Distill Inception-v4 to classify fine-grained aircraft images

https://github.com/anavuongdin/FGVC-Aircraft-Inception-v4

Paper ID *****

Abstract

Traditionally, a wide spectrum of computer vision tasks is based on transfer learning. The procedure usually starts with a pre-trained model on an enormous image dataset and then fine-tunes these weights on some datasets with fewer classes. In light of L2SP regularization, we have a decent method to drive model weights to their original values or toward zero. To further avoid catastrophic forgetting while training on a new task, elastic weight consolidation can be applied. This work studied all these principles and also proposed a Fisher information matrix-based transfer learning method to retain the knowledge of the pre-trained model. A compact experiment is carried out to testify whether the proposed approach is efficient or not.

1. Introduction

For the last decade, CNNs have gained attention due to resounding success [2]. It is constantly indicated that with rich-resource datasets and bottomless neural networks, CNNs can produce fruitful results [5]. Nevertheless, in most real-world applications, such costly calculations and inferences are extravagant because of the limited amount of time. Therefore, one of the most common approaches is to utilize an emerging method, which is transfer learning. The idea of transfer learning is to reuse neural networks which were immensely trained on some benchmark datsets, such as ImageNet 1K, and then fine-tune on specific datasets. Thanks to this utilization, effort for improving the performance of tasks, where the number of labels is much fewer, is economical [7].

When transferring knowledge, the standard methodologies are either setting initial hyperparameters for fine-tuning each layer or frozen lower layers. Due to the common sense that lower layers usually learn general characteristics while higher layers tend to focus on individual traits [8], binary classification seems to be inappropriate because neither a strong representation of lower layers nor highly discrimination of higher layers is exploited. L2SP regularization [7] is generally used for preventing a flawed transfer by decaying weights toward either their pre-trained values or zero. We can also exploit L2SP-Fisher [7] regularization to keep models away from neglecting old knowledge.

This work conducted experiments to study assumptions on non-binary classifications and the family of L2SP to understand their impacts. In this paper, a combination of L2SP and L2SP-Fisher is proposed to solve an image classification problem by utilizing transfer learning.

2. Related work

The original idea to improve performance in transfer learning was to study the impact of growing number of transferred layers [1]. After the remarkable breakthrough of residual networks [2], the techniques for enhancing performance have varied. More recently, the widely applied strategy is to freeze all weights absent from final classification layer, then impose searches for the best hyperparameters. Some hyperparameters searching was proposed in the literature, weight decay and learning rate in [4], momentum in [3] and L2SP in [7]. The outcome of this trend is quite promising, smaller models are more capable of reaching optimal objectives thanks to combinations of optimal learning rates and other hyperparameters.

Understanding the relationship between the source domain and target domain is also a great concern in transfer learning. Firstly, It was shown that relationship between target and source datasets may have greater impact on results than the volume of source dataset [4]. Secondly, a study on L2SP regularization indicates that driving model weights towards their pre-trained values rather than 0 results in a better performance when the two domains are closely overlapped [7]. Thirdly, in [3], the authors showed that high momentum may lead to poor results in closely related domains.

3. Methodology

- Dataset: experiments are performed on a single dataset, namely FGVC Aircraft. There are 10,000 images from the dataset, each image is a $3 \times 299 \times 299$ input. This dataset is proved to be difficult to handle by traditional transfer learning. The original authors provided a standard split, specifically, each of train set, valid set and test set has approximate 3333 images. This work uses ImageNet 1K as the source domain. Specially, the target dataset is fine-grained, which includes subordinate labels from a specific superordinate label.

- Model: This work reuses Inception-v4 model as well as the code by the authors in [6]. Learning is transferred from Image 1K domain to our concerned one, which is to classify variants of aircraft.
- Evaluation metric: top-1 accuracy is utilized.

4. Experiment

The calculation relies on Google Colab server, which provides a Tesla T4 GPU. The model is trained in 80 epochs. Figure 1 shows the performance of a run with the proposed methodology. The accuracy has been constantly improved and achieved the best score of 80.02% after 80 epochs. Meanwhile, the methodology proposed by [6] scored 78.94% within the same number of epochs. This work presents a short experiment to learn what weights of L2SP-Fisher will work best. The L2SP-Fisher loss function has the form of:

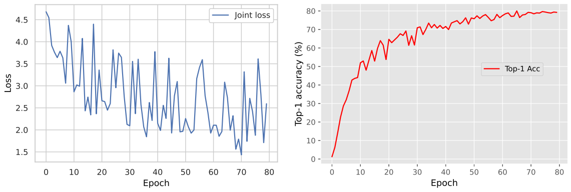

Figure 1. Loss function and accuracy over the dataset during a training by proposed methodology

$$\text{L2SP-Fisher}(\cdot,\theta) = \alpha\|\theta\|_2^2 + \beta\sum_{j\in\mathcal{S}}\mathbf{F}_{jj}(\theta_j - \theta_j^0)^2 \quad (1)$$

Where $\mathbf{F}$ is the Fisher information matrix which indicates the dependence on the pre-trained model θ_j^0.

4.1. The impact of weights of L2SP-Fisher

This short work desires to study what combinations of α and β will work best. The top-1 accuracy is shown in Table 1. This work selected 0.0005 for α because 0.0005 was the optimal value for L2 in paper [6]. 0 is also chosen because the main goal of this work is to study the impact of L2SP-Fisher in case there is no existence of L2 in Equation (1).

		α	
		0.0005	0
β	0.00004	77.83	78.94
	0.0004	78.16	**80.02**

Table 1. Top-1 accuracy (%) of several combinations

Overall, without the existence of L2 (i.e. $\alpha = 0$), we can obtain higher top-1 accuracy. When L2 is completely ignored, the produced result is 1.11% better in case of $\beta = 0.00004$, while 1.86% better in the second case.

Furthermore, this table suggests that an appropriate β can also lead to more successful models. 0.0004-β produces top-1 accuracy averagely 0.71% better than 0.00004-β. Combining these two observations, it is reasonable that the setting with 0.0004-β and no L2 provides the best accuracy of 80.02%.

5. Conclusion and future work

In conclusion, an improvement was proposed on the paper by [6] by studying the impact of L2SP-Fisher.

In the future, the author will study more about relationships between L2SP-Fisher, the number of frozen layers, and other hyperparameters.

References

[1] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. In *European conference on computer vision*, pages 329–344. Springer, 2014. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Hao Li, Pratik Chaudhari, Hao Yang, Michael Lam, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Rethinking the hyperparameters for fine-tuning. *arXiv preprint arXiv:2002.11770*, 2020. 1

[4] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. 1

[5] Jiquan Ngiam, Daiyi Peng, Vijay Vasudevan, Simon Kornblith, Quoc V Le, and Ruoming Pang. Domain adaptive transfer learning with specialist models. *arXiv preprint arXiv:1811.07056*, 2018. 1

[6] Jo Plested, Xuyang Shen, and Tom Gedeon. Rethinking binary hyperparameters for deep transfer learning for image classification. *arXiv preprint arXiv:2107.08585*, 2021. 2

[7] LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, pages 2825–2834. PMLR, 2018. 1

[8] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 1