

GENERATIVE ADVERSARIAL NETWORK WITH AUTOENCODER FOR CONTENT BASED IMAGE RETRIEVAL

Subhra Samir Kundu

Student, Amity Institute of Information Technology, Amity University Kolkata,
Kolkata, 700135, India
subhrasamirk@gmail.com

Ambar Dutta

Associate Professor, Amity Institute of Information Technology, Amity University Kolkata,
Kolkata, 700135, India
adutta@kol.amity.edu

Abstract

The internet generates a huge amount of information for a query, but not all of it is useful because it contains some misinformation and some manipulated data. Content-Based Image Retrieval (CBIR) is a state-of-the-art process that is employed by major IT companies all over the world. The research is nearly complete, and they are now being utilized to break the system rather than improving or classifying the right misinformation utilizing state-of-the-art adversarial networks. The major goal of this research is to classify a given misinformation and identify all of the images that were used to create it. A simple general adversarial network (GAN) is utilized in conjunction with an autoencoder to calculate the latent vector. Using the nearest neighbor computation metric, the latent vector is then used to obtain all of the closely matching images. Using the nearest neighbor computation metric, the latent vector is then used to obtain all of the closely matching images. The proposed study has demonstrated that it can retrieve images with much less distance than the current ones and those with a single component than using both in a collaboration. The proposal can lower the same in one-third of the cases already in use.

Keywords: Content-Based Image Retrieval, General Adversarial Network, Autoencoder, Nearest Neighbor, Convolutional Neural Network.

1. Introduction

Content-based Image Retrieval (CBIR) is a collection of approaches for retrieving semantically relevant photographs from a database using naturally inferred picture alternatives. Visual aspects are typically depicted at a low level in CBIR frameworks. They are essentially rigid numerical estimations that have no bearing on the subjectivity and haziness that characterize people's understandings and insights. As a result, there is a distinction to be made between low-level highlighting and unmistakable level semantics. One tends to observe a period of massive data processing where registering assets becomes the most significant bottleneck in dealing with such massive datasets. Because of the large dimensionality of data and the great spatiality of each perspective on it, feature selection is critical for improving clustering and classification outcomes.

Due to the introduction of less expensive storage devices and, more importantly, the internet, extremely large collections of pictures are fast growing. Finding an image among a large number of images is a difficult undertaking. Physically naming images is one solution to this problem. Regardless, it is prohibitively expensive, time-consuming, and impractical for only a few applications. Furthermore, the naming interaction is dependent on the semantic accuracy of the image being depicted. As a result, many content-based image retrieval frameworks have been developed to extract low-level elements for displaying image content.

Deep convolutional neural networks [Lai et al (2011)] have recently advanced the cutting edge in image categorization significantly, attracting a lot of attention in the computer vision field. The topic of image retrieval, for example, the task of identifying images that contain a comparable item or scene as in an inquiry picture, is related to the image classification problem. It has been suggested that features appearing in the upper layers of the CNN that figure out how to group photographs can serve as excellent descriptors for image recovery. The paper [Krizhevsky et al. (2012)] have demonstrated some subjective proof for the aforementioned.

The biggest issue that has arisen as a result of the increase in pictorial data available on the internet is the problem of fraudulent images. These false images can be fully synthetic or man-made, or they can be created by

combining multiple images. The goal of this research is to locate the photos that were utilized to make the false image in the first place. This can be useful in a variety of scenarios where it is necessary to obtain the photos that were utilized to produce the fictitious image in issue. This can assist in obtaining the n-original (three in the instance of this paper) photos that were utilized to create the false image.

Until now, all CBIR research has used full images or images as a whole, but no one has considered that there could be fake images, or that adversarial networks could be used to understand the features of the images and develop a new model based on those features, making the current CBIR system much more powerful. The use of Generative Adversarial Networks (GAN) [Goodfellow et al (2014)] has been incorporated into the model, as has the idea of learning the latent variable in the image using the autoencoder [Baldi (2011)]. CBIR's image retrieval work has become much more efficient as a result of this.

The paper is structured as follows. Section 2 provides an extensive literature survey of CBIR approaches which is followed by the proposed approach in section 3 which includes the proposed solution, methodology and architecture. The implementation details using software, hardware and different datasets used are provided in section 4. Section 5 deals with the results and discussion. Finally, conclusion and future scope is presented in section 6.

2. Related Works

Work on color, texture shape, and generally queries is now being done in the area, however there are no visible studies on the progress of adversarial nets for retrieval in CBIR frameworks. This is supported by the current literature. However, some research is being done in the field to break the system using adversarial networks. After their introduction in 2014, generative adversarial networks, or GANs for short, have not seen much use in the area. Some works have used the autoencoder, but the effects have not been properly realized.

A study has been conducted that focuses on three important calculations: color histogram, texture, and moment invariants. By considering the three distinct elements of the image as well as similarity assessments based on the Euclidean measure, the suggested image retrieval approach ensures that the results produced are extremely relevant to the content of an image query. A color histogram is used to separate a picture's shading highlights. The texture highlights are separated using the Gabor filter, and the shape elements of a picture are removed using the moment invariant [Iqbal et al (2012)].

There is a study that focuses on the HSV shading space, which is generally examined. Frame feature vectors are created by extracting color histogram and texture data from a co-event network [Yue et al (2011)]. Then there are publications [Malik and Baharudin (2013)] that investigate the difficulties of efficient feature extraction and effective image coordination in a packed space. Using the large energy of the DC and the first three AC coefficients of the squares, they extract quantized histogram quantifiable texture information from the DCT squares of the image.

While the shortest path concept originates from a paper that used the shortest path computation over a weighted chart, they finish the recovery cycle by using another similarity measure, context-sensitive similarity measure, between the requested image and each information base image [Ma et al (2017)]. Then there are the calculations, which include preparing feature vectors after segmentation, which will be used to compare query and data set pictures for similarity [Singh et al (2012)]. Authors [Torres and Falco (2006)] attempted to highlight the challenges and concerns associated with the development of CBIR systems by presenting state-of-the-art research in the area rather than just describing existing solutions and applications. On the contrary, the article [Muller et al (2004)] provides an overview of the available literature and technology in the field of content-based access to medical image data.

The authors of [Noli et al (2017)] suggest DELF, a mindful neighborhood feature descriptor suitable for large-scale picture retrieval (DEep Local Feature). The new feature is based on convolutional neural networks that are built using image level comments on a landmark image dataset. Then there are studies that suggest Content-based image retrieval (CBIR) [Zhou et al (2017)], which uses visual content representation to identify relevant images and has gotten a lot of attention over the previous two decades. On the other hand, the authors of [Magliani and Prati (2018)] propose a few improvements to the R-MAC descriptor generating process to make the recently proposed R-MAC+ descriptors more representative than previous ones. In the literature, there are also articles that look at the recovery effects of different color spaces such as RGB, YUV, HSV, YCbCr, and Lab [Giveki et al (2015)].

Paper [Iizuka et al (2017)] describes a revolutionary picture completion process that produces pictures that are both locally and internationally dependable. They can finish pictures of arbitrary resolutions using a totally convolutional neural network by filling in lacking parts of any shape. They use global and local context discriminators to prepare this picture completion network to be stable. These discriminators are ready to distinguish real photos from completed ones. On small, basic photographs from CIFAR10, the authors [Burlin et al (2017)] presents alternative interpretation of a few image inpainting methods. They enhanced encoder setup by using a few key GAN training methods, as well as modifying the network to WGAN. Authors [Babenko et al (2014)] investigate the use of neural codes in image retrieval applications. On the other hand, the work [Krizhevshkey and Hinton (2011)] explains how to acquire many layers of features from color images and how

to use these features to introduce deep autoencoders. They then use autoencoders to convert photos into short binary codes. The publication lays out a general numerical framework for studying both direct and non-direct autoencoders. The structure enables the most non-straight autoencoder, the Boolean autoencoder, to be given a scientific treatment.

In paper [Lindbo et al (2016)], an autoencoder is presented that uses learned representations to evaluate similarity better accurately in information space. They can use learnt feature representation in the GAN discriminator as justification for the VAE reconstruction objective by combining a variational autoencoder with a generative adversarial network. The Autoencoding Generative Adversarial Network (AEGAN) is a four-network model that learns bijective planning between a predetermined inactive space and a given example space by applying an adversarial loss and a reconstruction loss to both the created pictures and the produced dormant vectors, according to paper [Lazarou (2020)]. Autoencoder networks [Pidhorski et al (2020)], are unassisted approaches for combining generative and authentic qualities while modifying an encoder-generator map. The question of whether they have a similar creative force to GANs or learn unravelling representations has not been fully addressed, even though it has received a lot of attention. They introduce an autoencoder called Adversarial Latent Autoencoder that deals with both concerns at the same time (ALAE). Even though they are very noble and have filled many gaps, the gap of extracting images from fake images still exists.

3. Proposed Approach

The objective of the study is to locate similar images or images that were used to create a deepfake image. Because there are many fake images circulating across the multimedia consortium now, the goal was to locate all the images used to create a deepfake or entirely fake image, i.e., can we find the top n images that were used to create the image given a fake image? This is very important for distinguishing genuine from fake faces. Because we started with random noise to generate images, this model can also be utilized for partial faces (explained in the methodology section). This model will prevent all adversarial assaults since it will use an adversarial network to create the original images and then use the same to determine its latent variable in order to predict the top n closest images from the dataset.

3.1. Proposed Methodology

This idea is realized using the concept of CBIR where we have trained a model using the concept of generative adversarial network followed by finding the latent variable using autoencoder which is then used to search images from the database using the “Euclidean” distance measure. The starting point of the models is the random noise which we have used as the input to our generator of the GAN architecture and the images that are the input to our discriminator.

The model shown in Figure 1 is the model that we used to conduct our experiment. The random noise is fed into the GAN's generator part (as proposed in paper [Goodfellow et al (2014)]), which uses it to generate fake images, while the original images are fed into the discriminator part, which classifies the fake images produced by the generator as fake or real. In the process, the generator predicts values for the real image and thus modifies it. These weights, which are learned from the generator, are applied to calculate the latent vector using the autoencoder, and this latent vector enables for the quickest search of the query image from the database using the “Euclidean” distance of nearest neighbor. The GAN is trained for an n-number of epochs (20000 in this case), which enables the generator generate images that are almost identical to real ones and allows for good weight modulation. The value of proximity using simply the autoencoder is more than 5 units (Figure 2), however our model using GAN plus autoencoder is significantly better, giving us a distance of less than 2 units (Figure 3).

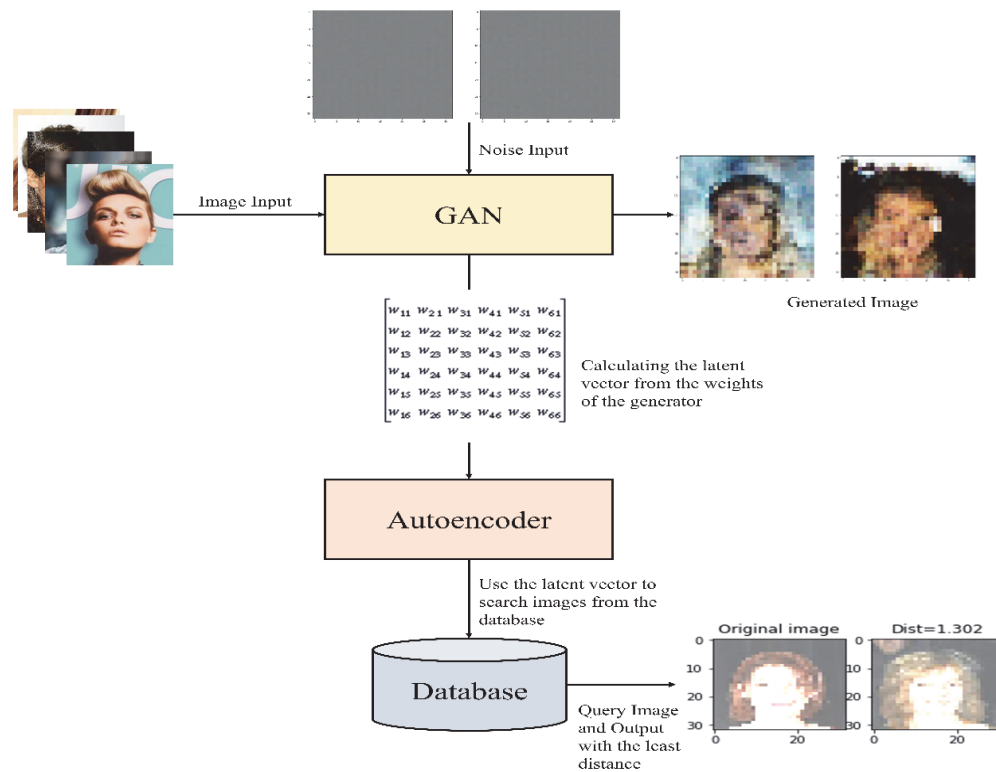


Fig. 1: Proposed Methodology.

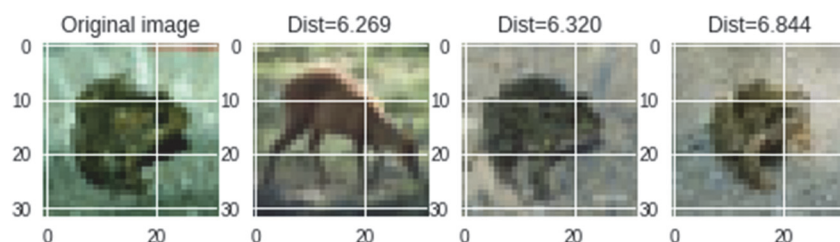


Fig. 2: Image generated using only Autoencoder with distances of around 6 units

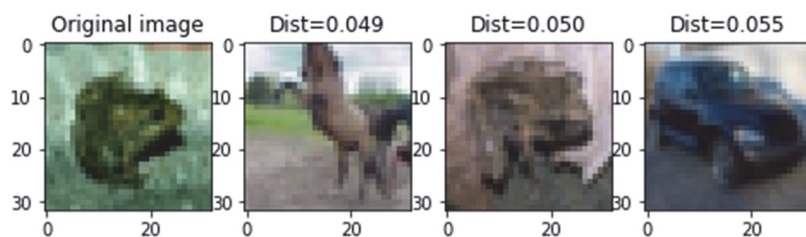


Fig. 3: Image generated by proposed model with distances less than 1 unit

3.2. Architecture

The architecture split of basic GAN, which is employed for the model, is illustrated in Figure 4. The architecture of the generator is depicted in Model G in the diagram, in which the input is pure noise prepared from the normal distribution, and that noise is used to predict values of each pixel and thus produce fake images, which are then classified as real or fake by the discriminator, which is labelled Model D in the diagram. The CNN structures in both the Model G and Model D are used to construct and deconstruct the images in this example. When the Model G is confronted with Model D, the weights learned by the former are updated every time the latter is classified. These weights are used in the autoencoder to learn the latent vector, which is then used to find the closest image of the query image in the database or images that can help in the creation of the false image.

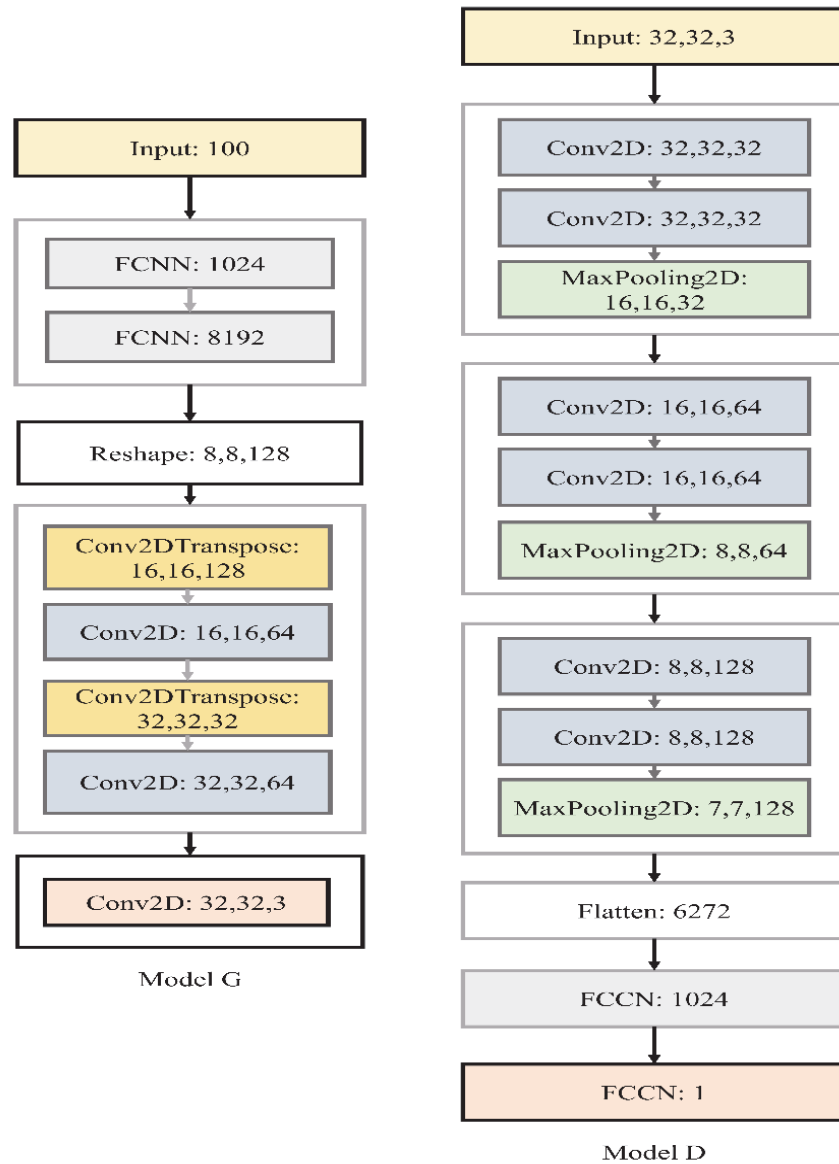


Fig. 4: Architecture of the Generator (named as Model G) and Discriminator (named as model D), the components of GAN

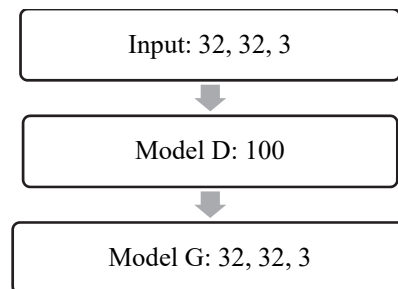


Fig. 5: Architecture of autoencoder used to calculate the latent variable of the weights generated by generator

The architecture of the autoencoder shown in Figure 5 is used to learn the latent vector using the weights learned by the generator in the GAN producing the fake image and compare it to the real image. Here, the GAN's Model D serves as the encoder for the GAN, while the GAN's Model G serves as the decoder for the autoencoder. The encoder and decoder structure produces the latent vector, which is then used to retrieve the query image's n-nearest neighbor from the database. When given a fake image, it can recognize the images that were used to create it. When the structures are used independently, they provide ambiguous outcomes, but when they are combined, they produce promising results.

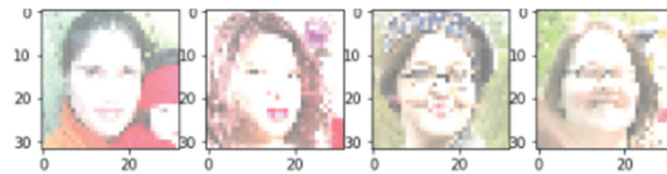


Fig. 8: FFHQ Samples.

4. Implementation Details

4.1. Software and Hardware used

We used the Tensorflow [Abadi et al (2016)] Library for Python and Python 3.7.7 as our prototyping programming language. Tensorflow is a multi-platform open-source software library for high-performance numerical computation that provides sequential and functional ways to generating artificial neural networks (ANNs) and deep neural network variants such as CNNs and RNNs. It also includes optimizers, classifiers for neural network output layers, weight matrix initializers, and other tools.

The image dataset was pre-processed on a workstation with an i7-9700K octa-core CPU with 8 threads and a maximum turbo frequency of 4.8GHz, as well as a 13MB smart cache. The 8GB GDDR6 NVIDIA GeForce RTX 2070 Super graphics processor unit has a boost clock speed of 1950MHz and 2560 CUDA cores. The workstation was equipped with 16GB of DDR4 RAM. The operating system was chosen based on performance and simplicity, and Linux Debian20.4 with Ubuntu desktop was chosen. On the aforementioned GPU, CUDA 10.0 was used to reliably complete the training and testing.

4.2. Datasets used

4.2.1. CIFAR-10

The dataset contains 50,000 train models and 10,000 test models and is made up of 32x32x3 images. CIFAR10 is a subset of the Tiny Images dataset, which contains 80 million images. Each of the 10000 images in the dataset is divided into five training batches and one test batch (an example is given in Figure 6). The test batch has 1000 randomly selected pictures from each class. Excess photographs from irregular requests are stored in training batches; nevertheless, some training batches may contain a greater quantity of pictures from one class than another. The training batches each contain exactly 5000 photos from each lesson.

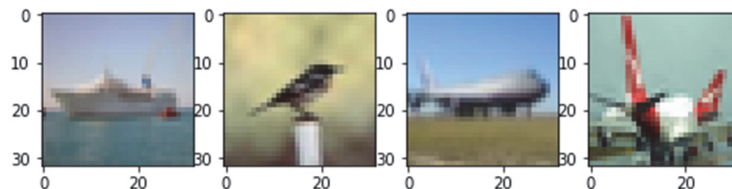


Fig. 6: Cifar-10 Samples

4.2.2. CelebA

The CelebFaces Attributes Dataset (CelebA) is a large-scale face attributes dataset with over 200K VIP images and 40 attribute annotations per image. The images in this set cover a wide range of poses as well as backdrop clutter. CelebA offers a lot of variety, a lot of data, and a lot of explanations, with 10,177 characters, 202,599 face photographs (an example is shown in Figure 7), and 5 landmark areas and 40 binary characteristics annotations for each picture.

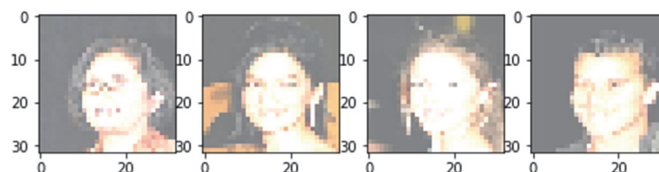


Fig. 7: CelebA Samples

4.2.3. Flickr Faces High Quality

The Flickr-Faces-HQ (FFHQ) collection of human appearances was created as a baseline for generative adversarial networks (GAN). The collection contains 70,000 high-quality PNG images with a resolution of 1024x1024 pixels, with a wide range of age, identity, and picture background. It also provides a lot of accessories like eyeglasses, shades, caps, and so on (an example given in Figure 8).

We employed a variety of other datasets to validate the proposed model, which should be highlighted which includes JAFFE, Cohn-Kanade, FER2013, Google Face Expression, and LFW. They all gave us results that were inside our range, thus the outcome was calculated.

5. Results and Discussion

The datasets used in this study are listed above in the datasets section. The starting point for all is pure random noise, which is employed in GANs, as proposed in [Goodfellow et al (2014)]. This provides the software with a wide range of values from which to forecast the fake images, classify them, and eventually learn the genuine weights. The experiment employs the architecture of a GAN as well as an autoencoder (both structures are discussed in the architecture section). The major goal was to use the “Euclidean” distance measure for the Nearest Neighbor algorithm to run CBIR on the fake images and retrieve the most nearby images. The goal was to acquire the shortest distance possible using the proposed model with GAN and autoencoder and compare it to a model that simply used autoencoder. Our model performed considerably better with less than 2-units distance compared to the 6-unit distance when only auto encoder was employed, as indicated in the results presented with images below.

The random noise was sampled from the standard normal distribution i.e. $N(\mu=0, \sigma=1)$, indicating that the generator in the GAN starts with random noise, which it uses to predict values for the generated image, which the discriminator with the original images classifies as real or fake, and the generator learns the weights and tries to reproduce the fake images as closely as possible.

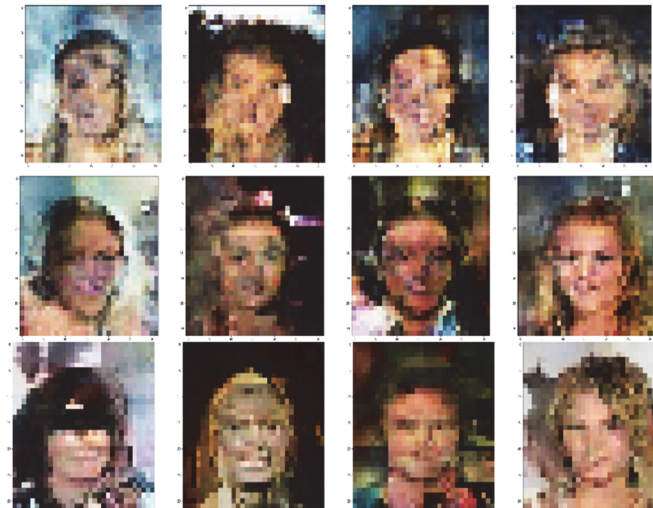


Fig. 9: The Result after 5000, 10000 and 20000 epochs respectively

The result in Figure 9 shows that as the number of epochs increases, the generator is quite capable of predicting the weights of the real image. In this feature of GAN two different Convolutional Neural Network (CNN) models, one for the generator and the other for the discriminator, run in parallel and compete with each other, where the generator's job is to create fake images and the discriminator's job is to classify them as fake or real, and the generator improves at producing the weights of the r in the process. As a result, after a given period, the created images are nearly identical to the originals. The results given here were trained using the CelebA Dataset, which contains over 200000 pictures.

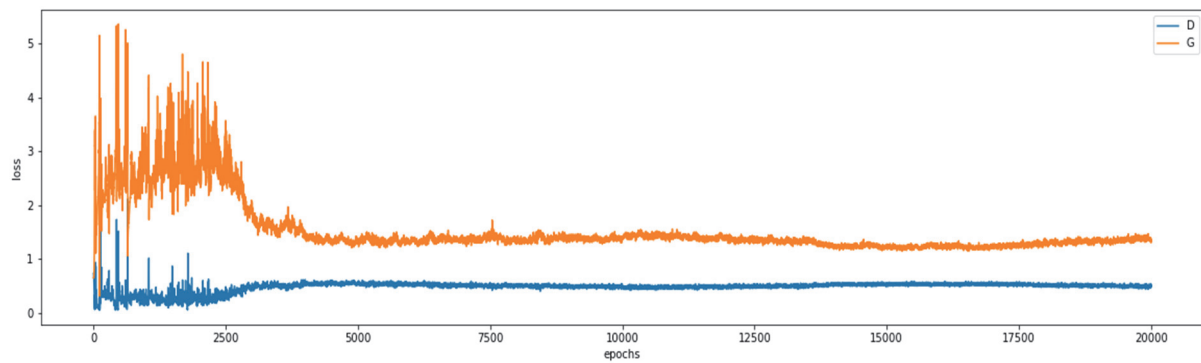


Fig. 10: The Loss of Discriminator (in blue) and the loss of Generator (in orange) over epochs

The graph in Figure 10 illustrates the discriminator and generator losses as a function of the number of epochs. As can be seen in the graph, the losses are not stable at first, but after about 2500 epochs, they begin to stabilize and yield a result that is considerably more accurate than using merely the auto encoder.



Fig. 11: The image before going in the autoencoder or the original image and the image coming out of the autoencoder or the encoded image

The autoencoder creates these images after learning from the generator's weight: the first labelled original image is the image from the database, and the second labelled encoded image is the image after encoding with the autoencoder (shown in Figure 11). Part (a) of Figure 12 shows the query picture and the obtained images from the data set using only the autoencoder, whereas part (b) shows the query image and the retrieved image with less distance than using simply the autoencoder. The one that only uses the autoencoder can reveal images with a nearest distance of about or more than 6-units, however the ones recovered by our model are less and around 2-units, providing us an advantage over the one that only uses the autoencoder. While doing the experiment on multiple datasets, it was discovered that datasets with some similarity or co-relation between them produced significantly better results than photos with no co-relation. JAFFE, Cohn-Kanade, FER2013, Google Face Expression, and LFW have all experienced this.



Fig. 12: Set (a) is the images with only autoencoder having nearest neighbor to be around or more than 6-unit distance while the set (b) is the image from the proposed model is around and less than 2-unit distance.

6. Conclusion and Future Scope

In this paper, a model for CBIR was presented that used a generative adversarial network to determine the latent variable, which was then used to search images from the database using an autoencoder. The random noise that we used as the input to our GAN architecture's generator and the images that we utilized as the input to our discriminator are the model's starting points. The proposed model was able to anticipate images at significantly closer distances than a single model could. The observation was that it was providing us a result of 2-units distance at most using Euclidean measure of distance for nearest neighbor, which was significantly less than the autoencoder-only results. This also validated our claim that we can obtain the images used to construct fake images, and that obtaining the images that aid in the same would assist us in classifying a variety of things, ranging from security to the medical area and beyond.

We are now working on improving the design of the GAN employed in this study, as well as the architecture of the autoencoder. The next stage is to make the model considerably more robust and faster to implement so that it can be done in real-time.

References

- [1] Lai, K., Bo, L., Ren, X., Fox, D. (2011). A large-scale hierarchical multi-view rgb-d object dataset. In: Neural Information Processing Systems
- [2] Krizhevsky, A., Sutskever, I., Hinton, G.E.(2012). Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems
- [3] Ian J. G., Jean P. A., Mehdi M., Bing X., David W. F., Sherjil O., Aaron C. and Yoshua B. (2014). Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14). MIT Press, Cambridge, MA, USA, pp 2672–2680.
- [4] Pierre B. (2011). Autoencoders, unsupervised learning and deep architectures. In Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop - Volume 27 (UTLW'11). JMLR.org, pp. 37–50.
- [5] Kashif I., Michael O. O., Anne J. (2012). Content-based image retrieval approach for biometric security using colour, texture and shape features controlled by fuzzy heuristics. In: Journal of Computer and System Sciences 78, 1258–1277.
- [6] Jun Y., Zhenbo L., Lu L., Zetian F. (2011). Content-based image retrieval using color and texture fused features. In: Mathematical and Computer Modelling 54, pp 1121–1127.
- [7] Fazal M., Baharum B (2013). Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain. In: Journal of King Saud University – Computer and Information Sciences 25, 207–218
- [8] Ling M., Xiabi L., Yan G., Yanfeng Z., Xinming Z., Chunwu Z. (2017) A new method of content based medical image retrieval and its applications to CT imaging sign retrieval. In: Journal of Biomedical Informatics 66 pp. 148–158.
- [9] Nidhi S., Kanchan S., Ashok K. S. (2012) A Novel Approach for Content Based Image Retrieval. In: Procedia Technology 4, 245 – 250.
- [10] Ricardo da S. T., Alexandre X. F. (2006) Content-Based Image Retrieval: Theory and Applications. RITA. 13. pp.161-185.
- [11] Henning M., Nicolas M., David B., Antoine G. (2004). A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. In: International Journal of Medical Informatics, 73, pp. 1-23
- [12] Noh H., Araujo A., Sim J., Weyand T., Han B. (2017). Large-Scale Image Retrieval with Attentive Deep Local Features," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 3476-3485
- [13] Zhou, W., Li, H., Jian, Q.. (2017). Recent Advance in Content-based Image Retrieval: A Literature Survey.
- [14] Magliani, F., Prati, A. (2018). An accurate retrieval through R-MAC+ descriptors for landmark recognition. 1-6.
- [15] Giveki, D., Soltanshahi, A., Shiri, F. and Tarrah, H. (2015) A New Content Based Image Retrieval Model Based on Wavelet Transform. Journal of Computer and Communications, 3, 66-73..
- [16] Satoshi I., Edgar S. S., Hiroshi I.. (2017). Globally and Locally Consistent Image Completion. ACM Trans. Graph. 36, 4, Article 107
- [17] Burlin, C., Calonnec, Y.L., Duperier, L (2017) Deep Image Inpainting. In: <http://cs231n.stanford.edu/reports/2017/pdfs/328>
- [18] Babenko A., Slesarev A., Chigorin A., Lempitsky V. (2014) Neural Codes for Image Retrieval. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham.
- [19] Krizhevshkey, A., Hinton, G.E.: Using Very Deep Autoencoders for Content-Based Image Retrieval. In: ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
- [20] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. In Proceedings of the 33rd International Conference on Machine Learning - Volume 48 (ICML'16). JMLR.org, 1558–1566.
- [21] Lazarou, C.: Autoencoding generative adversarial networks. In: arXiv:2004.05472v1 [cs.LG] 11 Apr 2020
- [22] Pidhorskyi, S., Adjeroh, D.A., Doretto, G.: Adversarial Latent Autoencoders. In: arXiv:2004.04467v1 [cs.LG] 9 Apr 2020
- [23] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In 12th {USENIX} symposium on operating systems design and implementation, pp. 265-283.

Authors Profile



Subhra Samir Kundu, is a post graduate student in Computer Application and currently associated with Amity Institute of Information Technology Kolkata. His research interests include social network analysis, image processing, computer vision, natural language processing and quantum machine learning. He is focused to contribute in deriving meaningful insights from data both in graphical and numerical forms. He has few SCOPUS publications to his name till date.



Dr. Ambar Dutta, did his B.Sc. (Honors) in Mathematics from Presidency College, Kolkata and Masters and Ph.D. from Jadavpur University, Kolkata. After serving in the department of Computer Science and Engineering, Birla Institute of Technology, Mesra for 15 years, he is at present working as Associate Professor in Amity Institute of Information Technology, Amity University, Kolkata. Dr. Dutta authored a book and has published more than 50 papers in reputed national/international journals/conferences. His research interest includes Image and Video Processing, Data Analytics, Machine Learning, Information Retrieval, Network Security. He is an active reviewer of many reputed journals like Pattern Recognition Letters (Elsevier), Multimedia Tools and Applications (Springer), IET Image Processing etc. He is also senior life member of various professional bodies.