Covid-19 search engine

Marco Ghezzi, Matteo Limoncini

October 19, 2022

Abstract

The aim of this project is to create an information retrieval system being able to search and make queries over a large dataset of medical images. This system will be able to find the most similar images to a given one. We applied two different machine learning techniques in order to build our retrieval system: supervised learning, using an artificial neural network to build a classification system, and clustering. These different techniques were compared by analyzing different performance metrics and computational time; employing a Convolutional Neural Network (supervised learning) produces the best results in terms of precision and time efficiency.

1 Introduction

Content-based image retrieval (CBIR) is an image search technique designed to find images that are most similar to a given query: it measures the similarity of two images based on the similarity of the properties of their visual components.

This project aims to create a search engine being able to respond to a given query.

This project tries to solve the problem of searching and making queries over a large dataset of images. The main goal is to build a content-based medical image retrieval system using different approaches and comparing their results based on classification metrics and computational time.

2 Research question and methodology

Two different Machine Learning approaches were used: the first method is performing unsupervised learning with clustering, the second one is using supervised training with binary classification.

The most important stage before applying one of the clustering algorithm is the feature extraction stage in which a visual concept is converted to a numerical form: these features could be in the form of global features (i.e. color, shape, texture ...) or local features (like corners blobs or edges) that are invariant against scale, translation and rotation changes.

After the feature extraction stage, clustering is performed. Clustering is an unsupervised learning algorithm that gathers image descriptors into a single group that semantically differs from other groups: the most used clustering algorithm in CBIR is K-Means, even though it fails in handling outliers and noisy data.

In this work we also decided to address the problem of classification since we had prior knowledge of the labels of the images: in order to classify the images in a supervised learning way, was developed a Convolutional Neural Network.

The final step is the similarity measurement between the extracted features from the query image and all other images in the dataset to retrieve the most relevant images. There are different metrics that could be used: in this work were considered the Euclidean distance [1] and the cosine distance [2].

$$d(a,b) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2} \tag{1}$$

$$d(a,b) = \frac{A \cdot B}{||A|| \cdot ||B||} \tag{2}$$

2.1 Data

The data collected in this dataset are computed tomography scans taken from real patients in São Paulo, Brazil: every image is paired with a label to specify if the person associated to the image was tested positive for Covid-19.

The dataset consists of 2482 images of different pixel sizes; the labels are approximately equally divided.

2.2 Image Preprocessing

Every image is associated with a label and they've been divided in two directories: the first containing all the images labeled as covid and the second containing all the images labeled as non covid. For the classification task the labeled images need to be divided in three different datasets: train, validation and test sets. It was used a split ratio of 0.8, 0.1 and 0.1 respectively. Starting from the images in these directories, three different datasets were created: one for training data, one for validation data and the last one for test data. While creating these datasets, it was specified the size of the images in order to find the best trade-off between result accuracy and training time: in fact this resize operation is necessary because all the images in the dataset have different sizes. Then all pixel values was normalized between 0 and 1 since the computation of high numeric values may become more complex.

In order to create a system that is able to answer the first query with the most precision and less time consuming (identifying the computed tomography scan images with covid) different Machine Learning approaches was evaluated.

2.3 Classification

Since this is an image binary classification problem, a good choice can be work with an Artificial Neural Networks, in particular with a Convolutional Neural Network. A Convolutional Neural Network (CNN) is an artificial neural network suited for images, with a sequence of layers, and every layer transforms one volume of activations to another through a differentiable function. Three main types of layers was used to build our neural network: Convolutional Layer [1], Pooling Layer [2] and Fully-Connected Layer.

Using this CNN, interesting results was obtained: this network is able to correctly classify a new image with an accuracy of 89% with less than a minute of training time over 1986 images.

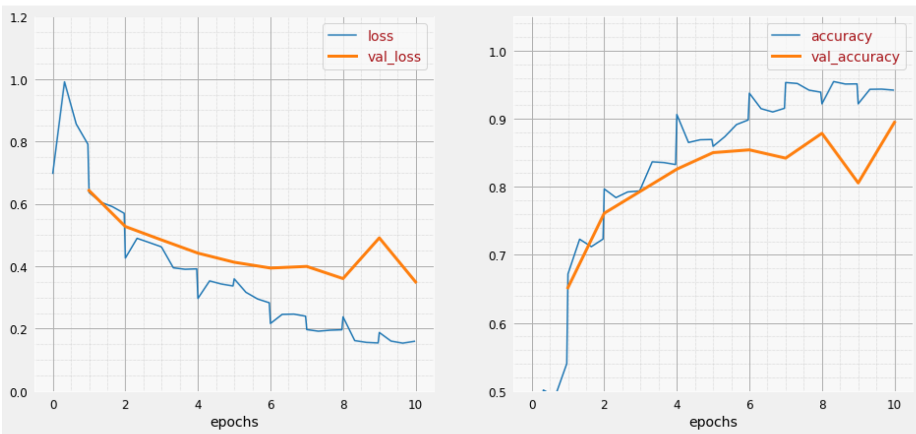


Figure 1: Performance of a CNN for binary classification

2.4 Clustering

The second method used in this project was clustering. First of all, since the dataset has images with multiple dimensions, it's important to choose the correct feature extraction method.

2.4.1 Feature extraction methods

- Principal component analysis(PCA)

  It allows us to extract features and at the same time reduce the dimensions of our dataset to any number less than current number of features, always preserving the maximum amount of information. Another application of PCA is to compress the data and hence reduce the computational time. In this project, this technique was used to tackle both problems.

[1] `https://www.tensorflow.org/api\_docs/python/tf/keras/layers/Conv2D`
[2] `https://www.tensorflow.org/api\_docs/python/tf/keras/layers/MaxPool2D`

- VGG16

  Another idea is to use a Convolutional Neural Networks as feature extraction method. Vgg16 is a Convolutional Neural Network (CNN) model proposed by Karen Simonyan and Andrew Zisserman at the University of Oxford. Vgg16 has two targets. The first is to detect objects within an image coming from 200 classes, which is called object localization. The second is to classify images, each labeled with one of 1000 categories, which is called image classification. This CNN was choosen to perform the feature extraction stage, so it was removed the last layer of the network and the last dense layer was used as a feature vector. This means that the new final layer is a fully-connected layer with 4,096 output nodes. This vector of 4,096 numeric values is the feature vector used to cluster the images. Since this layer has over 4000 dimensions it's important to reduce its dimensions using PCA. PCA allow us to reduce the dimensionality keeping as much information as possible, after this step we have a smaller feature vector thus it was applied a classic cluster algorithm.

2.4.2 Clustering algorithm

After feature extraction we need to apply the cluster algorithm. Clustering is an unsupervised learning technique (no labels needed) that has the goal of grouping objects in such a way that objects in the same group (*cluster*) are more similar, according to a certain distance measure, to each other than objects in different groups.

There are different approaches and algorithms to perform clustering tasks which can be divided into three sub-categories:

- Partition-based clustering: E.g. k-means
- Hierarchical clustering: E.g. Agglomerative
- Density-based clustering: E.g. DBSCAN

K-Means

K-Means is a simple, relatively fast and easy scalable partition based algorithm, its objective is to minimize the average squared Euclidean distance of images from their cluster centers. K-Means needs to know in advance the number of clusters, it's sensitive to outliers and can consider only linear boundaries, so it's not optimal for clusters with non linear structure. Another problem is that K-Means has a random initialization so can generate different clusters in different run.

Agglomerative clustering

Agglomerative clustering, also called bottom-up clustering, treats each image as a singleton cluster and after that it merges clusters until a single cluster will contain all the images. It aggregates items starting with the most similar, each cluster that is defined this way, substitutes the corresponding items in the dataset. Then is evaluated the similarity between different clusters in order to define their similarity. It is an easy to implement algorithm and always generate same clusters but it's slow for large datasets, complexity is at least $O(n^2)$.

DBSCAN

DBSCAN views clusters as areas of high density separated by areas of low density. It has good performance with arbitrary shapes clusters and it is robust to outliers; on the other hand it has some parameters difficult to determine (like the threshold under which two points are considered neighbors) and it doesn't behave very well if clusters are very different in term of in-cluster densities.

In this project K-Means was used because in this case it's important to use an algorithm able to scale to large dataset and able to produce a result in not so much time. It's not a problem fixing a number of clusters predetermined because the number of clusters is 2 (images with covid and images without covid).

3 Experimental results

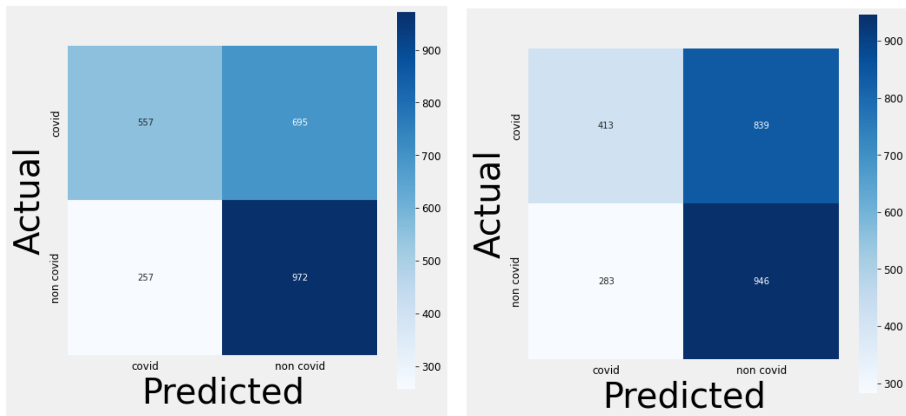
In this work different approaches was used to solve the same queries.

The different approaches with different execution times and precision was summarized in figure 2.

	Computational time
Classification with CNN	0:00:51.235565
Clustering with PCA - KMeans	0:03:40.649903
Clustering with VGG - KMeans	0:02:44.163987
Image similarity - Euclidean	0:02:18.823460
Image similarity - Cosine	0:05:50.546088

Figure 2: Comparison of different execution times

In the figure [3] the confusion matrix of the clustering with PCA and the confusion matrix of the clustering with VGG16 was compared.



(a) Confusion matrix clustering PCA (b) Confusion matrix clustering VGG16

Figure 3: Comparison between PCA and VGG16

From a more generic point of view with clustering there was a great number of false negative, but the system was able to recognize pretty well the true negative with a small number of false positive. There's not a great difference between the two approaches of clustering with different feature extraction methods.

4 Conclusion and future work

In conclusion the classification task, using a Convolutional Neural Network, produces better results using less training time than all the other methods; this is due to the fact that this supervised learning technique makes use of labeled images, so it is easier to classify two different images. It is also quite useful to note that even the clustering technique, an unsupervised learning method, produces good results, correctly classifying more than 60% of the given images; it can be an alternative way to analyze images when its respective labels are unknown.

One of the first things to do in order to improve the quality of the final results is to enhance the image preprocessing; it is possible to apply data augmentation in order to find the correct pattern that properly classifies an image.

Another option is to improve the explainability of the process, in particular how a method achieves its corresponding result. For example some explainable AI tools can be used, a set of processes and methods that allows human users to comprehend and trust the results and output created by the machine learning algorithms. Explainability can help developers ensure that the system is working as expected and in this case in particular helps us to understood why one image is chosen instead of another.

References

[1] Kumar, A., Kim, J., Cai, W., Fulham, M., & Feng, D. (2013). Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data. Journal of digital imaging, 26(6), 1025-1039.

[2] Yan, K., Wang, X., Lu, L., & Summers, R. M. (2018). DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of medical imaging, 5(3), 036501

[3] Ibtihaal M. Hameed, Sadiq H. Abdulhussain & Basheera M. Mahmmod — D T Pham (Reviewing editor) (2021) Content-based image retrieval: A review of recent trends, Cogent Engineering, 8:1, DOI: 10.1080/23311916.2021.1927469

[4] Ling Ma, Xiabi Liu, Yan Gao, Yanfeng Zhao, Xinming Zhao, Chunwu Zhou, A new method of content based medical image retrieval and its applications to CT imaging sign retrieval, Journal of Biomedical Informatics, Volume 66, 2017, Pages 148-158, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2017.01.002. (https://www.sciencedirect.com/science/article/pii/S1532046417300023)

[5] Afshan Latif, Aqsa Rasheed, Umer Sajid, Jameel Ahmed, Nouman Ali, Naeem Iqbal Ratyal, Bushra Zafar, Saadat Hanif Dar, Muhammad Sajid, Tehmina Khalil, "Content-Based Image Retrieval and Feature Extraction: A Comprehensive Review", Mathematical Problems in Engineering, vol. 2019, Article ID 9658350, 21 pages, 2019. https://doi.org/10.1155/2019/9658350

[6] Very Deep Convolutional Networks for Large-Scale Image Recognition Karen Simonyan, Andrew Zisserman