




# Image Document Optimization - SBI

Powered By - Microsoft  
Corporation Pvt Ltd.

**TEAM KALAM**

- Abha Porwal
  - Sambit Kumar Mishra
- 

## Problem Being Solved

1. The customers of SBI are increasing on a daily basis. This increase in the number results in increase in the size of the database required to store the documents of the customers.
2. The documents are generally in the image format uploaded by the customers themselves.
3. The raw format of the documents are typically of sizes of few MBs.
4. Storing the raw format of the uploaded data in the database is an inefficient approach because it increases the POST and FETCH latency. It also increases the cost of maintaining the database.
5. Scaling is very inefficient if such inefficient storage methods are implemented.

# Approach taken to create the model

The major focus is kept on reducing the size and white space of the document without compromising with the quality much.

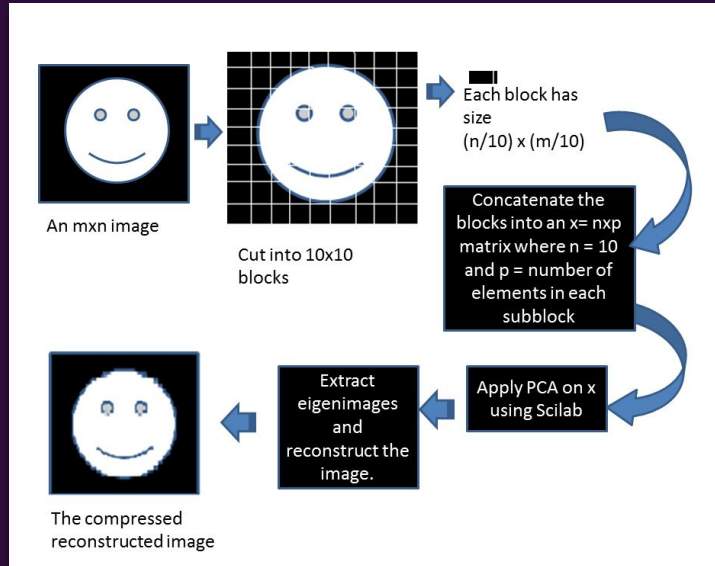
There are various options available for image compression like -

- K-Means Clustering
- PCA
- GAN
- Computer Vision
- Azure CDN

## Approach taken to create the model

- Among the various possible approaches, the implementation of the solution is currently done using PCA.
- **Principal Component Analysis** or **PCA** is a method of reducing the dimensions of the given dataset while still retaining most of its variance.
- Eg. In one of the iterations, the original image of an ID Card of size 1.5 MB was compressed to 45 KB without significant loss of resolution of text and image in the ID.

# PCA Model Visualization



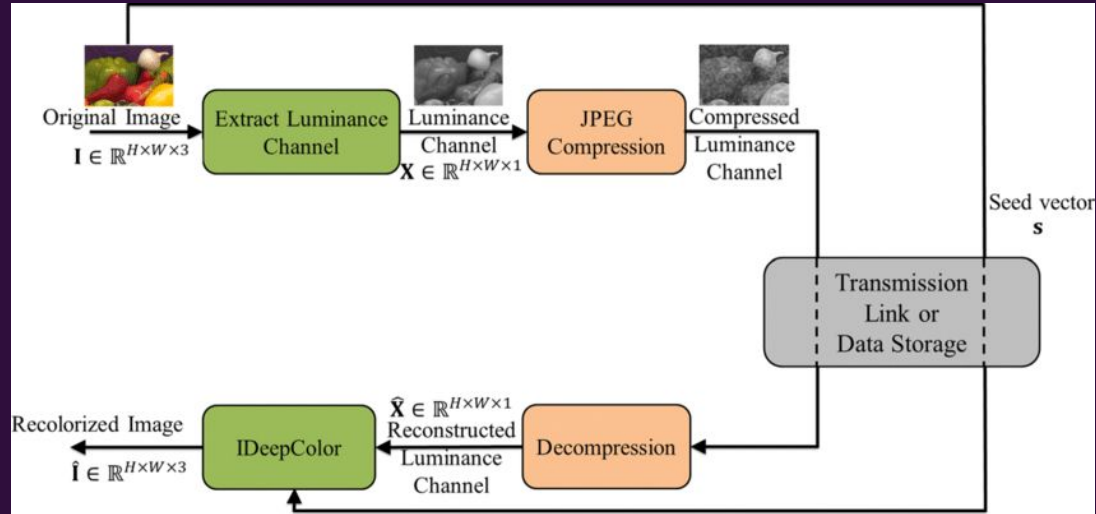
## Approach for improvement of model

- Although PCA would help to achieve the desired goal, but there are advanced methods which can optimize the compression.
  - GAN (Generative Adversarial Network)
  - Azure CDN (Content Delivery Network)
- GAN is one of the advanced methods.
- Azure CDN (Content Delivery Network) can also be leveraged to achieve significant image compression without losing the quality significantly.

# GAN (Generative Adversarial Network)

- GANs are *generative* models: they create new data instances that resemble your training data.
- GANs achieve this level of realism by pairing a generator, which learns to produce the target output, with a discriminator, which learns to distinguish true data from the output of the generator. The generator tries to fool the discriminator, and the discriminator tries to keep from being fooled.

# GAN (Generative Adversarial Network)

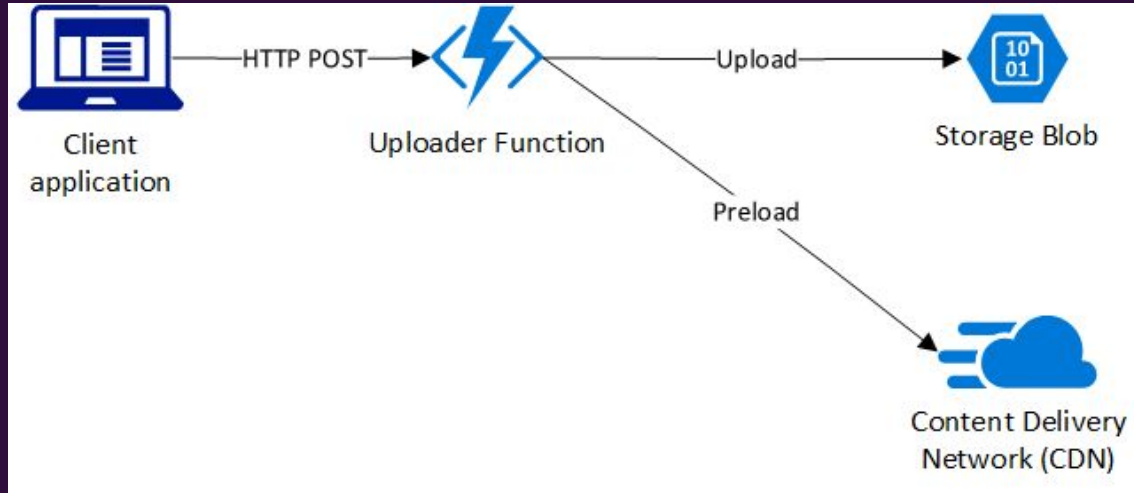




# Azure CDN (Content Delivery Network)

- Enable compression on your origin server. In this case, Azure CDN passes along the compressed files and delivers them to clients that request them.
- Enable compression directly on the CDN POP servers (*compression on the fly*). In this case, the CDN compresses the files and serves them to the end users, even if they were not compressed by the origin server.

# Azure CDN (Content Delivery Network)



## FR (Functional Requirements)

- Reduced latency
- Response time should be less than 1 minute
- Minimum system overhead
- High compression ratio
- No significant reduction in image resolution
- POST and FETCH operations should be available for multiple concurrent users without latency.
- ACID/BASE principle should be followed
- Highly scalable database

## NFR (Non-Functional Requirements)

- Improved UI and UX
- Hints and instructions in the webpage for guiding the new customers
- Timely removal of documents of ex-users from the database to free up space
- Success/Failure message to the customer after uploading the documents

## Reason why the solution should be considered

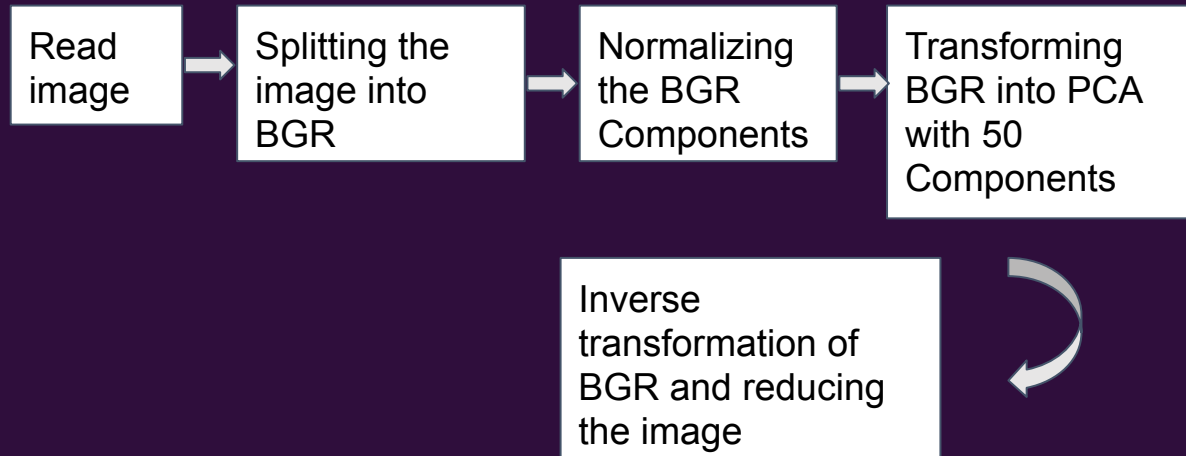
- Most optimized model selected for the implementation of the solution.
- PCA implemented as of now, but the final model would be based on GAN which would improve the compression efficiency.
- Entire database size will reduce significantly.
- This method is applicable to all the types of document images uploaded by the user.

## Source code (Github Repository Link)

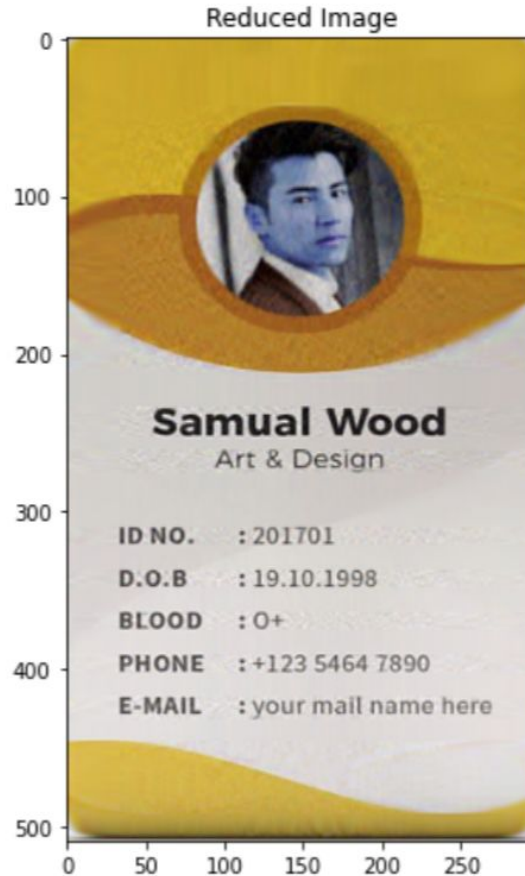
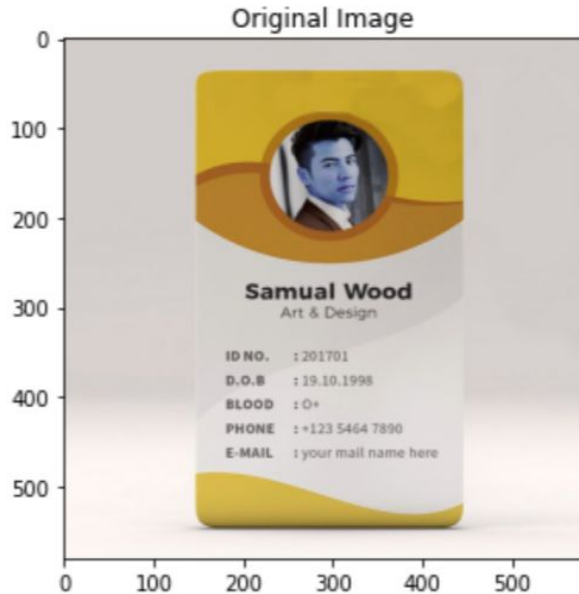
<https://www.github.com/MAVIN-07/sbi-hackathon/>

## Working of the solution.

The proposed PCA model works in the following manner -



## OUTPUT OF THE PROPOSED MODEL





# THANK YOU

Submitted By : Abha Porwal , Sambit Kumar Mishra

Email : [abhaporwal12@gmail.com](mailto:abhaporwal12@gmail.com) , [sambitmishra1968@gmail.com](mailto:sambitmishra1968@gmail.com)

Mobile No: 9111027570 , 9630185553

**TECHGIG**