High-Performance Object Recognition by Employing a Transfer Learned Deep Convolutional Neural Network

Md. Mehedi Hasan*, Azmain Yakin Srizon†, Abu Sayeed‡ and Md. Al Mehedi Hasan§
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh
Email: *mmehedihasann@gmail.com, †azmainsrizon@gmail.com, ‡abusayeed.cse@gmail.com, §mehedi_ru@yahoo.com

***Abstract*—Researchers have been conveying and producing various models for the precise identification of objects for a long period now. Distinguished contributions are notable in the field of object identification. Numerous benchmark datasets are now available for composing an effective model that can discover a large number of target classes without a tremendous loss in performance. In this study, we began with a benchmark dataset Caltech-101 which consists of 102 classes. Being an extremely imbalanced dataset and having a blend of RGB and Gray images, the recognition is extremely challenging. We proposed and implemented a modified DenseNet-201 design after the preprocessing steps and produced an overall accuracy of 99.73% which exceeded all the previous results or achievements by a noteworthy boundary.**

***Index Terms*—Object Recognition, Caltech-101, Transfer Learning, Modified DenseNet-201, Augmentation**

I. Introduction

Object detection is a combined piece of computer vision and image processing that operates with recognizing appearances of semantic aims of a distinct group e.g. persons, buildings, or vehicles in digital photos and videos [1]. Fully reviewed fields of object detection cover face detection and pedestrian revealing. Object identification has utilization in diverse realms of computer vision, including photo retrieval and video inspection. Object discovery is largely employed in computer vision jobs e.g. pictorial explanation [2], movement study [3], face detection, face classification, video target segmentation. It is additionally studied in tracing targets, for example tracking a ball while a football or soccer match, tracking the movement of a cricket or baseball bat, or following a person in a video.

Every target group has its private sole characteristics that serve in recognizing the group – for example, each circle is round. Target group identification employs these unique properties. For example, when looking for circles, targets that are at a particular range from a point (the center) are investigated. Furthermore, while browsing for squares, targets that are orthogonal to ridges and have the same side measures are demanded. A similar method is exercised for face classification where eyes, nostrils, and mouth can be discovered and characteristics like skin condition and range separating eyes can be achieved. Beforehand, comprehensive efforts have been done on object discovery. For our research, we examined the Caltech-101 dataset. To precisely classify or identify the target classes, firstly, some preprocessing steps were performed. After that, the dataset was divided into the train set, validation set, and test set. We implemented a modified DenseNet-201 model on the training data. Test data was employed to achieve the performance of our proposed model. At the end of the study, our proposed architecture achieved an overall accuracy of 99.73% which outperformed all the previous studies by a distinguished margin.

II. Literature Review

Object detection has been a region of concern for the researchers for a decade now [4], [5]. Both machine learning [6], [7] and deep learning [8], [9] concepts have been applied for the recognition of targets so far. In this study, we've analyzed a benchmark dataset Caltech-101. Previously, many works have been accompanied by this dataset. In 2009, Lee et al. suggested a CNN based strategy to identify the classes and produced an overall accuracy of 65.4% [10]. In 2018, Song et al. proposed the principal component analysis technique on SIFT features to achieve 83.9% overall accuracy [11]. Another research during the same year proposed an EL+YCbCr based method to identify the classes with an accuracy of 78% [12]. A research effort of Pan et al. recommended a k-mean reduction design on deep CNN features achieving 85.78% accuracy [13]. In 2019, a deep CNN and SIFT feature-based method produced an accuracy of 89.7% by practicing entropy-based determination and deep fusion [14]. Throughout the same year, a study was proclaimed concerning augmentation during identification obtaining an accuracy of 86.9% [15]. Later that year, another study presented an overall accuracy of 91.8% [16]. An research of 2020 achieved an overall accuracy of 94.38%, 91.13%, 92.07%, 89.5% and 87.77% for VGG-16, ResNet-50, MobileNet, DenseNet-121 and NASNetMobile respectively [17]. Furthermore, during the same year, another study recently achieved a classification accuracy of 90.1% [18].

III. Materials and Methods

A. Dataset Description

In this study, we examined the Caltech 101 benchmark dataset involving 102 classes [19]. There were various num-

978-0-7381-1102-5/20/$31.00 ©2020 IEEE

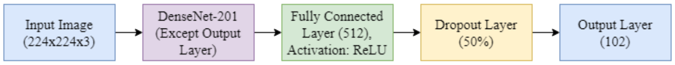
Input Image (224x224x3)
DenseNet-201 (Except Output Layer)
Fully Connected Layer (512), Activation: ReLU
Dropout Layer (50%)
Output Layer (102)

Fig. 1: Modified DenseNet-201 Architecture

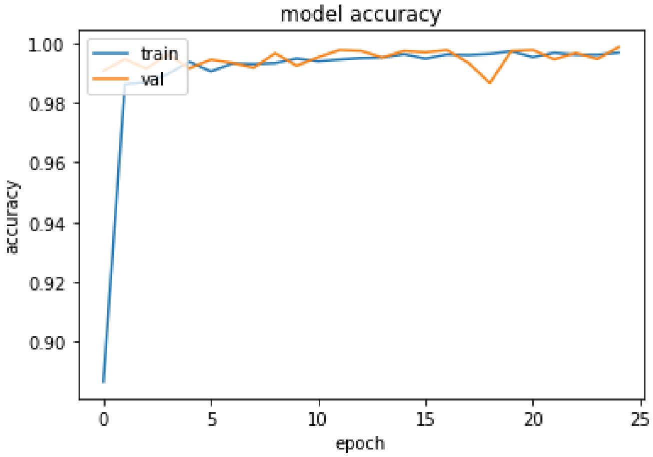

Fig. 2: Training accuracy and validation accuracy of our proposed architecture

bers of samples per class; hence, it was an imbalanced dataset. The dataset includes both RGB and gray pictures which makes the identification method more complex. There was a total of 9145 pictures in the dataset.

B. Transfer Learning

Transfer learning focuses on gathering knowledge collected while determining one obstacle and performing it to another but similar predicament [20]. For example, the knowledge gathered while learning to distinguish cars could utilize during the perception of trucks. This field of research shows an exceptional connection to the enduring history of cerebral investigation on the transfer of learning, though certain similarities among the two fields are incompetent. From a pragmatic perspective, transferring or dispatching knowledge from earlier succeeded assignments for the training of new jobs can dramatically improve the individual performance of an agent.

C. Modified DensNet-201 Architecture

The widespread foundation of Convolutional Neural Network (CNN) can be located in [21]. Nevertheless, in conventional CNN, all layers are constantly correlated which makes the network difficult to stretch wider and deeper, as it may evolve beyond difficulties of either collapsing or gradient missing conditions. After that, ResNet proposed an approach to engaging the shortcut attachment by jumping at least two layers. Then, DenseNet additionally developed the model by concatenating all the characteristic graphs sequentially instead of summation of the output characteristic charts from all former layers. In this study, we've introduced a modified DenseNet-201 architecture which is represented in Figure-1. After the DenseNet-201 basic layers, we implemented a fully connected layer of size 512 supported by a dropout layer of 50%. Finally, an output layer is attached having a size of 102. While training by the design, no layer was held frozen.

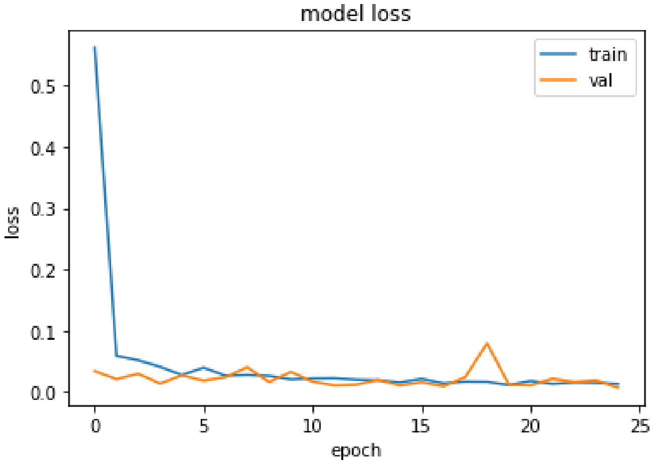

Fig. 3: Training loss and validation loss of our proposed architecture

D. Augmentation

Inadequate data has forever been a remarkable restriction while performing deep learning structures like convolutional neural networks. Furthermore, imbalanced data in terms of labels can be a supplementary obstacle. While there may be adequate data for some groups, uniformly significant, but the under-sampled groups will undergo ineffectual class-specific performance or suitability. This aspect is compatible. If the model learns from a few examples or events of a presented class, it is less feasible to prophesize the group label and test label. Image augmentation artificially produces training photographs through various methods of processing or association of multiple processing, such as random rotation, transfers, shear, and flips, etc. [22].

IV. Experimental Analysis

In this section, firstly, preprocessing will be discussed. After that design of the experiment and result analysis will be presented.

A. Preprocessing

Because of rendering pictures to a convolutional neural network, a heavy preprocessing of the images was jumped as CNN is a compelling network that can extract important characteristics from raw pictures. Nevertheless, some preprocessing measures were needed. The input photographs were in various aspects, therefore, they were reshaped to 224x224x3. For more proper identification, augmentation was implemented with the compensation of the Augmentor Library [34]. While implementing the procedure of augmentation, the max left rotation, the max right rotation, and the probability of rotation of the rotation function was fixed to 3, 3, and 0.4 respectively. The values of grid width, grid height, probability,

TABLE I: Class-wise Precision, Recall and F1-Score

Classes	Precision	Recall	F1-Score	Classes	Precision	Recall	F1-Score
Background_Google	0.99	0.85	0.91	helicopter	0.99	1.00	1.00
Faces	1.00	1.00	1.00	ibis	1.00	1.00	1.00
Faces_easy	1.00	1.00	1.00	inline_skate	1.00	1.00	1.00
Leopards	1.00	0.96	0.98	joshua_tree	0.96	1.00	0.98
Motorbikes	1.00	1.00	1.00	kangaroo	1.00	1.00	1.00
accordion	1.00	1.00	1.00	ketch	0.99	1.00	1.00
airplanes	1.00	0.99	0.99	lamp	1.00	1.00	1.00
anchor	1.00	1.00	1.00	laptop	0.98	1.00	0.99
ant	1.00	1.00	1.00	llama	1.00	1.00	1.00
barrel	1.00	1.00	1.00	lobster	1.00	1.00	1.00
bass	1.00	1.00	1.00	lotus	0.97	1.00	0.99
beaver	0.99	1.00	1.00	mandolin	1.00	1.00	1.00
binocular	1.00	1.00	1.00	mayfly	1.00	1.00	1.00
bonsai	1.00	0.97	0.98	menorah	1.00	1.00	1.00
brain	1.00	1.00	1.00	metronome	1.00	1.00	1.00
brontosaurus	1.00	1.00	1.00	minaret	1.00	1.00	1.00
buddha	1.00	1.00	1.00	nautilus	0.99	1.00	1.00
butterfly	1.00	0.98	0.99	octopus	0.99	1.00	1.00
camera	1.00	1.00	1.00	okapi	1.00	1.00	1.00
cannon	1.00	1.00	1.00	pagoda	1.00	1.00	1.00
car_side	1.00	1.00	1.00	panda	1.00	1.00	1.00
ceiling_fan	1.00	1.00	1.00	pigeon	1.00	1.00	1.00
cellphone	1.00	1.00	1.00	pizza	0.99	1.00	1.00
chair	1.00	1.00	1.00	platypus	1.00	1.00	1.00
chandelier	1.00	1.00	1.00	pyramid	1.00	1.00	1.00
cougar_body	1.00	1.00	1.00	revolver	1.00	0.99	0.99
cougar_face	0.99	1.00	1.00	rhino	1.00	1.00	1.00
crab	1.00	1.00	1.00	rooster	1.00	1.00	1.00
crayfish	1.00	1.00	1.00	saxophone	1.00	1.00	1.00
crocodile	0.97	1.00	0.99	schooner	1.00	0.99	0.99
crocodile_head	1.00	1.00	1.00	scissors	1.00	1.00	1.00
cup	1.00	1.00	1.00	scorpion	1.00	1.00	1.00
dalmatian	1.00	1.00	1.00	sea_horse	0.99	1.00	1.00
dollar_bill	0.99	1.00	1.00	snoopy	1.00	1.00	1.00
dolphin	1.00	1.00	1.00	soccer_ball	1.00	1.00	1.00
dragonfly	1.00	1.00	1.00	stapler	1.00	1.00	1.00
electric_guitar	1.00	1.00	1.00	starfish	1.00	1.00	1.00
elephant	1.00	1.00	1.00	stegosaurus	1.00	1.00	1.00
emu	1.00	1.00	1.00	stop_sign	0.99	1.00	1.00
euphonium	1.00	1.00	1.00	strawberry	1.00	1.00	1.00
ewer	1.00	1.00	1.00	sunflower	1.00	1.00	1.00
ferry	0.99	1.00	1.00	tick	1.00	1.00	1.00
flamingo	1.00	1.00	1.00	trilobite	1.00	1.00	1.00
flamingo_head	1.00	1.00	1.00	umbrella	1.00	1.00	1.00
garfield	1.00	1.00	1.00	watch	1.00	1.00	1.00
gerenuk	1.00	1.00	1.00	water_lilly	1.00	1.00	1.00
gramophone	0.99	1.00	1.00	wheelchair	1.00	1.00	1.00
grand_piano	1.00	1.00	1.00	wild_cat	0.99	1.00	1.00
hawksbill	1.00	1.00	1.00	windsor_chair	1.00	1.00	1.00
headphone	1.00	1.00	1.00	wrench	0.99	1.00	1.00
hedgehog	1.00	1.00	1.00	yin_yang	1.00	1.00	1.00

and magnitude of the random distortion function were set to 4, 4, 0.4, and 4 respectively. Moreover, the percentage regions and the probability of the zoom random function were set to 0.9 and 0.2 respectively. After the augmentation scheme, we had 400 pictures per group, a total of 40,800 images.

B. *Experimental Settings*

The model was trained for 25 epochs with a batch size of 24 as after that the validation loss became approximately constant for the rest of the epochs. 'Adam' optimizer [23] with the learning rate of 0.0001 was employed to maximize the error

TABLE II: Comparison between our proposed work and notable previous works

Classifier or Model Name	Overall Accuracy
Lee et al [10]	65.40%
Song et al [11]	83.90%
Li et al [12]	78.00%
Pan et al [13]	85.78%
Rashid et al [14]	89.70%
Cubuk et al [15]	86.90%
Sawada et al [16]	91.80%
Basha et al [17]	94.38%
Hussain et al [18]	90.10%
Proposed	**99.73**%

function. Categorical cross-entropy function was employed for the loss or error function. For bypassing overfitting, the dropout method was utilized.

C. Result Analysis

Firstly, the augmented dataset was divided into the train set and test set. 80% of the data were stored in the train set and the rest of the 20% data was stored to test set. Then, the proposed modified DenseNet-201 architecture was applied to the train set. Figure-2 represents the training accuracy and validation accuracy of our proposed design. On the contrary, Figure-3 represents the training loss and validation loss of our proposed design. TABLE I represents the precision, recall and f1-score of each of the classes under consideration. Our proposed design produced an overall accuracy of 99.73% for the Caltech-101 dataset. TABLE II represents the comparison among our proposed work and distinguished former works. From Table-2 it can be remarked that our proposed design outperformed all the previous strategies by a notable border, hence, our model is competent in recognizing the respected classes more perfectly.

V. Conclusion

In this study, we examined a benchmark dataset, Caltech-101 which involves 102 classes. It is a challenging job to produce high accuracy for all the groups under consideration as despite having a large number of groups, the dataset is extremely imbalanced and the incorporation of RGB and Gray pictures made the job more challenging. We implemented augmentation first to balance the dataset. Next, modified DenseNet-201 architecture was employed. After that, we evaluated the performance on the test set and obtained an overall accuracy of 99.73% which is the highest obtained accuracy till now. From the result analysis, we settled that our approach outperformed all the previous works by a notable margin.

References

[1] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V.-K. Papastathis, and M. G. Strintzis, "Knowledge-assisted semantic video object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1210–1224, 2005.

[2] L. Guan, Y. He, and S.-Y. Kung, *Multimedia image and video processing*. CRC press, 2012.

[3] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *2007 IEEE 11th international conference on computer vision*. IEEE, 2007, pp. 1–8.

[4] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[5] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[6] S. Varma, M. Shinde, and S. S. Chavan, "Analysis of pca and lda features for facial expression recognition using svm and hmm classifiers," in *Techno-Societal 2018*. Springer, 2020, pp. 109–119.

[7] S. Ahlawat and A. Choudhary, "Hybrid cnn-svm classifier for handwritten digit recognition," *Procedia Computer Science*, vol. 167, pp. 2554–2560, 2020.

[8] G. Melotti, C. Premebida, and N. Gonçalves, "Multimodal deep-learning for object recognition combining camera and lidar data," in *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE, 2020, pp. 177–182.

[9] N. Wang, Y. Wang, and M. J. Er, "Review on deep learning techniques for marine object recognition: Architectures and algorithms," *Control Engineering Practice*, p. 104458, 2020.

[10] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 609–616.

[11] J. Song, G. Yoon, H. Cho, and S. M. Yoon, "Structure preserving dimensionality reduction for visual object recognition," *Multimedia Tools and Applications*, vol. 77, no. 18, pp. 23 529–23 545, 2018.

[12] M. A. Khan, T. Akram, M. Sharif, M. Awais, K. Javed, H. Ali, and T. Saba, "Ccdf: Automatic system for segmentation and recognition of fruit crops diseases based on correlation coefficient and deep cnn features," *Computers and electronics in agriculture*, vol. 155, pp. 220–236, 2018.

[13] Y. Pan, Y. Xia, Y. Song, and W. Cai, "Locality constrained encoding of frequency and spatial information for image classification," *Multimedia Tools and Applications*, vol. 77, no. 19, pp. 24 891–24 907, 2018.

[14] M. Rashid, M. A. Khan, M. Sharif, M. Raza, M. M. Sarfraz, and F. Afza, "Object detection and classification: a joint selection and fusion strategy of deep convolutional neural network and sift point features," *Multimedia Tools and Applications*, vol. 78, no. 12, pp. 15 751–15 777, 2019.

[15] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 113–123.

[16] Y. Sawada, Y. Sato, T. Nakada, S. Yamaguchi, K. Ujimoto, and N. Hayashi, "Improvement in classification performance based on target vector modification for all-transfer deep learning," *Applied Sciences*, vol. 9, no. 1, p. 128, 2019.

[17] S. Basha, S. K. Vinakota, S. R. Dubey, V. Pulabaigari, and S. Mukherjee, "Autofcl: Automatically tuning fully connected layers for transfer learning," *arXiv preprint arXiv:2001.11951*, 2020.

[18] N. Hussain, M. A. Khan, M. Sharif, S. A. Khan, A. A. Albesher, T. Saba, and A. Armaghan, "A deep neural network and classical features based scheme for objects recognition: an application for machine inspection," *Multimed Tools Appl. https://doi. org/10.1007/s11042-020-08852-3*, 2020.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[20] J. West, D. Ventura, and S. Warnick, "Spring research presentation: A theoretical foundation for inductive transfer," *Brigham Young University, College of Physical and Mathematical Sciences*, vol. 1, no. 08, 2007.

[21] Y.-D. Zhang, C. Pan, X. Chen, and F. Wang, "Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling," *Journal of computational science*, vol. 27, pp. 57–68, 2018.

[22] M. D. Bloice, C. Stocker, and A. Holzinger, "Augmentor: an image augmentation library for machine learning," *arXiv preprint arXiv:1708.04680*, 2017.

[23] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. IEEE, 2018, pp. 1–2.