Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering

Basılı/Printed ISSN: 1302-9304. Elektronik/Online ISSN: 2547-958X

A Deep Learning-Based Hotel Image Classifier for Online Travel Agencies

Çevrimiçi Seyahat Acenteleri için Derin Öğrenme Yöntemleri Kullanarak Otel Görüntülerinin Sınıflandırması

Fatma Bozyiğit [1*], Alperen Taşkın [2], Kadir Akar[3], Deniz Kılınç[1]

[1] İzmir Bakırçay University, Department of Computer Engineering, İzmir, Turkey
[2,] Metglobal , İstanbul, Turkey
[3] Tatilbudur Seyahat Acenteliği ve Turizm A.Ş, İstanbul, Turkey
*Sorumlu Yazar / Corresponding Author **: fatma.bozyigit@bakircay.edu.tr

Geliş Tarihi / Received: 25.03.2020 *Araştırma Makalesi/Research Article*
Kabul Tarihi / Accepted: 28.06.2020 *DOI:10.21205/deufmd.2021236722*
Atıf şekli: BOZYIĞIT,F., TAŞKIN, A., AKAR, K., KILINÇ, D. (2021). A Deep Learning-Based Hotel Image Classifier for Online Travel Agencies. DEUFMD, 23(67), 257-264.

Abstract

Along with the extensive application of information technologies, there has been increased usage of online travel agencies (OTA) to search holiday alternatives. Hotel images demonstrated by OTAs plays a critical role in providing information to ease selection process. Although serving the images in the right context is an important task to clearly reflect hotel properties, there has been no attempt in previous studies to organize hotel images into appropriate category. For this reason, we aim to conduct a study to organize and classify 20,000 hotel images using Convolutional Neural Networks (CNN), a prominent deep learning method widely applied in the field of computer vision. Due to the limited training data, we experiment transfer learning to train experimented models. In this phase, we choose a widely applied CNN models, VGG-16, VGG-19, and Inception-v3 which are trained on over one million images. The results demonstrate that experimented models achieve effective categorization of hotel images with the considerable accuracy scores. We believe that our study can help improve OTAs performance in competitive tourism market.

Keywords: *Online travel agencies, hotel images, image classification, Convolutional Neural Networks, VGG-16, VGG-19, Inception-v3, transfer learning.*

Öz

Bilgi teknolojileri uygulamaların yaygınlaşması ile birlikte, tatil alternatifleri arayışında çevrimiçi seyahat acentelerinin kullanımı artmıştır. Çevrimiçi seyahat acenteleri tarafından sunulan otel görüntüleri, tatilcilere bilgi sağlayarak seçim sürecini kolaylaştırma aşamasında kritik bir rol oynamaktadır. Görüntülerin doğru bağlamda sunulması, otel özelliklerini açıkça yansıtmak için önemli bir işlem olduğu halde otel görüntülerini düzenleme ve uygun kategorilere yerleştirme girişiminde bulunan bir çalışmaya rastlanılmamaktadır. Bu sebeple, çalışmamıza bilgisayarlı görü alanında yaygın olarak uygulanan önemli bir derin öğrenme yöntemi olan Konvolüsyonel Sinir Ağları kullanılarak 20.000 otel görüntüsünü sınıflandıran bir yaklaşım gerçekleştirimi hedeflenmiştir. Sınırlı eğitim verileri nedeniyle, önerilen modelimizin eğitimi aşamasında transfer öğrenme yöntemi uygulanmıştır. Bu bağlamda, önerilen yaklaşımımız için bir milyondan fazla görüntü üzerinde eğitilmiş, yaygın olarak uygulanan CNN modelleri olan VGG-16, VGG-19 ve Inception-v3 tercih edilmiştir. Sonuçlar, test ettiğimiz modellerin otel görüntülerini göz ardı edilemeyecek doğruluk skoru ile etkin bir şekilde sınıflandırılmasını sağladığını göstermektedir. Çalışmamızın rekabetçi turizm pazarında çevrimiçi seyahat acentelerinin performansını artırmaya yardımcı olabileceğine inanmaktayız.

Anahtar kelimeler: *Çevrimiçi seyahat acentaları, görüntü sınıflandırma, Konvolüsyonel Sinir Ağları, VGG-16, VGG-19, Inception-v3, transfer öğrenme.*

1. Introduction

The World Wide Web (WWW) has drastically increase the amount of electronically exchanged information. These changes bring new forms to the supply-demand model in tourism industry.

Nowadays, travellers generally prefer booking through online travel agencies (OTAs) serving as a hotel reservation channel owing to provided price comparisons and the visualization of the holiday destinations [1].

The content of the visuals is one of the most important parameters while deciding the holiday alternatives. If the images do not reflect well the hotel information, a traveller may not be sure about booking. Thus, the hotel images in OTAs must be categorized with the most appropriate content in order to give reliable information to the travellers.

OTAs use more than one global distribution system (GDS) that provides a single point of access for thousands of travel agents to reach a broad base of customers [2]. In OTA, hotel images which is rendered from Global Distribution Systems' (GDS) databases, is generally labelled with human involvement. As the number of GDS used increases, certainly, classification of images has become a much more difficult to manually perform. According to our view, there is a need to develop efficient methods to automatically label hotel images with relevant tags.

Recently, deep learning models such as convolutional neural networks (CNN) [3] have been extensively used for a wide range of visual perception tasks, such as object detection/classification, action/activity recognition, and so on [4]. However, a solution utilizing image classification task to organize visual contents of OTAs is still lacking. According to our view, this is one of the critical research gaps in tourism marketing that must be filled.

The overall objective of our study is to design, implement and evaluate CNN based models that automatically finds relevant tags for hotel images. Our proposed approach is performed on the dataset including 20,000 hotel images on the website of the "Make my trip". All the photographic images in the dataset are categorized into 215 sub-categories based on the tags of ImageNet [5] dataset by utilizing three state-of-the-art deep models VGG-16 [6], VGG-19 [7], and Inception-v3 [8]. Finally, we compare the results of each experimented methods and we provide the corresponding results in related sections. To the best of our knowledge, this study is the first attempt to label the hotel images with relevant content using deep learning models. We believe that this paper will bring new insights into future works on the relevant issues in tourism industry.

The rest of the paper is organized as follows: In Section 2, the literature review is introduced. Section 3 presents the details materials and methods. In Section 4, the experimental setup is described and obtained experimental results are discussed. At last, the conclusion and future works takes part in Section 5.

2. Related Works

It is generally believed that images of holiday resorts are the most critical factor affecting the hotel selection process. Some researchers have presented studies to control relevance of the hotel images in OTAs with the reality be. Mackay and Couldwell conduct a study to check whether the images presented in promotional materials correspond to those generated by visitors [9]. They aim to provide realistic and consistent information to the customer. As a result, they state that data from the Visitor-Employed Photography study provide valuable information confirming the visitors' experiences and image of the site. In the other study, Phelps categorizes destination images into primary and secondary regarding information sources [10]. While primary images are formed through internal data such as past experiences, secondary images are obtained from some external sources (i.e., first-time visitors). He concludes that analyses on the primary and secondary data show deviations between the reality and preconceived images.

The content of the visuals is also necessary parameter while choosing holiday destination by the potential travellers. Thus, the hotel images must be categorized with the relevant tags in order to give reliable information to the customers. With the development of Deep Learning, it became motivating to use such approaches to classify hotel images. Zhang et al. state that there is significant value in optimizing images in e-commerce settings [11]. They perform a CNN model on 16-month Airbnb panel dataset to classify the aesthetic quality for each image in the training sample. Experimental results contribute insights for housing and lodging e-commerce managers (of Airbnb,

hotels, realtors, etc.) to optimize product images for increased demand. In another study [10], the researcher investigates the economic impact of visuals and lower-level image factors that influence Airbnb's demand. He performs difference-in-difference analyses on a dataset including 13,000 accommodation images in 7 U.S. cities, from January 2016 to April 2017. VGG-16 transfer learning is used for training proposed CNN model. Evaluation results show that units with verified photos (taken by Airbnb photographers) generate 8.9% more demand per year on average.

3. Materials and Methods

3.1. Dataset

The dataset on the Kaggle contains data of approximately 20,000 hotels having features such as the address of the hotel, the number of stars and the links of hotel images. Although there are approximately 20,000 hotel data in the dataset, only 8,600 of them have image links. Some of these records have only 1 image link, while others contain 6-7 image links. Therefore, approximately 35,000 image links are obtained. With a program developed using the Python programming language, 7,209 of them are visually controlled and 5,000 of them are downloaded.

Since the downloaded 5,000 images have no category, they are categorized using the VGG16 model pre-trained with the ImageNet data set. While predicting the category of a hotel picture, the category with the maximum label score is selected from the most likely 5 categories. At the end of this process, there are 215 unique categories from 1,000 categories. The names of some of these categories and the number of pictures they contain are as follows: 'four-poster', 'restaurant', 'studio_couch' and 'patio' categories contain 647, 420, 253, and 246 images respectively. Among the 215 categories, the ones having the most pictures are selected. For example, the 'patio' category for the 'patio' label, the 'four-poster' and 'studio_couch' categories for the 'bedroom' label, and the 'restaurant' category for the restaurant label are selected.

After making the category selections, there are 6 labels left to classify; '0-bathroom', '1-bedroom', '2-patio', '3-restaurant', '4-building' and '5-others'. 'others' is removed from the dataset because it includes pictures from more than one category. As a result, 4,691 pictures are left behind.

3.2. Convolutional Neural Network (CNN)

CNNs consist of multiple layers of neurons that learn to derive the essential features of an image to perform the image classification task. The types of layers are categorized into three types; convolutional, pooling, and fully connected layers. The model receives the 2D image as an input. Each layer takes the output of the previous layer as its input. The depth of the output depends on the filter size of is convolutional layers for the generation of feature maps. When the features of the image are available in the convolutional layer, the pooling layer joins similar features, and the network's performance becomes robust to image deformation [12]. The pooling layer also provides dimension reduction of the feature map. Consequently, densely connected convolutions at the top of the stacked layers ultimately connect to the output units.

A CNN trained for image classification with the ImageNet dataset is regarded as also performing well with the other experimental datasets.

In this study, we experiment VGG-16, VGG-19, and Inception-v3 which are the CNN models pre-trained with ImageNet dataset.

VGG-16: The name of the model originates from the fact that it has 16 layers. Its layers are convolutional, max pooling, activation, and fully-connected. A convolutional layer includes a set of filters whose weights need to be learned. After each convolution layer, it is common to add a pooling layer. Activation functions enable the model to learn complex functional mappings between the inputs and response variables by introducing non-linearity. The fully connected layer subdivides the image into features and analyses them separately to generate the output. The network has 41 layers: 1 input, 13 convolution, 5 max-pooling, 15 rectified linear unit (relu), 2 dropout, 3 fully connected, 1 softmax, and 1 output. Sixteen out of total layers have learnable weights: 13 convolutional layers, and 3 fully connected layers. Conv 1, Conv 2, Conv3 have 64, 128, and 256 filters respectively. Conv 4 and Conv 5 have 512 filters.

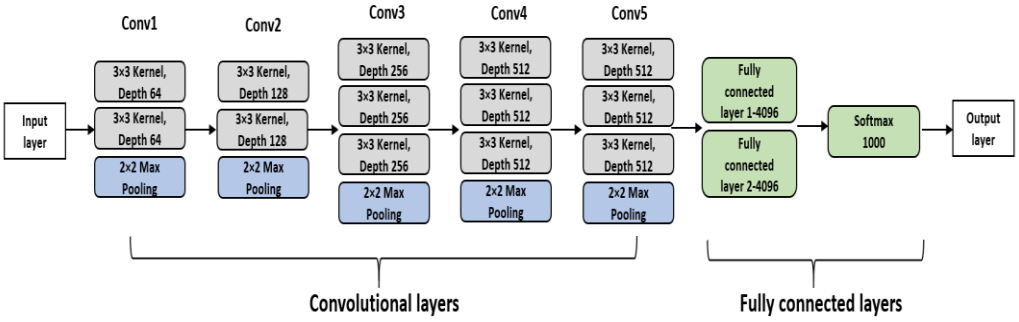

Figure 1. VGG16 architecture.

VGG-19: VGG-19 is the most recent version of the VGG models and its architecture is similar to VGG16. It has 19 layers with extra convolution layers in the last three blocks. The network has 47 layers: 1 input, 13 convolutional, 5 max pooling, 15 relu, 2 dropout, 3 fully connected, 1 softmax, and 1 output. There are 19 layers with learnable weights: 16 convolutional layers, and 3 fully connected layers. Conv 1, Conv 2, Conv3 have 64, 128, and 256 filters respectively. Conv 4 and Conv 5 have 512 filters.

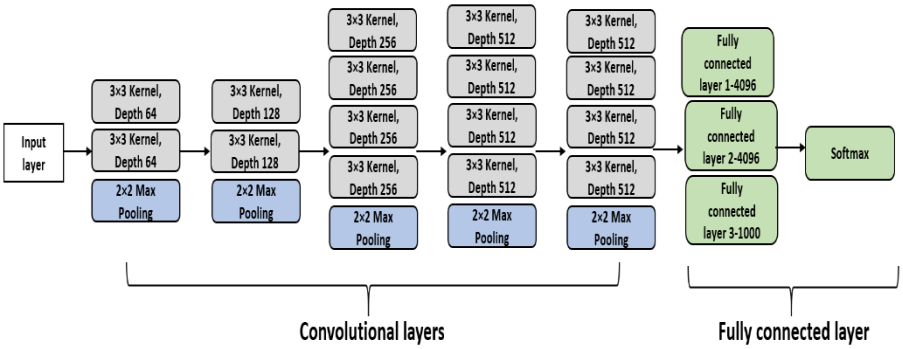

Figure 2. VGG19 architecture.

Inception-v3: The main objective of Inception is to make the CNN model wider by utilizing a parallel connection of various layers having different filters. Then all of those parallel paths are joined to convey pass to the next layers. Inception-v3 has 22 layers in where fully learned filters are available. The Inception-v3 architecture has an average pooling layer with 5×5 filter size, 1×1 layer with 128 filters for dimension reduction and rectified linear activation, a fully connected layer with 1024 units and rectified linear activation, dropout layer with 70% ratio of dropped outputs.

InceptionNets are preferable as they are not only deeper, but also wider and use less amount of computation. A traditional convolutional layer tries to learn filters (weights) in a 3D space, with width, height, and channel dimensions. Thus, a single convolution is associated with concurrently planning cross-channel correlations and spatial correlations.

3.3 Transfer learning

Transfer learning is a method of reusing a pre-trained model information for another deep learning task. It can be used for all types of machine learning categories: i) classification, ii) regression and iii) clustering problems. As the main idea behind the transfer learning is to obtain identified data from another related problem, performing transfer learning significantly improve the performance of proposed approach [13]. Figure 3 illustrates the general differences of learning process between conventional machine learning and transfer learning. As it can be seen from the figure, conventional machine learning tries to learn

each task separately while transfer learning tries to obtain the knowledge from previous trained model. With the advent of transfer learning, a prominent way to deal with the absence of training data or computational cost of training such models is to pre-train models on images from another domain. In this study, a typical example of transfer learning is the use of VGG-16, VGG-19, and Inception-v3 which have been pre-trained on the ImageNet database. Including a very large collection of images with annotations to be experimented in academic researches conducting image classification tasks. As the models are trained on a huge dataset, representations of low-level features like spatial, edges, rotation, lighting, shapes are learned, and these features are transmitted to provide the knowledge transfer for new image classification problems.

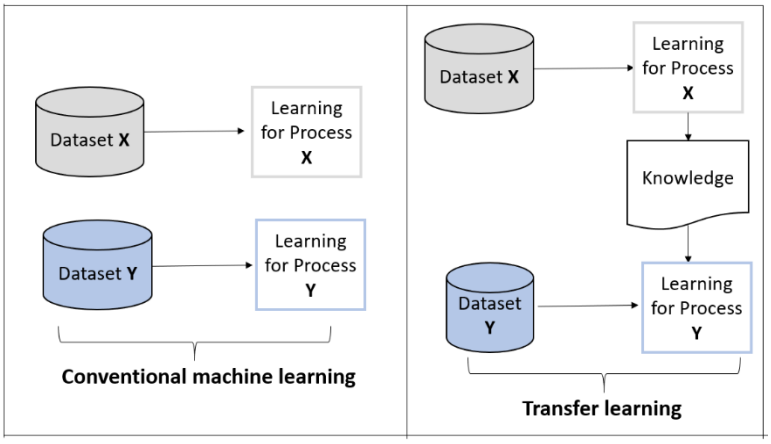

Figure 3. Comparison of conventional and transfer learnings.

4. Experimental Study

In this section, implementation details, experimental studies and evaluation of the experimented CNN models are introduced.

4.1 Pre-processing

In pre-processing phase, first we utilize data augmentation because of class sampling problem in our dataset. We use Keras ImageDataGenerator [14] and imbalanced-learn python package to balance amount of image samples per-class. In this way, our model does not see twice the exact same image and overfitting problem is eliminated.

After data augmentation, images in the dataset are resized to a size of 255x255pixels to achieve efficient computational results. Then, we apply the 'One hot encoding' to make dataset suitable for Keras CNN Library. One hot encoding is a process by which categorical variables are converted into binary vectors that enable data science algorithms to perform better in classification.

4.2. Experimental results

After the pre-processing steps are utilized, three baseline CNN methods, which are commonly used to classify the textual data, are implemented. In this study, the experimented models are performed using scikit-learn and Keras Python libraries in the GPU-supported Google Colaboratory [15] service since the model's training takes a long time.

In this study, F-measure is used to evaluate the performances of the experimented models. F-measure metric is calculated based on confusion matrix outcomes. In other words, F-measure is calculated with the use of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) outcomes. A TP is a result where classifier correctly predicts the positive label. And similarly a TN is a result of the classification if the algorithm predicts the negative label correctly. FP is the case where the classifier predicts negative class as positive. The last confusion matrix term, i.e. FN, is the prediction of positive label as negative. The precision in

terms of TP, FP, TN is calculated with the Equation 1.

$$\textbf{Precision (Pr)} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}} \qquad (1)$$

Similarly recall is calculated with the use of Equation 2.

$$\textbf{Recall (Re)} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}} \qquad (2)$$

In order to calculate the accuracy of the proposed model, the harmonic mean of the precision and recall values are obtained and the F-measure is calculated according to the equation given in Equation 3.

$$\textbf{F}_{\textbf{measure}} = \frac{\textbf{2(Pr} \times \textbf{Re)}}{\textbf{Pr} + \textbf{Re}} \qquad (3)$$

The evaluation results of each methods are obtained with the use of 10-fold cross validation. Overall results of the experiments are given in Table 1.

Table 1. Evaluation results of the experimented methods (lr=0.0005, batch_size=32, epochs=20)

Model	Precision	Recall	F-measure
VGG16	84.86%	84.09%	84.31%
VGG19	82.68%	81.25%	81.38%
Inception3	76.21%	72.51%	72.57%

Considering the evaluation results of the experimented CNN models, it is obviously seen that VGG-16 has better performance scores than VGG-19 and Inception-v3."

To further improve previous evaluation results of VGG-16, we utilize "fine-tune" on the last convolutional block of the VGG-16 model with the training samples from the initial pre-trained model for generic object recognition in Keras [16]. Accordingly, learning rate is determined as 0.0005 and batch size is set to 32. The last 3 layers of the VGG-16 model are pre-trained with 35 epochs for the transfer learning process. Figure 4 shows the loss and acc (accuracy) values for each epoch.

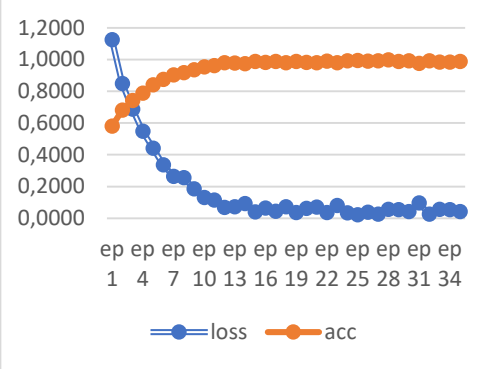

Figure 4. The loss and acc values of fine-tuned VGG16 model

When the confusion matrix in Figure 5 is examined, it is seen that the optimized VGG-16 model achieves good accuracy results in the categories of 'restaurant', and 'bathroom'. It is also seen that, the number of errors is higher in the 'patio' and 'building' categories. Actually, this result is not surprising because it is easy to confuse 'building' and 'patio' categories due to having similar visual contents like shape of objects, colours, etc.

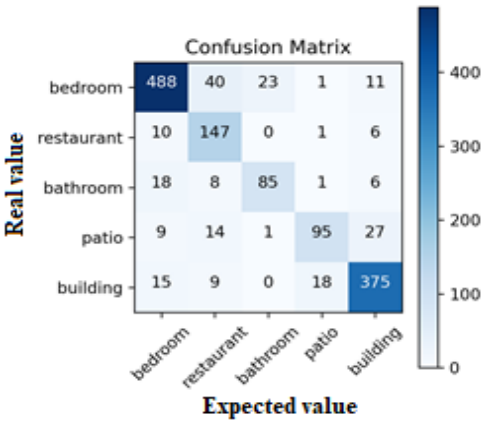

Figure 5. Confusion matrix of the fine-tuned VGG-16 model.

Figure 6 shows the corresponding ROC curve. Derived indices, such as the area under the entire curve (AUC), the True Positive Rate at a specific False Positive Rate, or the partial area corresponding to a labelled with relevant tag range of False Positive Rate, are the most commonly used to measure model accuracy [17].

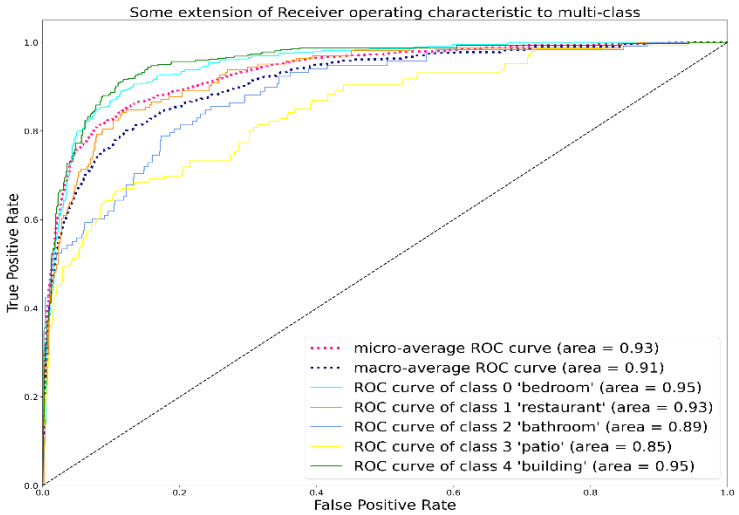

Figure 6. Area under the ROC curve for prediction of image label

The loss is the value attempted by a deep neural network to minimize and the accuracy is the percentage of instances which are categorized correctly. In a well-designed deep neural network, the value of acc increases while the loss decreases. Figure 7 shows the val_loss and val_acc values for each epoch. The loss and val_loss are the metrics obtained from training and test datasets respectively. Similarly, acc is the accuracy result on the training and val_acc on the test dataset. It's best to rely on val_acc for a fair representation of model performance. Therefore, in the study, 0.84 val_acc value is selected as the final performance of the deep neural network VGG-16.

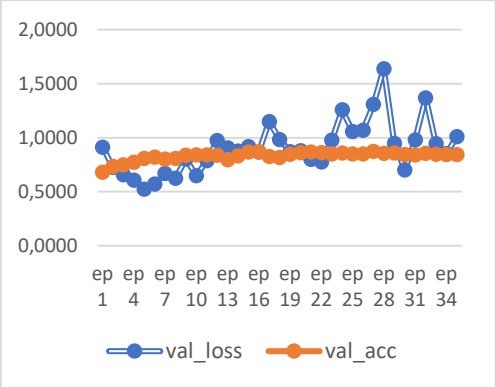

Figure 7. The val_loss and val_acc values of fine-tuned VGG16 model

5. Conclusion and Future Works

Internet has become an important distribution channel in the tourism industry. Due to increasing number of online services in travel marketing, more people tend to use OTAs that provide platforms to access more contents such as images and videos.

It is important to organize the hotel images in accordance to the relevant categories for online booking sites. The other considerable point is the order in which hotel images is displayed. For example, images such as bathrooms and toilets should appear after then building of the hotel, the facilities such as pool and restaurant. Since organizing of hotel images is difficult and time consuming process to manually perform, there is a need to develop automatic hotel image classifier/organizer.

In this paper, we investigate the success of three CNN methods, VGG-16, VGG-19, and Inception v3, in image classification problems. The overall aim of this study is to label hotel images with the most appropriate tags and set the order of images. In this direction, we evaluate and compare the results of the experimented deep learning methods. The result of the optimized VGG-16 is superior to the VGG-19 and Inception-v3 algorithms with the 84.53% F-measure value. As a future work, higher amount of tagged image data can be used to improve system

performance. Also, we plan to use a deeper network with enhanced training.

Acknowledgements

Funding for this work was partially supported by the Research and Development Center of Tatilbudur.com accredited on Turkey Ministry of Science.

References

[1] Ling, L., Dong, Y., Guo, X., Liang, L. 2015. Availability management of hotel rooms under cooperation with online travel agencies, International Journal of Hospitality Management, vol. 50, pp. 145-152.

[2] Hatton, M. 2004. Redefining the relationship: The future of travel agencies and the global agency contract in a changing distribution system, Journal of Vacation Marketing, vol. 10, no. 2, pp. 101-108.

[3] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826.

[4] Hyungtae, L, Heesung, K. 2017. Going deeper with contextual CNN for hyperspectral image classification", IEEE Transactions on Image Processing.

[5] Deng, J., Dong, W., Socher, R., Li, J., Li, K., Fei, L. 2009. ImageNet: A large-scale hierarchical image database. IEEE conference on computer vision and pattern recognition, pp. 248-255. IEEE.

[6] Qassim, H., Verma, A., Feinzimer, D. 2018. Compressed residual-VGG16 CNN model for big data places image recognition. In IEEE 8th Annual Computing and Communication Workshop and Conference, pp. 169-175. IEEE.

[7] Mateen, M., Wen, J., Song, S., Huang, Z. 2019. Fundus image classification using VGG-19 architecture with PCA and SVD.Symmetry, vol. 11, no. 1.

[8] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence.

[9] MacKay, K. J., Couldwell, C. M. 2004. Using visitor-employed photography to investigate destination image. Journal of Travel Research, vol. 42, no. 4, pp. 390-396.

[10] Phelps, A. 1986. Holiday destination image the problem of assessment: An example developed in Menorca. Tourism management, vol. 7, no. 3, pp. 168-180.

[11] Zhang, S., Lee, D., Singh, P. V., Srinivasan, K. 2017. How much is an image worth? Airbnb property demand estimation leveraging large scale image analytics. Airbnb Property Demand Estimation Leveraging Large Scale Image Analytics.

[12] Zhang, S. 2019. A structural analysis of sharing economy leveraging location and image analytics using deep learning. Carnegie Mellon University, PhD Thesis.

[13] Simonyan, K., Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[14] Pan, S.J., Yang, Q. 2009. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345-1359.

[15] Bisong, E. 2019. Google Colaboratory. Building Machine Learning and Deep Learning models on Google Cloud platform, pp. 59-64. Apress, Berkeley, CA.

[16] Gulli, A., Pal, S. 2017. Deep learning with Keras. Packt Publishing Ltd.

[17] Walter, S. D. (2005). The partial area under the summary ROC curve. Statistics in medicine, vol. 24, no. 13, pp.2025-2040.