turnitin

TIC-322006 - Improving Accuracy of Multi-Class Image Classification using Ensemble Learning of Convolutional Neural Networks.doc

Jan 12, 2022

2812 words / 15366 characters

# TIC-322006 - Improving Accuracy of Multi-Class Image Classifi...

## Sources Overview

**8%**

**OVERALL SIMILARITY**

| | | |
|---|---|---|
| 1 | Sri Lanka Institute of Information Technology on 2019-10-07<br>SUBMITTED WORKS | <1% |
| 2 | lirias.kuleuven.be<br>INTERNET | <1% |
| 3 | University of London External System on 2019-05-07<br>SUBMITTED WORKS | <1% |
| 4 | journals.plos.org<br>INTERNET | <1% |
| 5 | National College of Ireland on 2021-02-01<br>SUBMITTED WORKS | <1% |
| 6 | University of Wales Swansea on 2019-05-13<br>SUBMITTED WORKS | <1% |
| 7 | www.mdpi.com<br>INTERNET | <1% |
| 8 | acervodigital.ufpr.br<br>INTERNET | <1% |
| 9 | National College of Ireland on 2020-12-16<br>SUBMITTED WORKS | <1% |
| 10 | University of Aberdeen on 2020-11-03<br>SUBMITTED WORKS | <1% |
| 11 | University of Wollongong on 2021-03-21<br>SUBMITTED WORKS | <1% |
| 12 | Liverpool John Moores University on 2021-08-24<br>SUBMITTED WORKS | <1% |
| 13 | digital.lib.washington.edu<br>INTERNET | <1% |
| 14 | www.hindawi.com<br>INTERNET | <1% |
| 15 | www.shangyexinzhi.com<br>INTERNET | <1% |
| 16 | Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh. "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, 2006<br>CROSSREF | <1% |

17 University of Northumbria at Newcastle on 2020-08-26
SUBMITTED WORKS

<1%

**Excluded search repositories:**

None

**Excluded from document:**

Bibliography

Quotes

**Excluded sources:**

None

# Improving Accuracy of Multi-Class Image Classification using Ensemble Learning of Convolutional Neural Networks

**G M Oktavian* and H Santoso**

Department of Information Technology, Universitas Pradita

*Corresponding author: grady.matthias@student.pradita.ac.id

**Abstract.** Multi-class image classification is one of the tasks of computer vision in which convolutional neural networks are widely used. In this paper, we create a novel ensemble learner by stacking InceptionV3, Xception, MobileNetV2, and ResNet50 convolutional neural network in an attempt to outperform the best performance of each individual models. In an ensemble learning algorithm, by combining multiple different models with different architectures, an error of prediction from a single model can be corrected by other models, thus reducing the likelihood of selecting a wrong prediction. The dataset that we used to do benchmarking is the Intel Image Classification dataset which consists of 6 classes that displays images of fully colored natural scenes. The training set consists of 14034 images, while the test set consists of 3000 images, all having 150x150 as their default size. The training process on all models involves reducing their learning rate when the performance plateaus, and early stopping if there aren't any significant improvements from an epoch to the next. The result is that each individual models have an average of 83% accuracy, yet our ensemble model managed to reach 88% on the test set.

## 1. Introduction

Computer vision is a scientific and technological field in which artificial intelligence is leveraged to find meaningful information and insights from digital photographs, images, videos, or even real-life visual inputs such as web/security cameras [1]. The mining of information from visual data is then used to make classifications, give recommendations, or perform a specific task tailored to the project at hand. Therefore, computer vision enables machines to see and obtain an understanding of what it is seeing [2]. In contrast to human vision where we are adjusted to see through our eyes since the moment we're born, computer vision requires a lot of data and a lot of learning process to enable the machine make inferences and predictions correctly. This is where deep learning comes into the picture. Deep learning is a type of machine learning that uses algorithmic neural networks that allows computer to learn from streams of input data that it reads and teach itself the context of the visual data [3].

Since computers can only read and calculate numerical values, during the process of deep learning, each image is broken down into pixels, and each pixel carries its own unique numerical values (as well as information of colors). In classic machine learning, an image classification process will require every pixel to be treated as features, but this process is computationally exhaustive. This is why most computer vision tasks uses a special type of neural network called Convolutional Neural Network (CNN). The main advantage of CNN is that it can automatically detect the important features of an images (outline, shades, shapes) and learn distinctive features for each class by itself.

There have been several ways that can be done to increase the accuracy of deep learning model to do prediction. In this paper, the novelty lies in conducting an ensemble learning of four Convolutional Neural Network to improve the prediction accuracy of each individual model. In the subsequent parts, we will briefly learn about the architecture of each model, and evaluate the performance of each individual model on a benchmark dataset provided by Intel. After that, we will stack the models together to form an ensemble learning and investigate the performance of this ensemble learner. By using different models, our ensemble learner is more robust and should give an improved accuracy compared to just using an individual model

## 2. Literature Review

In this section, theories and related works about deep learning, image classification, and convolutional neural networks are discussed.

### *2.1. Computer Vision*

Computer Vision is an interdisciplinary field of interest in the study of artificial intelligence that investigates how computers can achieve understanding when looking at visual data [4]. It's an engineering attempt at replicating and automating what human visual system can do.

### *2.2. Convolutional Neural Network*

Convolutional Neural Networks (CNN) are special type of Neural Network which were inspired by the visual system's architecture. A CNN has three main types of neural layers [5]. The first layer is the convolutional layers in which CNN uses different types of kernels to convolve the image to generate feature maps. Processing the image this way enables the algorithm to not process all pixels at once, and all with the same importance. Through kernels, a CNN can learn the pattern carried within a group of pixels, highlighting the relationship of a pixel and its surroundings, and extract different features of the image. The second layer is known as a pooling layer, which is used to reduce spatial complexity of the input data for the next layers. A reduction in dimension, when done right, can lead to reducing computational necessities in the next layers, as well as preventing against overfitting. The last layer is called a fully connected layer. This layer converts 2D feature maps which has gone through convolutions and pooling into 1D feature vector, in which each neuron is connected with one another. The output is this layer is used to classify images into several different classes [6].

#### *2.2.1. Inception*

In this paper, the first CNN that we use to do the image classification task is the InceptionV3 [7]. This CNN's first iteration originates from researchers in Google who theorizes that some sparsity would be beneficial to the network's performances. The novelty that Inception brings to the table is having parallel processing done in the same layer, namely 1x1, 3x3, 5x5 convolutions, as well as max pooling. The 1x1 convolution blocks are used for depth reduction, the 3x3 convolution is used to capture distributed features, while the 5x5 convolution captures global features. This allows each feature of an image to be extracted at each level before they are fed into the next layer. The global architecture of the network is highlighted in figure 1 [8].

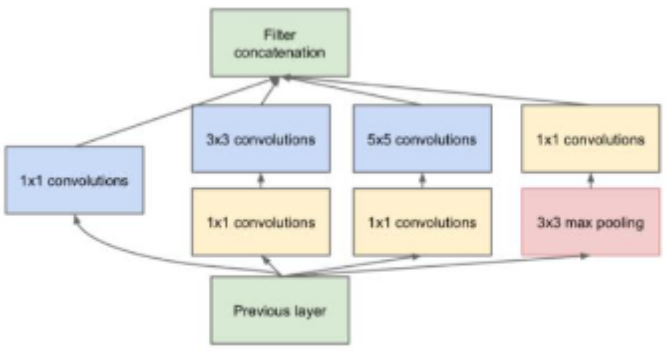


**Figure 1.** Architecture of Inception Convolutional Neural Network

#### *2.2.2. MobileNet*

The second CNN that we explore is MobileNetV2. This CNN's specialty is being smaller and faster in order to tackle real world applications that prioritizes speed more than anything, while attempting to preserve as much accuracy as possible [9]. The novel idea of MobileNet is using depthwise separable convolutions instead of regular convolution layers stacked on top of each other. Depthwise separable convolutions work by separating a convolution process into several smaller kernels, doing each convolution process separately, before concatenating them together in the end. By doing more smaller convolutions instead of less large convolutions, computational intensity is reduced due to the multiplicative characteristic of convolutions, and the additive nature of concatenations.

As an illustration, the 3x3 kernel in figure 2 can be reduced to 2 smaller kernels using the properties of matrix multiplication. If an image is passed through the 3x3 kernel, it will have to go through 9 multiplications, but if we passed the image to 2 smaller kernels, it will only do 3 multiplications each for each kernel, bringing the total number of multiplication operation to only 6.

$$\begin{bmatrix} 3 & 6 & 9 \\ 4 & 8 & 12 \\ 5 & 10 & 15 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

**Figure 2.** Illustration of separating a kernel into two smaller kernels

The architecture of MobileNet can be seen in figure 3.

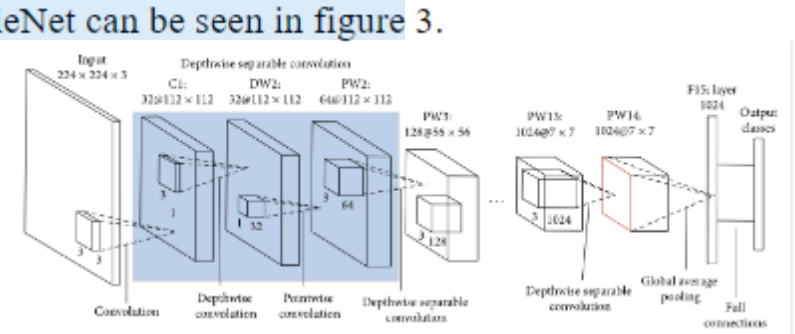


**Figure 3.** Architecture of MobileNet Convolutional Neural Network

*2.2.3. ResNet*

The third CNN that we explore is called ResNet50. ResNet challenges the notion that "deeper neural network is better". In the experiments of the ResNet researchers, they find that attaching more layers to a network does not always achieve a better accuracy than the accuracy of the overall model. Therefore, they introduced a method called "shortcut connection" to deal with this problem, partly due to vanishing gradient problem [10]. Shortcut connection allows the output of an earlier layer to be passed as another input to the next 2-3 layers, thus allowing the later layers to still remember the output of its earlier layers [11]. An illustration of such process is shown in the flowchart depicted in figure 4, in which the shortcut mapping is denoted with the teal colour on the right image.

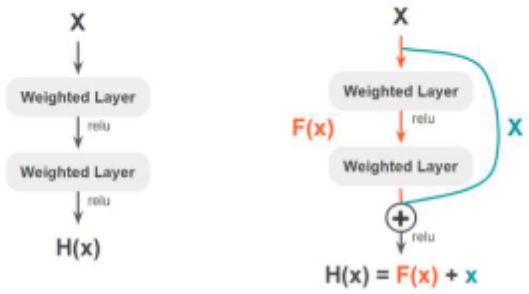


**Figure 4.** Illustration of shortcut connection in ResNet

The full comparison of a ResNet CNN against a regular CNN can be seen in figure 5. ResNet CNN is the CNN on top, while a regular CNN is the model depicted below the ResNet CNN.

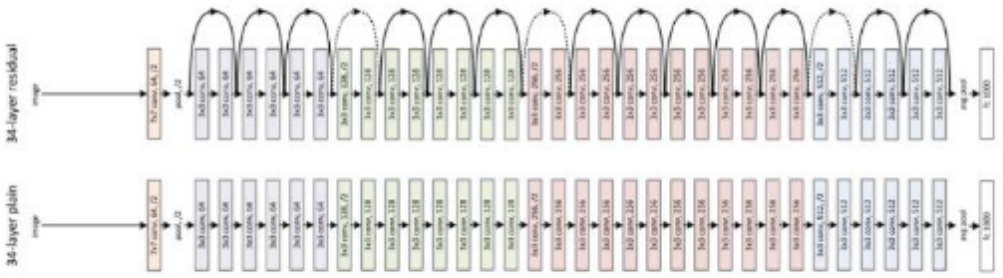


**Figure 5.** Architecture of ResNet Convolution Neural Network

*2.2.4. Xception*

The final CNN that is explored in this paper is the Xception. This modernization of the Inception architecture asks the question about if we should do convolution and processing on the image and channel space at the same time [12]. Image space refers to the shape of the image, while the channel space refers to colors of the image (in an RGB setting, there are 3 channels, the red, green, and blue).

Xception divides each input based on each channel before conducting parallel convolutions like what Inception does, while regular CNN usually conducts convolution without separating each input into different channels according to their color. The architecture of Xception is represented in figure 6 [13].

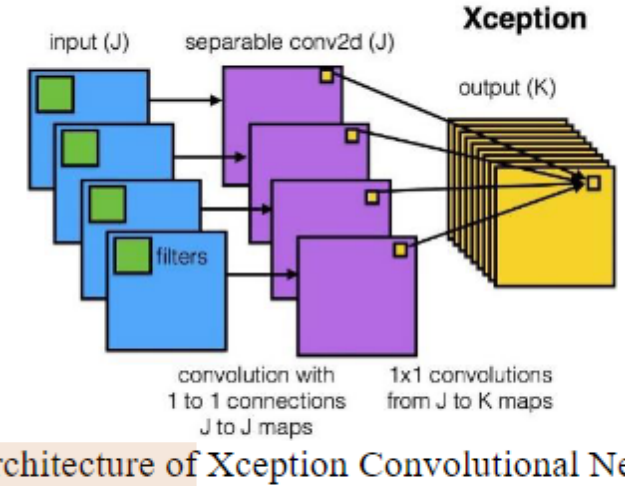


**Figure 6.** Architecture of Xception Convolutional Neural Network

*2.3. Ensemble Learning*

Ensemble learning is the process where multiple models are combined in one way or another in the hopes of achieving a better accuracy than each individual models [14]. In this paper, stacking generalization is performed on four of our CNN – Inception, MobileNet, ResNet, and Xception. Each neural network is trained individually before their output is fed into a generalizer classifier that will give us the final output. The weakness of individual models can be offset by other models, which will help improve the prediction accuracy.

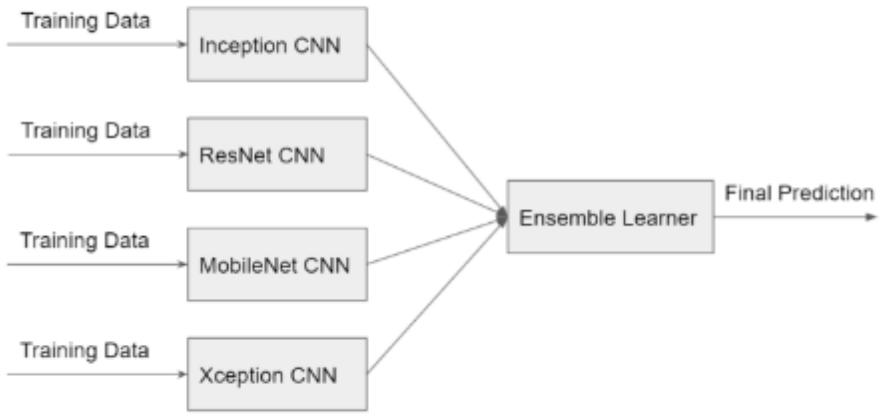


**Figure 7.** Architecture of Ensemble Model

*2.4. Related Works*

In 2019, Wang [15] proposed to use InceptionV3 Convolutional Neural Network to do a pulmonary image classification task. The research indicates that InceptionV3 performed the best compared to other classifiers which were tried during the experiment (softmax, logistic, SVM). During the training process, data augmentation is also done to provide researchers with more robust images. The study concludes that conducting a transfer learning approach to train an InceptionV3 model for pulmonary image classification yields a 86.40% accuracy, which is noted to outperform other models.

In a separate study, Pan (2019) explores the usage of MobileNet to conduct an image recognition and classification method to detect welding defects [16]. Pan describes the MobileNet to be lightweight, having less parameters, and easily implemented in mobile devices in order to perform near-real time image classification. This study proves that MobileNet is one of the key state-of-the-art CNN model to use for lightweight image classification. Since the Intel Image Classification is only 150 x 150 pixels, it would be worth to try to incorporate MobileNet as one of the members of our ensemble learner.

The final study to note is one authored by Jinsakul and Tsai in 2019 [17] where they explore Xception with swish activation function for colorectal polyp preliminary screening. In this study, Xception proves to be the best model which achieves up to 98% for the task that they gave, which is yet another image classification inside the medical field. The highlight of this study is that using different activation functions turns out to be effective in fine tuning the model. In summary, all studies

[15,16,17] use architectures of CNN that we have discussed to do image classification tasks with great success. MobileNet is popular in a more practical environment in which speed is its priority, while Inception and Xception are more popular in medical settings.

A novelty that this study brings to the discussion is that we go further than just fine tuning a single model – we build individual models and then train them together in an ensemble system. Our benchmark image classification dataset is also more robust as it has a wide selection of scenery, so this experiment could explore the performance of models when faced with a more general task (instead of specific ones as we see earlier in medical problems).

## 3. Result

The benchmark dataset that we use consists of 17034 images of natural sceneries provided by Intel, which is divided further into 6 classes: mountains, streets, forests, buildings, sea, and glacier [18]. The distribution of each class in training and testing set are displayed in table 1.

**Table 1.** Number of images per class in training and test set

| Class | Train | Test |
|---|---|---|
| **Buildings** | 2191 | 437 |
| **Forest** | 2271 | 474 |
| **Glacier** | 2404 | 553 |
| **Mountain** | 2512 | 525 |
| **Sea** | 2274 | 510 |
| **Street** | 2382 | 501 |
| **Total** | **14034** | **3000** |

Training is conducted using GPU in a TensorFlow and Keras framework, utilizing the Adam optimizer [19] for a maximum of 60 epochs.

**Table 2.** Accuracy of different convolutional neural networks

| Model | Accuracy |
|---|---|
| **InceptionV3** | 83% |
| **MobileNetV2** | 85% |
| **ResNet50** | 81% |
| **Xception** | 84% |
| **Ensemble Learner** | 88% |

Table 2 shows that creating an ensemble learner that combines four of our convolutional neural networks yield a better accuracy. While each individual models struggle to score an accuracy above 85%, our ensemble learner reaches 88%.

## 4. Conclusion

In this paper, we demonstrate that convolutional neural networks can enable computer to understand and extract information within images, especially in the field of computer vision. Each convolutional neural networks proposed by researchers have its own unique characteristics attempting to solve a particular issue that they find in previous iterations of the network in order to maximize accuracy and efficiency. We can create a more robust neural network that can reduce errors by creating an ensemble learner over multiple neural networks. By using the Intel Image Classification dataset, we have demonstrated that our ensemble learner managed to get a better accuracy than each of the individual model.

## References


[1] Distante, A., Distante, C. (2020). Handbook of Image Processing and Computer Vision: Volume 3: From Pattern to Object. Germany: Springer International Publishing

[2] Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python. (2019). (n.p.): Machine Learning Mastery.

[3] Campesato, O. (2020). Artificial Intelligence, Machine Learning, and Deep Learning. (n.p.): Mercury Learning & Information.

[4] Andres, E., Planche, B. (2019). Hands-on Computer Vision with TensorFlow 2: Leverage Deep Learning to Create Powerful Image Processing Apps with TensorFlow 2.0 and Keras. India: Packt Publishing.

[5] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni-tion, pages 3431–3440, 2015

[6] Dawani, J. (2020). Hands-On Mathematics for Deep Learning: Build a Solid Mathematical Foundation for Training Efficient Deep Neural Networks. United Kingdom: Packt Publishing.

[7] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 2818–2826.

[8] Zeiser, F. A., da Costa, C. A., Ramos, G. de O., Bohn, H., Santos, I., & Righi, R. da R. (2021). Evaluation of Convolutional Neural Networks for COVID-19 Classification on Chest X-Rays.

[9] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

[10] Martens, J., Ballard, A., Desjardins, G., Swirszcz, G., Dalibard, V., Sohl-Dickstein, J., & Schoenholz, S. S. (2021). Rapid training of deep neural networks without skip connections or normalization layers using Deep Kernel Shaping.

[11] Wightman, R., Touvron, H., & Jégou, H. (2021). ResNet strikes back: An improved training procedure in timm.

[12] Lin, H., Luo, W., Wei, K., & Liu, M. (2021). Improved Xception with Dual Attention Mechanism and Feature Fusion for Face Forgery Detection

[13] Chollet, F. (2016). Xception: Deep Learning with Depthwise Separable Convolutions. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-January, 1800–1807

[14] Sarkar, D., Natarajan, V. (2019). Ensemble Machine Learning Cookbook: Over 35 Practical Recipes to Explore Ensemble Machine Learning Techniques Using Python. India: Packt Publishing.

[15] Wang, C., Chen, D., Hao, L., Liu, X., Zeng, Y., Chen, J., & Zhang, G. (2019). Pulmonary image classification based on inception-v3 transfer learning model. IEEE Access, 7, 146533–146541.

[16] Pan, H., Pang, Z., Wang, Y., Wang, Y., & Chen, L. (2020). A New Image Recognition and Classification Method Combining Transfer Learning Algorithm and MobileNet Model for Welding Defects. IEEE Access, 8, 119951–119960.

[17] Jinsakul, N., Tsai, C.-F., Tsai, C.-E., & Wu, P. (2019). Enhancement of Deep Learning in Image Classification Performance Using Xception with the Swish Activation Function for Colorectal Polyp Preliminary Screening. Mathematics 2019, Vol. 7, Page 1170, 7(12), 1170.

[18] Bansal, P. (2018). Intel Image Classification | Kaggle. https://www.kaggle.com/puneet6060/intel- image-classification

[19] Theis, F., Kůrková, V., Karpov, P. (2019). Artificial Neural Networks and Machine Learning – ICANN 2019: Deep Learning: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part II. Germany: Springer International Publishing.