Scenery Classification Using Convolutional Neural Network Towards Indonesia Tourism

Nana Ramadijanti, Tita Karlita, Achmad Basuki, Ulima Inas Shabrina, Feri Afrianto, Andro Aprila Adiputra, Muhammad Dzalhaqi
Department of Informatics and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
{nana; tita; basuki}@pens.ac.id, {uishabrina; feriafrianto; androaprila12}@it.student.pens.ac.id, dzalhaqi@ds.student.pens.ac.id

***Abstract*—Indonesia is a country that possesses natural wonders and historical buildings, which made Indonesia become one of the popular tourist destination. The scenery classification is a challenging task where the feature distribution from each image may spread. In addition, Indonesian tourism spots are plentiful. In this research, we proposed to utilize deep learning for tourism spot classification using Convolutional Neural Network (CNN) as feature extraction. The dataset consists of different tourism varieties, from man-made objects such as monuments to natural objects such as mountains or beaches. The dataset was self-gathered from the internet and various social media with different angles and does not include any images that dominantly contain people. In addition, in the context of CNN as the basis of feature extractor, we also compared the result with pre-trained CNN architecture trained with Place-365 and ImageNet dataset. The first test was conducted with shallow CNN achieving 48% for the non-augmented dataset and 51% for the augmented dataset. The second test performed with VGG16 and ResNet, combining data augmentation and a pre-trained network. The result reveals data augmentation improves the validation accuracy. Pre-trained VGG16 with Place-365 achieved the highest validation of 90% compared to the other combination. A pretrained network with an augmentation combination improves the model performances significantly by a rough margin.**

Keywords*— *Convolutional Neural Network; Scenery Classification; Pretrained Network

I. INTRODUCTION

Tourism is one of the huge sectors for Indonesia. Based on The Travel & Tourism Competitiveness Report by World Economic Forum (WEF), in 2019, Indonesia's tourism sector placed at 40th position among 140 countries and 12th among 22 pacific countries [1]. Indonesia has diverse tourist spots. However, some of them are overshadowed by famous sites and remain unheard of by most tourists.

In today's era, people share ideas, photos, videos, and posts with others to maintain their social relationships. We can find news and information through social networking services [2]. As the number of users connected to network platforms has increased exponentially, social networking services can be used as the primary data source in various fields. The development of social media services contributes to the increasing amount of information about tourism spots represented as images rather than text [3]. As a result, tourists who are interested in a particular tourist spot shown in the image may not know how to do a text search for more information about that tourist spot. This research is based on these problems and to improve the tourism market's competitiveness. This research proposes an innovative tourism spot identification and recognition mechanism, based on deep learning-based object detection technology, for automatic detection and identification of tourist objects by taking pictures at the location or taking pictures from the internet [4] [5] The efficient solution is to utilize technologies by applying scenery classification based on images. CNN is one of the popular and robust methods to extract various features from images.

Previous research [5] uses the Convolutional Neural Network model to detect locations in photos. The author collects a dataset in the form of imagesfrom 15 cities in Turkey. For photo location detection, the author compared different CNN models(VGG16, ResNetv50, and Inceptionv3). The results obtained from the experiment achieved the best results for VGG16 models with Place365 pre-trained models.

This research examines the technology of face detection recognition systems [6]. The model used in this research is a CNN-based architecture to design a facial expression recognition system using the CNN algorithm Visual Geometry Group 16 (VGG16). The dataset used in this system is FER2013, where this dataset has a large number of facial images, totaling 35,887 images with seven categories of emotions. The best performance results were obtained in the proposed model, a modified VGG16 model with the parameters using augmentation data, epoch 100, and learning rate 0.001, which reached a test accuracy of 70.63%. This accuracy is much better than previous studies using the VGG16 model base and the FER2013 dataset.

This research proposes a system to classify landscape images into different groups with several classes (sunset, desert, mountain, tree, and sea) [7]. A single model that predicts different label probabilities has used a probabilistic threshold value for each label to change the probabilities in the presence and absence of classes/labels. This research proposes a ConvNet model for the classification of natural landscape images. This method results in higher accuracy and requires less time than other methods.

A. Azizah and E. D. Ariyani (Eds.): iCAST-SS 2022, ASSEHR 719, pp. 656–660, 2022.
https://doi.org/10.2991/978-2-494069-83-1_114

This research proposes a new deep architecture for calculating scene categories by deriving stable templates hierarchically using a generative model [8]. The researchers use subspace embedding to create a semantic space by incorporating image-level labels. The researchers extracted a group of superpixels from a large number of scene pictures to represent objects and their parts. From scene images with contaminated labels, a probabilistic model hidden stability analysis then learns stable templates for each scene category. Aggregation Deep Network combines them based on the learned per stable category templates that describe local scene composition to capture the global scene composition. Finally, an image kernel for scene classification receives these learned deep representations. Extensive testing has shown that this method works. Empirical research of 33 SIFT-flow categories reveals that the stable learned templates remain nearly unchanged even when image label contamination rates reach almost 36%.

This research proposes a method for automatically classifying tourist photos by tourist attractions using image feature vector clustering and a deep learning model [9]. The datasets were compiled by searching TripAdvisor for photos and reviews posted by foreign tourists. Individual images were embedded as 512-dimensional feature vectors using the VGG16 network pretrained with Places365 and reduced to two dimensions using t-SNE (t-Distributed Stochastic Neighbor Embedding). This research used the Siamese Network to remove noise from photos within the cluster and classify them according to category. As a result, which visual elements of tourist attractions appeal to tourists could be identified. This method has the advantages of not requiring the creation of a classification category in advance, extracting categories for each tourist destination flexibly, and improving classification performance even with a relatively small dataset volume.

All the researchers mentioned [6] [5] [7] [9] [7] [8] used an ideal data set with a maximum of 15 classes and got excellent results. However, whether this model can classify the condition of imbalanced datasets with a total of 30 classes is still a question. Identification is needed to answer the underlying problem. This research will focus on finding out which CNN model has the most optimal results with unbalanced dataset conditions. However, the traditional CNN model alone is certainly not strong enough. End-to-end models are needed to enhance the dataset context with various preprocessing or augmentation approaches. In this research, we consider an experiment with a deep learning architectural model based on a CNN using an imbalanced dataset with a total of 30 classes. We experiment with an efficient model to overcome the data shortage complications for imbalanced datasets.

II. Materials And Method

The focus in this research is to identify the performance of CNN based model toward the underlying task. In addition, we will compare the result of shallow CNN with the pre-trained model to identify the potential of the CNN-based model to tackle the underlying problem.

We will oversample the dataset due to the deep learning characteristic of demanding more data and identify the model performances.

Based on Fig 1, the dataset will be split into two training and testing. Only the training set will be augmented, and the purpose is to differentiate the distribution between the train set and the test set. In the desired result, the model will not be biased.

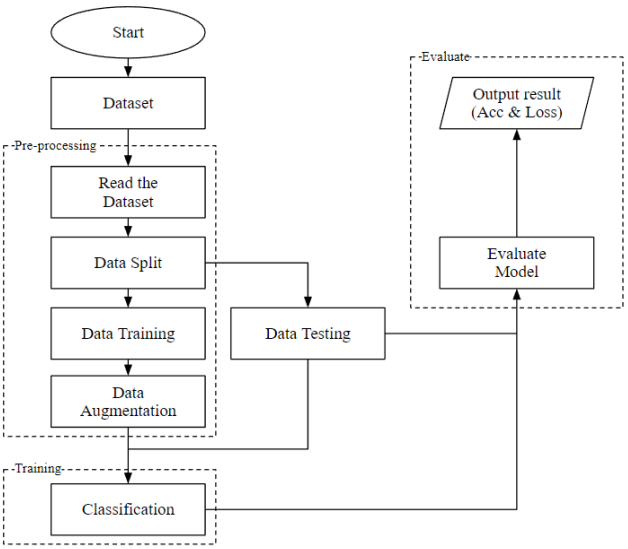

Fig. 1. Working flow

A. Dataset Characteristic

The dataset should be prepared before using it as input for the convolutional neural networks as part of deep learning. The prediction dataset and the training dataset should be the same as possible. For example, your workout data should consist of low-resolution photos if your use case contains images taken using your phone's camera. In general, it is worth considering providing multiple perspectives, resolutions, and backgrounds for training photos. Data augmentation techniques can be used to expand the sample size, including resizing, translation, blurring, and orientation modification. The images in the dataset are saved in Joint Photographic Experts Group (JPEG) and Portable Network Graphics (PNG) format. The objects in the dataset image must be photos of real-world objects, and the maximum image size must be 256 pixels. If the size is more than 256 pixels, some image quality may be lost during normalization.

Based on Fig. 2, the selection of tourist objects used in this research is iconic tourist sites in Indonesia. Tourist objects are not limited to only one type of tourism. This research will use photos of monuments, buildings, or natural objects.

Based on Fig. 3, the selection of photos from the dataset used is photos with different angles and lighting, affecting the training results on the CNN model. The learning process in this research carried out supervised learning, so the images used for training data must be managed first before the learning process. The datasets were taken using scraping techniques on the Google Images platform. The photo taken is an iconic tourist location. There are several specifications for dataset management: maintaining photos of iconic objects at tourist sites and discarding pictures that are not ideal.

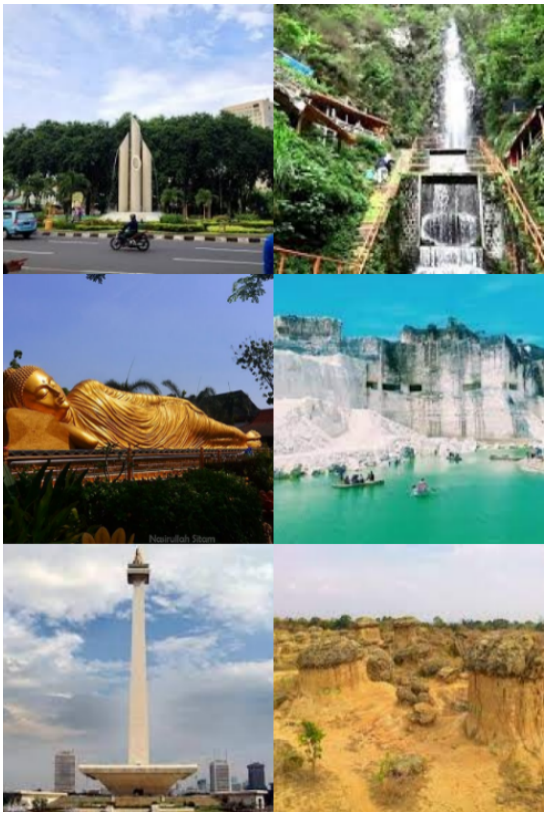

Fig 2. Tourism Place Photos for Dataset

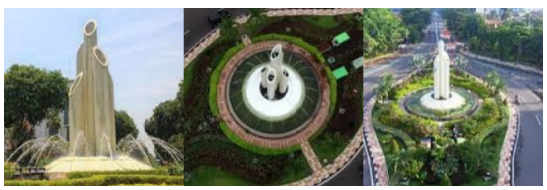

Fig 3. Dataset Characteristic

The dataset was collected with a total of 4256 images. The total dataset is divided into two parts for training and validation. The total number of images for training is 844, and the rest are validation datasets. In the dataset development process, a problem called dataset imbalance causes misclassification; in other words, misclassification makes some classes undetectable. To solve the problem, the number of each class in the dataset must be equal to the others, the image size must be the same, and the distribution of the dataset must be spread out.

30 class of tourist attraction were used in this research. As we can see on Fig 4, the dataset was randomly selected from tourist attractions in Indonesia and was limited to only 30 classes. Collecting a dataset rangin from 90 to 180 images in each class to minimize the inequality of large margin differences.

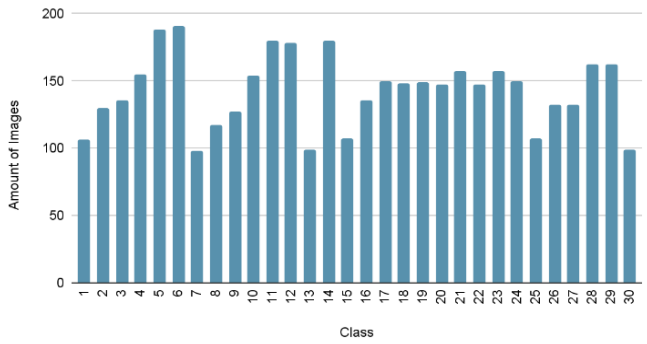

Fig 4. Dataset Distribution

B. *Convolutional Neural Network (CNN)*

Convolutional neural networks are part of the deep learning architecture after artificial neural networks. An artificial neural network inspired by how the human brain works. Each neuron receives input and operates points with weights additions and bias additions. The result will be an activation function parameter that will be the neuron's output. The mathematical model of the output will look like Equation (1).

$$f(\sum_1^\infty \omega_i x_i + b) \quad (1)$$

The architecture of the neural network is shown in Fig. 5.

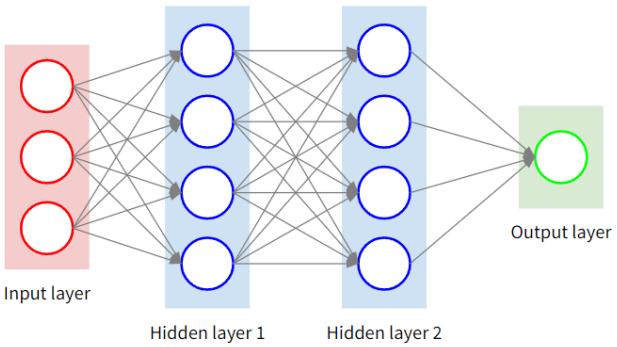

Fig 5. Neural Network Architecture

From the architecture in Fig. 5, the first architecture has three input layers and an output layer of 2 nodes. Weight is the relationship between nodes. The convolutional neural network is an architecture developed based on artificial neural networks. Certain convolutional neural networks it has one or more layers of convolutional units. The convolution unit receives its input from several previous layer units. The weights on the convolutional neural network are shared. Convolutional neural network architecture with a convolutional unit layer reduces computational complexity, so this architecture is very good for image processing [10].

As part of the deep learning method, Convolutional Neural Network requires a collection of datasets in the form of images. The machine must learn each data set to determine the output

value found in the real-time image. Deep learning consists of feature extraction and classification.

C. Training Parameter

TensorFlow with Jupyter notebook is used in this research. Parameters are needed to test the accuracy of the model. This parameter includes 100 epochs used in each experimental model, has not used dropout or augmentation, and the learning level used is 0.0001.

D. Proposed Model

This research uses the shallow CNN, VGG16, and ResNet v50 models for feature extraction and classification in Indonesia's tourism location recognition system. Fig 6 is the shallow CNN model architecture used in this research.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 128, 128, 128)     9728

conv2d_1 (Conv2D)            (None, 64, 64, 128)       409728

activation (Activation)      (None, 64, 64, 128)       0

conv2d_2 (Conv2D)            (None, 64, 64, 128)       409728

conv2d_3 (Conv2D)            (None, 32, 32, 128)       409728

activation_1 (Activation)    (None, 32, 32, 128)       0

conv2d_4 (Conv2D)            (None, 32, 32, 128)       409728

conv2d_5 (Conv2D)            (None, 32, 32, 128)       409728

activation_2 (Activation)    (None, 32, 32, 128)       0

dropout (Dropout)            (None, 32, 32, 128)       0

flatten (Flatten)            (None, 131072)            0

dense (Dense)                (None, 256)               33554688

dense_1 (Dense)              (None, 256)               65792

activation_3 (Activation)    (None, 256)               0

dense_2 (Dense)              (None, 30)                7710

=================================================================
Total params: 35,686,558
Trainable params: 35,686,558
Non-trainable params: 0
_________________________________________________________________
```

Fig 6. Shallow Model Structure

III. Experiments

A. Abbreviations and AcronymTesting Scenario

Tests are carried out using one parameter value for one operation on the system. These parameters are used on the system whose output is the model's accuracy with certain parameters. Parameter 100 epoch and learning rate used is 0.0001. The form should be completed and signed by one author on behalf of all the other authors.

B. Test Results

From all the tests carried out on the system that has been built, the results obtained are the accuracy of the model used. Table 1 shows the results of the accuracy of the test data from the shallow model with certain parameters.

Table 1. Results of Shallow Model

Augmentation	*Acc*	*Val Acc*	*Loss*	*Val Loss*
Yes	*0.9668*	*0.5156*	*0.1245*	*2.9784*
No	*0.9961*	*0.4824*	*0.0391*	*3.4868*

This research experimented with the transfer learning model; the default epoch parameter was 100. The experiment used a dense layer of 128 and a dropout of 0.5. We compared two transfer learning models, VGG16 and ResNet. The experiments were compared with the differences in the pre-trained datasets using ImageNet and Place365. It was found that the pre-trained dataset with optimal results was ImageNet rather than Places365. Then the experiment was continued again to know whether the ResNet model or VGG16 was more optimal.

Table 2. Results of Transfer Learning Model

Model	*Augmen-tation*	*Dataset Pre-Trained*	*Acc*	*Val Acc*	*Loss*	*Val Loss*
VGG16	No	Place365	0.9789	0.8808	0.2979	0.8302
ResNet	No	Place365	0.9531	0.6699	0.0689	0.4495
VGG16	No	ImageNet	0.9844	0.0664	0.0642	9.8936
ResNet	No	ImageNet	0.9931	0.6699	0.0689	1.2495
VGG16	Yes	Place365	0.9857	0.9062	0.1028	0.3613
ResNet	Yes	Place365	0.9583	0.8962	0.0619	0.3821
VGG16	Yes	ImageNet	0.9853	0.8862	0.2853	0.7891
ResNet	Yes	ImageNet	0.9783	0.8962	0.0619	0.3621

The experiment results show that the VGG16 model with pre-train uses the Place365 dataset, which is the most optimal learning to recognize the Indonesian tourism dataset. Accuracy and validation accuracy reached 98% and 90%, while Loss and loss validation reached 10% and 36%, respectively.

The learning process is carried out in 100 epochs. The results of each step tested are recorded, and the loss of the compared image must be close to zero or zero to represent the learning process's accuracy. Furthermore, if the total loss value is getting further away from zero, it means the learning process is getting worse. The total data loss recorded after learning is shown in Fig 7.

The total loss value shows that the error from the learning process decreases. The more steps are taken, the smaller the loss value. And also, based on Fig 7, there is a decrease in validation which means the model results are not overfitting or underfitting.

The accuracy graph in Fig 8 shows the accuracy and validation line increases. Accuracy shows the ratio of how correctly the model learns from the dataset being studied, while validation shows how well the model sees the distribution of different images.

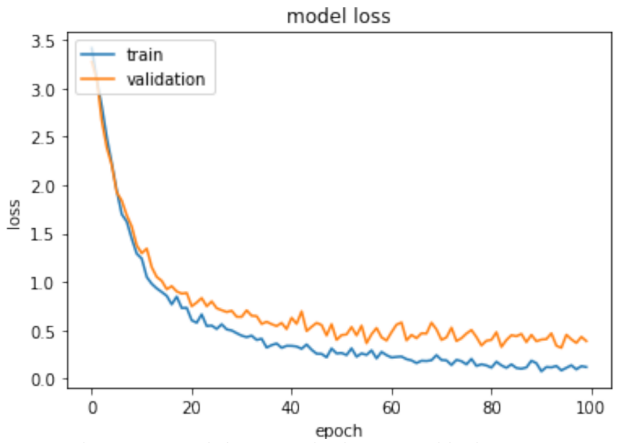

Fig 7. Loss Graph for Pre-Trained VGG16 with Places365

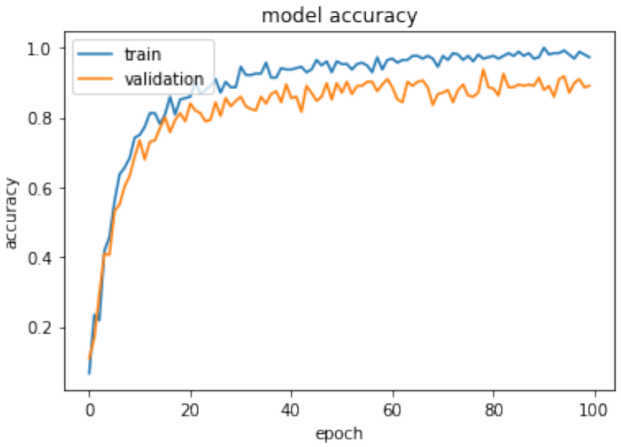

Fig 8. Accuracy Graph for Pre-Trained VGG16 with Places365

IV. Conclusion

The CNN-based model for this scenery classification of Indonesia tourism achieved an excellent result on the pre-trained model. However, the shallow CNN model validation accuracy reaches 51% with augmentation and 48% with non-augmentation, which means the model still cannot generalize the features inside the dataset. With augmentation, the model may be able to figure out more about the important feature due to some similar pattern. In pretrained model, the highest possible validation accuracy result could achieve 90% with VGG16-Place365 and augmentation. The pre-trained model that did not utilize data augmentation tends to have validation accuracy lower than 80%, yet the VGG16-Place365 non-augmentation achieves 88% validation accuracy.

Overall, CNN is a robust model for this task. The consideration that properly is taken is data augmentation is significant to easier the model to extract the essential features. The feature taken from Place365 helps the model extract the features easily and significantly improves the performance.

Acknowledgment

We would like to express our sincere gratitude to Knowledge Engineering Laboratory - Politeknik Elektronika Negeri Surabaya for supporting this research.

References

[1] N. Ramadhani, J. Hendryli and D. E. Herwindiati, "PENCARIAN OBJEK WISATA BERSEJARAH DI PULAU JAWA MENGGUNAKAN CONVOLUTIONAL NEURAL NETWORK," *JURNAL ILMU KOMPUTER DAN SISTEM INFORMASI,* p. Vol. 7 No. 1, 2019.

[2] Y. Kang, N. Cho, J. Yoon, S. Park and J. Kim, "Transfer Learning of a Deep Learning Model for Exploring Tourists' Urban Image Using Geotagged Photos," *ISPRS International J. Geo-Inf.,* pp. 10(3), 137, 2021.

[3] Y.-C. Chen, K.-M. Yu and T.-H. Kao, "Deep Learning Based Real-Time Tourist Spots Detection and Recognition Mechanism," *Sage Jorunals,* p. Vol. 104, 2021.

[4] T. Hirotsu, M. Hirota, T. Araki, M. Endo and H. Ishikawa, "Tourism application with CNN-Based Classification specialized for cultural information," *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services,* 2019.

[5] Y. E. Ozkose, T. A. Yilikoğlu, L. Karacan and A. Erdem, "Finding Location of A Photograph with Deep Learning," *IEEE,* 2018.

[6] R. J. Gunawan,, B. Irawan and C. Setianingsih, "Pengenalan Ekspresi Wajah Berbasis Convolutional Neural Network Dengan Model Arsitektur Vgg16," *eProceedings of Engineering,* p. Vol. 8 No.5, 2021.

[7] A. R. Rout and S. B. Bagal, "Natural Scene Classification Using Deep Learning," *International Conference on Computing, Communication, Control and Automation (ICCUBEA),* 2017.

[8] L. Zhang, X. Ju, Y. Shang and X. Li, "Deeply Encoding Stable Patterns From Contaminated Data for Scenery Image Recognition," *IEEE Transactions on Cybernetics,* 2019.

[9] J. Kim and Y. Kang, "Automatic Classification of Photos by Tourist Attractions Using Deep Learning Model and Image Feature Vector Clustering," *International Journal of Geo-Information,* 2022.

[10] F.-F. Li, J. Johnson and S. Yeung, "Convolutional Neural Networks for Visual Recognition," in *CS231n*, San Francisco, 2018.