Analyzing Generated Images using Machine Learning

Notebook for ImageCLEF Lab at CLEF 2024

Diana Miranda*, Aparna Rane† and Bipin Naik†

Department of Information Technology, Goa College of Engineering, Farmagudi, Ponda, Goa, India, 403401

Abstract

This paper provides a detailed explanation of the approaches proposed by the Biomedical Imaging Goa Lab for the ImageCLEFmedical GANs tasks. The proposed approaches use feature vectors extracted from the penultimate layer of a pre-trained CNN to represent CT scan images. For the first task, k-means clustering is first performed on the extracted features. Then, 1-nearest neighbour classifiers are used to determine if an input real CT scan image is used to generate the synthetic CT scan images. The best-performing model using this approach produced an F_1-score of 0.5315. For the second task, a Gaussian Mixture Model is used to perform clustering of the input deep features extracted from the CNN. This produced an Adjusted Rand Index of 0.63812.

Keywords

k-nearest neighbour classifier, Gaussian Mixture Model, Generative Adversarial Networks, synthetic medical images

1. Introduction

Artificial intelligence has brought about significant improvements in the fields of medicine and health care. The use of machine learning in automating diagnosis, content-based image retrieval (CBIR), and treatment recommendation systems has attempted to lessen the burden on existing healthcare professionals. However, the development of such automated systems requires enormous amounts of training data. Such training data, especially that involving medical images is often difficult to obtain. This could be due to privacy concerns among patients as well as a lack of homogeneity in the techniques used to capture the images. In the absence of such data, it is difficult to train machine learning models to give accurate results.

This issue could be solved by creating synthetic medical images using Generative Adversarial Networks (GANs), in which synthetic medical images can be generated from real ones. If this generation process could be performed without affecting the privacy of patients, it could be ideal for producing large amounts of medical data with ease. However, if the original medical image could be recovered from the synthetic image, it could raise privacy concerns. The ImageCLEF 2024 challenge is an initiative that tackles such issues [1]. As part of this challenge, the ImageCLEFmedical GAN track invites researchers to explore two tasks related to computed tomography (CT) scan images [2]. Task 1 Identify training data 'fingerprints' involves investigating if the original CT scan images can be detected from given generated CT scan images. This is a classification problem in which the generated CT scan images are provided along with two types of images, those that are used for the generation process and those that are not used. Participants are also given another set of synthetic images and are then required to predict if unknown test images are used to generate these synthetic images. Task 2 Detect generative models' 'fingerprints' is a clustering problem that aims to produce clusters of synthetic CT scan images so that it can be investigated if image fingerprints can be identified.

In this paper, we have proposed to use deep features extracted from a convolutional neural network (CNN) to represent the visual characteristics of the CT scan images. For the first task, we have assumed that the test images are similar to the training images. Based on this assumption, we propose to use

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

†These authors contributed equally.

diana@gec.ac.in (D. Miranda); aparna@gec.ac.in (A. Rane); bn@gec.ac.in (B. Naik)

a 1-nearest neighbour classification approach in which the test image is compared to cluster centers obtained after performing k-means clustering on the training data. The second task aims to find image fingerprints in the synthetic CT scan images. To do this, we propose to extract deep features from these CT scan images and then use a Gaussian mixture model to perform clustering. This is based on the assumption that the data follows a Gaussian distribution in a GMM. This is a preliminary study on possible approaches to solve these problems and we will investigate other solutions in the future.

2. Proposed Approach

Each CT scan image is given as input to a CNN and the output of one of the layers of the CNN is considered as the feature vector for the input image. Identifying whether or not, a test image is used to generate synthetic CT scan images is done using a 1-nearest neighbour-based classification approach using the feature vectors obtained from the CNN. For the clustering task, a GMM is used to perform clustering on the extracted feature vectors.

In this section, a description of the method used for feature extraction is provided. This is followed by an explanation of the technique used for the classification of the CT scan images. Lastly, the clustering approach used to identify the fingerprints in the generated images is elucidated.

2.1. Feature Extraction

A CNN performs two main functions: (1) feature extraction from the input, and (2) classification of the extracted features to certain classes. A CNN consists of different layers that each perform a distinct operation. An image when passed through a CNN undergoes a series of transformations across these layers. The final output layer of the CNN performs the classification task. Once an image is passed through a CNN, the output of any of the internal layers can be used to represent the given image. For a CNN to effectively capture the visual characteristics of the input images, the CNN must be trained on a large number of images. However, in the absence of sufficient training images, a CNN that is pre-trained on images from a similar dataset can be used. As the number of CT scan images available for both tasks was not adequate to train a CNN from scratch, we have used a CNN that is pre-trained on natural images from the ImageNet dataset [3]. The work proposed in [4], shows the effectiveness of features extracted from the ResNet50 CNN in performing medical image modality classification for datasets that contained a large number of radiographic images. Therefore, we propose to use the features extracted from the penultimate layer of the ResNet50 CNN to represent the CT scan images [5]. Each CT scan image I is passed through a pre-trained ResNet50 CNN and the output from the average pooling layer after the last convolutional layer is extracted. This is a 2048-dimension vector I' that is used to represent the features of the input image I. These features are then used for classifying the CT scan images as well as for clustering to identify the fingerprints in generated images.

2.2. Classification Approach for Task 1 Identify Training Data 'Fingerprints' to Detect Real Images Used to Generate Synthetic Images

The proposed approach to determine if a given CT scan image is used to generate a set of synthetic CT scan images is based on the assumption that the images used for generation in the training and test datasets are similar. As shown in Figure 1, each image I in the training dataset is passed through the pre-trained ResNet50 CNN, and the 2048-dimension output I' of the average pooling layer after the last convolutional layer is extracted as the input feature vector. If the training dataset contains n training images $U = \{U_1, \ldots, U_n\}$ that were used to generate the synthetic images, the corresponding feature vectors are $U' = \{U'_1, \ldots, U'_n\}$. If the training dataset contains n images $O = \{O_1, \ldots, O_n\}$ that were not used to generate the synthetic images, then their corresponding feature vectors are $O' = \{O'_1, \ldots, O'_n\}$. Both sets of feature vectors are separately clustered using the k-means clustering approach where the value of k is varied [6]. Upon performing clustering on the set U', k cluster

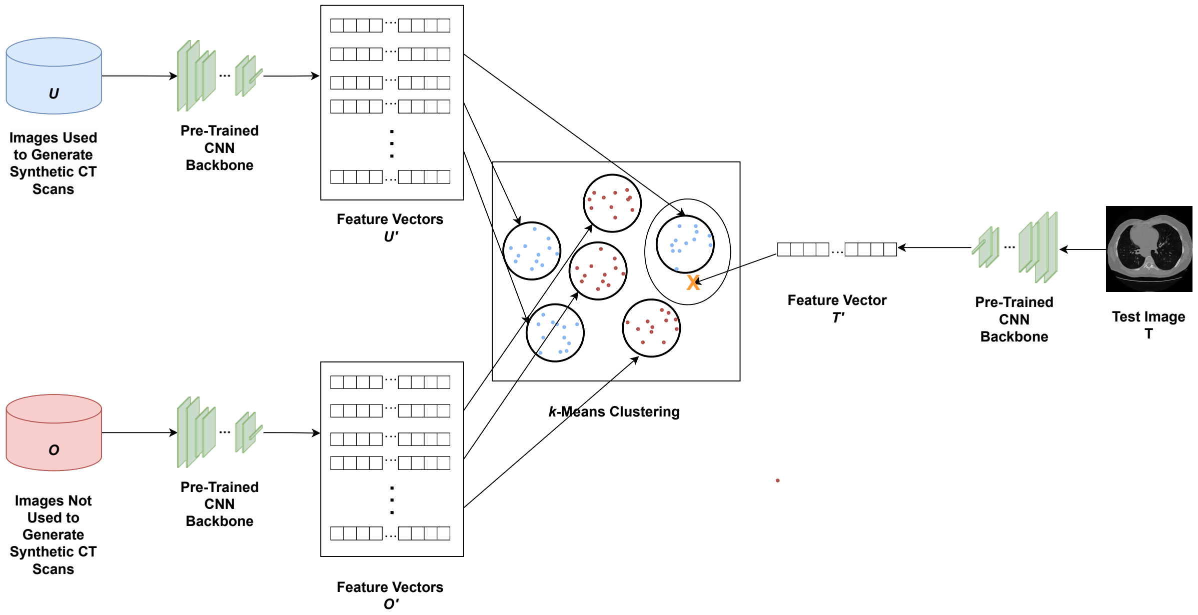

Figure 1: Proposed approach for Task 1 Identify training data 'fingerprints'.

centers $M^U = \{m_1^U, \ldots, m_k^U\}$ are obtained. Clustering the feature vectors present in O' results in the formation of k clusters with cluster centers $M^O = \{m_1^O, \ldots, m_k^O\}$.

For the testing phase, each test image T is also passed through the same pre-trained ResNet50 CNN, and the 2048-dimension output is extracted as T' to represent the test image T. To identify the closest cluster center, the Manhattan distance is used as a similarity measure [6]. The Manhattan distance between two d-dimensional vectors $\boldsymbol{a} = [a_1\ a_2\ \ldots\ a_d]$ and $\boldsymbol{b} = [b_1\ b_2\ \ldots\ b_d]$is calculated as follows:

$$D(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{d} |a_i - b_i| \tag{1}$$

The Manhattan distance between T' and each of the k cluster centers in M^U and M^O is calculated as $D_{T'}^U$ and $D_{T'}^O$, respectively. If the smallest distance in $D_{T'}^U$ is less than the minimum distance in $D_{T'}^O$, the test image T is classified as used and is considered to be used to generate the synthetic images. Otherwise, the test image T is considered not to be used to generate the synthetic CT scan images. In this way, the class label of the 1-nearest neighbour cluster center is assigned to the test image T.

2.3. Clustering Approach for Task 2 Detect Generative Models' 'Fingerprints' to Identify Image Fingerprints

For this task, the feature vectors for all the images in the training and test datasets are extracted from the penultimate layer pre-trained ResNet50 CNN that is pre-trained on the ImageNet dataset. The 2048-dimension vectors are then clustered using a Gaussian Mixture Model (GMM) [7] as shown in Figure 2. GMM is effective when the data is heterogeneous. This is useful when a data point is close to multiple clusters making it difficult to assign that data point to a single cluster. In the proposed approach using GMM, the data is distributed into k components that are assumed to follow Gaussian distributions where each i^{th} component has its mean μ_i and covariance matrix Σ_i where $i = 1, \ldots, k$. The mean μ_i and covariance matrix Σ_i are initialized using four different methods. The expectation maximization (EM) algorithm is then used to estimate the mean μ_i and covariance matrix Σ_i for each of the k components.

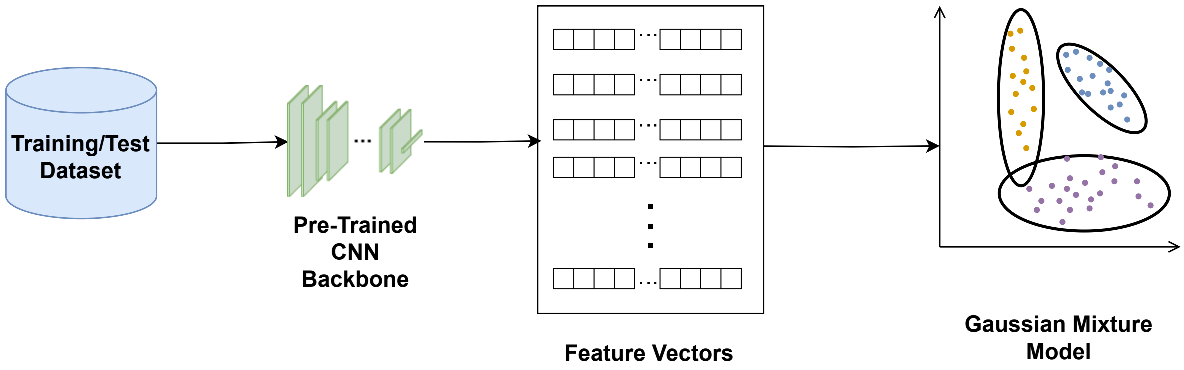

Figure 2: Proposed approach for Task 2 Detect generative models' 'fingerprints'.

3. Experimental Results

This section consists of a description of the datasets used for both tasks followed by a detailed explanation of the experiments conducted on the datasets.

3.1. Datasets

Both datasets consist of axial slices of three-dimensional CT scan lung images. These images were of 256×256 pixels in size. Since the first task involves detecting which real CT scan images were used to generate the synthetic CT scan images, this dataset consists of a mix of real and synthetic CT scan images. The details of the dataset used for Task 1 Identify training data 'fingerprints' are as follows:

- The training dataset consists of two sets of image data. The first set contains 100 real images used for the generation task, 100 images not used for the generation task, and $10,000$ generated images that are generated from the real CT scan images. The second set consists of $3,000$ real images used for the generation task, $3,000$ images that are not used for the generation task, and $10,000$ CT synthetic CT scan images generated from the real images.
- The test dataset also consists of two sets of image data. The first set contains $5,000$ synthetic CT scan images and $4,000$ unknown real CT scan images. The second set consists of $7,200$ synthetic CT scan images and $4,000$ unknown real CT scan images. The main aim of the task involving this dataset is to determine if the unknown real CT scan images are used to generate synthetic CT scan images.

The goal of the second task is to identify the image fingerprints in the generated CT scan images in the dataset. The details of the dataset used for Task 2 Detect generative models' 'fingerprints' are as follows:

- The training data contains three sets of synthetic CT scan images with each set having 200 images each.
- The test dataset consists of $3,000$ synthetic CT scan images.

3.2. Details of the Submitted Runs for Task 1 Identify Training Data 'Fingerprints'

The training data consists of images U that are used to generate the synthetic CT scan images and images O that are not used to generate the synthetic images. Once the feature vectors U' and O' are extracted after passing the images through the CNN, they are separately clustered using the k-means clustering method to obtain k cluster centers each in M^U and M^O, respectively. During testing, each test image is compared with the $2 \times k$ cluster centers in M^U and M^O, and the label of the closest cluster is assigned to the test image. For this task, we have submitted 6 runs for different values of k

Table 1
Performance of the proposed approach for Task 1 Identify training data 'fingerprints' on the test dataset

Submission ID	Number of clusters (k)	F_1-score
876	1	0.4295
877	2	0.3855
898	4	**0.5315**
875	8	0.5145
874	16	0.499
873	32	0.4975

Table 2
Performance of the proposed approach for Task 2 Detect generative models' 'fingerprints' on the test dataset

Submission ID	Method of GMM initialization	Adjusted Rand Index
307	k-means	**0.638117**
321	k-means++	0.434414
323	random	0.000047
324	random from data	0.272976

used for k-means clustering. The values of k are powers of 2, i.e. $1, 2, 4, 8, 16, 32$. The results of this task are provided in Table 1.

The results indicate that the F_1-scores vary with the value of k. The best result of an F_1-score of 0.5315 from the submitted runs was obtained by the model where $k = 4$. The second best result was obtained by the model with $k = 8$ with an F_1-score of 0.5145. The submitted run where $k = 2$ gave the worst performance. This shows that an optimal value of k must be chosen.

3.3. Details of the Submitted Runs for Task 2 Detect Generative Models' 'Fingerprints'

Once the features are extracted from the CNN, they are clustered using a Gaussian Mixture Model (GMM). During the training process, we experimented with GMMs that had 3 components. This was based on our observations of the training data. During the test phase, we assumed that the GMMs have 4 components. The 4 submitted runs differ based on the methods used to initialize the means of the components in the GMMs. The 4 methods used for initializing are as follows:

1. k-means: Uses the k-means clustering algorithm for initialization [6]
2. k-means++: Uses the k-means++ clustering algorithm in which the first clusters are chosen randomly while the cluster centers in the subsequent iterations are chosen based on the maximum squared distance [8]
3. random: Chooses the component means randomly
4. random from data: Chooses random data points as component means

The results of this task are provided in Table 2. The performance of the proposed approach indicates that initialization of the GMM components using the k-means algorithm gives the best results while random initialization of the GMM components is not effective. However, the proposed approach is based on the assumption that the data follows a Gaussian distribution. If this is not true, then the proposed approach may not be an ideal solution to this problem.

4. Conclusion

This paper describes the methods used by the Biomedical Imaging Goa group for the ImageCLEFmedical GANs task. The proposed approach uses features extracted from the penultimate layer of a CNN that is pre-trained on natural images. For the classification task, the test image is compared with the cluster centers formed from the training data and labeled according to a 1-nearest neighbour approach. For the clustering task, the feature vectors are clustered using a GMM-based approach.

Since this is a preliminary study conducted by our group, in the future we could explore other possible methods to improve efficiency. This could involve using feature vectors from other CNNs as well as different classification and clustering methods. Further, the proposed approaches are based on certain assumptions that could be investigated to provide a better explanation of other possible solutions.

References

[1] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. B. Abacha, A. G. S. de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[2] A. Andrei, A. Radzhabov, D. Karpenka, Y. Prokopchuk, V. Kovalev, B. Ionescu, H. Müller, Overview of 2024 ImageCLEFmedical GANs Task – Investigating Generative Models' Impact on Biomedical Synthetic Images, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[4] D. Miranda, V. Thenkanidiyoor, D. A. Dinesh, Detecting the modality of a medical image using visual and textual features, Biomedical Signal Processing and Control 79 (2023) 104035.

[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[6] R. O. Duda, P. E. Hart, D. G. Stork, Pattern classification, John Wiley & Sons, 2006.

[7] E. Patel, D. S. Kushwaha, Clustering cloud workloads: K-means vs gaussian mixture model, Procedia computer science 171 (2020) 158–167.

[8] D. Arthur, S. Vassilvitskii, et al., k-means++: The advantages of careful seeding, in: Soda, volume 7, 2007, pp. 1027–1035.