International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Image Classification Using Deep Learning

Aman Ashutosh[1], Shubham Kumar[2], Aman Kumar[3], Aryan Dev[4], Neeraj Singh[5], Er. Sandeep Kaur[6]

[1]*Aman Ashutosh Computer Science & Engineering Chandigarh University*
[2]*Shubham Kumar Computer Science & Engineering Chandigarh University*
[3]*Aman Kumar Computer Science & Engineering Chandigarh University*
[4]*Aryan Dev Computer Science & Engineering Chandigarh University*
[5]*Neeraj Singh Computer Science & Engineering Chandigarh University*
[6]*Er. Sandeep Kaur(Assistant Professor) Computer Science & Engineering Chandigarh University*

Abstract - Image classification is an important topic of study in the field of image processing nowadays and is a popular area of research. By providing the computer with data to learn from, image categorization was created to close the gap between computer vision and human vision. In this paper, the methods for categorising images using traditional machine learning and deep learning are compared and investigated. This study employs a tensor flow framework and convolutional neural networks to classify images. This paper implements CNN in binary classification and multi-class classification for object identification and analyses the performance of well-known convolutional neural networks (CNNs). We built five unique image datasets on our own for multiclass classification, using the dog vs. cat dataset for binary classification. For our investigation, we classified the photos using a separate machine learning model and then classified them again using CNN because evaluating CNN's performance on a single data set hides its actual potential and limits. Additionally, trained CNNs perform very differently across various categories of objects, and we will thus talk about some potential causes.

Key Words: Image classification, python, Deep Learning, Tensor flow, Convolutional Neural Network, Open CV, NLP

1. INTRODUCTION

Particularly with the expansion of information in a number of areas, such as online commerce, transportation, health care, and gaming, picture categorization is evolving and becoming more popular among technology designers. When working with large photographs, finding valuable visual information quickly is essential. As a result, the results of picture classification are somewhat influenced by the excellent performance of the image classification method.

There are numerous picture classification algorithms. The typical method for classifying pictures is machine learning. Particularly with the expansion of information in several areas, such as online commerce, transportation, health care, and gaming, picture categorization is evolving and becoming more popular among technology designers. When working with large photographs, finding valuable visual information quickly is essential. As a result, the results of picture classification are somewhat influenced by the excellent performance of the image classification method.

Major developments have been made in the fields of scene categorization, object recognition, and image labelling, according to numerous researchers from around the globe. This makes it possible to develop solutions for issues involving object identification and scene classification. This work focuses on choosing the optimum network for this function since artificial neural networks, particularly convolutional neural networks (CNN), have demonstrated a performance breakthrough in the areas of object detection and scene classification.

Feature extraction constitutes a crucial phase within these algorithms. In the context of images, feature extraction entails the derivation of a concise feature set that encapsulates a substantial amount of object or scene information, thus effectively discerning distinctions among the object categories under consideration. Several conventional techniques for extracting features from images encompass methods such as Scale-invariant feature transform (SIFT), histogram of oriented gradients (HOG), Local binary patterns (LBP), and Content-Based Image Retrieval (CBIR).

Following the feature extraction phase, the subsequent task involves classifying these features by recognizing objects within an image. Multiple classification methods are at one's disposal, such as Support Vector Machine (SVM), Logistic Regression, Random Forest, and decision trees, among others. The effective presentation of CNNs has significantly improved image identification, segmentation, detection, and retrieval. CNNs have established themselves as highly capable models for comprehending the information contained within images, thus finding successful application in a diverse array of pattern and image recognition tasks. These include tasks like gesture and face recognition, item categorization, and scene description.

Convolutional neural networks (CNN) are widely applied in numerous applications, delivering exceptional performance across various tasks. The initial successful deployment of the CNN architecture was in recognizing handwritten digits. Since its inception, CNNs have seen continuous advancements, characterized by the introduction of new layers and the incorporation of diverse computer vision algorithms. The ImageNet Challenge predominantly leverages convolutional neural networks, along with a variety of sketch dataset combinations. A limited number of studies have conducted comparisons between the detection capabilities of trained networks and human subjects using image datasets. These comparative results indicate that humans achieve a 73.1% accuracy rate on the dataset, whereas trained networks exhibit a 64% accuracy rate.

2. LITERATURE REVIEW

Image classification refers to the task of placing an image into predefined categories. This is an aspect of machine vision with applications in areas such as autonomous driving, medical diagnosis, and remote sensing. Deep learning is a technique within machine learning that uses networks to learn from data. These networks are inspired by the structure and functioning of the brain, so they can recognize patterns from input data.

In recent times, Deep Learning has witnessed notable progress in the domain of image classification. Among the various approaches, Convolutional Neural Networks (CNNs) have risen to prominence as a specific type of Deep Neural Network tailored for this purpose. CNNs are explicitly engineered to identify features within images, a fundamental necessity for precise image recognition.

Krizhevsky et al. (2012)[1]: The authors propose a new deep learning architecture called AlexNet that performs best on the ImageNet benchmark for image classification.

Simonyan and Zisserman (2014)[2]: The authors propose a new deep learning architecture called VGGNet, which is deeper than AlexNet and performs even better on the ImageNet benchmark for image classification.

He et al. (2016)[3]: The authors propose a new Deep Learning architecture called ResNet that uses residual connections to improve the performance of Deep Learning models. ResNets have become the standard architecture for deep learning in image classification and other computer vision tasks.

Huang et al. (2016): The authors introduce a novel deep learning framework named Spatial Transformer Networks (STNs), which is capable of learning image transformations to enhance classification performance. STNs have demonstrated their effectiveness across a range of computer vision tasks, encompassing image classification, object recognition, and image segmentation.

Vaswani et al. (2017)[3]: The authors introduce an innovative deep learning framework known as Transformers, which leverages self-awareness to acquire knowledge about long-range dependencies within images. Transformers have consistently delivered state-of-the-art performance across a diverse range of computer vision tasks, encompassing image classification, object recognition, and image segmentation

Xie et al. (2017)[4]: The authors introduce a novel loss function, termed "focal loss," designed to enhance the performance of deep learning models in the context of dense object detection. Focal loss assigns lower weights to well-classified examples, compelling the model to prioritize challenging and misclassified instances.

Oquab et al. (2014)[5]: The authors propose a deep learning model that can learn representations of images at different levels of abstraction. These representations can be used for a variety of tasks, including image classification, object recognition, and image segmentation.

Romero-Paredes et al. (2015)[6]: The authors propose a deep learning model that can predict the affordances of objects, that is, the possible interactions that can be performed with an object. This model can be used to develop robots and other intelligent systems that can interact with the world around them in a more natural and efficient way.

Chen et al. (2015)[7]: The authors propose a deep learning model that can be trained to recognise objects in images with only weak surveillance. This means that the model does not need to be trained using a large dataset of images with bounding box annotations. Instead, it can be trained on a dataset of images with labels that indicate whether or not an object is present in the image.

Redmon and Farhadi (2016)[8]: The authors propose a Deep Learning model called YOLO that can perform real-time object detection in videos. This model is able to achieve high accuracy and speed by using a unified framework for object detection and localization.

Hu et al. (2018)[9]: The authors propose a Deep Learning model that uses attention mechanisms to learn long-range dependencies in images. Attentional mechanisms allow the model to focus on important parts of an image and ignore irrelevant parts. This can lead to improved performance in image classification tasks.

Zhu et al. (2017)[10]: The authors propose a deep learning model that can translate images from one domain to another without requiring paired training data. This model can be used to generate synthetic training data for image classification tasks or to translate images from one language to another.

3. METHODOLOGY

Deep Learning has emerged as a pivotal technology for addressing self-perception challenges, including the interpretation of visual data, speech recognition, and enabling robots to navigate and interact with their surroundings.

Our goal is to use the Convolutional Neural Network principle to recognise photos.

A. The Proposed Model's structure

Individuals embarking on the journey of learning deep learning and neural networks quickly discover that convolutional neural networks (CNNs) stand out as one of the most prominent supervised deep learning strategies. CNNs are specifically engineered to directly extract visual patterns from pixel images with minimal preprocessing. Most CNN architectures share common design principles, involving the sequential application of convolutional layers to the input data, occasional spatial dimension reduction through techniques like Max pooling, and an increase in the number of feature maps. Additional components include fully connected layers, activation functions, and loss functions such as cross-entropy or softmax. Nevertheless, the core CNN operations revolve around convolutional layers, pooling layers, and fully connected layers. Therefore, we will provide a concise overview of these layers before introducing our proposed model.

The initial layer where image features are extracted is the convolutional layer. Convolution allows us to preserve the interconnections among different parts of an image since pixels are influenced primarily by their nearby and adjacent counterparts. Convolution is a method that scales down an image's dimensions while retaining the inherent pixel relationships. For instance, applying a 3x3 filter with a 1x1 stride meaning a one-pixel shift at each step on a 7x7 image results in a 5x5 output through convolution.

To train an image classifier using TensorFlow, the process involves the following sequential steps:

- Data Loading: Acquire and prepare the dataset.
- CNN Architecture Definition: Establish the structure of a Convolutional Neural Network (CNN).
- Loss Function Specification: Define the appropriate loss function for the model.
- Model Training: Train the model using the provided training data.
- Network Evaluation: Assess the model's performance by testing it on the separate test dataset.

The proposed methodology aims to understand and leverage Convolutional Neural Networks (CNN) for image recognition systems. By employing filters, CNN extracts feature maps from 2D images. Unlike a fully connected layer of neurons, CNN explores the relationship of visual pixels within their local vicinity. Extensive evidence has demonstrated the remarkable effectiveness and promise of Convolutional Neural Networks in image processing. CNN has outperformed previously utilized technologies in the field of computer vision, encompassing tasks such as handwriting recognition, natural object classification, and image segmentation.

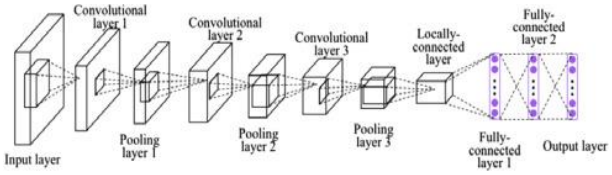

Fig 1. **Structure of CNN [13]**

The Convolutional Neural Network employs various layers, each serving distinct purposes:

- *Input Layer:* In the commencement of each CNN, the "input layer" is tasked with receiving and resizing images before transmitting them to the following layers for the purpose of feature extraction.
- *Convolution Layer:* Following the input layer are the "Convolution layers," which act as image filters. They reduce image dimensions while preserving pixel relationships through filtering using smaller pixel-sized matrices, such as a 3x3 filter with a 1x1 stride, which results in a 5x5 output when applied to a 7x7 image. These layers are pivotal in feature extraction and also play a role in matching feature points during testing.
- *Pooling Layer:* The "pooling layer" receives the extracted feature sets. Pooling layers are often added after each convolution layer to diminish spatial representation size. This, in turn, reduces parameter counts and computational complexity. Pooling layers also assist in mitigating overfitting. The choice of maximum or average values within these pixels helps reduce parameter numbers.

1) *Max Pooling:* This technique involves selecting the pixel with the highest value within the receptive field as it scans the input and then passing this selected value to the output array. It's worth noting that max pooling is a more frequently employed method compared to average pooling.

2) *Average* Pooling: In this approach, the filter calculates the average value within its receptive field as it progresses across the input, transmitting this calculated average to the output array.

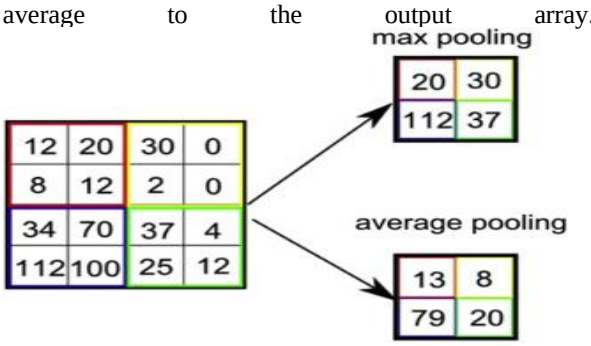

Fig 2. **Types of pooling layer[13]**

- *ReLU Layer:* The "Rectified Linear Unit" or ReLU layer follows the pooling layer and replaces every negative integer with 0. This action prevents learned values from stagnating near 0 or diverging towards infinity, ensuring the CNN maintains mathematical stability.
- ReLU computation is straightforward, involving a simple comparison of the input to the value 0.
- This simplicity carries notable consequences for the backpropagation process during training, as it signifies that the computation of a neuron's gradient is highly computationally efficient method.

$$<\mathrm{ReLU'}(x) = \begin{cases} 0, & \text{for } x < 0 \\ 1, & \text{for } x \geq 0 \end{cases}$$

Fig 3. **ReLU Activation Function**

- *Fully Connected Layer:* The final layer, known as the fully connected layer, is responsible for transforming the high-level filtered images into labeled categories.
- *Dropout Layer:* The Dropout layer functions as a mask, selectively deactivating the contributions of certain neurons while leaving others unaffected. It can be applied to the input vector, in which case it suppresses specific features, or it can be applied to a hidden layer, deactivating particular hidden neurons.

4. IMPLEMENTATION

The first phase in this research is data preparation. One must prepare training and testing data, merge data, combine labels, and reshape into the proper size before one can begin to develop the network. One keep the labels, other (meta) data, and the dataset of normalised data (single precision and zero mean).

The second step is to construct and compile the model. Tensorflow must be initialised before one can define crucial initialization parameters for the CNN, such as the batch size, number of epochs, learning rate, etc.

```
cnn = models.Sequential([
    layers.Conv2D(filters=32, kernel_size=(3,3), activation='relu', input_shape=(32,32,3)),
    layers.MaxPooling2D((2,2)),

    layers.Conv2D(filters=64, kernel_size=(3,3), activation='relu'),
    layers.MaxPooling2D((2,2)),

    layers.Conv2D(filters=128, kernel_size=(3,3), activation='relu'),
    layers.MaxPooling2D((2,2)),

    # Dense Layers
    layers.Flatten(),
    layers.Dense(128, activation='relu'),
    layers.Dense(64, activation='relu'),
    layers.Dense(32, activation='relu'),
    layers.Dense(10, activation='softmax')
])
```

```
cnn.compile(optimizer = 'adam',
            loss = 'sparse_categorical_crossentropy',
            metrics = ['accuracy'])
```

```
cnn.fit(X_train, y_train, epochs = 16)
```

Fig 4. **CNN model**

The batch size dictates the number of samples used in the training phase of the CNN. While the CNN processes the entire training dataset, it does so in batches of the chosen size. Adjusting the batch size can be leveraged to enhance computational performance, with the optimal choice often contingent on the available hardware resources. An epoch represents a complete cycle of both forward and backward passes through the network. When satisfaction is achieved with the model's convergence at a specific stage (selected epoch), it is typically recommended to start with a higher value for epochs and subsequently reduce it. In this particular experiment, a standard batch size of 32 was employed over the course of 16 epochs.

5. Results and Discussion

This model misclassifies a total of 90 photos out of 313 test cases after 16 epochs, which is a 71.5% recognition rate. The findings are not too bad for a model with only 16 epochs, CPU training, and a brief training period of 13 minutes.

```
loss: 1.0235 - accuracy: 0.7155

[1.0235388278961182, 0.715499997138977]
```

Fig 5. **Test Accuracy**

This model will be able to correctly classify photographs, despite the fact that some of them are challenging to identify. For instance, algorithm can identify the following image and categorise it as an "aeroplane".

```
plot_sample(X_test, y_test,3)
```

Fig 3. **Test data**

6. Conclusion and Future Work

This example demonstrates the model's capability to identify and categorize images. Furthermore, it can be extended to support real-time object recognition, as well as object and character recognition. Within the extensive domain of artificial intelligence and computer vision, image recognition holds a pivotal position as an initial stage. The experiment's results clearly indicate the substantial superiority of CNN over other classifiers. Enhanced results can be attained by incorporating additional convolutional layers and hidden neurons. By employing our methodology, individuals can effectively discern objects within blurred images. Utilizing image recognition as a model problem offers valuable insights into understanding neural networks, thus facilitating the development of more advanced deep learning techniques. Looking ahead, we are committed to the development of a real-time image recognition system

References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

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

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.

[4] Huang, Z., Liu, W., Anguelov, D., Krizhevsky, A., Berg, T. L., & Fei-Fei, L. (2016). Convolutional neural networks with spatial transformer networks for scene recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2227-2236.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Llionart, L., Gomez, A. N., ... & Joulin, A. (2017). Attention is all you need. Advances in neural information processing systems, 30, 5998-6008.

[6] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, 312-320.

[7] Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 481-488.

[8] Romera-Paredes, J., Torr, P. H. S., & Zisserman, A. (2015). Seeing beyond what is seen: A weakly-supervised deep network for predicting object affordances. Proceedings of the IEEE international conference on computer vision, 1651-1658.

[9] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2015). Fast R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the IEEE international conference on computer vision, 3642-3650.

[10] Redmon, J., & Farhadi, A. (2016). Yolo: Real-time object detection. arXiv preprint arXiv:1506.02640.

[11] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 7132-7141.

[12] Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, 2617-2625.

[13] https://www.sciencedirect.com/topics/mathematics/pooling-layer