Mobile-PIRL: A Mobile Version of Self-Supervised Learning of Pretext-Invariant Representations

İzzet Emre Küçükkaya
Boğaziçi University
Istanbul, Turkey
izzet.kucukkaya@boun.edu.tr

Berkcan Üstün
Boğaziçi University
Istanbul, Turkey
berkcan.ustun@boun.edu.tr

***Abstract*—Supervised learning techniques of the Deep Neural Networks which are the main component of various applications in the domain of computer vision such as classification, segmentation or object detection are in need of the large amount of semantically annotated data in order to perform efficiently. However, it is a compelling situation to obtain a great quantity of semantically annotated data. The Self-Supervised learning techniques are introduced to overcome this obstacle. On the other hand, plenty of these techniques uses convolutional neural networks that have a great deal of complexity, parameters and inference time. In terms of performance, the recent mobile networks are considered to be comparable to complex network architectures despite their simple structure, few number parameters and less inference time. By this sense, the usage of newly introduced mobile networks in the domain of self supervised learning is an interesting topic of research. Regarding these facts, the performance of the mobile networks in a self supervised learning architecture can be examined. Our choice as the baseline architecture is PIRL (Pretext-Invariant Representation Learning). In addition, the number of transformations used in the PIRL can be increased along with the opportunity that they can be combined together.**

***Index Terms*—Mobile Networks, Data Augmentation, Self-Supervised Learning**

I. Introduction

Image representations are learned by modern image-recognition algorithms using massive datasets of photos and their semantic annotations. Class labels, hashtags, bounding boxes, and other types of annotations can be used to offer these annotations. Pre-defined semantic annotations do not scale well to the large tail of visual concepts, obstructing future image recognition advancements. Building more intelligent generalist models that can execute numerous applications and learn new abilities without vast quantities of labeled data is hampered by supervised learning. Self-supervised learning attempts to overcome these drawbacks by learning image representations from the pixels themselves rather than depending on those pre-defined lexical annotations. Many of the self-supervised learning techniques includes a pretext task which conducts a transformation to the input image with the need of the estimation of the properties of the transformation from the transformed image [1]–[5].

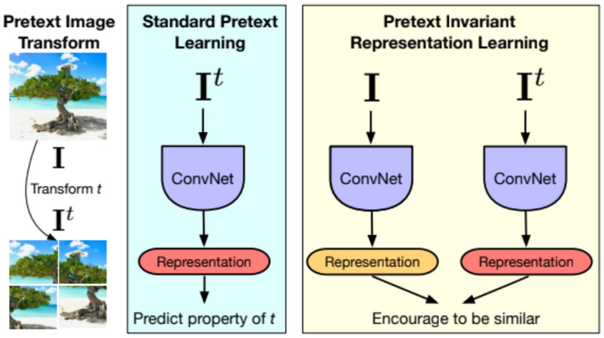
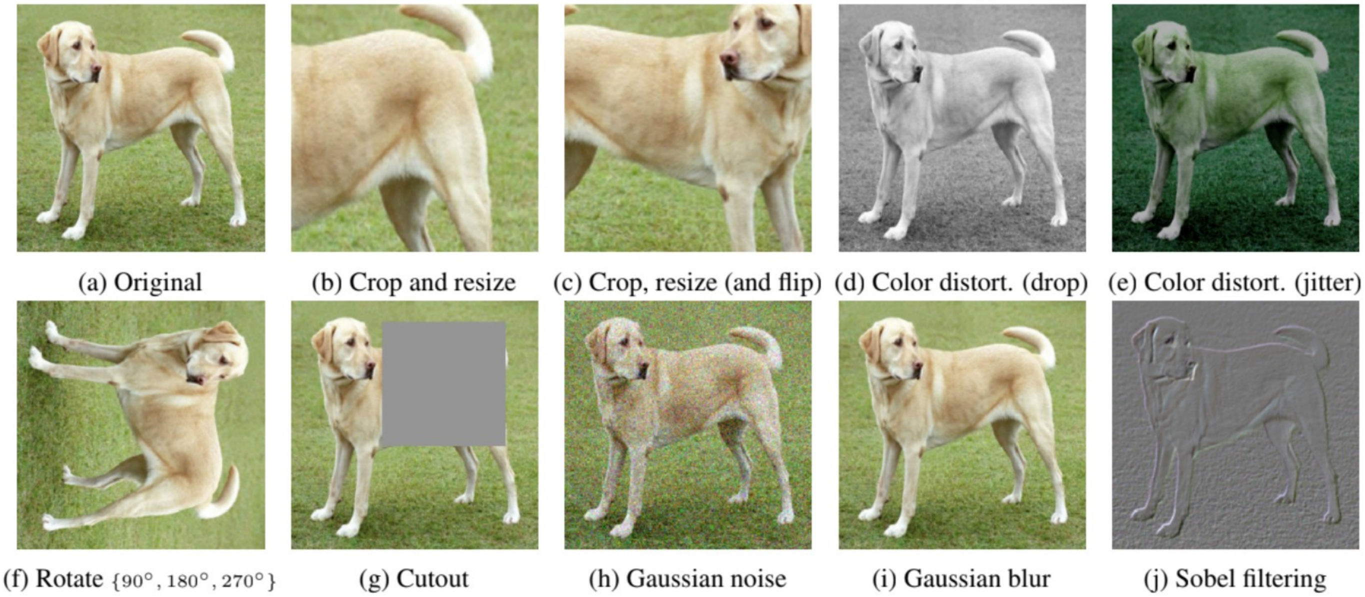


Fig. 1. Comparison of the Standard Pretext Learning and Pretext Invariant Representation Learning [1]

Although, often these techniques result on learning visual representations that are co-variant to the transformations, Pretext-Invariant Representation Learning (PIRL), which is one of the most successful methods, states the fact that learning co-variant representations is not useful for most of the visual recognition applications. In consideration of this, the PIRL underlines the necessity of learning invariant representations since the transformations do not change the semantics of the visual inputs. In order to obtain invariant representations, PIRL tries to create image representations that are close to other representations which are resulted in the transformation of the same image and far from the representations of other images.

PIRL uses primarily the ”Jigsaw” pretext task which is based on the division of the input image to nine patches and shuffling them. In addition, rotation pretext task and the combination of those two tasks presented along with the Jigsaw. However, as mentioned in the original article [1], the number of transformations can be enhanced in order to improve the performance. Bunch of transformations can be seen in the Figure 3. Those transformation can be engaged one by one, in the form of combination of two, three etc.

PIRL uses the ResNet-50 [6] architecture in order to obtain the image representations of both the original image and the transformed image. In spite of the fact that ResNet is an evolutionary architecture that accomplishes state-of-the-art performance with relatively small number of parameters and floating point operations, we think that there is still room for progress in the area of decreasing the parameters and floating point operations while keeping the performance at a some standard thanks to newly introduced concept mobile networks. The introduction of the mobile networks or so-called



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90^\circ, 180^\circ, 270^\circ\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Fig. 2. Examples of data augmentations [3]

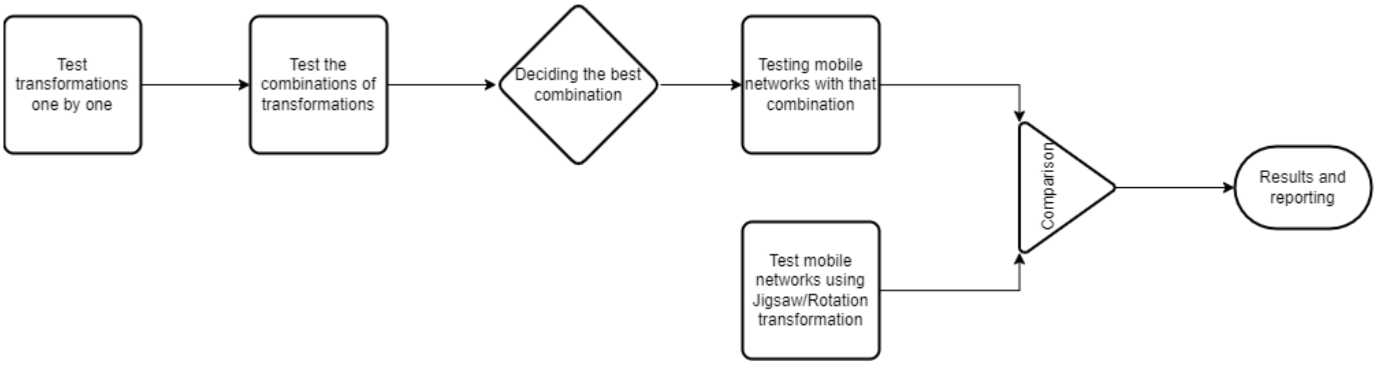

Test transformations one by one
Test the combinations of transformations
Deciding the best combination
Testing mobile networks with that combination
Comparison
Results and reporting
Test mobile networks using Jigsaw/Rotation transformation

Fig. 3. Contingency Table

”Lightweight-CNNs” makes understand their efficacy in the real world applications such as augmented reality glasses. In real world applications, the problem of the memory usage can be arisen due to the great amount of parameters of the deep network architectures. Besides, the floating point operations and hence the inference of the standard very deep networks can be considerably grandiose giving rise to an obstacle to use these networks in the applications that the timing is essential. In consideration of these facts, various mobile networks has been constructed [7]–[9]. An example comparison of the mobile networks with the regular deep networks can be seen in the Table .. The table is taken from the original paper of the EfficientNet [9]. As can be seen from the table, the EfficientNet-B0 structure overwhelms the other deep networks with only small number of parameters. As a result of this example comparison, it is a promising idea that mobile networks might be performing properly with fewer number of parameters in the domain of self-supervised learning.

II. Proposed Approach

Our proposed idea can be seen in the category of improving an already proposed idea since we are desiring to improve the Pretext-Invariant Representation Learning architecture by 1) increasing the number of transformations 2) changing ResNet-50 which is can be considered as the backbone of the architecture with various mobile networks. On the other hand, since the performance of the mobile networks have never been examined in the self-supervised learning architectures, this work might be considered as applying an already proposed method to a different domain.

A. Mobile Networks with Jigsaw Transformation

The first thing to do is the test the performance of the mobile networks. In order to create the controlled experiment environment, the mobile networks must be integrated to the PIRL with the Jigsaw pretext task same as the [1]. The same experiments must be conducted in order to get a healthy

TABLE I
EFFICIENTNET PERFORMANCE RESULTS ON IMAGENET [9]

Model	Top-1 Acc.	Top-5 Acc.	#Params	Ratio-to-EfficientNet	#FLOPs	Ratio-to-EfficientNet
EfficientNet-B0	**77.1%**	**93.3%**	**5.3M**	**1x**	**0.39B**	**1x**
ResNet-50	76.0%	93.0%	26M	4.9x	4.1B	11x
DenseNet-169	76.2%	93.2%	14M	2.6x	3.5B	8.9x

comparison of the mobile networks with the ResNet [6] in the self-supervised learning.

B. Deciding the Best Combination of Transformations

In order to decide the best combination of the transformations, various combinations of the transformations must be tried with the same network. This network can be the ResNet itself or it can be the best performing mobile network.

C. Mobile Networks with the Best Transformation

In this part, the best transformation can be examined with the mobile networks.

D. Overall

The best transformation combination-mobile network duo is going to be tried to find. The number of transformations and the number of mobile networks are not be determined yet. Given the fact that training convolutional neural networks takes time, those numbers must be chosen efficiently.

REFERENCES

[1] I. Misra and L. van der Maaten, “Self-Supervised Learning of Pretext-Invariant Representations,” *arXiv e-prints*, p. arXiv:1912.01991, Dec. 2019.

[2] C. Doersch and A. Zisserman, “Multi-task Self-Supervised Visual Learning,” *arXiv e-prints*, p. arXiv:1708.07860, Aug. 2017.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” *arXiv e-prints*, p. arXiv:2002.05709, Feb. 2020.

[4] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised Representation Learning by Predicting Image Rotations,” *arXiv e-prints*, p. arXiv:1803.07728, Mar. 2018.

[5] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, “FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence,” *arXiv e-prints*, p. arXiv:2001.07685, Jan. 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *arXiv e-prints*, p. arXiv:1512.03385, Dec. 2015.

[7] M. Sandler, A. Howard, B. Chen, M. Tan, R. Pang, V. Vasudeyan, and Q. V. Le, “Mnasnet: Platform-aware neural architecture search for mobile,” in *Proceedings of the IEEE/CVF CVPR*, 2019.

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE/CVF CVPR*, 2018.

[9] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proceedings of the ICML*, 2019.