ANKARA ÜNİVERSİTESİ

MÜHENDİSLİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

COM4061 PROJECT REPORT

VQA(Visual Question Answering)

Selman YILDIZ

18290133

Doç.Dr.Gazi Erkan Bostancı

Aralık / 2022

ABSTRACT

The problem addressed in this report is to design a system that works by understanding the content of that image as human-like, in order to create a system that tries to understand correctly and answer the questions we ask of the image. In this report, I presented the Visual Question Answering system, a system that is constantly improving itself and prepared with artificial intelligence methods, using effective technologies.

ABSTRACT

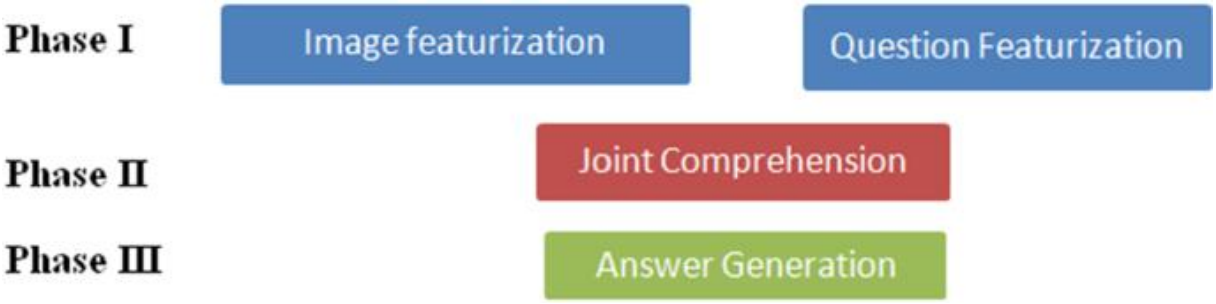
1. INTRODUCTION 4
    1.1. Image Featurization 5
        1.1.1. CNN (Convolutional Neural Network) 5
    1.2. Question featurization 6
        1.2.1. RNN (Recurrent Neural Network) 6
    1.3. Joint Representation 7
    1.4. Answer Generation 8
2. DATASETS 8
    2.1. ImageNet 8
    2.2. COCO (Common Objects in Context) 9
    2.3. MNIST 10
    2.4. FashionMNIST 10
    2.5. PASCAL VOC 11
    2.6. The VQA Dataset 11
3. STAGES FOR TRAINING MODEL 12
    3.1. Preprocessing 13
    3.2. Chosing Training Algorithm And A Model 14
        3.2.1. Models 14
4. TECHNOLOGY 14
    4.1. KAGGLE 14
    4.2. Tensorflow 16
5. CONCLUSION 19
6. BIBLIOGRAPHY 20

1. INTRODUCTION

Visual question answering (VQA) is a task in the field of artificial intelligence that involves building machine learning models that can understand and reason about visual information and generate natural language answers to questions about images. The goal of VQA projects is to create models that can understand and reason about visual information in a way that is similar to how humans do. This requires the model to have a deep understanding of both language and image content, and to be able to integrate this understanding in order to generate appropriate responses to questions. VQA is a challenging task that has many potential applications, including image captioning, image search, virtual assistants, and educational tools. It is an active area of research, and there have been numerous successful VQA projects developed in recent years. To build a VQA model, researchers typically start by collecting and annotating a large dataset of images and associated questions and answers. They then use this dataset to train and evaluate machine learning models that are designed to generate appropriate answers to the questions. These models may be based on a variety of machine learning techniques, including deep learning, natural language processing, and computer vision. To give you an idea of the subproblems the task of visual question answering entails:

Task for VQA	VQA questions
Object recognition	What is in the image?
Object detection	Are there any dogs?
Attribute classification	What color is in the umbrella?
Scene classification	Is it raining?
Counting	How many people are there in the image?
Activity Recognition	Is the child crying?
Relationships among objects	What is between cat and sofa?
Knowledge-base reasoning	Is this vegetarian pizza?

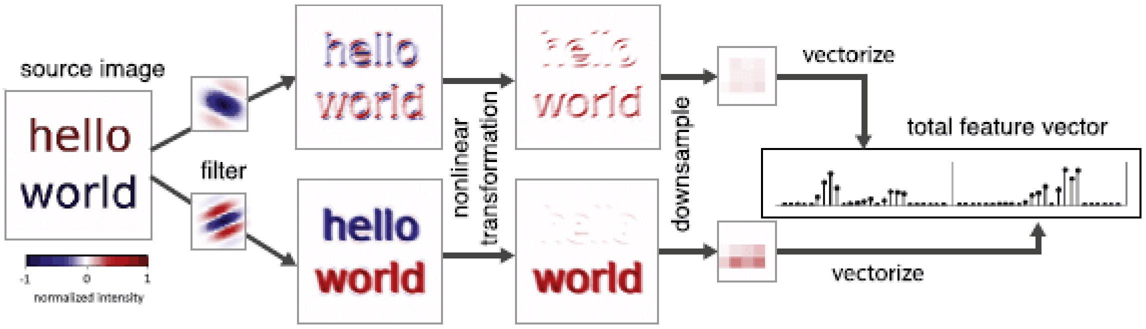
Solutions to these problems involve four major steps:



- **Image featurization:** converting images into their feature representations for further processing.
- **Question featurization:** converting natural language questions into their embeddings for further processing.
- **Joint feature representation:** ways of combining image features and the question features to enhance algorithmic understanding.
- **Answer generation:** utilizing the joint features to understand the input image and the question asked, to finally generate the correct answer.

1.1. Image Featurization

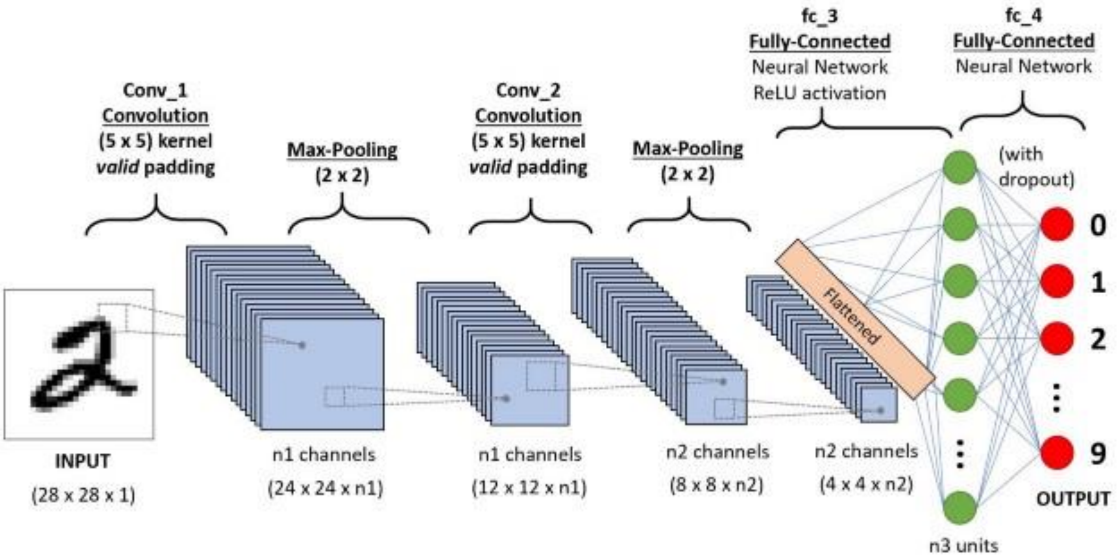
Convolutional neural networks have become the gold-standard for pattern recognition in images. After an input image is passed through a convolutional network, it gets transformed into an abstract feature representation. Each filter in a CNN layer captures different kinds of patterns, such as edges, vertices, contours, curves, and symmetries



.

1.1.1. CNN (Convolutional Neural Network)

A Convolutional Neural Network (CNN) is a Deep Learning algorithm that can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image, and be able to differentiate one from the other.



1.2. Question featurization

There are several methods to create embeddings. Older approaches include count-based, frequency-based methods like count vectorization and TF-IDF. There are prediction-based methods like continuous bag of words and skip grams as well. Pretrained models for the Word2Vec algorithm are also available in open source tools like Gensim. Deep learning architectures like RNNs, LSTMs, GRUs, and 1-D CNNs can also be used to create word embeddings. In VQA literature, LSTMs are used most frequently.

1.2.1. RNN (Recurrent Neural Network)

RNN works on the pirinciple of saving the output of particular layer and feeding this back to input in order to predict the output of the layer.Below is how you can convert a Feed-Forward Neural Network into RNN:

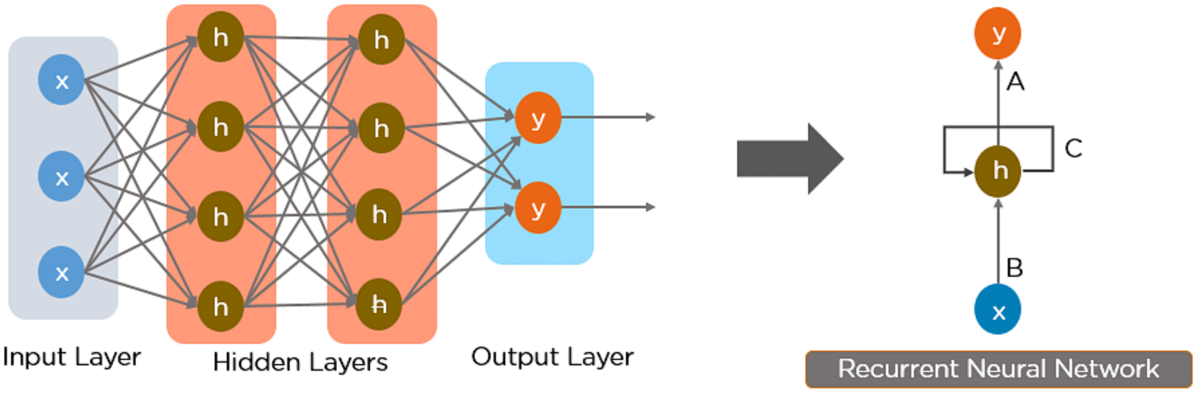
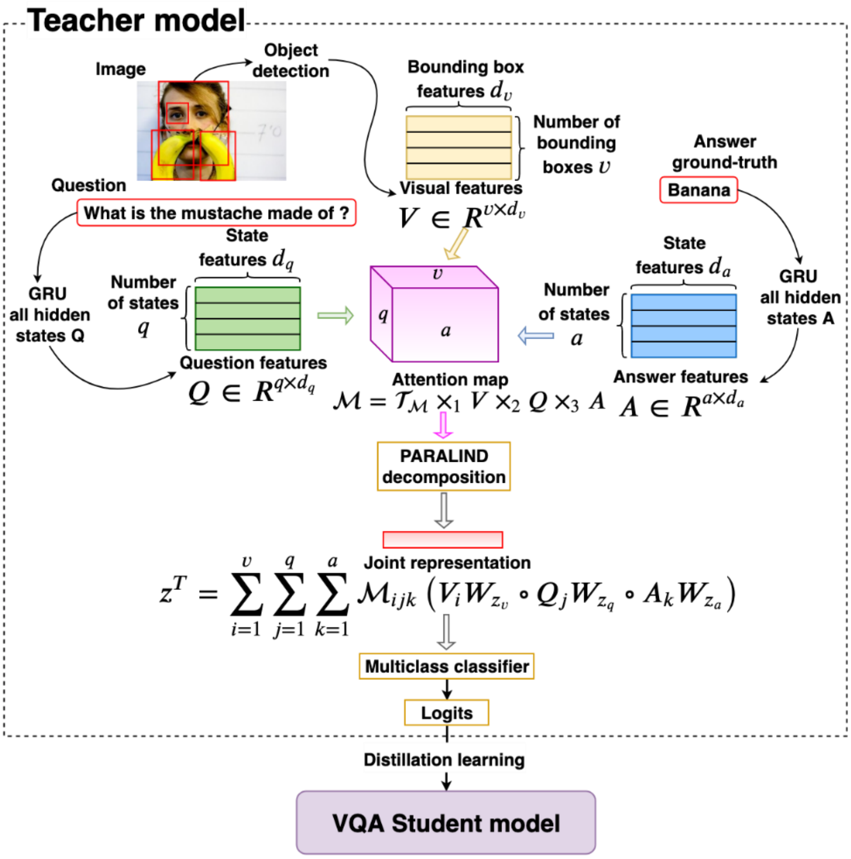


Figure1. Simple Recurrent Neural Network

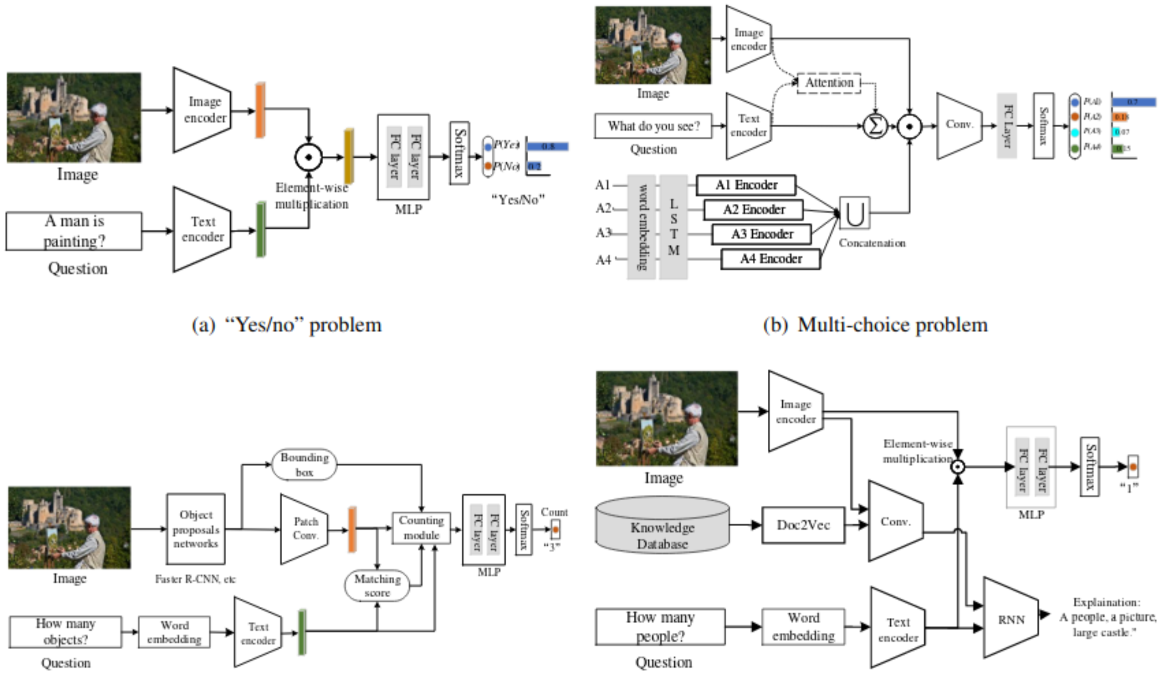
1.3. Joint Representation

In current VQA systems, the joint modality component plays an essential role since it would learn meaningful joint representations between linguistic and visual inputs by applying the attention mechanism. There are many works that learn the interaction between question and image. For instance, a novel trilinear interaction model which simultaneously learns high level associations between image, question and answer information.



1.4. Answer Generation

Binary questions and multiple choice questions often utilize a sigmoid layer at the end. The joint representations are passed through one or two fully-connected layers. The output is passed through a single neuron layer which functions as the classification layer.



(a) "Yes/no" problem

(b) Multi-choice problem

(c) Number counting problem

(d) Open-ended problem

Figure 2 Common types of visual question answering. "Yes/No" problem and multi-choice problem can be regarded as a classification problem, while number counting problem and open-ended problem can be viewed as a caption generation problem.

Figure.

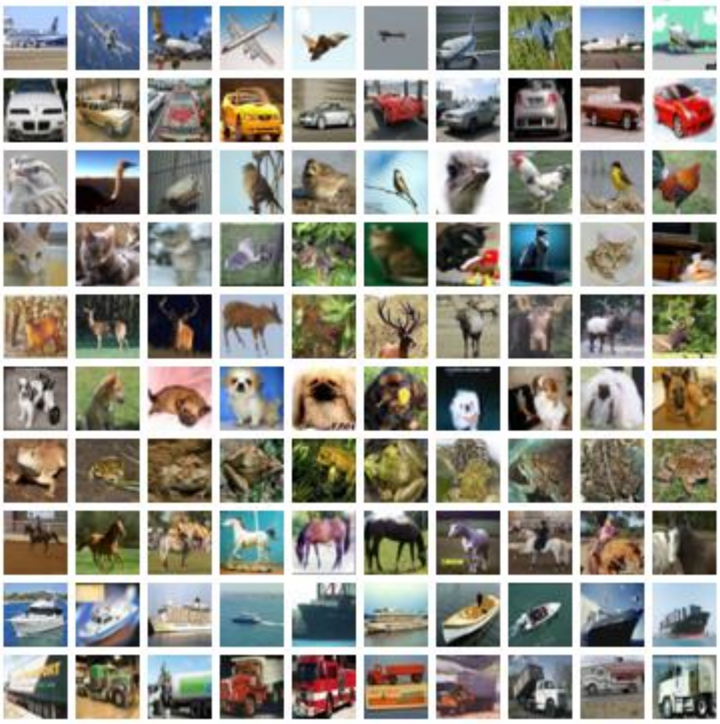
We talked briefly about algorithms in the previous sections. Now in this section, I will give information about how we should choose the dataset and about each dataset.

2. DATASETS

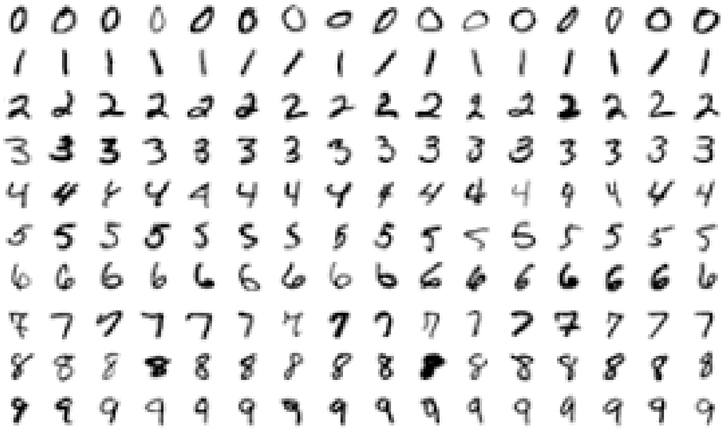
There are many datasets available that are suitable for image processing tasks, and the most suitable dataset will depend on the specific task and characteristics of the data. Some popular datasets for image processing tasks include:

2.1. **ImageNet:** ImageNet is a large dataset of over 14 million images that has been widely used for image classification and other image processing tasks. It contains images of various objects and scenes from a wide range of categories, and it has been instrumental in the development of many successful machine learning models for image processing tasks.

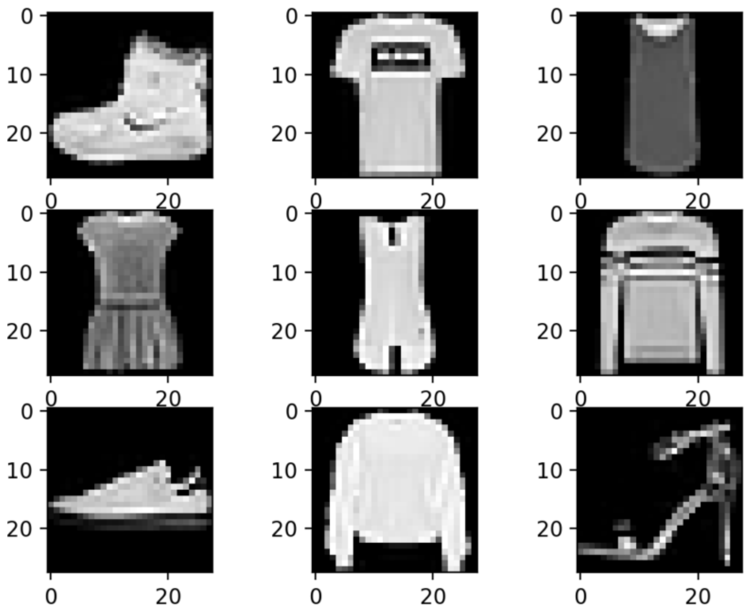
2.2. **COCO (Common Objects in Context):** COCO is a dataset of over 200,000 images that contains a wide variety of objects and scenes, along with rich annotations for object detection, segmentation, and other tasks. It is often used in conjunction with the ImageNet dataset for image processing tasks.



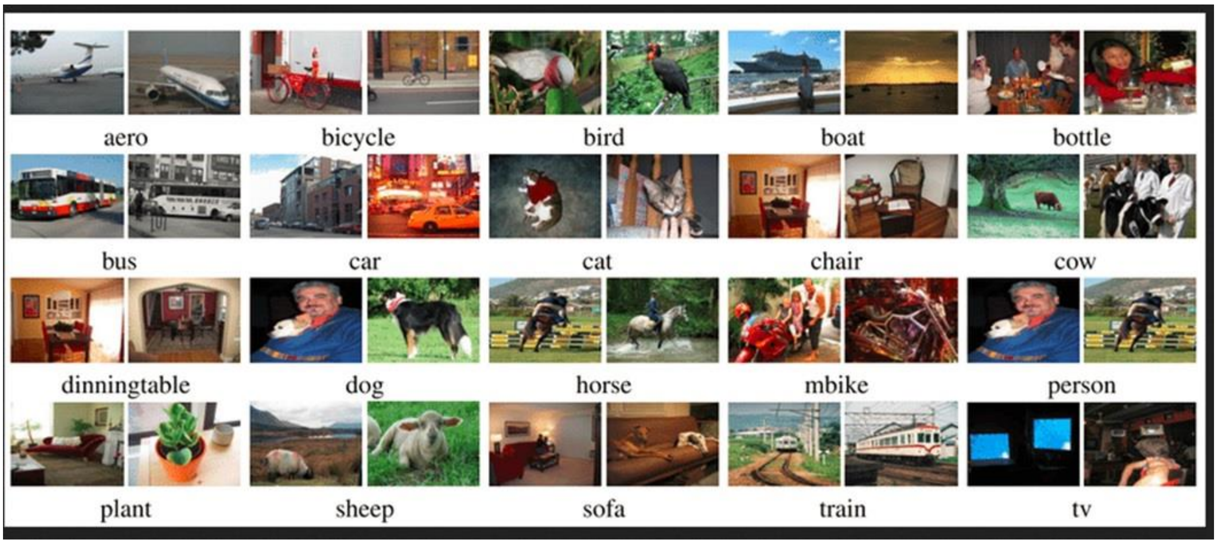
2.3. MNIST: The MNIST dataset is a dataset of handwritten digits that is commonly used for image classification and other image processing tasks. It contains 60,000 training images and 10,000 test images of handwritten digits, and it is often used as a benchmark for evaluating the performance of machine learning models on image classification tasks.



2.4. FashionMNIST: FashionMNIST is a dataset of images of clothing and accessories that is similar to the MNIST dataset. It is often used as a benchmark for evaluating the performance of machine learning models on image classification tasks, and it is a good choice for tasks that involve clothing or fashion-related images.



2.5. **PASCAL VOC:** The PASCAL VOC (Visual Object Classes) dataset is a dataset of images that contains rich annotations for object detection, segmentation, and other tasks. It is often used as a benchmark for evaluating the performance of machine learning models on image processing tasks.



In addition, the VQA dataset, which is the dataset that will make our work easier and contains the questions and images, is very important. Let's talk about the VQA dataset now.

2.6. **The VQA Dataset:** Typically consists of a large number of images, each of which is accompanied by a set of questions and answers. The questions and answers are typically provided in natural language, and the answers may be free-form text or a selection from a predefined set of answers. The VQA dataset is often used to evaluate the performance of machine learning models that are designed to understand and reason about visual information. It is a challenging task that requires a model to have a deep understanding of both language and image content. There are several versions of the VQA dataset available, including the original VQA dataset and the VQA v2 dataset.

Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

3. STAGES FOR TRAINING MODEL

To train a machine learning model using a dataset, there are generally several steps that you will need to follow:

1.Preprocessing: Before training a model, it is often necessary to preprocess the data to ensure that it is in a suitable format. This may involve tasks such as cleaning and formatting the data, normalizing numerical values, or encoding categorical values.

2.Splitting the data: It is generally a good practice to split the dataset into a training set, a validation set, and a test set. The training set is used to train the model, the validation set is used to evaluate the model during training, and the test set is used to evaluate the model after training.

3.Choosing a model and training algorithm: Next, you will need to choose a machine learning model and an algorithm to train it. There are many different types of models and algorithms available, and the choice will depend on the specific task and characteristics of the data.

4.Training the model: Once you have chosen a model and training algorithm, you can begin training the model using the training set. This usually involves feeding the training data to the model and adjusting the model's parameters based on the errors made during training.

5.Evaluating the model: After training the model, it is important to evaluate its performance on the validation set and test set. This will give you an idea of how well the model generalizes to new data and can help you identify any overfitting or underfitting issues.

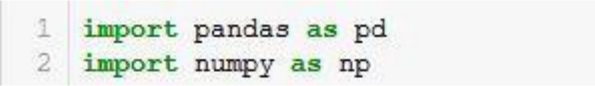
6.Fine-tuning and optimizing the model: If the model's performance is not satisfactory, you may need to fine-tune the model or try different techniques to improve its performance. This may involve adjusting the model's architecture, changing the training algorithm, or using techniques such as regularization or early stopping

3.1. Preprocessing

Data Preprocessing techniques:

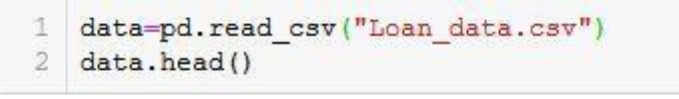
1.Importing the libraries:

At first, all we need is to import all the necessary python libraries which are being used in all the steps. It is important to import *pandas* as it helps to read files of different formats(csv, excel, json etc).Pandas provide variety of options to manipulate data and analyzing it. Similarly, *numpy* is also very helpful to perform mathematical operations on our data. All other libraries can also be imported when required.



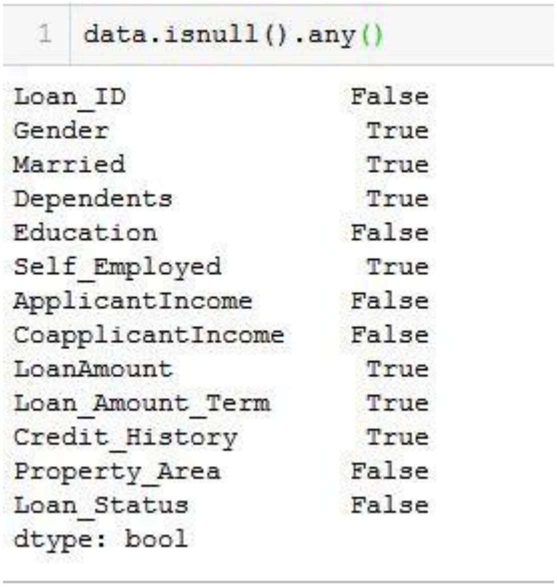
```
import pandas as pd
import numpy as np
```

2.Reading the dataset: Pandas helps to read files of different formats into a pandas data frame: this is a constructor which creates a 2 dimensional table (rows and columns), mutable and can contain heterogeneous data. Here I’m taking a Loan dataset(make sure the data file is in the same path as the working notebook). Doing data.head() will display top 5 rows(starting from index 0).



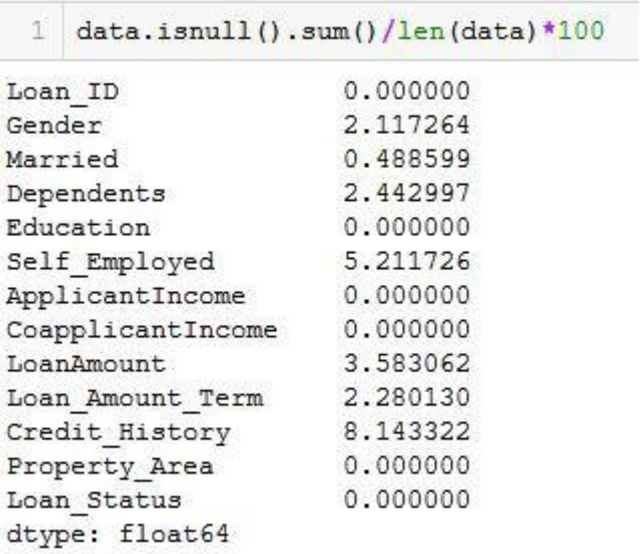
```
data=pd.read_csv("Loan_data.csv")
data.head()
```

Checking For Missing Values: Once the data is loaded, the first thing is to check for null values or any missing values. All machine learning algorithm fails to work with datasets having null values. Below snippet tells which columns contains missing values.



```
data.isnull().any()
```

```
Loan_ID              False
Gender                True
Married               True
Dependents            True
Education            False
Self_Employed         True
ApplicantIncome      False
CoapplicantIncome    False
LoanAmount            True
Loan_Amount_Term      True
Credit_History        True
Property_Area        False
Loan_Status          False
dtype: bool
```



```
data.isnull().sum()/len(data)*100
```

```
Loan_ID              0.000000
Gender               2.117264
Married              0.488599
Dependents           2.442997
Education            0.000000
Self_Employed        5.211726
ApplicantIncome      0.000000
CoapplicantIncome    0.000000
LoanAmount           3.583062
Loan_Amount_Term     2.280130
Credit_History       8.143322
Property_Area        0.000000
Loan_Status          0.000000
dtype: float64
```

3.2. Chosing Training Algorithm And A Model

3.2.1. Models

The most suitable model for image processing tasks will depend on the specific task and characteristics of the data. Some popular models for image processing tasks include:

Convolutional neural networks (CNNs): CNNs are a type of neural network that are particularly well-suited for image processing tasks. They are composed of multiple layers of filters that operate on the input image, and they are able to learn features at multiple scales and levels of abstraction. CNNs are commonly used for tasks such as image classification, object detection, and image segmentation.

Fully convolutional networks (FCNs): FCNs are a type of neural network that are designed specifically for image segmentation tasks. They are similar to CNNs, but they are fully convolutional, which means that they are able to process images of any size and output a segmentation map of the same size. FCNs are often used for tasks such as image segmentation and medical image analysis.

Transfer learning: Transfer learning is a technique that allows you to use a pre-trained model as a starting point for training a model on a new task. This can be a useful approach if you have a small dataset or if you want to leverage the knowledge learned by a pre-trained model. Transfer learning can be applied to a variety of image processing tasks, and it is often used in conjunction with CNNs or FCNs.

Autoencoders: Autoencoders are a type of neural network that are used for unsupervised learning tasks, such as dimensionality reduction and feature learning. They can be used for image processing tasks, such as image denoising and image generation, and they are often used as a preprocessing step for other machine learning tasks.

Overall, the most suitable model for image processing tasks will depend on the specific task and characteristics of the data, and it may be necessary to try several different models and approaches to find the best model for your specific needs.

4. TECHNOLOGY

4.1. KAGGLE

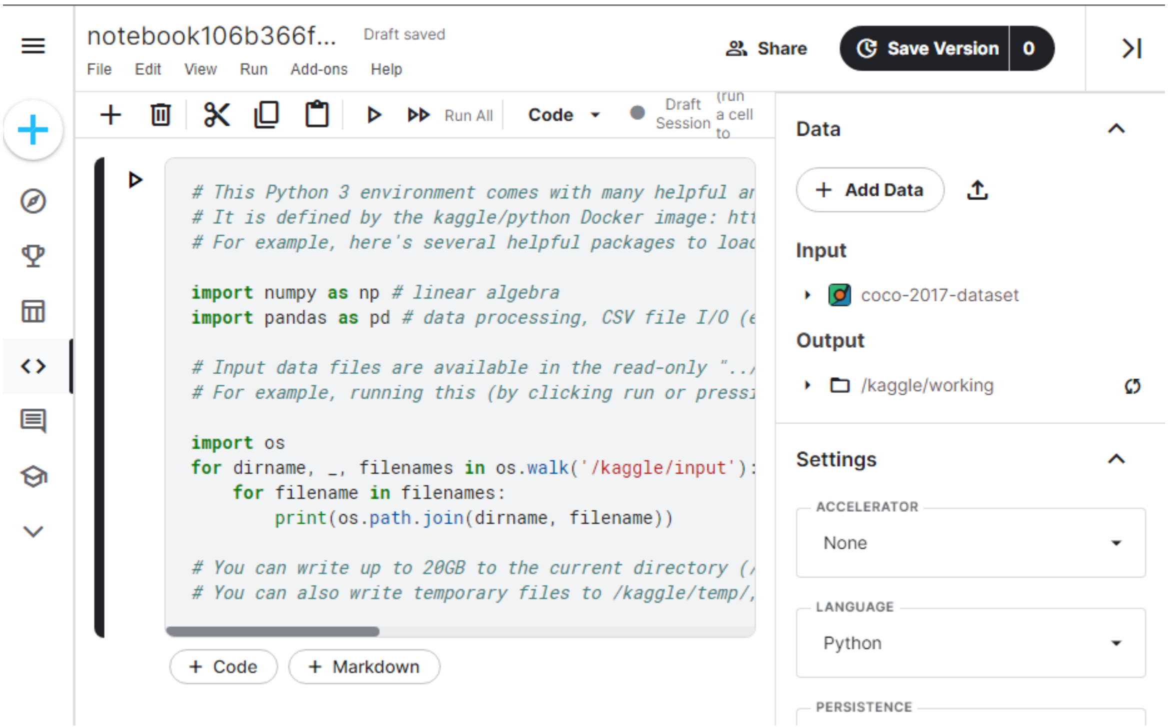
Kaggle is a platform for data science and machine learning competitions, as well as a resource for finding and sharing datasets and tools for data analysis. It was founded in 2010 and is currently owned by Google.Kaggle offers a number of resources and tools for data scientists, including:

Data: Kaggle hosts a large number of datasets on a variety of topics, which can be used for data analysis, machine learning, and other tasks.

Notebooks: Kaggle provides a cloud-based Jupyter notebook environment that allows users to develop and share code, documentation, and visualizations.

Competitions: Kaggle hosts a number of data science and machine learning competitions, in which participants compete to build the best models for a specific task. These competitions often provide prize money and other incentives for the top performers.

Learning resources: Kaggle provides a range of learning resources, including online courses, tutorials, and blogs, that can help users learn about data science and machine learning.Kaggle is a popular platform among data scientists and machine learning professionals, and it has been used to host a number of high-profile competitions, including the Netflix Prize and the Heritage Health Prize.



```
# This Python 3 environment comes with many helpful ar
# It is defined by the kaggle/python Docker image: htt
# For example, here's several helpful packages to loac

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e

# Input data files are available in the read-only "../
# For example, running this (by clicking run or pressi

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/
# You can also write temporary files to /kaggle/temp/,
```

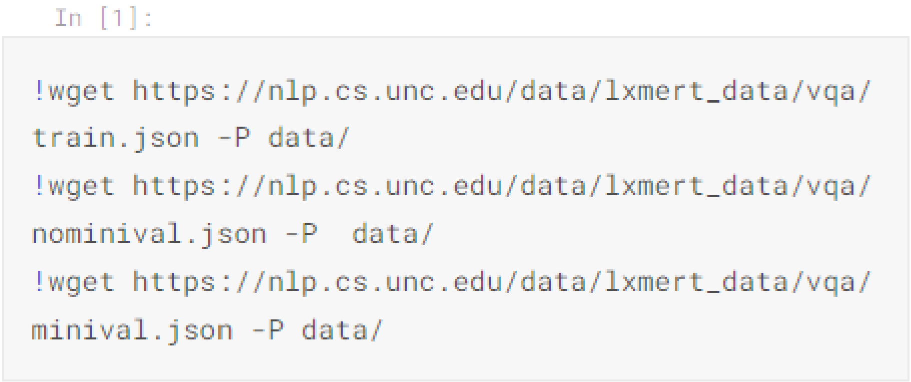
ACCELERATOR: None

LANGUAGE: Python

PERSISTENCE

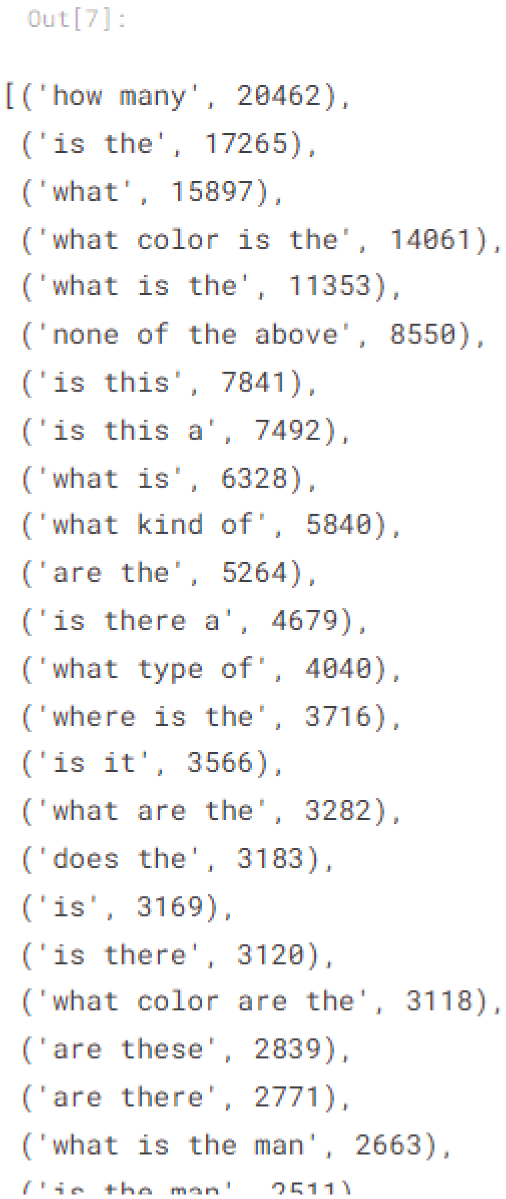
Figure Kaggle Environment

After reviewing multiple datasets, I decided to use the COCO dataset. But since there are no questions to be directed to the images in the COCO dataset, I had to use a question dataset separately. To use this question dataset, I had to download it from the web address. I used this code block for this:



```
In [1]:
!wget https://nlp.cs.unc.edu/data/lxmert_data/vqa/
train.json -P data/
!wget https://nlp.cs.unc.edu/data/lxmert_data/vqa/
nominival.json -P  data/
!wget https://nlp.cs.unc.edu/data/lxmert_data/vqa/
minival.json -P data/
```

Total validation questions like this:



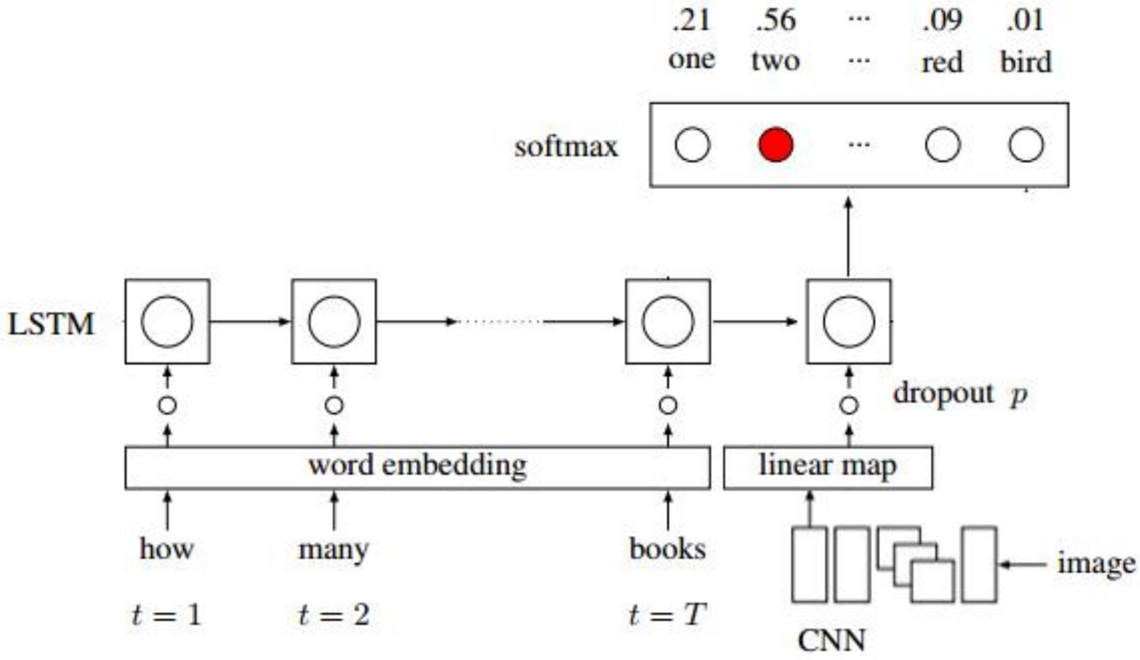
```
Out[7]:

[('how many', 20462),
 ('is the', 17265),
 ('what', 15897),
 ('what color is the', 14061),
 ('what is the', 11353),
 ('none of the above', 8550),
 ('is this', 7841),
 ('is this a', 7492),
 ('what is', 6328),
 ('what kind of', 5840),
 ('are the', 5264),
 ('is there a', 4679),
 ('what type of', 4040),
 ('where is the', 3716),
 ('is it', 3566),
 ('what are the', 3282),
 ('does the', 3183),
 ('is', 3169),
 ('is there', 3120),
 ('what color are the', 3118),
 ('are these', 2839),
 ('are there', 2771),
 ('what is the man', 2663),
 ('is the man', 2511)
```

4.2. Tensorflow

TensorFlow is an open-source software library for machine learning and artificial intelligence. It was developed by Google and is widely used for training deep learning models, implementing machine learning algorithms, and performing other tasks related to artificial intelligence and machine learning.

One of the key features of TensorFlow is its ability to create and execute computational graphs, which are structures that represent the flow of data through a series of operations. TensorFlow uses data flow graphs to represent mathematical computations as a series of interconnected nodes, where each node represents a mathematical operation and the edges between nodes represent the data being passed between them. This allows TensorFlow to perform computations efficiently, using parallelism and hardware acceleration where available.

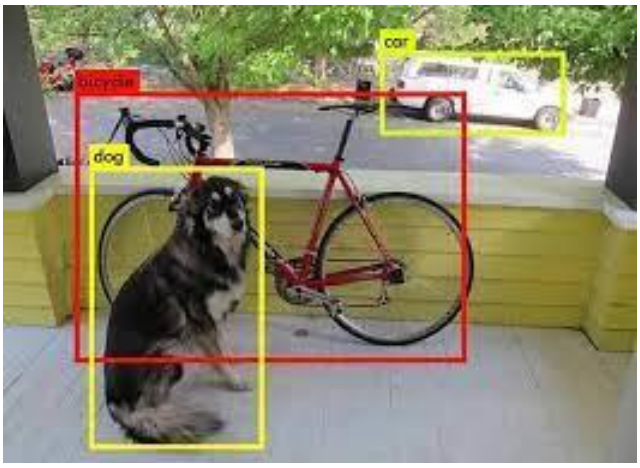
TensorFlow is a powerful tool for developing and training machine learning models, and it has been used to build and deploy numerous successful applications in a wide range of fields, including natural language processing, computer vision, and robotics. It is available for use on a variety of platforms, including desktop, mobile, and cloud, and it can be used with a variety of programming languages, including Python, C++, and JavaScript.



TensorFlow is a popular choice for building and training machine learning models for image processing tasks. Some common image processing tasks that can be implemented using TensorFlow include:

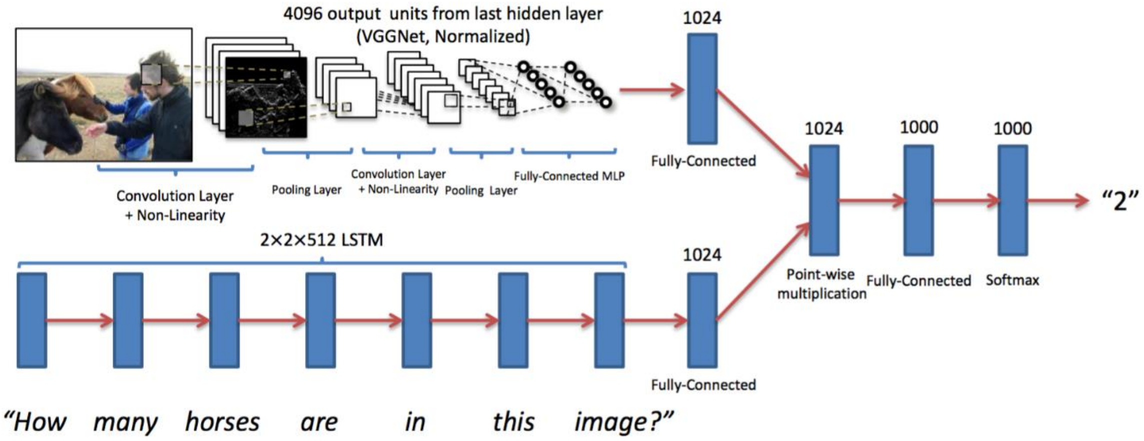
1. Image classification: This involves identifying the class or label of an input image. For example, a model trained for image classification might be able to recognize and classify images of different types of animals.

2. Object detection: This involves identifying and localizing objects in an image. For example, a model trained for object detection might be able to identify and draw a bounding box around objects in an image, such as cars or pedestrians.



3. Image segmentation: This involves dividing an image into different regions or segments, each of which corresponds to a different object or background. For example, a model trained for image segmentation might be able to separate the foreground and background in an image or identify different parts of an object.

TensorFlow provides a number of tools and libraries that can be used to build and train models for image processing tasks, including the Keras API, which provides a high-level interface for creating and training deep learning models. There are also many pre-trained models available that can be fine-tuned or used as a starting point for developing custom models.



5. CONCLUSION

I think making a system that behaves like a human and responds to the questions asked of the images can be done by using artificial intelligence or deep learning techniques. Respectively, I take a step towards the AI-complete task of Visual Question Answering. Specifically, I tackle the problem of answering binary questions about images. I balance the existing abstract binary VQA dataset by augmenting the dataset with complementary scenes, so that nearly all questions in the balanced dataset have an answer “yes” for one scene and an answer “no” for another closely related scene. For an approach to perform well on this balanced dataset, it must understand the image. I will make our balanced dataset publicly available. I propose an approach that extracts a concise summary of the question in a tuple form, identifies the region in the scene it should focus on, and verifies the existence of the visual concept described in the question tuple to answer the question. Our approach outperforms the language prior baseline and a state-of-the-art VQA approach by a large margin on the balanced dataset. We also present qualitative results showing that our approach attends to relevant parts of the scene in order to answer the question.

6. BIBLIOGRAPHY

1. https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn
2. https://blog.paperspace.com/introduction-to-visual-question-answering/
3. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53
4. https://ai.aioz.io/research/vqa-intro/
5. https://blog.paperspace.com/introduction-to-visual-question-answering/
6. https://paperswithcode.com/dataset/clevr
7. https://visualqa.org/download.html
8. https://www.kaggle.com/code/manwithaflower/coco-vqa-questions-eda/notebook
9. https://chat.openai.com/chat
10. https://www.anaconda.com/products/distribution
11. https://arxiv.org/pdf/1612.00837.pdf