the guide for model selection. Such metrics are the most popularly used.

We consider 10 cases, each being associated with one dataset presented above, namely MNIST-Dᵢ or FashionMNIST-Dᵢ, i ∈ {0, . . . , 4}. In each case, we select N samples from the testing set of the corresponding dataset, 10 ≤ N ≤ 10,000. Given these samples, we see whether the model selection methods can select the right model from the candidates. Note that Mᵢ is the ground truth answer for samples drawn from the i-th dataset, i = 0, . . . , 4, since it is built based on the training data of that dataset.

In each case, we consider different choices of sample size N. The purpose is to evaluate the efficiency of our approach. For each case, we plot the experimental result on it with a figure. Thus we get 10 figures in total.
Here we only show 2 representatives of them in Figs.5-6, due to the space limitation. The other figures are deferred to the Appendix Section. In both cases shown in Figs.5-6 and those presented in the Appendix, one can see that the CRC score performs as well as the other label-data-based metrics for identifying the right model to select, even when the sample size $N$ is small. Note that a higher CRC score indicates a better model and vice versa. For the other performance metrics, a smaller value indicates a better model and vice versa.

**5. CONCLUSIONS AND FUTURE WORKS**

In this paper, we gave the CRC principle for the first time. Based on this principle, we developed an approach to label-free CNN model selection. We demonstrated its performance in image classification experiments. Results showed that, while being simple, our approach is effective and highly efficient. In particular, it works as well as existent methods that depend on labeled data to make decisions.

In principle, this approach can be considered as a choice for many scenarios that lack labeled data. For example, it can be easily extended to work in an iterative manner to search parameter values that maximize the CRC score. In this way, it becomes a self-supervised learning method. As the CRC score is not differential, to maximize it one may choose derivative-free algorithms, e.g., evolutionary methods [11–14] or particle filter optimization [15–17]. The biggest obstacle to implement the above idea may come from the computation, as such optimization methods are much more computationally expensive than gradient-based ones. Besides, if one could find a way to translate the CRC score to an appropriate weight or probability of the associated model, then model averaging techniques, see e.g., in [18–23], can be employed, which may result in more accurate decisions, while at a cost of an increase in computing burdens.
6. REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.

[3] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, “Continual lifelong learning with neural networks: A review,” *Neural Networks*, vol. 113, pp. 54–71, 2019.

[4] J. Chen and X. Ran, “Deep learning with edge computing: A review,” *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

[5] L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2020.

[6] J. Xie, R. Girshick, and A. Farhadi, “Unsupervised deep embedding for clustering analysis,” in *Int’l Conf. on Machine Learning (ICML)*. PMLR, 2016, pp. 478–487.

[7] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *Proc. of the European Conference on Computer Vision (ECCV)*, 2018, pp. 132–149.

[8] G. R. Steele, “Understanding economic man: psychology, rationality, and values,” *American Journal of Economics and Sociology*, vol. 63, no. 5, pp. 1021–1055, 2004.

[9] D. J. Power, S. L. Meyeraan, and R. J. Aldag, “Impacts of problem structure and computerized decision aids on decision attitudes and behaviors,” *Information & management*, vol. 26, no. 5, pp. 281–294, 1994.

[10] D. Landsbergen, D. H. Coursey, S. Loveless, and RF Shangraw Jr, “Decision quality, confidence, and commitment with expert systems: An experimental study,” *Journal of Public Administration Research and Theory*, vol. 7, no. 1, pp. 131–158, 1997.

[11] K. O. Stanley, J. Clune, J. Lehman, and R. Miikkulainen, “Designing neural networks through neuroevolution,” *Nature Machine Intelligence*, vol. 1, no. 1, pp. 24–35, 2019.

[12] S. Fujino, T. Hatanaka, N. Mori, and K. Matsumoto, “Evolutionary deep learning based on deep convolutional neural network for anime storyboard recognition,” *Neurocomputing*, vol. 338, pp. 393–398, 2019.

[13] H. Iba and N. Noman, *Deep Neural Evolution: Deep Learning with Evolutionary Computation*, Springer, 2020.

[14] E. Galván and P. Mooney, “Neuroevolution in deep neural networks: Current trends and future challenges,” *IEEE Trans. on Artificial Intelligence*, 2021.

[15] B. Liu, S. Cheng, and Y. Shi, “Particle filter optimization: A brief introduction,” in *Int’l Conf. on Swarm Intelligence*. Springer, 2016, pp. 95–104.

[16] B. Liu, “Posterior exploration based sequential Monte Carlo for global optimization,” *Journal of Global Optimization*, vol. 69, no. 4, pp. 847–868, 2017.

[17] B. Liu, “Particle filtering methods for stochastic optimization with application to large-scale empirical risk minimization,” *Knowledge-Based Systems*, vol. 193, pp. 1–9, 2020.

[18] B. Liu, Y. Qi, and K. Chen, “Sequential online prediction in the presence of outliers and change points: an instant temporal structure learning approach,” *Neurocomputing*, vol. 413, pp. 240–258, 2020.

[19] P. Domingos, “Bayesian averaging of classifiers and the overfitting problem,” in *Int’l Conf. on Machine Learning (ICML)*, 2000, vol. 747, pp. 223–230.

[20] W. Maddox, T. Garipov, P. Izmailov, D. Vetrov, and A. G. Wilson, “Fast uncertainty estimates and bayesian model averaging of dnns,” in *Uncertainty in Deep Learning Workshop at UAI*, 2018.

[21] B. Liu, “Robust particle filter by dynamic averaging of multiple noise models,” in *Proc. of the 42nd IEEE Int’l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4034–4038.

[22] Y. Qi, B. Liu, Y. Wang, and G. Pan, “Dynamic ensemble modeling approach to nonstationary neural decoding in Brain–computer interfaces,” in *Advances in neural information processing systems (NeurIPS)*, 2019, pp. 6087–6096.

[23] B. Liu, “Instantaneous frequency tracking under model uncertainty via dynamic model averaging and particle filtering,” *IEEE Trans. on Wireless Communications*, vol. 10, no. 6, pp. 1810–1819, 2011.
7. APPENDIX

Here we provide supplementary experimental results that do not appear in Section 4.2.

Fig. 7: CRC score and label-based performance metrics calculated over testing samples included in MNIST-D0

In Fig. 7, one can see that it fails to select the right model \( M_0 \) either using the CRC score or any of the other label-data-based performance metrics.

Fig. 8: CRC score and label-based performance metrics calculated over testing samples included in MNIST-D2

In the case shown in Figs. 8 and 10, the right model can be selected by using the CRC score or any other label-data-based performance metric.

The cases shown in Figs. 9 and 13 are similar to that in Fig. 7, in which the right model can not selected by using any performance metric.

For the case shown in Fig. 11, \( M_1 \) is selected according to the CRC score, which is inconsistent with the ground truth \( M_0 \). However, \( M_1 \) is a better choice in terms of the error rate, as shown in the 3rd sub-figure from top to down.

For the case shown in Fig. 12, the CRC score performs as well as the other metrics.

Finally, the CRC score performs surprisingly better than the other metrics for the case shown in Fig. 14.

Fig. 9: CRC score and label-based performance metrics calculated over testing samples included in MNIST-D3

Fig. 10: CRC score and label-based performance metrics calculated over testing samples included in MNIST-D4

Fig. 11: CRC score and label-based performance metrics calculated over testing samples included in FashionMNIST-D0
Fig. 12: CRC score and label-based performance metrics calculated over testing samples included in FashionMNIST-D1

Fig. 13: CRC score and label-based performance metrics calculated over testing samples included in FashionMNIST-D3

Fig. 14: CRC score and label-based performance metrics calculated over testing samples included in FashionMNIST-D4