Advances in Consumer Research
ISSN (Print): 0098-9258
ISSN(Online): 3079-1766

Vol. 2, Issue 4 (2025) *https://acr-journal.com/*

Cartoon Retrieval using Deep Learning Approaches

Prem Singh M[1] and Sharath Kumar Y H[2]
[1]Department of Computer Science, Government College (Autonomous) Mandya, Karnataka, India
Email- premmsingh1973@gmail.com
[2]Department of Information Science and Engineering, Maharaja Institute of Technology Mysore, Karnataka, India
Email- sharathyhk@gmail.com

Cite this paper as: Prem Singh M and Sharath Kumar Y H, (2025) Cartoon Retrieval using Deep Learning Approaches. *Advances in Consumer Research*, 2 (4), 3442-3453

KEYWORDS
Cartoon Image Retrieval, Content-Based Image Retrieval (CBIR), Convolutional Neural Networks (CNN), Image Indexing and Scalability, Feature Extraction and Similarity Matching.

ABSTRACT
The Odisha Forest Department (OFD) has also been quite proactively involved in these challenges. Some of their activities include encouraging sustainable management practices in the Odisha Forest Department (OFD) in 2018, which helped NTFP dealers and collectors to get better market access and income. Also, there are some government programs, such as the "Prime Minister's Employment Generation Program (PMEGP)", extending financial assistance to NTFP-based enterprises. The Marketing of Minor Forest Produce scheme, launched by the Ministry of Tribal Affairs in 2013, covers fair prices to indigenous producers while promoting sustainable working mechanisms. The Pradhan Mantri Vanbandhu Kalyan Yojana (PMVKY), launched in 2014, aims at the overall development of the tribal populace at the village level.

1. INTRODUCTION

Cartoon is one of the most popular artificial arts in the history. It is well-liked by everyone for its art and the comical characters. Cartoons are one of the most effective means of communication which can convey the message more quickly than a written notice. The cartoon images can be used powerfully in the advertising industries. The skill of the effective cartoonist is not just the skill to draw well, it involves the ability to distill an idea in the form of images. This communication shortcutting makes the cartoon image to carry message effectively. The cartoon images as secondary products are now all over the internet and hence the customers have a strong demand to find the cartoon image by retrieval. Due to the huge growth in the amount of cartoon images the necessity for cartoon images retrieval is increased. It would be impossible to cope with the extension of cartoon images unless those data could be retrieved effectively and efficiently. Unfortunately, most cartoon image retrieval systems are text-based methods. The text-based retrieval methods could be retrieved if the images are well annoted. In other words, cartoon images without annotation make them incapable of being retrieved. Hence content-based cartoon image retrieval, retrieves with content based instead text based, plays an important role in multimedia system. The cartoon-based image retrieval systems can be defined as: For a given query cartoon image, finding similar cartoon images stored in database. The search process relies on how faster the images can be retrieved from the large database. To achieve a fast retrieval speed and to make the retrieval system truly scalable to the large size of the image collections, an effective indexing structure is a paramount part of the whole system. There are mainly two problems lying on the cartoon image retrieval. Firstly, the appearance of the cartoon image changes from time to time, when used in different applications, such as advertisement, poster and operating system themes and what is main content in the cartoon image is still not clear. Secondly, most of the global features used in the Content Based Image Retrieval (CBIR) will fail to localize the similar image patch in the local part of the whole image, even when the image patch is the same as the query. It is well known that cartoon was colourless in history and nowadays there are also lots of colourless cartoon, besides a vast number of colourful cartoons. In this work, we developed an effective system using CNN architectures. In the evolved work, the content-based music facets of songs are inserted as input and redressed them into vectors using the 2D Fourier transform. Further, projecting the songs into a LeNet-5, AlexNet, VGG-D, GoogLeNet and ResNet. is employed to compare the similarity of songs and retrieve the most resembling songs from the large-scale database.

[1] In this study, the words "app" and "platform" are used interchangeably.

2. RELATED WORKS

Haseyama and Matsumura [1] have proposed a cartoon image retrieval system, in which the partial features are referred as Regions and Aspects [1]. These partial features were used to compute cartoon image similarities. The cartoon images in the database that are most similar to the query image are returned as results. This system proposed can only be used to retrieve the face of the cartoon images. It limits the application of this system to the character design reuse rather than the cartoon sequence synthesis which is a setback. Yang et al., [2] proposed a Retrieval-based cartoon Clip Synthesis (RCCS) method that is used to combine edge and motion direction to represent a cartoon character. It uses an unsupervised distance metric learning algorithm for the cartoon image retrieval. The algorithm is formulated by a trace ratio maximization problem, which can be optimized by the iterative approach [2]. Sykora et al., [3] proposed a counter algorithm for cartoon image detection called Curve structure extraction for cartoon images. In their work, a novel counter detection algorithm is used to get counters with a very good connectivity based on the assumption that foreground parts of cartoons are surrounded by bold dark contours. Curve structure extraction for cartoon images was not suitable for most modern cartoons as the assumption were too strict. Tiejun Zhang et al., [4] proposed feature that uses the Harris-Laplace corner to localize all the key points and corresponding scale in the cartoon image. Then, the local shape was described by the shape context. The feature point matching is achieved by a weighted bipartite graph matching algorithm and the similarity between both the query and the indexing image is presented by the match cost. The curves in the cartoon images, and described as the main content and introduces a local feature named Scalable Shape Context (SSC) based on the SC feature. Jun Yu et al., [5] proposed a boundary curve extraction methods for all the general images: A summary of boundary curve extraction methods [5], it is also called edge detection. These methods define curves as sharp changing image region and detect them by finding curve points with a maximal gradient magnitude along the curve profile. Though it works well for boundary curves, it is not suitable for decorative curves. Two parallel curves on either side of decorative curve will be produced. This brings difficulties to further processing like vectoring and editing. H. Kang et al., [6] proposed decorative curve extraction methods for general images: Decorative curve extraction methods define curves as lines with small but with a finite width [6]. These algorithms analyze curves by modelling the curves and their surroundings as well. However, their models of curve profiles are suitable only for decorative curves. Significant bias will arise for boundary curve. Moreover, as not introducing regular feature of cartoons, noise in orientation information can't be reduced well enough especially in weak curve area. Belongie [7] proposed a feature called Shape Context (SC), which allows measuring the shape similarity between all the curvilinear structures, and it is used in recognition of digits, silhouette similarity-based retrieval. The basic idea of SC is to selext a pixel and model the distribution to other curve pixels. Shape Context will not describe the object path for a more common scene with background and when scale of the object is not predefined. K bozas et al., [8] this method each image is divided into pool of patches. For each patch Histogram of Oriented Gradients (HOG) is applied to find the object in that particular patch. Later binarization is applied for the extracted feature. Then Min-hash tuples are calculated and it is the final patch representation. For each tuples created a lookup in hash table is performed. The sketch provided from the user also under goes the same process and finally the hash tables of the sketch and the hash in database are compared on the voting basis. The database consists of 31 user drawn queries outlining objects and sceneries. Each sketch query is associated with 40 photos assigned with a value between 1 and 7. The final benchmark score is the average correlation value across 31 queries. The reader can verify the spatial coherence between query and the acquired photos achieved with patch voting scheme. it consumes more memory for storing the hash keys in the hash table and comparing the hash keys in hash table consumes more time if the number of hash keys increases. R Zhou et al., [9] took advantage of the two types of regions that can be found in an image ie. Main region: defined by weighted centre image feature, with this feature no one can retrieve objects in images regardless of their size and positions . Region of Interest (ROI): it is used to find the most salient part of an image and it is helpful to retrieve images with objects similar to the scene in a complicated scene. Feature is extracted using series of operations such as Hierarchical Orientation Combination, Orientation Refinement, Candidate Region Estimation and Multi-Scale Feature Extraction. A hierarchical database index is created using these regions and it helps in the easy and fast retrieval of images. Since the images are retrieved regardless of position of ROI the system is more prone to the retrieved images which are of less interest to the user. Saavedra et al., [10] proposed a novel method based on the edge orientation which gets a global representation in the name of both sketch and the test image. This method is focuses on estimating local edge orientations and forming the global descriptor name HELO (Histogram of Edge Local Orientation). This has two stages in which the first stage performs the pre processing tasks which gives an abstract representation of both the query and test images in second stage the histogram is made. To get the rotation invariance this method uses the fourier transform after the feature extraction. In this method the query images need to be drawn in continuous strokes. Juan and Bodenheimer [11] developed Graph based cartoon clip synthesis (GCCS) that combines similar characters into user-directed sequence based on dissimilarity measure in edges. GCCS builds a graph and generates a new cartoon sequence by finding the shortest path between them. It performs well on the simple cartoon characters whereas for the complex colour and gesture pattern will not generate a smooth pattern because the edges can encode neither the colour information nor the gesture pattern and the synthesis result is fixed by the initial specification. R Hu et al., [12] proposed that each image is represented by a bag of regions derived from a region tree. The contours are extracted by region maps containing various level of details and capture local structure in contour map using the Gradient Field Histogram Oriented Gradients (GF-HOG) descriptor. After the shape descriptors are clustered from the Bag of Words (BoW) codebook via k-means and resulting histogram is constructed for the matching. BoW model has not been extensively tested from the view point of scale invariance and the performance is unclear. Houssem et al., [13] proposes two content based image retrieval algorithms. The first algorithm extracts the shape features by using support region and the second algorithm uses the shape context descriptor to make it a scale invariant and enhances its performance in the presence of the noise. Both the algorithms calculate the feature extraction window according to the image size . From the literature

[1] In this study, the words "app" and "platform" are used interchangeably.

survey, it is clear that very few works has been carried out on cartoon based image retrieval. Some of the limitations include that there is no proper solution for the optimization as it requires performing eigen decomposition matrix during each iteration. GCCS fails to generate smooth clips because edges can encode neither the color information nor the gesture pattern. Shape Context fails to describe the object path for a more common scene with background and when scale of the object is not predefined. A decorative curve extraction method leads to too many sporadic curves in their results. Curve structure extractions for cartoon images assumptions are too strict and hence it is not suitable for most modern cartoons. The performance of the BoW model was unclear as it ignores the spatial relationship. In the edge orientation method the images needs to be drawn in the continuous strokes. The images retrieved using the ROI method was regardless images are retrieved regardless of position of ROI. Retrieving the images using the hash key consumes more memory and time with the increase in the number of hash key. The boundary curve extraction method was not suitable for decorative curves as there is sharp change in the image regions. In [21], they propose a novel technique called facial emotion recognition using convolutional neural networks (FERC). The FERC is based on two-part convolutional neural network (CNN): The first part removes the background from the picture, and the second part concentrates on the facial feature vector extraction. In [22] the cartoon faces in the wild (IIIT-CFW) database and associated problems. This database contains 8,928 annotated images of cartoon faces of 100 public figures. It will be useful in conducting research on spectrum of problems associated with cartoon understanding. This database contains cartoon images of 100 international celebrities (politician, actor, singer, sports person, etc.). In [23], develop a system for recognizing cartoon caricatures of public figures. The proposed approach is based on the Deep Convolutional Neural Networks (DCNN) for extracting representations. The model is trained on both real and cartoon domain representations of a given public figure, in order to compensate the variations in the same class. In [24], propose a dynamic multi-task learning approach for cross-modal photo-caricature face recognition. The proposed dynamic multi-task learning network can learn the feature representations of images from different modalities by combining different modality-specific tasks in the network

3. PROPOSED MODEL

In this study, a CNN-based architecture is given for classifying Cartoon. The retrieved features are utilized to train a model with different classifiers such as LeNet-5, AlexNet, VGG-D, GoogLeNet, and ResNet. In addition, a reduced CNN was proposed for categorization purposes. The output label from classifiers determines the class label of the Script class. Figure 1 depicts the procedure for our work.

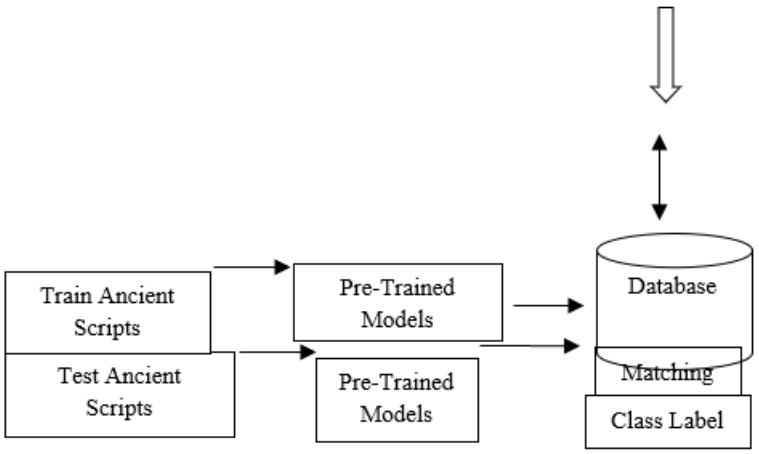


Figure 1: Proposed work

3.1 Common Architectures

There are numerous CNN architecture, such as LeNet-5, AlexNet, VGG-D, GoogLeNet and ResNet.

LeNet-5

The classic neural network architecture employs a pattern recognizer for MNIST handwritten digits. The LeNet-5 design follows a standard pattern of a single convolutional layer with tanH activation function, two pooling layers, and three fully connected layers. Whereas two completely connected layers are used to learn the non-linear combination of picture features, and the final fully connected layer is used to obtain the correct output.

[1] In this study, the words “app” and “platform” are used interchangeably.

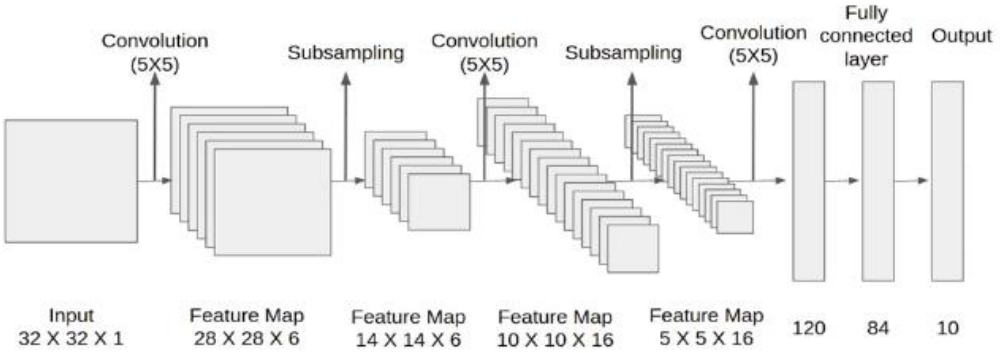


Figure 2. LeNeT-5 Architecture

AlexNet

Alex Krizhevsky proposed the AlexNet architecture, which learns by training enormous datasets. There are no fixed rules for building the architecture; instead, the number of convolutional layers to be employed is determined experimentally. It typically consists of five CL, three PL, and two FC layers, with around 60 million parameters. It employs the ReLU activation function as an alternative to Sigmoid or tanH functions. However, it can fix the vanishing gradient problem that happens when employing the sigmoid function, and ReLU performs five times faster with the same accuracy. Another issue addressed by this architecture is the reduction of overfitting by the use of a dropout layer. Dropout layers are added after each Fully Connected layer and are associated with probability (p). The activations are turned off at random with probability p, so the diverse group of nerve cells that are turned off represents a diverse architecture, and all of these diverse architectures are laterally trained with the given weights, with weight addition being one for each subset. If n neurons are coupled to a dropout layer, the number of subsets generated by the architecture will be 2n. As a result, it produces a regularized structural model, which aids in the prevention of overfitting. The other method is to select neurons at random, which helps to avoid co-adaptation and allows them to be independent in producing significant properties.

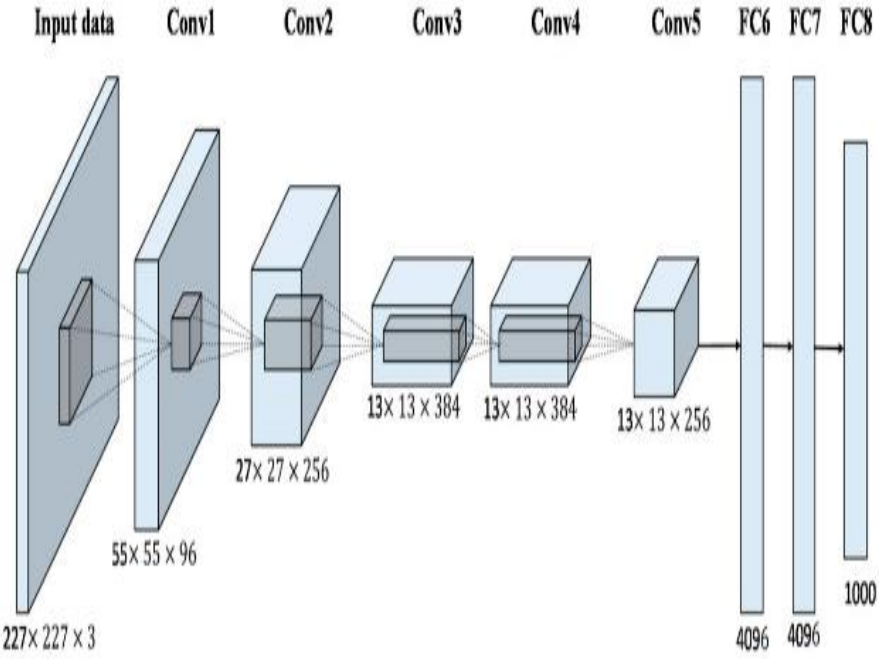


Figure 3. AlexNet Architecture

VGG- D

The Visual Geometry Group at Oxford is the source of VGG, which developed this architecture. It contains 16 learned layers and improves on the AlexNet architecture by utilizing a 3X3 kernel filter for the first and second convolutional layers, respectively. VGG-D is a basic architecture model because it will not need many hyperparameters. It commonly employs a 3 x 3 filter kernel with a stride value of 1 for CL and SAME padding for pooling layers with a stride size of 2 x 2. This network is built with two convolutional layers followed by a pooling layer; the group of two convolutional layers and a pooling layer is called a block, and the block building is repeated several times. These blocks are built using comparable filter sizes that are used repeatedly to extract additional complicated image attributes. Following VGG, the idea of constructing the block evolved into a generic approach for network creation. Even though VGG achieves the highest accuracy on the ImageNet dataset, meeting VGG's processing needs in terms of memory and time is quite hard. The large breadth of

[1] In this study, the words “app” and “platform” are used interchangeably.

the convolutional layers makes them inefficient.

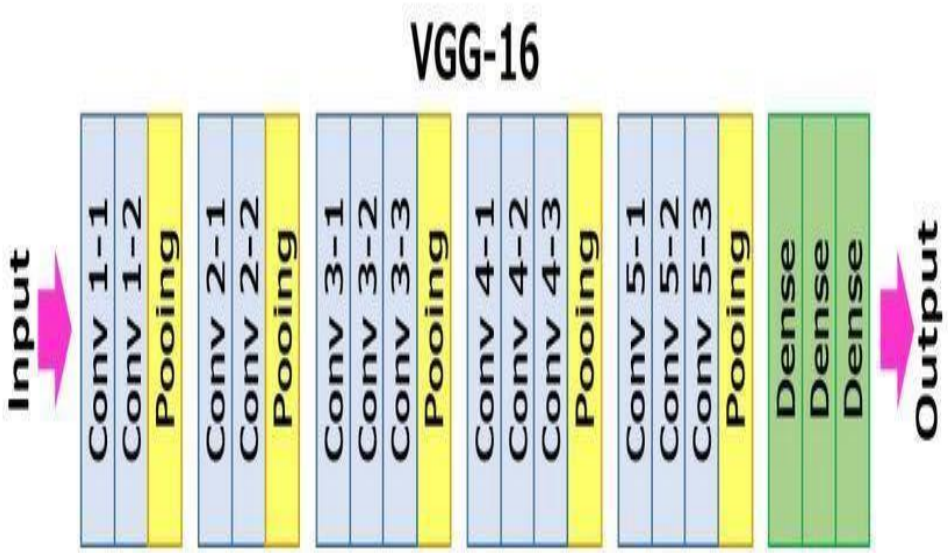
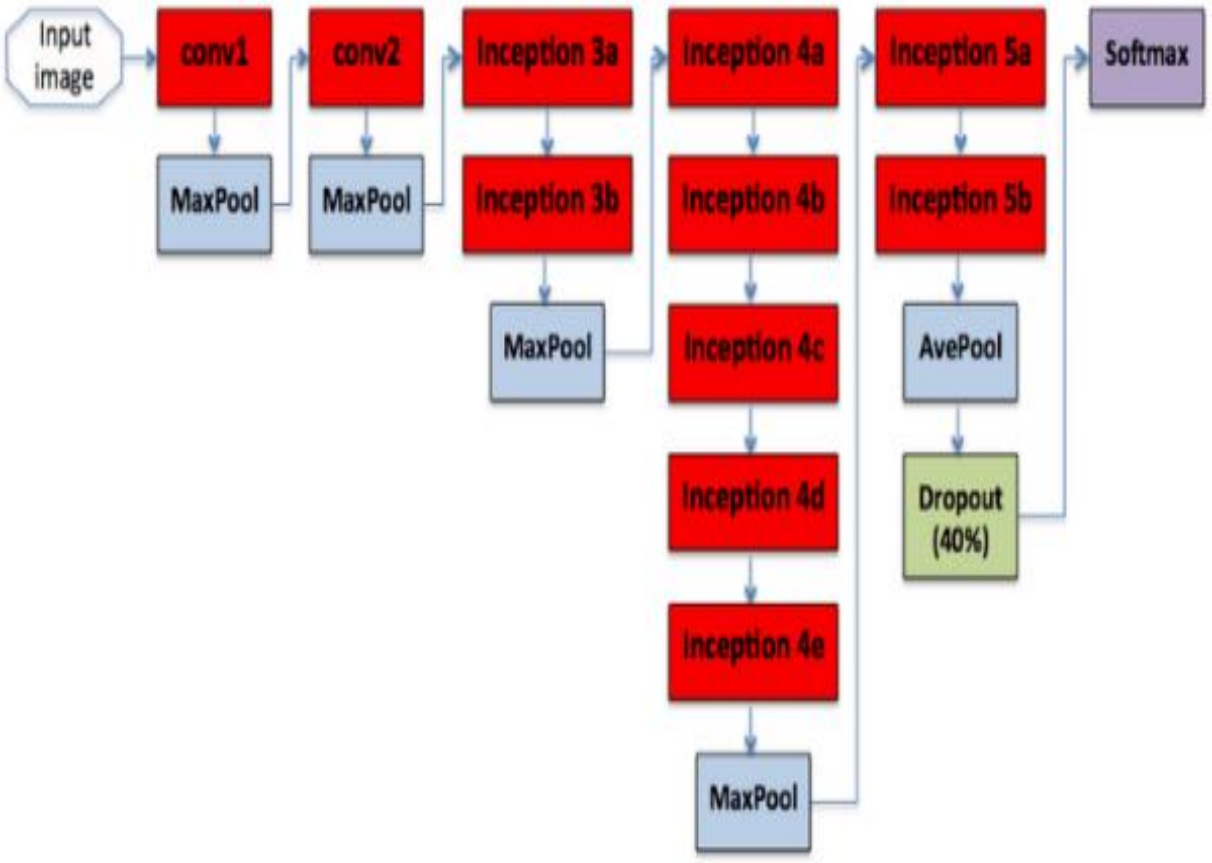


Figure 4. VGG-D Architecture

GoogLeNet

The GoogLeNet model is far deeper and more complicated than prior CNN architectures. To reduce the number of parameters, this model was built using a series of very small convolutions. Although the architecture consisted of 22 layers of DCNN, it employed just 4 million parameters, which were reduced from the 60 million parameters used in AlexNet. More specifically, it adds a module that can combine many filter sizes and dimensions into a single new filter. An inception module is the basic building component of the network, and it helps to minimize the computing requirements. The GoogLeNet underwent additional transformations to replace the fully linked layers with a simple layer that aggregated global averages. This helps to considerably reduce the number of parameters. As a result, the use of a large network enables GoogleNet to eliminate the FC layers without compromising accuracy.



[1] In this study, the words "app" and "platform" are used interchangeably.

Figure 5. GoogLeNet Architecture

ResNet

Kaiming's suggested Residual Neural Network architecture includes a concept known as skip connections. The ReLU activation function computes the input matrix for two linear transformations. Increasing the depth not only exacerbates the overfitting problem, but also enhances the network's accuracy rate. Applying negligible learning produces a vanishing gradient since it is more harder to tackle the challenge of increasing the depth of the layers required to adjust the weight from the end. The second issue is that training a deeper network with a wider parameter space causes a faster training rate. ResNet is a model that trains deep networks and creates networks with residual models.

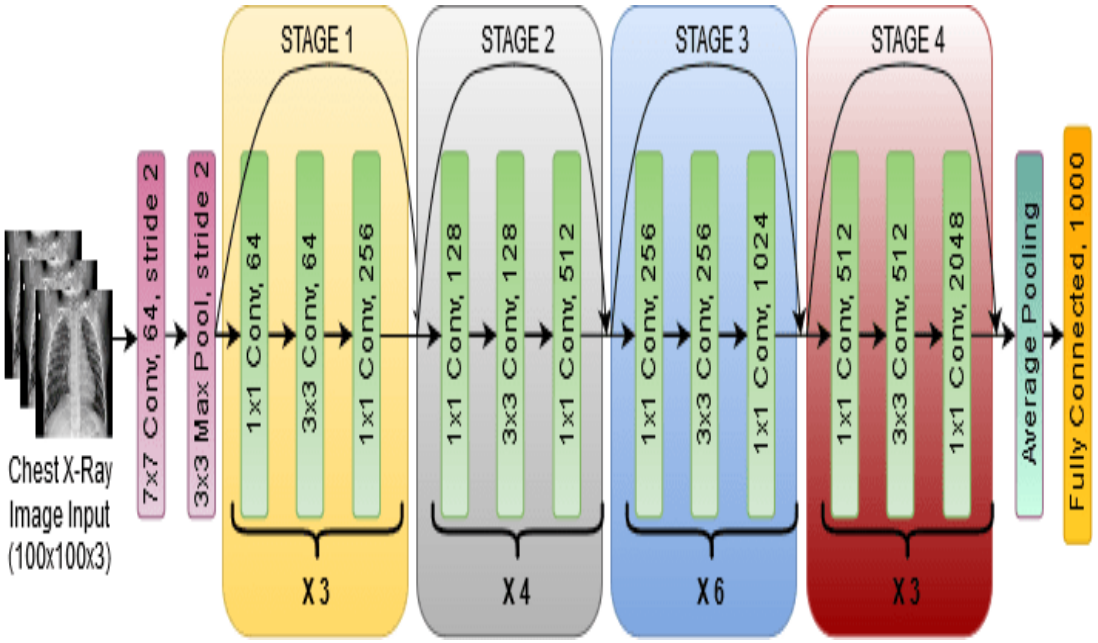


Figure 6. ResNet Architecture

3.2 CNN Classifier Model

The CNN Classifier processes and categorizes input pictures using convolutional and pooling layers, as well as the fully linked layer. Here are summaries of the strata.

Convolutional Layers

Given that the convolution layer's volume accepts sizes equal to W x H x D, it requires four hyperparameters, as shown below: Following the formula for the convolutional layer's output volume, K represents the number of filters, Fw and Fh are spatial extensions that reflect the filters' width and height, respectively, Sw and Sh represent the filters' stride width and height, and P denotes padding.

Pooling Layer

Consider the pooling layer's volume; it accepts sizes equal to W x H x D, thus it requires two hyperparameters, which are as follows: K stands for "K" filters, while S stands for "S" strides. The pooling layer's output volume can then be determined using the following formula: where OM is the output matrix, IM is the input matrix, and P, F, and S are padding, filtering, and stride, respectively. The output for the convolutional layer, pooling layer, and feature maps can be obtained by applying the computation formulas listed above. High-level reasoning in NN is carried out by entirely linked layers, which operate similarly to regular neural networks in that all neurons are associated with the activation of the previous layer.

3.3 Activation Functions

Rectified Linear Unit (ReLU) activation has recently gained popularity and effectiveness. This work uses an experiential evaluation of medical images with the best-observed activation function to illustrate the CNN classifier's classification accuracy and average loss. Activation functions are used to normalize data across layers. Based on the data, the sigmoid function determines the area of interest. Although it can map any real number to a limited range, it loses the advantage of down streaming processing. The stimulation function allows you to train for proper CNN creation. The logistic sigmoid activation function has simplified CNN training by correcting concerns with weight initialization and vanishing gradient, resulting in a finer model with the proposed rectified linear activation function. ReLU6 is one of the suggested minor ReLU variations. It helps to prepare networks for fixed-point inference. Softplus is a smoother version of ReLU. In compared to ReLU, it has been released with a few potential advantages. Softplus is an alternative to traditional functions since it is differentiable and has a simple derivative. The Exponential Linear Unit (ELU), an activation function, alleviates the problem

[1] In this study, the words “app” and “platform” are used interchangeably.

of vanishing gradients. It also has a negative value, which allows activations of units closer to zero in order to perform batch normalization and accelerate the learning rate. According to Clevert (2011), ELU has a positive extra-alpha constant. In contrast to the ReLU function, which sharply smoothes this advantage ELU to produce negative numbers that differ from ReLU, it smoothes out gradually until the output is a negative value. Furthermore, it avoids and resolves the vanishing gradient problem, while ReLU employs more straightforward mathematical operations to reduce the cost of lead calculation when compared to the sigmoid and tanH. Leaky ReLU is a variant of ReLU in which the LReLU allows for a small non-zero constant gradient that, while slower than the Sigmoid and hyperbolic tangent (tanH), aids in the solution of the dying ReLU problem by including a slight negative slope. A sigmoid accepts real values as input and returns values between 0 and 1. It is a nonlinear function with a fixed range of output, continuous differentiability, and monotonicity. It has a smooth gradient and supports analogue activation. It causes the problem of vanishing gradients, which saturates and eliminates gradients. The tanH function is another nonlinear function that can attain zero centering. However, in addition to the vanishing gradient problem, it has a higher gradient value than the sigmoid. When the classes are mutually exclusive in multi-class classification, softmax regression is a generalized variation of logistic regression. Similar to the sigmoid function, the Softmax function compresses each unit's output to a value between 0 and 1. However, it divides each output into smaller bits so that the aggregate of all outputs equals one. TanH's activation function outperforms the logistic sigmoid and is equivalent. It too has a sigmoid curve with values ranging from -1 to 1.

Experimentation

In this work, we experimented with various pre-trained models such as LeNet-5, AlexNet, VGG-D, GoogleNet, Resnet, and the proposed Reduced CNN. Even though the dataset was produced for experimentation, it is influenced by regional languages. The majority of songs have at least one cover version, but several have three or more. Similarly, the expanded real dataset covers80 (Ellis 2007), proposed at MIREX 2007, is used to evaluate cover song recognition systems. It features 80 sets of real and cover songs, totaling 166 tracks. In other words, covers80 primarily focuses on western music. Furthermore, for the experiment, covers50 tracks were generated by randomly selecting songs from existing datasets such as covers30 and 80. Figure 7 illustrates the accuracy of LeNet-5 on Cover 80. By examining the results, it was determined that the performance of all scripts was adequate, with a maximum accuracy of 0.93. Figure 8 depicts AlexNet's accuracy; on cover 80, it is noticed that the performance of all scripts is satisfactory, with a maximum accuracy of 0.90. Figure 9 depicts the accuracy of various scripts employing VGG-D; it is noticed that all of the scripts perform well, with a maximum accuracy of 0.94. Figure 10 depicts GoogleNet's accuracy on cover 80, with all scripts doing well and reaching a maximum accuracy of 0.92. Figure 11 depicts the accuracy of ResNet, and it can be seen that all of the scripts perform well, with a maximum accuracy of 0.89.

Furthermore, Figure 12 depicts CNN for the purpose of effective classification, and it is seen that the proposed CNN obtains the highest accuracy of 0.96. Furthermore, when comparing classifiers for ten script classes, CNN outperforms all other classifiers in terms of accuracy. Figure 13 compares the classifiers based on their average accuracy. The proposed CNN achieves maximal accuracy. To improve the effectiveness of the suggested method, we also tabulated the results using the confusion matrix displayed in Table 2. The figures and table show that CNN outperforms all other classifiers in terms of accuracy.

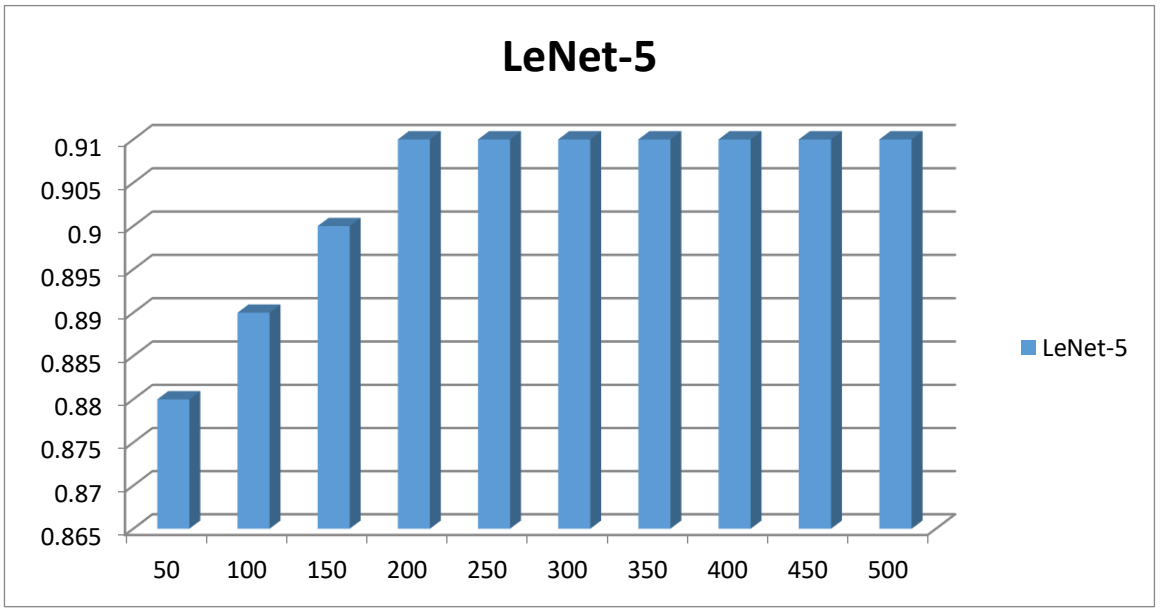


Figure 7: shows the accuracy of LeNET-5 on Cover-80

[1] In this study, the words "app" and "platform" are used interchangeably.

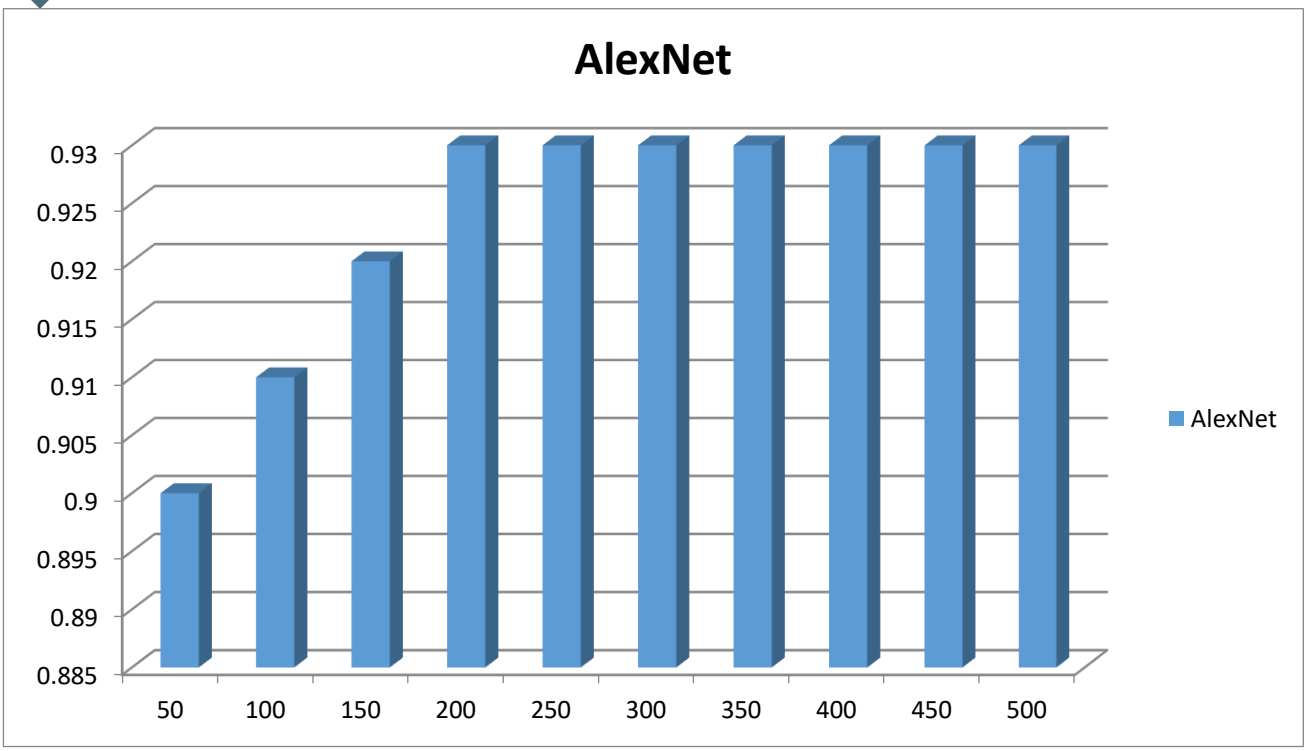


Figure 8: shows the accuracy of AlexNet-5 on Cover-80

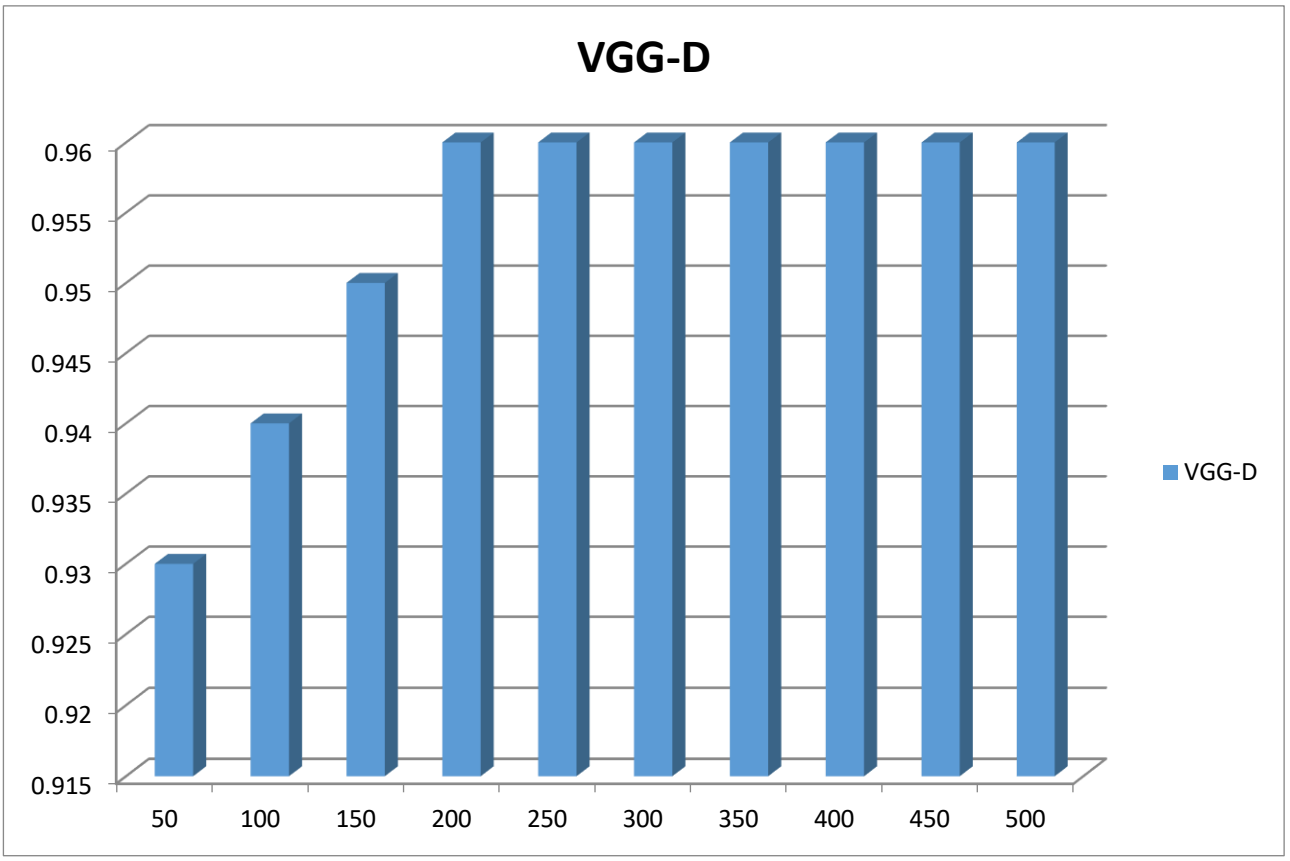


Figure 9: shows the accuracy of VGG-D on Cover-80

[1] In this study, the words "app" and "platform" are used interchangeably.

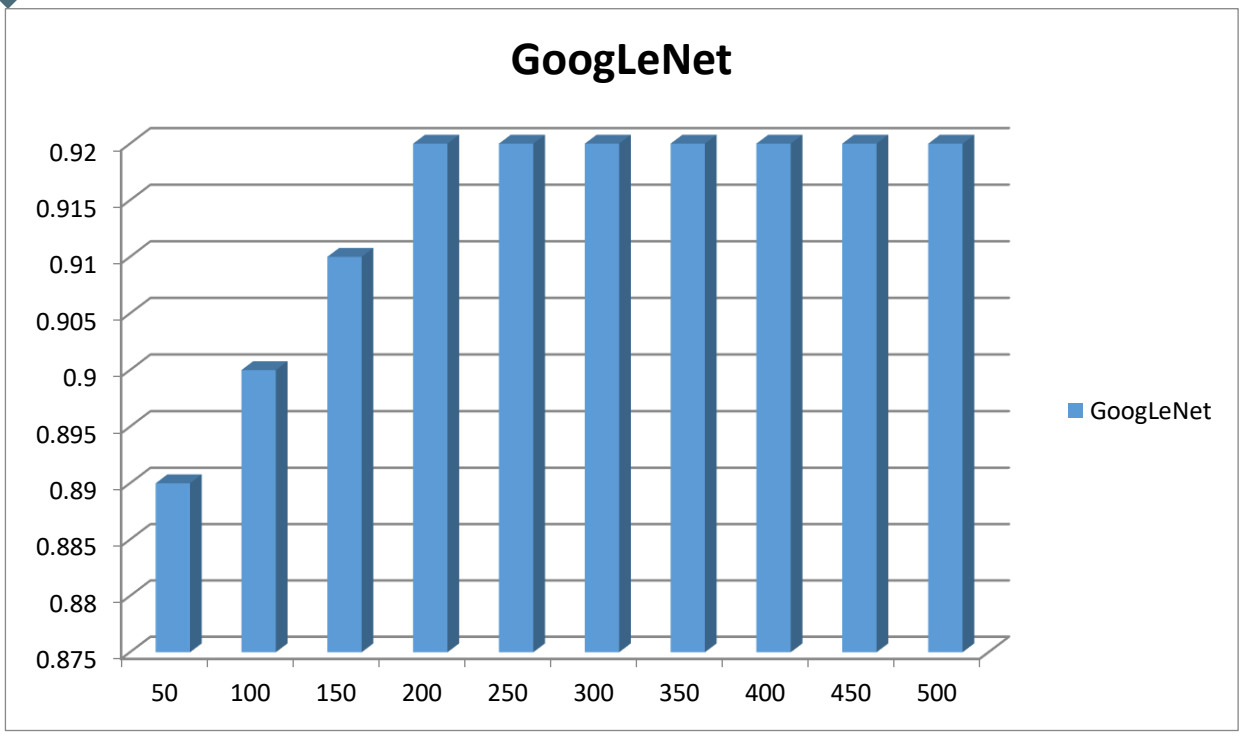


Figure 10: shows the accuracy of GoogleNet on Cover-80

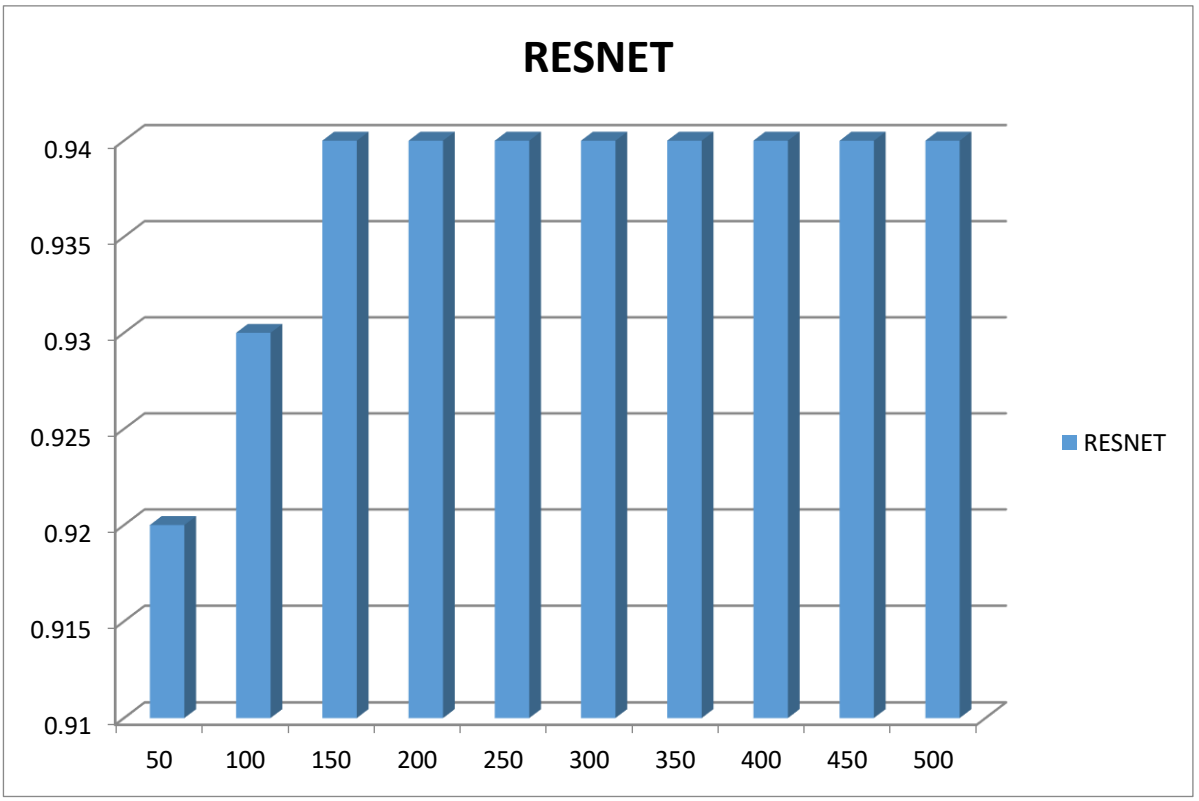


Figure 11: shows the accuracy of GoogleNet on Cover-80

[1] In this study, the words "app" and "platform" are used interchangeably.

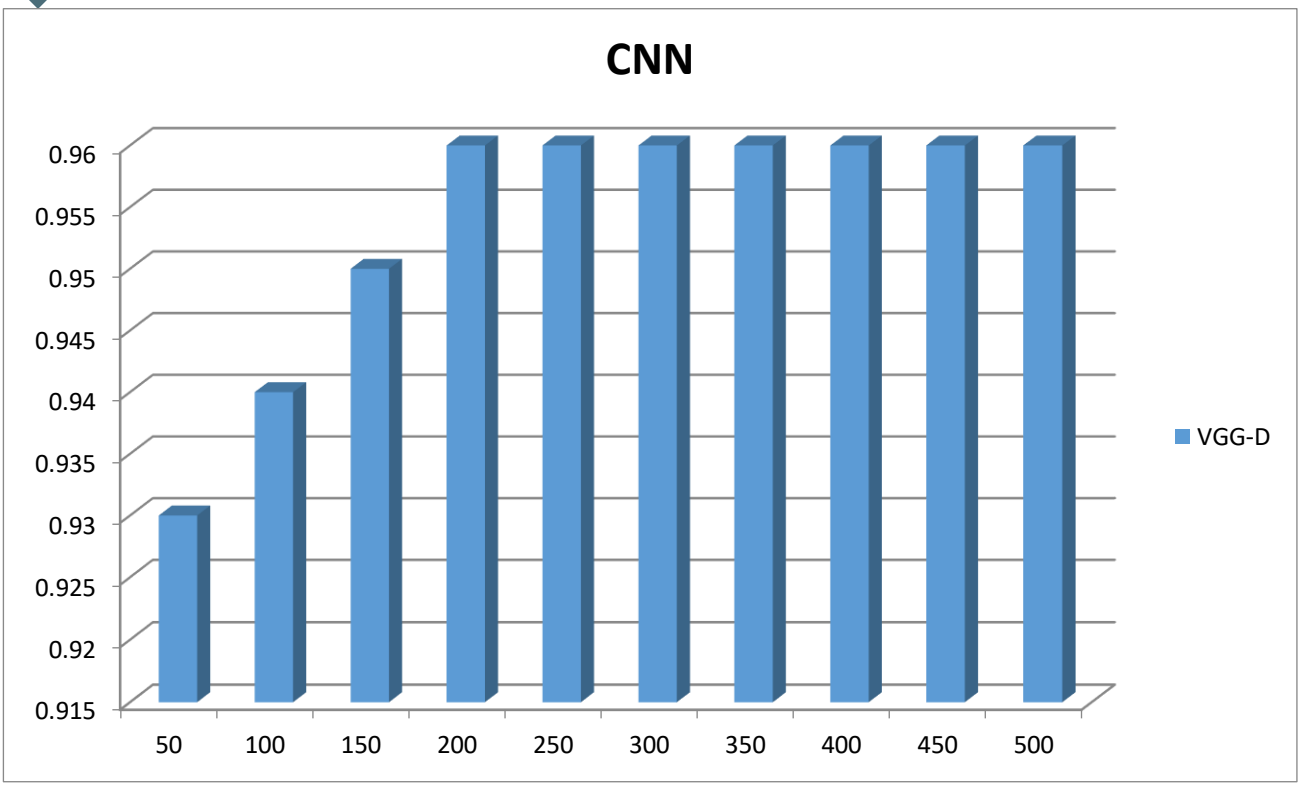


Figure 12: shows the accuracy of GoogleNet on Cover-80

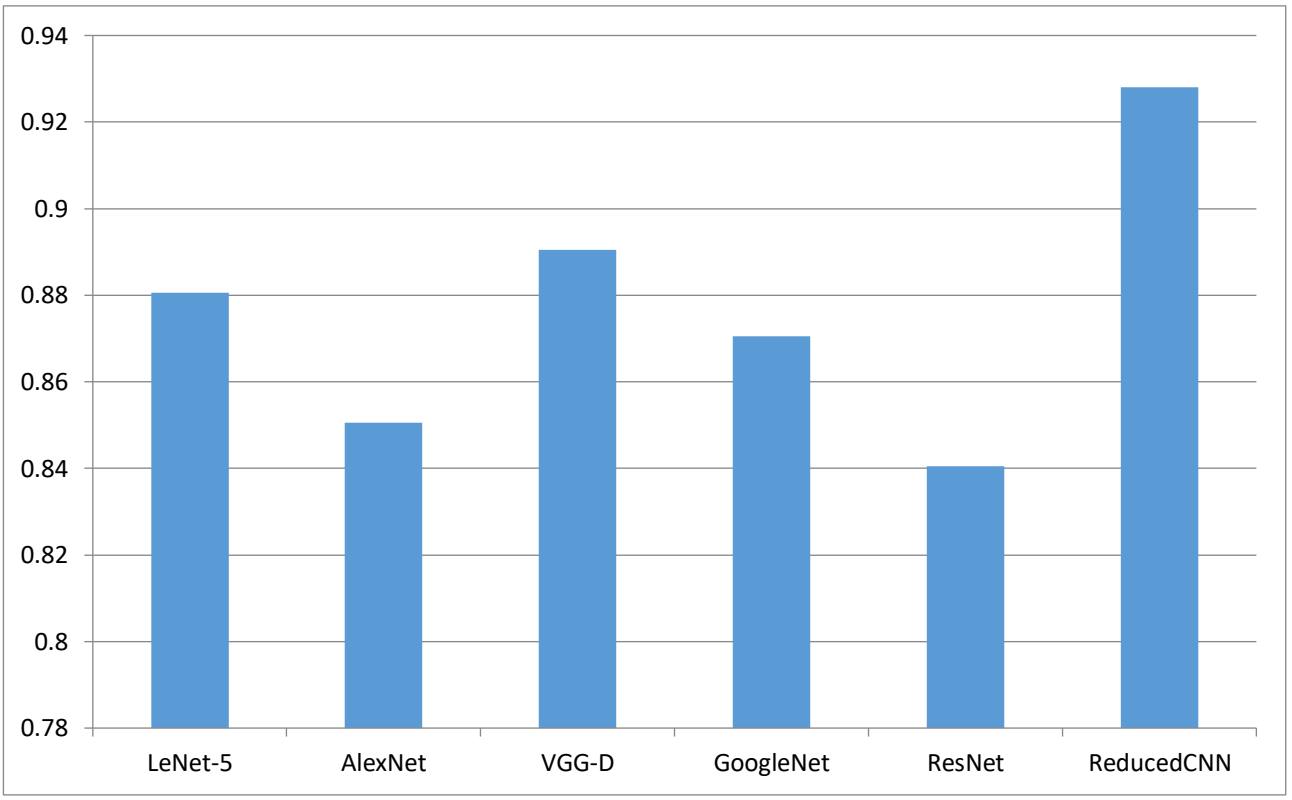


Figure 13: shows the accuracy of GoogleNet on Cover-80

4. CONCLUSION

The new structure of cover song retrieval using survey scores as features, practicing feature standardization, and preparing has resulted in a significant increase in execution. Furthermore, cover tune recovery was analyzed by displaying the highest score. Furthermore, feature normalization has shifted cover tune recovery from just determining high scores to a generic

[1] In this study, the words "app" and "platform" are used interchangeably.

recovery framework, while controlled preparation delivers a significant increase in execution.

In this work, we built an effective system with CNN architectures. In the advanced work, the content-based music components of songs are used as input and redressed into vectors using the 2D Fourier transform. The songs are then projected into LeNet-5, AlexNet, VGG-D, GoogLeNet, and ResNet to compare their similarity and retrieve the most comparable songs from the large-scale database.

5. REFERENCES

[1] M. Haseyama and A. Matsumura, "A trainable retrieval system for cartoon character images," in Proc. ICME, Jul. 2003, pp. 393–396.

[2] Y. Yang, Y. Zhuang, D. Xu, Y. Pan, D. Tao, and S. Maybank, "Retrieval based interactive cartoon synthesis via unsupervised bi-distance metric learning," in Proc. ACM Multimedia, 2009, pp. 311–320.

[3] D. Sykora, J. Burianek, J. Zara, Sketching cartoons by examples, In Proceedings of The 2nd Eurographics Workshop on Sketch-Based Interfaces and Modeling, 27–34, 2005

[4] Tiejun Zhang, D. Tao, D. Xu, J. Yu, and J. Luo," Recognizing cartoon image gestures for retrieval and interactive cartoon clip synthesis," IEEE Trans. Circuits Syst. Video Technol., vol. 20, no. 12, pp. 1745–1756, Dec. 2010.

[5] Jun Yu, Dongquan Liu, Dacheng Tao and Hock Soon Seah "On Combining Multiple Features for Cartoon Character Retrieval and Clip Synthesis", IEEE Trans. Cybernetics, Vol. 42, no. 5, pp.1413-1427, Oct. 2012

[6] H. Kang, S. Lee, C. K. Chui, Coherent line drawing, In Proceedings of the 5th International Symposium on Non-photorealistic Animation and Rendering, 43–50, 2007

[7] S. Belongie, J. Puzicha and J. Malik Shape Matching and Object Recognition Using Shape Contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24 (24): 509-521

[8] K bozas, Qi Han, Handan Hou, Xiamu Niu " Local Invariant Shape Feature for Cartoon Image Retrieval", IEEE Trans. 2013 Second International Conference on Robot, Vision and Signal Processing

[9] R Zhou ,Y. Yang, S. Xiang, and C. Zhang, "Neighborhood MinMax projections," in Proc. IJCAI, 2007, pp. 993–998

[10] D. A. Forsyth, J. Ponce. Computer Vision: A Modern Approach, Prentice Hall, 2002

[11] C. D. Juan and B. Bodenheimer, "Cartoon textures," in Proc. Symp.Comput. Animation, 2004, pp. 267–276.

[12] R Hu and S. Lee, "Human action recognition using shape and CLG-motion flow from multi-view image sequences," Patt. Recog., vol. 41, no. 7, pp. 2237–2252, 2008.

[13] Y. Houssem, D. Zhang, G. Lu, and W. Ma, "A survey of content-based image retrieval with high-level semantics," Patt. Recog., vol. 40, no. 1, pp. 262–282, 2007.

[14] T.Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," IEEE Trans. Known. Data Eng., vol. 21, no. 9, pp. 1299–1313, Sep. 2009.

[15] Dalal, N., Triggs B., 2005. "Histogram of Oriented Gradients for human detection" in : CVPR. pp. 886-893

[16] Lowe, David G. (1999). " Object recognition from local Scale invariant features". Proceedings of the international conference on computer vision 2. pp .1150-1157

[17] E. Oja. Subspace methods of pattern recognition, volume 6 of Pattern recognition and image processing series. John Wiley & Sons, 1993.

[18] D.Gleich[Online].Available:http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=12422&objectType=FILE

[19] S. Balakrishnama, A. Ganapathiraju, Linear Discriminant Analysis.

[20] Deng Liqiong, Chen Danwen, Yuan Zhimin, Wu Lingda. Attributebased Cartoon Scene Image Search System, Advanced Materials Research, v268-270, p1030~1035, 201

[21] SongHai Zhang, Tao Chen, YiFei Zhang, ShiMin Hu and Ralph R. Martin. Vectorizing Cartoon Animations. IEEE Transactions on Visualization and Computer Graphics, 2009,15 (4).

[22] Yuxiang Xie, Xidao Luan, Xin Zhang, Chen Li, Liang Bai. A Cartoon Image Annotation and Retrieval System Supporting Fast Cartoon Making. 2014 IEEE 17th International Conference on Computational Science and Engineering.

[23] Mehendale, N. Facial emotion recognition using convolutional neural networks (FERC). SN Appl. Sci. 2, 446 (2020).

[24] Mishra, A., Rai, S.N., Mishra, A., Jawahar, C.V. (2016). IIIT-CFW: A Benchmark Database of Cartoon Faces in the Wild. In: Hua, G., Jégou, H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science(), vol 9913. Springer, Cham.

[25] Shukla, P., Gupta, T., Singh, P., Raman, B. (2020). CARTOONNET: Caricature Recognition of Public Figures.

[1] In this study, the words "app" and "platform" are used interchangeably.

In: Chaudhuri, B., Nakagawa, M., Khanna, P., Kumar, S. (eds) Proceedings of 3rd International Conference on Computer Vision and Image Processing. Advances in Intelligent Systems and Computing, vol 1022. Springer, Singapore.

[26] Ming, Z., Burie, JC. & Luqman, M.M. Cross-modal photo-caricature face recognition based on dynamic multi-task learning. IJDAR 24, 33–48 (2021).

[27] W. Zheng, L. Yan, C. Gou, W. Zhang and F. -Y. Wang, "A Relation Network Embedded with Prior Features for Few-Shot Caricature Recognition," (2019) IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 2019, pp. 1510-1515, doi: 10.1109/ICME.2019.00261.

[1] In this study, the words "app" and "platform" are used interchangeably.