Cat v/s Dog Image classification using K Means Clustering, CNN, and Transfer Learning

Submitted by:
Abhijith Dameruppala (adameru@iu.edu)

Abstract:

The project is based on cat vs dog dataset, that has been long referenced used for image classification problems, we have used 3 methods, K-means clustering, CNN, and Transfer learning. The dataset needed augmentation for better generalization. After initial experimentation with K-means we found the accuracy to be around 50% for 2 clusters, and after using Elbow method we found that the accuracy increases when the k-clusters are increased. But we have analysed that CNN works well with the data. But the best results were found to be transfer learning technique with an accuracy of 97.75 %.

Introduction:

It's very important for a small kid to understand the difference between 2 animals from the early age itself, in the same way for intelligent machine to be created, it needs to learn from easy task and then build up on its confidence. Cats Vs Dogs dataset is that kind of dataset that's been created to analyse and study how certain data mining and deep learning tasks happen.

But why is data mining such an important part of human being?

"Data is the new oil" as quoted by British Mathematician Clive Humby, because of the new age evolution and rapid increase in computation power has exponentially increased the data generated, data may be in the form of text that a person sends to his/her friend or a celebrity sharing pictures on social media or creating a video log and uploading it on YouTube. These increased data generation has led everyone to think how and why to use them for better productivity and growth.

The process of extracting and finding patterns in massive volumes of data using methods from machine learning, statistics, and database systems is known as data mining. Data mining is a computer science and statistics multidisciplinary field that tries to abstract or extract information (through intelligent algos) from a data source and convert it into an understandable structure for future use. The analytical step of the "knowledge discovery in databases" (KDD) process is known as data mining. In addition to the raw analysis stage, it incorporates database and data administration features, data pre-processing, model and inference considerations, interestingness measurements, complexity considerations, post-processing of identified structures, visualization, and online updating.

Dataset:

Before we turn to using dataset, lets understand why we chose Cats Vs Dogs as the main reference dataset. It's very easy for us humans to identify some dogs and some cats, but what happens to us when a cat is looking like a dog or vice versa. This is a case of judgment for us humans but can a 64-bit CPU or a 4 GB GPU find the difference using maths. That's a difficult task but we find that it's not impossible.

Figure 1. Is it a Cat or a Dog?

About dataset: The dataset was used from Kaggle which was published by Microsoft research team.

ASSIRA Dataset:

Web services are usually secured by a challenge that is supposed to be easy for people to respond but difficult for computers to solve. This sort of challenge is also known as a CAPTCHA (Completely Automated Public Turing Test to Distinguish Between Computers and Humans) or HIP (Human Interactive Proof). HIPs are used for several purposes, including decreasing email and blog spam and blocking website brute-force login attempts.

ASIRRA (Animal Species Image Recognition for Restricting Access) that asks users to recognize photographs of cats and dogs. Although computers struggle with this activity, research demonstrate that humans can complete it fast and accurately.

Problem With dataset:

Few of the images in dataset were misclassified or some of them were having different targets such as Humans, No images, Horses etc. So, we had to do pre-processing and then apply few data augmentation techniques.

Data Augmentation Techniques:

Data augmentation techniques create numerous duplicates of a real dataset to increase its size. Computer vision and natural language processing (NLP) models use data augmentation methodologies to deal with data scarcity and insufficient data diversity.

Data-centric AI/ML development methodologies such as data augmentation can increase machine learning model accuracy. According to an experiment, a deep learning model with picture augmentation beats a deep learning model without augmentation in training loss (i.e., penalty for a false prediction) and accuracy, as well as validation loss and accuracy for image classification tasks.

Gaussian Noise:

The phrase Gaussian noise, named after Carl Friedrich Gauss, refers to a type of signal noise with a probability density function (pdf) equal to that of the normal distribution (which is also known as the Gaussian distribution). In other words, the noise's possible values are Gaussian-distributed.

$$p_g(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-u)^2}{2\sigma^2}}$$

Where, $P\ is\ probablity\ density\ function, z\ is\ Guassian\ Random\ variable, \mu\ is\ grey\ mean\ and\ \sigma\ is\ the\ grey\ Standard\ deviation$

The major causes of Gaussian noise in digital images occur during capture, such as sensor noise caused by insufficient illumination and/or high temperatures, and/or transmission, such as electrical circuit noise. In digital image processing, Gaussian noise may be reduced by using a spatial filter; however, while smoothing a picture, an undesired impact may occur in the blurring of fine-scaled image edges and features since they also correlate to blocked high frequencies. Mean (convolution) filtering, median filtering, and Gaussian smoothing are examples of traditional spatial noise reduction methods.

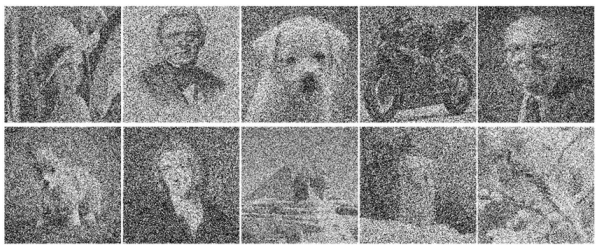


Figure 2. Applying Gaussian Noise

Random Crop:

Random Crop is a data enhancement technique that allows researchers to crop photos into a certain dimension, producing synthetic data. The cropping might result in any part of the image being cropped, hence the term "Random Crop."

Figure 3. Applying Random Crops

Image Flip:

The image has been flipped horizontally and vertically. Flipping rearranges the pixels while keeping the qualities of the image. Vertical flipping is not required for all photos, although it can be useful in cosmology and microscopy.

Figure 4. Applying Flipping

These are few data augumenation techniques that we used in our project.

K Means:

Now that we have an idea of the data set that we are using and the techniques of data augmentation, lets us dive deeper into how the Machine Learning algorithms are implemented and what are those that we have used in our project.

For all we know, K Means is an unsupervised divide and conquer machine learning algorithm that divides *n* datapoints into *k* clusters/groups. And this algorithm is a centroid based algorithm, which means that, in this method, we assign a centroid to a cluster and the whole algorithm tries to minimize the sum of the distances between the centroid of the cluster and the datapoints inside that cluster.

Algorithm:

- We need to define the number of clusters for the algorithm.
  In our case, it is 2(cats and dogs)
- Select k random points from the data as a center
- Associate each data point with the nearest center calculating the Euclidean Distance.
- Calculate the centroid and mean of all data points in the cluster.
- Repeat 2, 3, 4 steps until the stopping criteria

The Stopping Critirea:

- The Maximum number of iterations is reached.
- Centroid of the newly formed cluster does not change.
- Data points remain in the same cluster.

$$J = \sum_{J=1}^{K} \sum_{I=1}^{N} \left|\left| x_I^j - C_J \right|\right|^2$$

Where
J is the Objective Function or The Squared Error Function
K = Number of Clusters
N = Number of observations/cases
X_i^j = Case *I* belonging to cluster *j*
C_j = Centroid for cluster *j*

Since K Means is an unsupervised algorithm, it should classify the images by itself. Since we know that there can only be 2 kinds of images: One kind that belong to CATS and the other

belonging to DOGS, we can confidently define that the number of clusters in our project as 2. However there is a method called “Elbow Method” which is commonly used to determine the number of clusters of any cluster analysis.

The below graph is the result of Elbow Method for K Means Cluster analysis. We can see that the “sum of squared distance”, or the statistical measure of deviation from the mean, or in a word “Variance”. The higher the value of it indicates the higher variability and vice versa.

- Accuracy for k = 2: 0.531
- Accuracy for k = 16: 0.573
- Accuracy for k = 64: 0.629
- Accuracy for k = 100: 0.635
- Accuracy for k = 152: 0.656
- Accuracy for k = 256: 0.704
- Accuracy for k = 300: 0.718

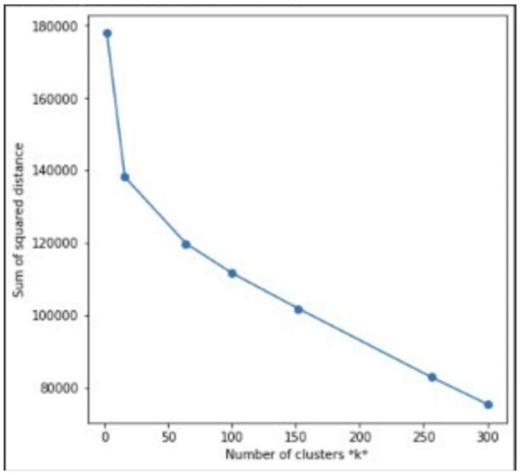


Figure 5. Elbow Method for Best K-Value

We can observe from the above results that the accuracy for $k = 2$ is around 50% where as the accuracy for $k = 300$ is around 71%. We can deduce from this that there are many images that classify into clusters other than CAT and DOG. This could be because of the color of the animal, or the lighting conditions of the image, there could be involvement of Humans in the images etc., However we cannot take all these conditions into account for analysis, because that could lead to overfitting model.

Convolution Neural Network(CNN):

In order to overcome the low accuracy, we have used another model called Convolution Neural Network(CNN).

What is CNN?

Convolutional neural networks (CNNs) are a form of deep neural network developed particularly to analyze input with a grid-like structure, such as an image. CNNs are made up of layers such as convolutional layers, pooling layers, and fully connected layers.

The convolutional layer is the foundation of a CNN. It is in charge of extracting features from input data using a collection of learnable filters. These filters slide across the input data, producing a dot product at each place between the filter's entries and the input data. This reduces the network's computational complexity and makes it more resistant to tiny changes in the input data.

At the network's end, fully connected layers are employed to create final predictions based on the combined output of the convolutional and pooling layers. The output of the fully connected layer is then passed into a final output layer, which generates a probability distribution across all possible classes using a softmax activation function.

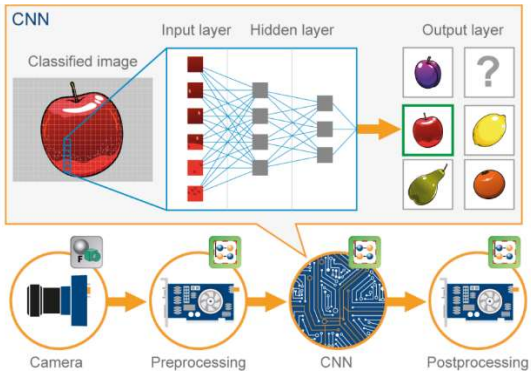


Figure 6. A end-to-end CNN Pipline

The Math Behind CNN:

Lets try some basic maths which is applied and used for a CNN layer to work

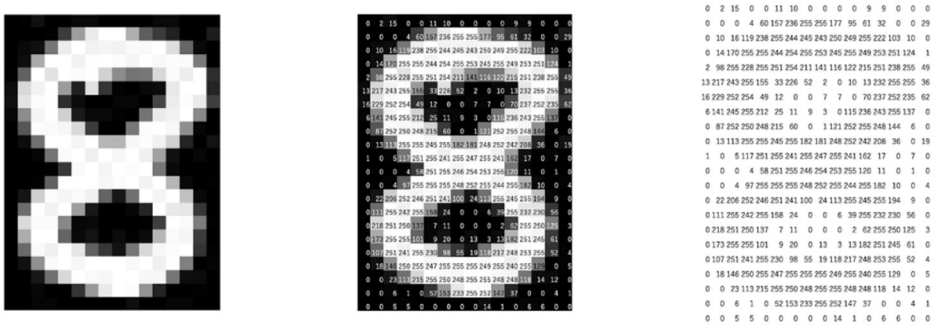


Figure 7. Picture depecting an image in terms of pixel values

As we see the above picture is depecting the pixel value which can be used for our kernel operation lets try some maths to understand the convolution operation.

$$Z = X * f$$

Where,

$$Z = Convolved\ Image$$
$$X = Input\ Image\ Matrix$$
$$f = Filter\ Matrix$$

Below is a small numerical calculation of CNN is performed:

$$\begin{matrix} 1 & 2 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 2 \end{matrix} * \begin{matrix} 1 & 1 \\ 0 & 1 \end{matrix} = \begin{matrix} 5 & 6 \\ 6 & 7 \end{matrix}$$

The above output matrix is the convolved matrix. The above matrix is only for our understanding, in-general the opertion happens on images of sizes 28x28 or more and is much complicated. Hence, Tensor proocessing or GPU's are prefered fot the operation.

Transfer Learning

Transfer Learing is used when we have small dataset and want a very rich source of knowldege, Transfer learning is like a university where students try to gain knowledge out from a very rich source of knowldege like professors,library etc.

Lets understand transfer learning an image,

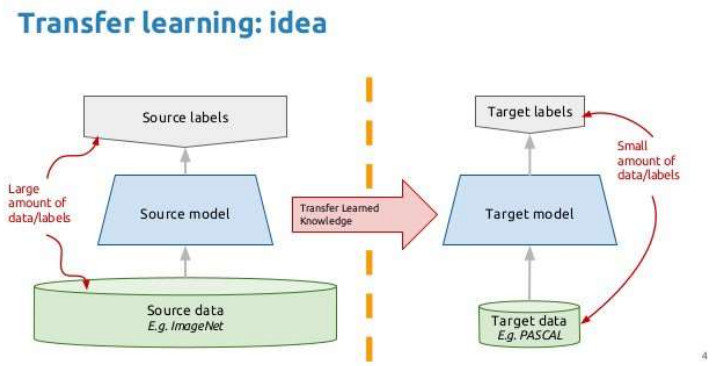


Figure 8. Transfer Learning basics

The basic idea of transfer learning is to use the rich knowldege of already trained, we have used MobilNet as our architecture. MobileNet is trained on ImageNet which is a very big dataset of having around 1000 labels and of around 14 million plus images.

What is MobilNet?

MobileNet is a family of convolutional neural network (CNN) models specifically designed for efficient on-device inference. MobileNet models are trained using the ImageNet dataset, which consists of 1.4 million labeled images from 1000 classes. This allows them to learn to extract useful features from images, such as edges and textures.

Once a MobileNet model is trained, it can be used as the starting point for a model on a different task, using transfer learning. Transfer learning is the process of using a pre-trained model on one task as the starting point for a model on a different, but related, task.

To perform transfer learning with a MobileNet model, you would first remove the final output layer of the pre-trained model. This output layer is specifically designed to predict one of the 1000 classes in the ImageNet dataset. You would then add a new output layer that is trained to predict the classes for your specific task. This new output layer would be trained using a dataset of labeled images for your task.

By using transfer learning with a MobileNet model, you can take advantage of the knowledge gained by the pre-trained model, and use it to train a more accurate model for your specific task. This can be especially useful when you have a limited amount of training data, as it allows you to leverage the knowledge gained from the much larger ImageNet dataset.

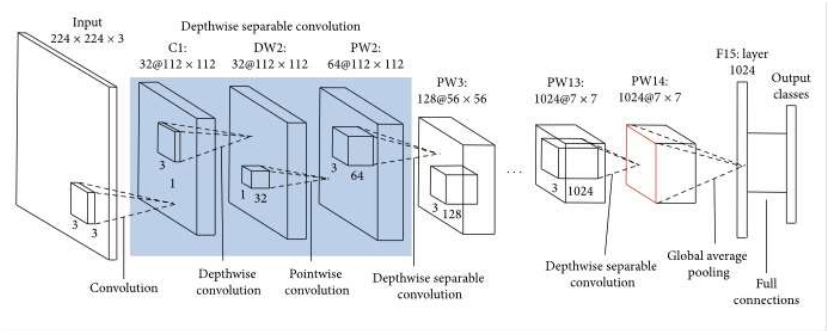


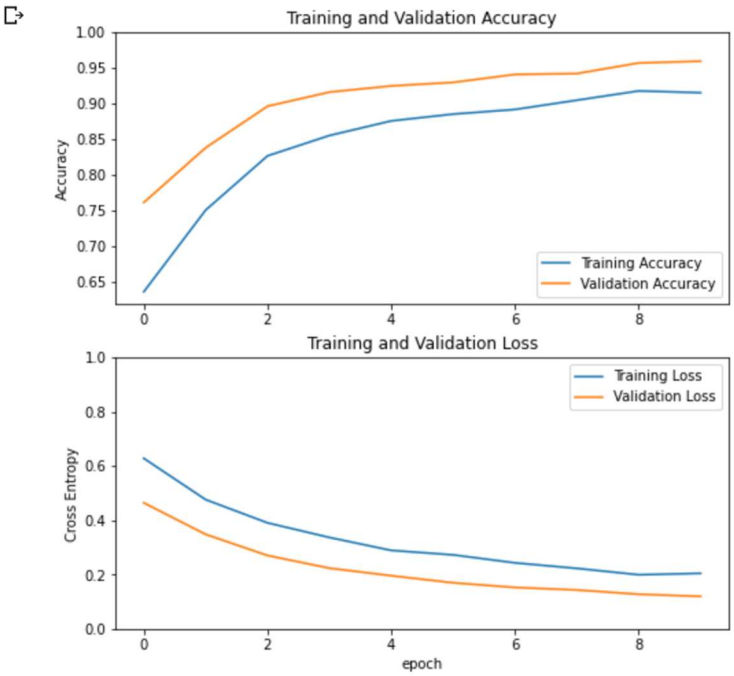
Figure 8. MobileNet Architecure

Results:

Machine Learning Model	Accuracy
KNN	53.1%
CNN	77.63%
Transfer Learning	97.75%

Table 1: Results after 10 epochs for 3 machine learning models

Graph 1: Accuracy v/s Loss for CNN Model



Graph 2: Accuracy v/s Loss for Transfer Learning Model

Future Scope:

We can always try to improve the model by using more data that is perfectly balanced, but it is known that having a perfect dataset is truly impossible. So, we try to use self-supervised learning methods that help us to create more better data without actually having one.

What is self-supervised learning?

Self-supervised learning is a type of machine learning in which the model learns from data as opposed to labeled examples. In contrast, supervised learning includes labeled data that has been tagged by a person being used to train the model. The model learns from input and generates labels autonomously using some form of unsupervised learning in self-supervised learning. This may be performed using a variety of ways, such as training the model using the underlying structure of the data or a challenge specific loss function. The goal of self-supervised learning is to create useable data representations that may be used for later tasks such as classification or regression.

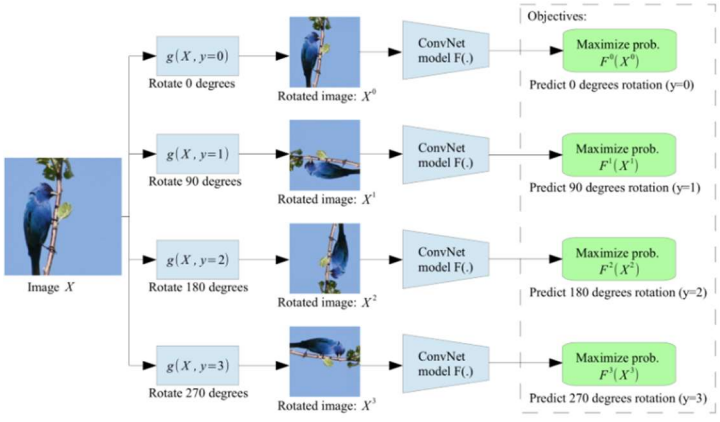


Figure 9. Self-Supervised Learning

With the use of self-supervised algorithms we can make a move without less data and have a much more better generalised model.

Another focus would be making a video based detection model that can be incorporated on users device and be used as a learning tool. For that sort of things we would require to use Yolo based frameworks.

Below is a small depiction of how can Yolo based framework be used.

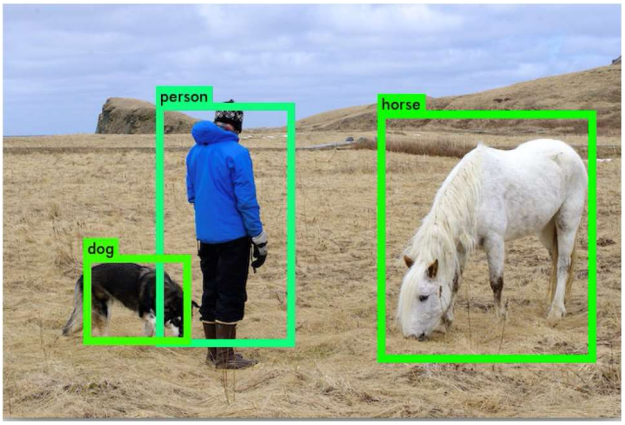


Figure 10. YOLO based model for detection

Conclusion:

In the project we try to check how the prediction varies from clustering algorithm compared with deep learning techniques and we can clearly see that deep learning techniques have the upper hand. We try to evaluate 3 models and can see that transfer learning of them all has the most accurate prediction. With this project we can help others evaluate and check how clustering and deep learning may help them for their respective work.

References:

[1] https://www.kaggle.com/c/dogs-vs-cats
[2] Deng L. The mnist database of handwritten digit images for machine learning research [best of the web] [J]. IEEE Signal Processing Magazine, 2012, 29(6): 141-142
[3] Jiang W. MNIST-MIX: A Multi-language Handwritten Digit Recognition Dataset[J]. IOPSciNotes, 2020, 1(025002).
[4] Sharma N, Jain V, Mishra A. An analysis of convolutional neural networks for image classification[J]. Procedia computer science, 2018, 132: 377-384.
[5] Lee, Youngjun. "Image Classification with Artificial Intelligence: Cats vs Dogs." *Https://Ieeexplore.ieee.org/Stamp/Stamp.jsp?Arnumber=9463236*, IEEE, 2021, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9463236.
[6] Ramya. A, Venkateswara Gupta Pola, Dr. Amrutham Bhavya Vaishnavi, Sai Suraj Karra, "Comparison of YOLOv3, YOLOv4 and YOLOv5 Performance for Detection of Blood Cells", April 2021
[7] Zhang, Hang, et al. "Resnest: Split-Attention Networks." ArXiv.org, 30 Dec. 2020, https://arxiv.org/abs/2004.08955.
[8] Kingma, Diederik P., and Jimmy Ba. "Adam: A Method for Stochastic Optimization." ArXiv.org, 30 Jan. 2017, https://arxiv.org/abs/1412.6980.
[9] Howard, Andrew G., et al. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." ArXiv.org, 17 Apr. 2017, https://arxiv.org/abs/1704.04861.