IMPROVING SELF-SUPERVISED LEARNING FOR OUT-OF-DISTRIBUTION TASK VIA AUXILIARY CLASSIFIER

Harshita Boonlia⋆†, Tanmoy Dam*, Md Meftahul Ferdaus†, Sreenatha G. Anavatti*, Ankan Mullick†

*† CSE Department, Indian Institute of Technology Kharagpur
* SEIT, University of New South Wales Canberra, Australia
† ATMRI, Nanyang Technological University, Singapore

ABSTRACT

In real world scenarios, out-of-distribution (OOD) datasets may have a large distributional shift from training datasets. This phenomena generally occurs when a trained classifier is deployed on varying dynamic environments, which causes a significant drop in performance. To tackle this issue, we are proposing an end-to-end deep multi-task network in this work. Observing a strong relationship between rotation prediction (self-supervised) accuracy and semantic classification accuracy on OOD tasks, we introduce an additional auxiliary classification head in our multi-task network along with semantic classification and rotation prediction head. To observe the influence of this addition classifier in improving the rotation prediction head, our proposed learning method is framed into bi-level optimisation problem where the upper-level is trained to update the parameters for semantic classification and rotation prediction head. In the lower-level optimisation, only the auxiliary classification head is updated through semantic classification head by fixing the parameters of the semantic classification head. The proposed method has been validated through three unseen OOD datasets where it exhibits a clear improvement in semantic classification accuracy than other two baseline methods. Our code is available on GitHub https://github.com/harshita-555/OSSL

Index Terms— out of distribution, self-supervised learning, auxiliary classifier

1. INTRODUCTION

In machine learning community, benchmarks like ImageNet [1], CIFAR [2] etc. are commonly used to know the generalization ability of classifiers, where we assume that the test time input distributions are the same as the training distribution. However, when classifiers are applied to real-world applications like product recommendation, medical diagnosis, autonomous driving, they may face complex and dynamic shifts in the data distributions. Besides, new objects can be exposed to the classifiers at any time. Such issues in out-of-distribution (OOD) datasets may lead to catastrophic failure of the classifiers. In addition, annotations of test samples are not provided in many cases. Under such environment, classifiers’ performance is usually evaluated by collecting new labeled test sets. Nonetheless, labeling adequate images in a novel scenario is very complex and highly expensive. To minimize such labeling cost, researchers have investigated various approaches for evaluating classifiers’ performance on unlabeled test sets. Some researchers have developed complexity measurements on model parameters to analyse generalization of the classifiers [3,4].

Researchers have proposed various methods to deal with OOD examples. For instance, probabilities from softmax distributions are utilized in [5] to detect wrongly classified and OOD examples. They have shown that OOD examples have a lower prediction probability than that of correct or in-sample examples. Researchers have also used self-supervision method [6,7] to handle OOD tasks by introducing an auxiliary task that supports to create labels from unlabeled samples. In [6], they classified four different rotation angles of an image at \{0°, 90°, 180°, 270°\} to pay attention to the pretext task of rotation prediction. They jointly trained their network for both classification task and pretext task using CIFAR-10, MNIST, Tiny-ImageNet, and COCO, where they studied the correlation between those two task’s accuracy. By considering many labeled test sets and plotting classification verses rotation prediction accuracy, a strong correlation (Pearson’s Correlation \(r > 0.88\)) is witnessed between those accuracies. Based on such findings, they learnt a linear regression model, which can predict classification accuracy on unseen test sets. They obtained ground truths from a given unlabeled test set by rotating images manually. It has been used to calculate the rotation prediction accuracy on the test images using the multi-task network. Afterwards, the linear regression model uses this rotation prediction accuracy to predict the semantic classification accuracy.

Unlike the earlier work, in this work, we mainly focus on improving rotation prediction accuracy (self-supervised learning) using an auxiliary classifier for out of distribution task, which ultimately improves the semantic classification accuracy. Therefore, our proposed approach can be for-
2. PROPOSED METHOD

In this paper, we have utilized a held-in training set and an unseen OOD test-set. We define the given training set as $\mathcal{D}^{train} = \{(x_k, y_k)\}_{k=1}^N \in (X \times Y)$ where, $x_k \in X$ is the $k$-th training image and $y_k = \{0, 1, ..., C-1\} \in Y$ corresponds to target variable spanning over $C$ classes. $N$ is the total number of samples present in the training dataset $\mathcal{D}^{train}$. The OOD test set is defined in a similar fashion, $\mathcal{D}^{test} = \{(x_i, y_i)\}_{i=1}^M$ where $y_i$ is target variable spanning over $C$ classes. Our objective is to design a robust classifier that can maximise classification accuracy for unseen OOD test-set.

2.1. Multi-task learning

Researchers have observed in [9] that a higher accuracy in rotation prediction indicates a model’s superiority in capturing representation of learned features in lower dimensional manifold. Though they followed a completely unsupervised learning strategy, they did not consider the OOD task. In recent times, similar method has been developed to deal with OOD problem [6], where a linearly proportional relationship between rotation prediction and semantic classification has been observed. However, their proposed two-stage method can not provide an end-to-end solution.

2.2. How to maximise the self-supervision (rotational-accuracy) task?

Network details: To attain this objective, a multi-task learning framework has been developed for semantic classification along with self-supervised (rotation prediction) task. We utilise the multi-head network along with the same base network. Utilization of such multitasking framework is not enforcing more complexity while improving the self-supervised task. To maximise the self-supervised performance on base-network, we introduce an auxiliary classifier along with semantic classifier head and self-supervised (rotation prediction) head. These minimal changes will not increase the burden on the base-network. For the base (feature extraction) network, we have taken a convolution neural network (e.g. densenet) followed by three fully connected layers for three different tasks. As depicted in Fig 1, the base feature extractor is parameterised by $\theta_b$. All the remaining task-specific classification head are described as follows,

- semantic-classification prediction head is parameterised by $\theta_{sc}$.
- rotation prediction head is parameterised by $\theta_{ss}$.
- auxiliary semantic classification head is parameterised by $\theta_{sa}$.

Rotational prediction head: We follow the similar rotation transformation as in [6, 9]. The four geometrical rotational transformations are applied to a train image($x$), $F = \{G_r(x)\}$, where $G_r$ is the geometrical rotation function with four rotation angles $r = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$. This geometrical transformation can not alter the invariant nature [10]. Therefore, the rotational head can predict rotational accuracy by 4-ways.

Loss functions: The proposed OSSL method is associated with three individual classifications losses for three different tasks. The semantic classification loss is defined as follows,

$$L_{ch} = CE(y_c, \theta_{sc}(\theta_b(x)))$$

where $CE = -\frac{1}{N} \sum_{y_c=1}^N y_c \log(\theta_{sc}(\theta_b(x)))$
The rotation prediction classification loss is defined as follows,

\[ L_{\text{rh}} = \frac{1}{4} \sum_{r \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}} CE(y_r, \theta_{ss}(\theta_b(G_r(x)))) \]  

where, \( y_r \) is represented as one-hot-encode labels for all four rotational.

The semantic auxiliary classification loss is defined as follows,

\[ L_{\text{ah}} = CE(y_s, \theta_{sa}(\theta_b(x))) \]  

We have utilized the above three losses into a bi-level optimisation problem to maximise the self-supervision performance. In upper-level optimisation, the semantic classification head and rotation classification head parameters are learnt simultaneously to update the base-network parameters as well as corresponding task specific class parameters, where the objective can be expressed as follows:

\[
\min_{\theta_b, \theta_{sa}, \theta_{ss}} L_{\text{upper}} \tag{4}
\]

where \( L_{\text{upper}} = (L_{\text{ch}} + L_{\text{rh}}) \). This upper level optimisation problem is solvable with the stochastic gradient descent (SGD) method where it first tunes the parameters of both the task specific classifiers:

\[ \{\theta_{sc}, \theta_{ss}, \theta_{b}\} = \{\theta_{sc}, \theta_{ss}, \theta_{b}\} - l_r \sum_{D_{\text{train}}} \nabla_{\theta_{sc}, \theta_{ss}, \theta_{b}} L_{\text{upper}} \] \n
where \( l_r \) is the learning rate of the upper-level loop.

Similarly, in lower-level optimisation, the objective can be expressed as follows:

\[
\min_{\theta_{sa}, \theta_{b}} L_{\text{lower}} \tag{6}
\]

where \( L_{\text{lower}} = (L_{\text{ch}} - L_{\text{ah}}) \). As similar to upper level, the SGD method is used to optimize only the parameters of the auxiliary head and base network:

\[ \{\theta_{sa}, \theta_{b}\} = \{\theta_{sa}, \theta_{b}\} - l_r \sum_{D_{\text{train}}} \nabla_{\theta_{sa}, \theta_{b}} L_{\text{lower}} \] \n
where, the same \( l_r \) is being used to nullify the effects of semantic classification head in the backward path. However, the semantic classification head parameters \( \theta_{sc} \) remain fixed.

The proposed OSSL framework’s is given in Algorithm 1.

**Algorithm 1 Learning Strategy of OSSL**

1. **Input:** training dataset \( D_{\text{train}} \), testing dataset \( D_{\text{test}} \), learning rates \( l_r \), iteration numbers \( n_{\text{epoch}} \)
2. **Output:** parameters of all the four networks \( \{\theta_b, \theta_{sc}, \theta_{ss}, \theta_{sa}\} \)
3. for \( p = 1 \) to \( n_{\text{epoch}} \) do
4. Update \( \{\theta_{sc}, \theta_{ss}, \theta_{b}\} \) parameters by using equation (5) /* upper level optimisation*/
5. Update \( \{\theta_{sa}, \theta_{b}\} \) parameters by using equation (7) when \( \theta_{sc} \) is fixed /* lower level optimisation*/
6. if \( p \geq 49 \& p \% 10 = 0 \) then
7. calculate testing accuracy for \( D_{\text{test}} \)
8. end if
9. end for

### 3. EXPERIMENTS AND VALIDATION

In this paper, our proposed OSSL method has been compared with two other baseline methods associated with two different losses, where the parameters of the first baseline is updated through semantic classification loss \( L_{\text{ch}} \) and the second one is updated through semantic classification with rotational head losses \( L_{\text{ch}} + L_{\text{rh}} \). For experimental validations, two popular classification benchmark data sets have been considered to train the model such as: digits (MNIST) and natural image (CIFAR-10) data set. For both data sets, three different unseen OOD test-sets have been utilised to evaluate the model prediction accuracy.

The LeNet-5 [11] model is a popular architecture for classifying the digit datasets (MNIST). Therefore, we consider LeNet-5 as a base feature extractor along with three classification heads. Original MNIST dataset is applied to train the model parameters, but, two different unseen OOD data sets namely USPS [12] and SVHN [13] are used to test it. Besides, both the unseen test-sets are having same number of classes(10) as in the training set. Therefore, it is practical to use these data sets as unseen OOD test-sets. On top of the backbone feature extractor i.e. LeNet-5, three tasks specific fully connected layers are being used. In addition, to analyse the effectiveness of the proposed method in a complex dataset, DenseNet - 40 (40 layers) architecture [14] is applied as a backbone feature extractor. In this case, CIFAR-10 is used to train the model, whereas CIFAR - 10.1 is utilized as an unseen test-set to evaluate the model performance. CIFAR-10.1 is a popular benchmark dataset for the OOD classification task where collected test-set samples distributional shift cannot vary too much compared with the original CIFAR 10 samples [15, 16].

**Table 1.** Quantitative classification performance for unseen OOD test-set

| Trainset       | MNIST       | CIFAR 10     |
|----------------|-------------|--------------|
| Unseen OOD test-set | USPS  | SVHN | CIFAR 10.1 |
| \( L_{\text{ch}} \) | 60.85 | 22.25 | 83.30       |
| \( L_{\text{ch}} + L_{\text{rh}} \) | 64.82 | 19.74 | 85.05       |
| OSSL           | 65.22       | 21.09        | 86.90       |
MNIST train-set is used to train the model. The obtained baseline classification performance (considering $L_{ch}$) for USPS testset is 60.85%. A significant improvement in performance has been observed when rotation prediction loss is considered along with semantic classification loss. However, our proposed OSSL method has obtained best classification accuracy compared with the above two methods. The obtained classification accuracy is 65.22%. However, for SVHN test-set, the best obtained accuracy is 22.25%, which is from baseline when using only classification head. A significant performance drop is observed while considering the rational prediction head along with classification head. However, OSSL has obtained better model accuracy than the rotational head prediction, but it cannot outperform the baseline. Large distributional shift in the unseen SVHN test set samples compared to the held-in train set [17] could be the main reason for such performance deterioration. Such large distributional shift in dataset can’t be represented by geometrical rotation with given held-in train-set. As a result, the classification performance is declined under multi-tasking learning framework.

In addition, we have considered a complex CIFAR-10 dataset, where more complex network architecture has been utilised as a baseline. It is clearly observed from the Table 1, the proposed OSSL has outperformed the other two methods. While considering only classification head, the obtained accuracy is 83.30%. A performance improvement in accuracy has been observed by considering both classification head and rotation prediction head, where the obtained classification accuracy is 85.05%. The proposed OSSL method has attained best classification accuracy among all the three methods and the accuracy is 86.90%.

3.1. tSNE analysis

In this work, tSNE analysis [18] is utilized to investigate the discriminative ability among different class distributions of our proposed methods and baselines as well. We considered test sets from CIFAR-10.1 dataset to project the original feature space to a two-dimensional space. In contrast to baselines, an effective separation among different classes is witnessed from OSSL as expected, where these classes are marked in different colors as shown in Fig. 2. Such outcomes are also confirming that our proposed method can extract discriminative information from OOD datasets.

4. LIMITATIONS

The OSSL approach maximises rotation prediction performance, improving semantic classification accuracy. For certain OOD datasets, combining a rotation prediction head with a semantic classification head cannot ensure satisfactory test-set performance. Rotation prediction head may adversely affect semantic classifier performance. Rotation prediction head-based OOD task must be well-defined and substantial [7]. Otherwise, rotational prediction head misses key details. Such as, SVHN dataset training and testing datasets have a large distributional shift. In such a case, though the proposed OSSL performs better than the rotation prediction head and single semantic classifier-based methods, it cannot ensure its improvement over the baseline.

5. CONCLUSION & FUTURE WORKS

This paper presents a joint learning strategy to improve classification performance for unseen OOD downstream tasks. To attain this objective, we formulated OSSL, where an additional auxiliary classifier was introduced to nullify the impact of the semantic classification. The proposed OSSL method performed better than the other two baselines for three unseen test sets. Finally, the following direction can be explored with the OSSL framework:

- To put a set of penalties for different dynamic test environments through invariant risk minimisation principle for handling large distributional shifts [19, 20].
- For handling class imbalance problem, a latent preserving GAN [21,22] can be used to generate minority class samples in dynamic tests environments.
6. REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009, pp. 248–255.

[2] Alex Krizhevsky, Geoffrey Hinton, et al., “Learning multiple layers of features from tiny images,” 2009.

[3] Ciprian A. Corneanu, Sergio Escalera, and Aleix M. Martinez, “Computing the testing error without a testing set,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2677–2685.

[4] Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio, “Predicting the generalization gap in deep networks with margin distributions,” arXiv preprint arXiv:1810.00113, 2018.

[5] Dan Hendrycks and Kevin Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” arXiv preprint arXiv:1610.02136, 2016.

[6] Weijian Deng, Stephen Gould, and Liang Zheng, “What does rotation prediction tell us about classifier accuracy under varying testing environments?,” in International Conference on Machine Learning. PMLR, 2021, pp. 2579–2589.

[7] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt, “Test-time training with self-supervision for generalization under distribution shifts,” in International Conference on Machine Learning. PMLR, 2020, pp. 9229–9248.

[8] Quang Pham, Chenghao Liu, Doyen Sahoo, and HOI Steven, “Contextual transformation networks for online continual learning,” in International Conference on Learning Representations, 2020.

[9] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, “Unsupervised representation learning by predicting image rotations,” arXiv preprint arXiv:1803.07728, 2018.

[10] Tanmoy Dam, Sreenatha G Anavatti, and Hussein A Abbass, “Improving clustergan using self-augmented information maximization of disentangling latent spaces,” arXiv preprint arXiv:2107.12706, 2021.

[11] Yann LeCun et al., “Lenet-5, convolutional neural networks,” URL: http://yann.lecun.com/exdb/lenet, vol. 20, no. 5, pp. 14, 2015.

[12] Jonathan J. Hull, “A database for handwritten text recognition research.” IEEE Transactions on pattern analysis and machine intelligence, vol. 16, no. 5, pp. 550–554, 1994.

[13] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng, “Reading digits in natural images with unsupervised feature learning,” 2011.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[15] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar, “Do cifar-10 classifiers generalize to cifar-10?,” 2018, https://arxiv.org/abs/1806.00451.

[16] Antonio Torralba, Rob Fergus, and William T. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1958–1970, 2008.

[17] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira, “Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10951–10960.

[18] Laurens Van der Maaten and Geoffrey Hinton, “Visualizing data using t-sne.,” Journal of machine learning research, vol. 9, no. 11, 2008.

[19] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz, “Invariant risk minimization,” arXiv preprint arXiv:1907.02893, 2019.

[20] Hao-Yun Chen, Jhao-Hong Liang, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan, “Improving adversarial robustness via guided complement entropy,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4881–4889.

[21] Tanmoy Dam, Md Meftahul Ferdaus, Sreenatha G Anavatti, Senthinath Jayavelu, and Hussein A Abbass, “Does adversarial oversampling help us?,” in Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 2970–2973.

[22] Tanmoy Dam, Sreenatha G Anavatti, and Hussein A Abbass, “Mixture of spectral generative adversarial networks for imbalanced hyperspectral image classification,” IEEE Geoscience and Remote Sensing Letters, 2020.