**Artificial Intelligence System for Humans**

**Capstone Project Report**

**Submitted by:**

**(101403022) Ajay Kumar Chhimpa**

**(101403023) Akash Gupta**

**(101403024) Akash Kumar Sikarwar**

**(101583005) Ayush Garg**

**BE Third Year, CSE**

**Lab Group: COE1, Project Team No. \_\_\_\_\_**

Under the Mentorship of

**Dr. Sanmeet Kaur**

**Assistant professor, CSED, Thapar University**

****

**Computer Science and Engineering Department**

**Thapar University, Patiala**

**May & 2017**

**Introduction**

**Aim**

This aim of this project is to develop a Digital assistant that can generate descriptive captions for images using neural language models. A Digital assistant help the user to provide answer to his questions which would be given in speech form as a command.

Intended audience

This project can act as vision for the visually impaired people, as it can identify nearby objects through the camera and give the output in audio form. The app provides a highly interactive platform for the specially abled people

## Project Scope

The goal is to design an android application which covers all the functions of image description and provides an interface of Digital assistant to the user. A Digital assistant help the user to provide answer to his questions which would be given in speech form as a command.

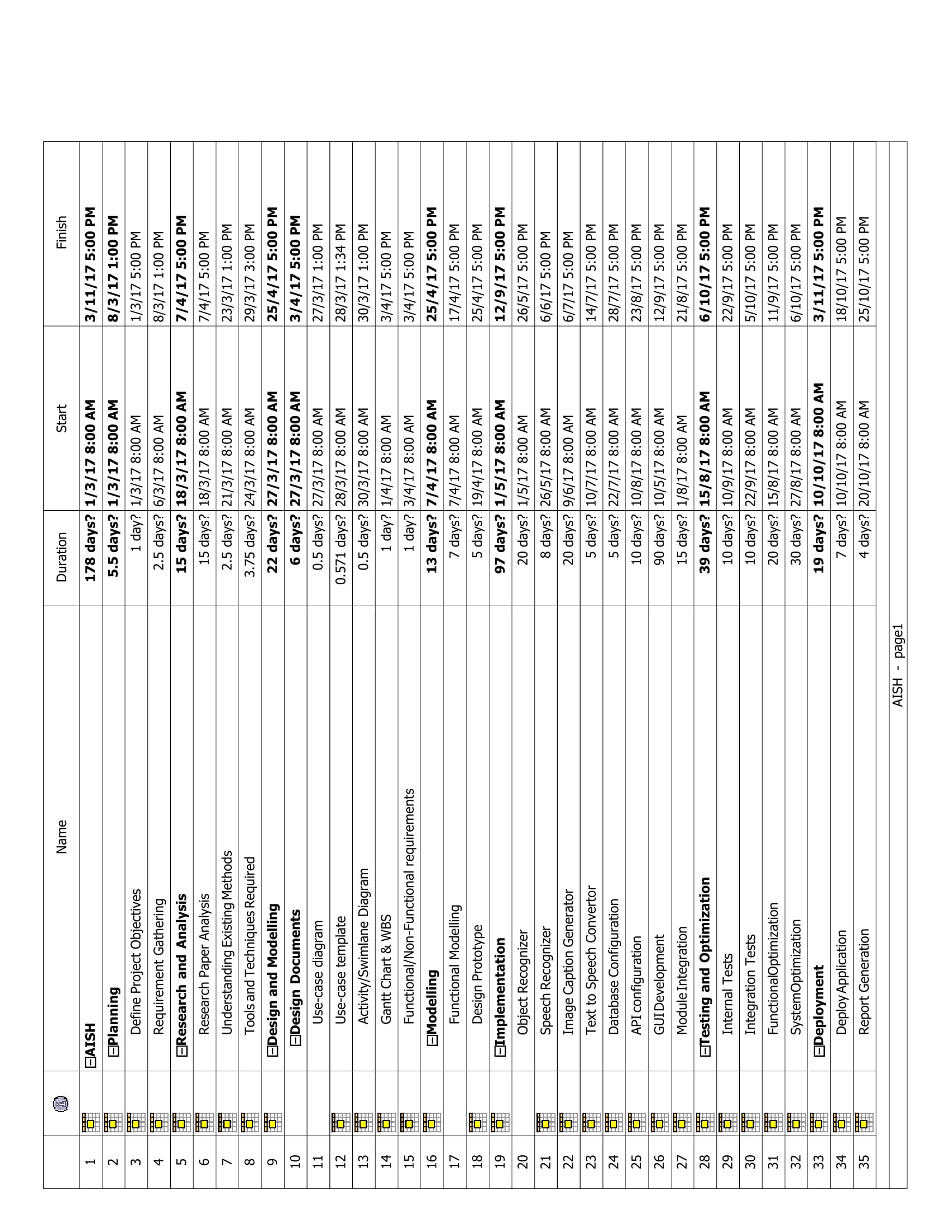
By using Deep learning techniques the project performes:

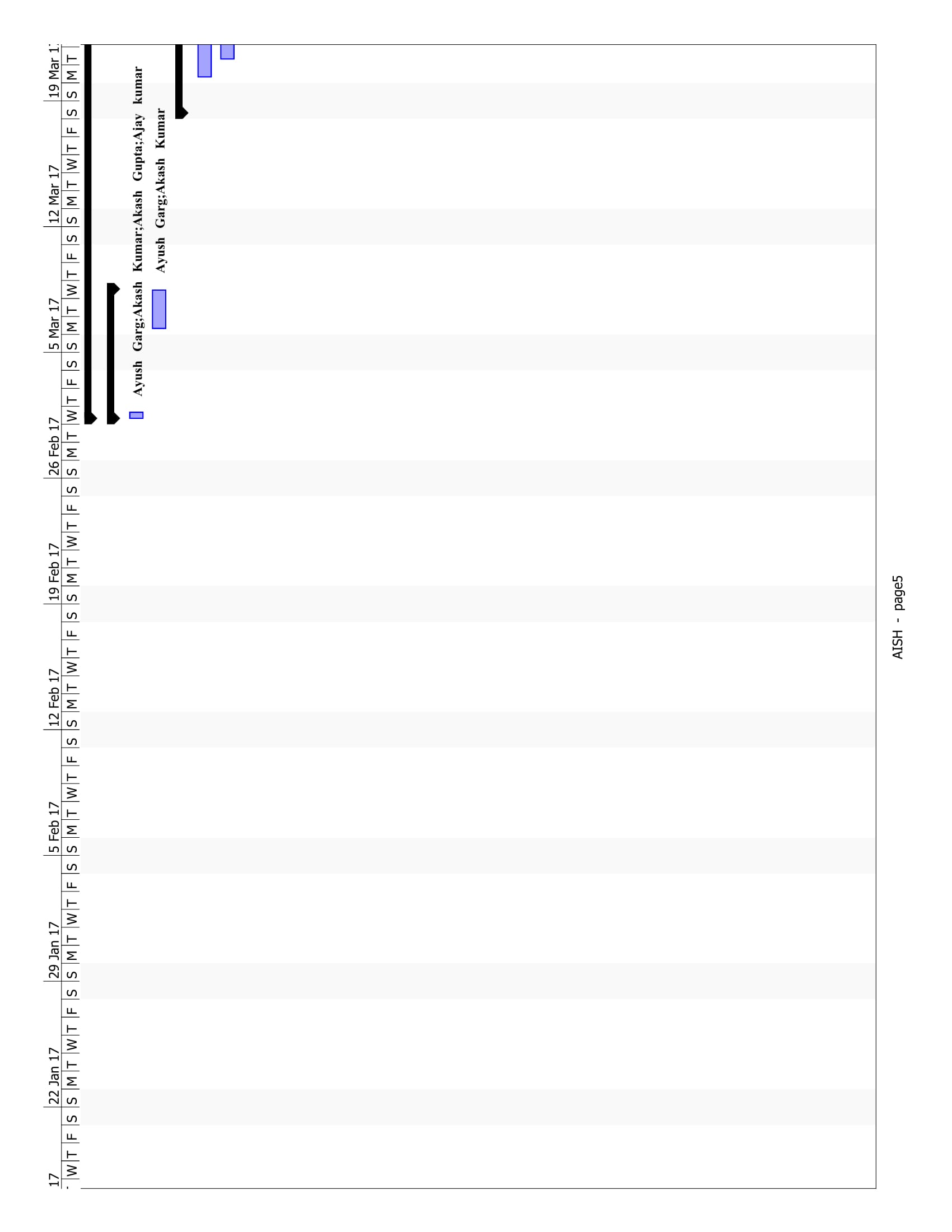
* Image Captioning**:** Recognising Different types of objects in an image and creating a meaningful sentence that describes that image to visually impaired persons.
* Text to speech conversion.
* Speech to text conversion and Identifying result for users query.

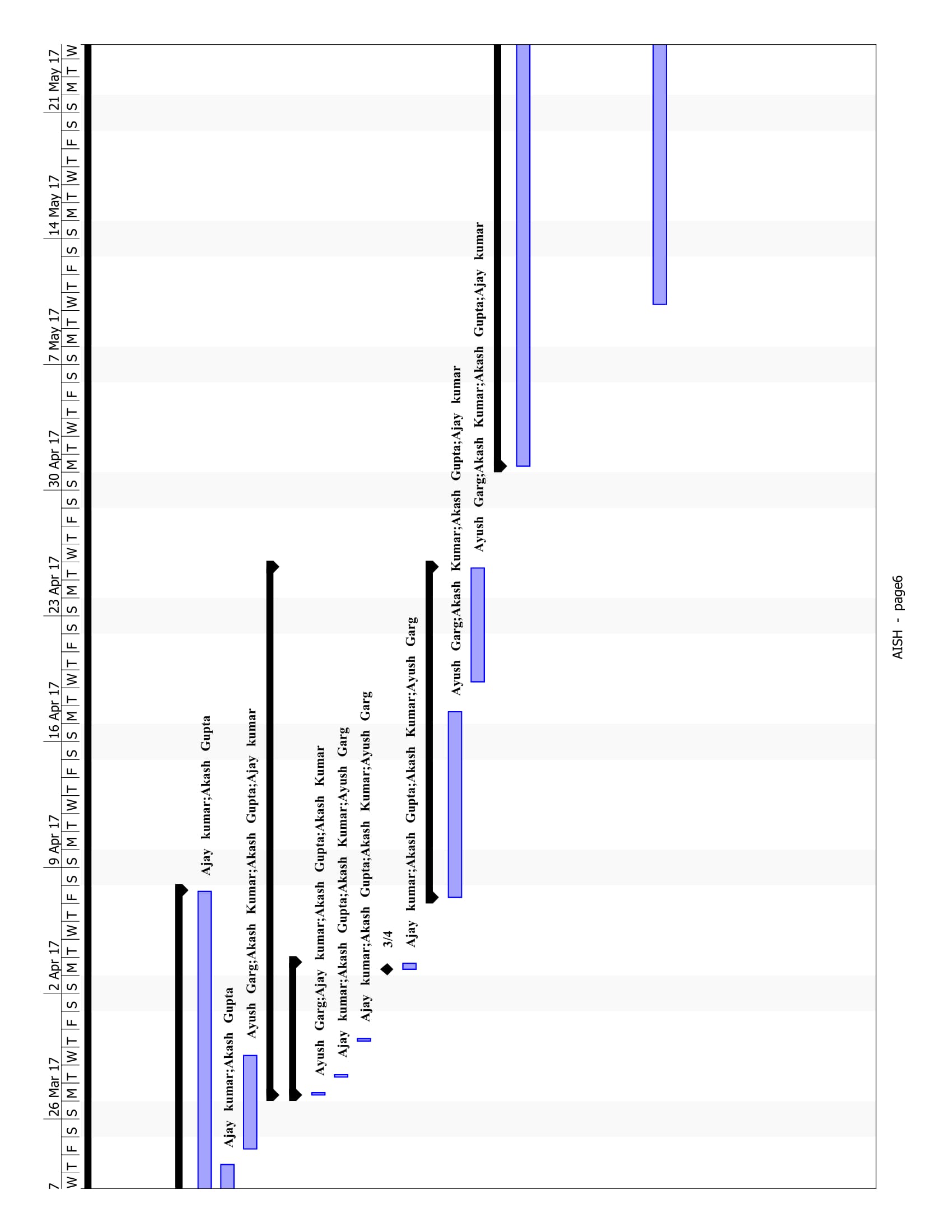
**the approach used in carrying out the project objectives**

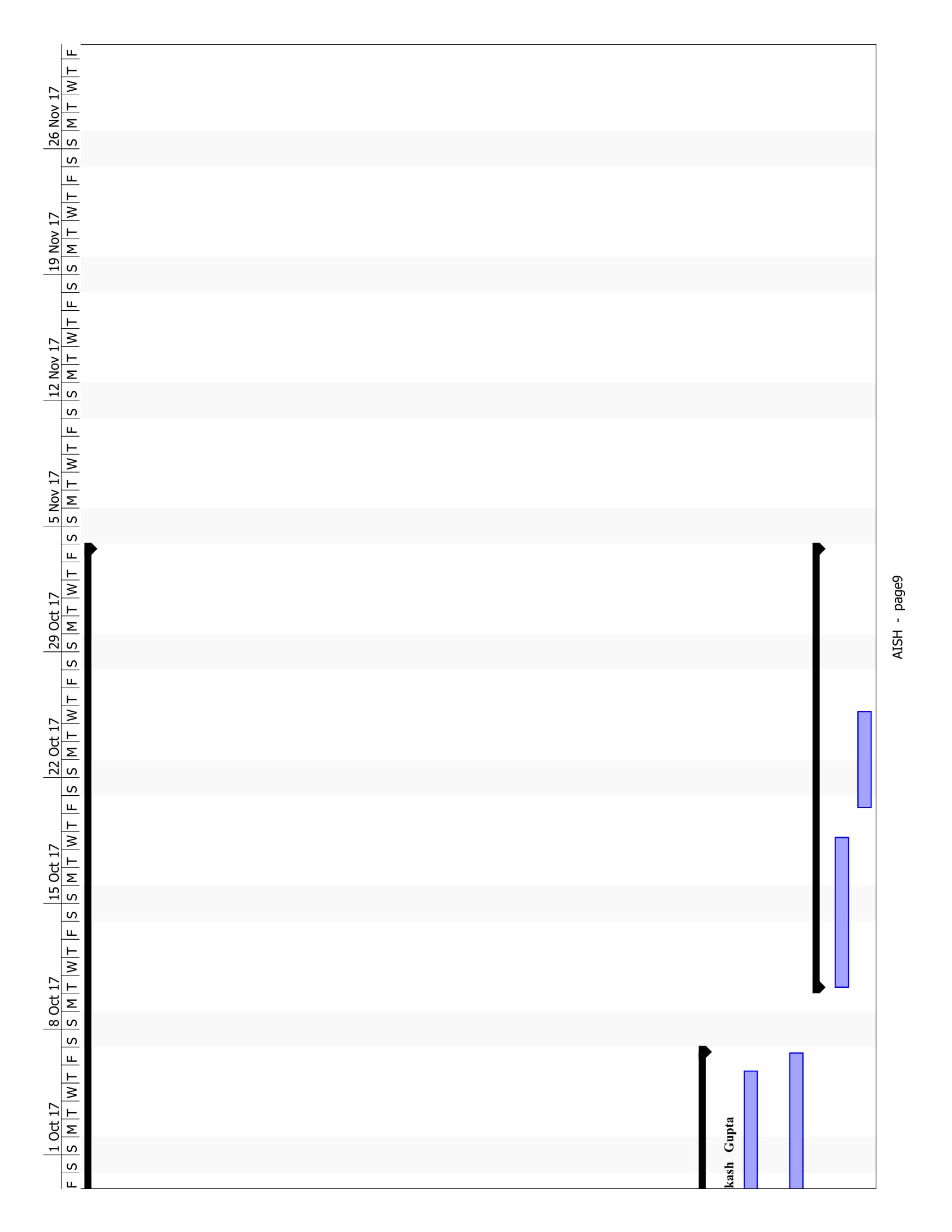
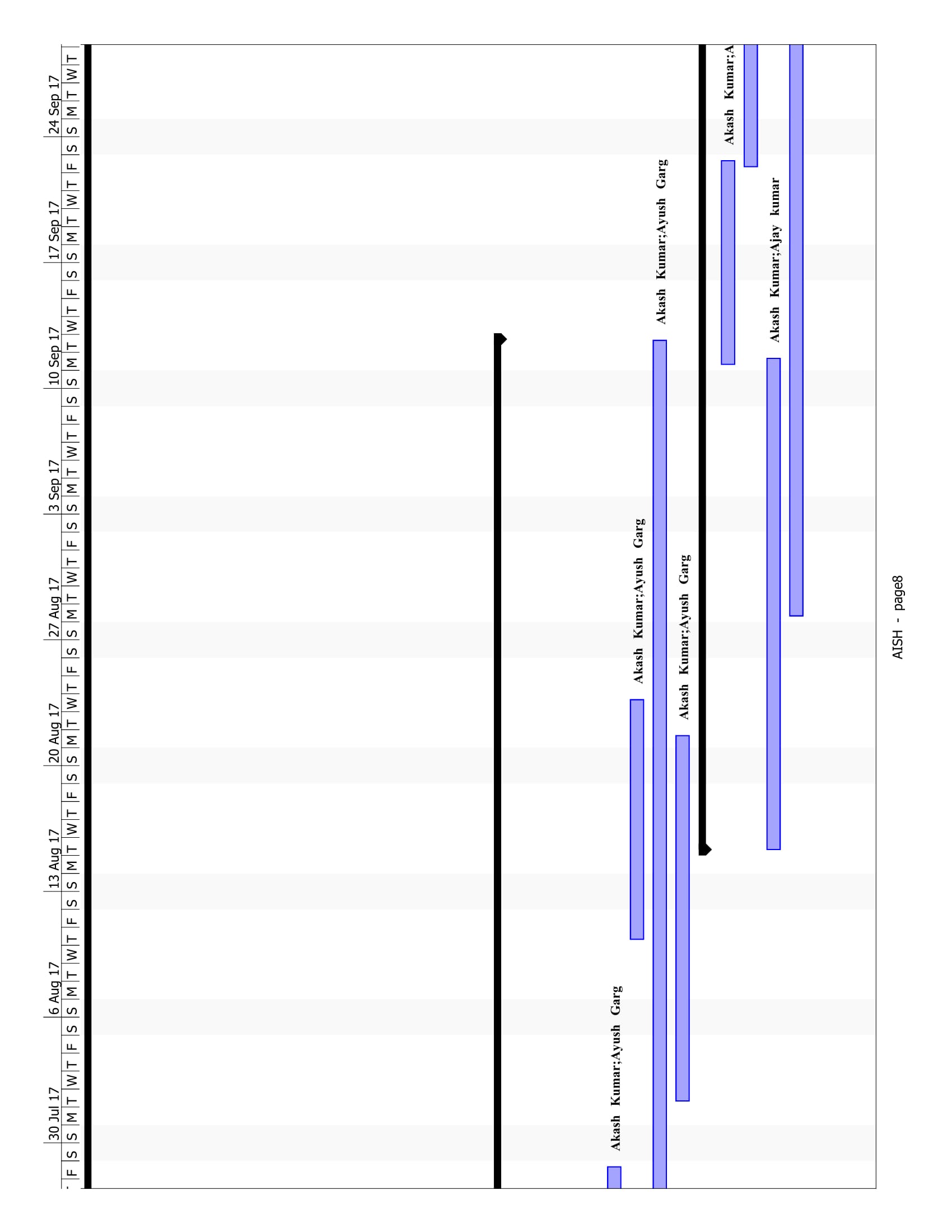
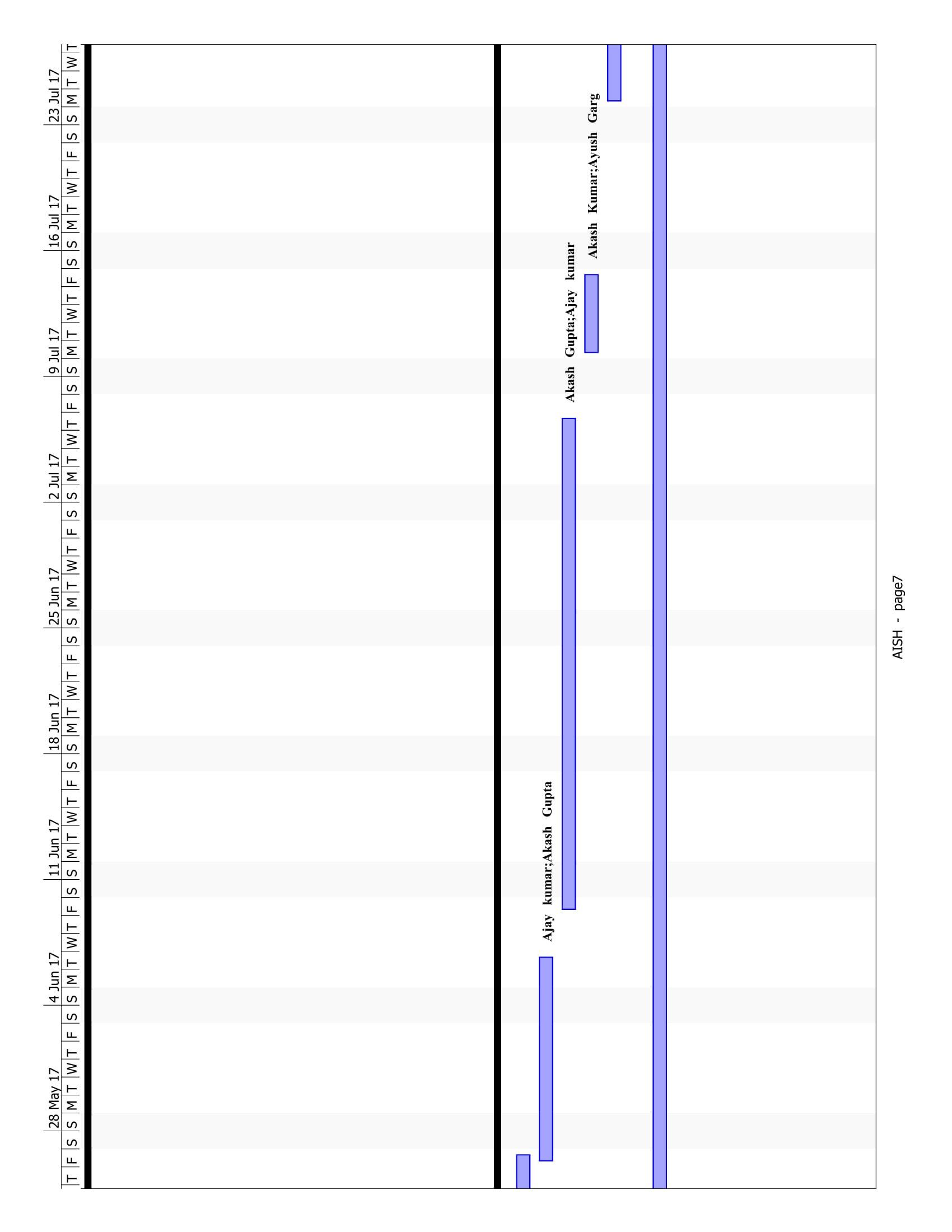
Deep learning is used extensively to recognize images and to generate captions. In particular, Convolutional Neural network is used to recognize objects in an image and a variation of Recurrent Neural network, Long short term memory (LSTM) is used to generate sentences.

**Gantt Chart:**

****





****

**Literature Review**

Generating captions for images is a very intriguing task lying at the intersection of the areas of Computer vision and Natural Language Processing. This task is central to the problem of understanding a scene.

The purpose of this model is to encode the visual information from an image and semantic information from a caption, into a embedding space; this embedding space has the property that vectors that are close to each other are visually or semantically related.  For a batch of images and captions, we can use the model to map them all into this embedding space, compute a distance metric, and for each image and for each caption find its nearest neighbors. If you rank the neighbors by which examples are closest, you have ranked how relevant images and captions are to each other.

**Previous work**

Traditionally, pre-defined templates have been used to generate captions for images. But this approach is very limited because it cannot be used to generate lexically rich captions.

The research in the problem of caption generation has seen a surge since the advancement in training neural networks and the availability of large classification datasets. Most of the related work has been based on training deep recurrent neural networks. The first paper that used neural networks for generating image captions was proposed by Kiros et al. [[6],](#page15) that used Multi-modal log bilinear model that was biased by the features obtained from input image.

Karpathy et.al [[4]](#page15) developed a model that generated text descriptions for images based on labels in the form of a set of sentences and images. They use multi-modal embeddings to align images and text based on a ranking model they proposed. Their model was evaluated on both full frame and region level experiments and it was found that their Multimodal Recurrent Neural Net architecture outperformed retrieval baselines.

In our project we have used a Convolutional Neural Network coupled with an LSTM based architecture. An image is passed on as an input to the CNN, which yields certain annotation vectors. Based on a human vision inspired notion of attention, a context vector is obtained as a function of these annotation vectors, which is then passed as an input to the LSTM.

**Methodology**

**Annotation vector extraction**

We use a pre-trained Convolutional Neural Network [(CNN)[8]](#page15) to extract feature vectors from input images. A CNN is a feed forward type of Artificial Neural Network which unlike fully connected layers, has only a subset of nodes in previous layer connected to a node in the next layer. CNN features have the potential to describe the image. To leverage this potential to natural language, a usual method is to extract sequential information and convert them into language.. In most recent image captioning works, they extract feature map from top layers of CNN, pass them to some form of RNN and then use a softmax to get the score of the words at every step.

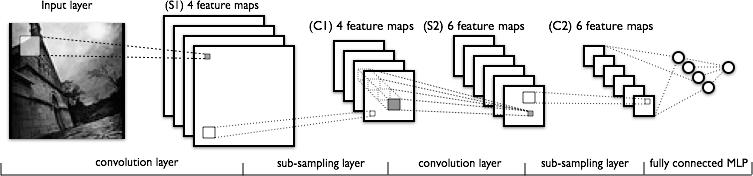
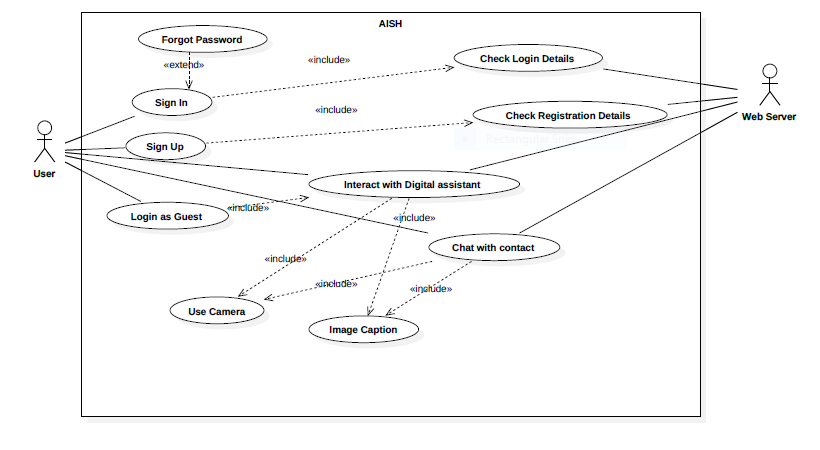


Figure 5 illustrates the process of extraction of feature vectors from an image by a CNN. Given an input image of size 24 24, CNN generates 4 matrices by convolving the image with 4 different filters(one filter over entire image at a time). This yields 4 sub images or feature maps of size 20\*20.These are then subsample to decrease the size of feature maps. These convolution and subsampling procedures are repeated at the subsequent stages. After certain stages these 2 dimensional feature maps are converted to 1 dimensional vector through a fully connected layer. This 1 dimensional vector can then be used for classification or other tasks. In our work, we will be using the feature maps (not the 1 dimensional hidden vector) called annotation vectors for generating context vectors.

For the image-caption relevancy task, recurrent neural networks help accumulate the semantics of a sentence. Strings of sentences are parsed into words, each of which has a GloVe vector representation that can be found in a lookup table. These word vectors are fed into a recurrent neural network sequentially, which captures the notion of semantic meaning over the entire sequence of words via its hidden state. We treat the hidden state after the recurrent net has seen the last word in the sentence as the sentence embedding.

**Requirement Analysis:**

**Use Case Diagram:**



**Use Case Templates:**

**Use Case:** **User Login**

**Id**:  UC- <UC-001>

**Description:**

User enter the username and password for authentication.

**Level:** Low Level

**Primary Actor:**

Application User

**Pre-Conditions:**

* User should be registered.
* User should have entered the username and password.

**Post Conditions:**

**Success end condition:**

Successfully authenticate the user.

**Failure end condition:**

* User’s username may be incorrect.
* User’s password may be incorrect.
* User may not be registered.

**Minimal Guarantee:**

User’s username and password is encrypted.

**Trigger:**

Unauthorized user opens the app.

**Main Success Scenario**

1. Open the app.
2. If not logged in,
   1. then enter username and password.
   2. hit login.

otherwise, user is automatically logged in.

**Frequency:**

Once,unless logged out.

**Use Case: User registration**

**Id**:  UC- <UC-002>

**Description:**

User makes account in the application.

**Level:** Low Level

**Primary Actor:**

Application User

**Pre-Conditions:**

* App is opened and any other user is not logged in.
* The user information is valid in registration form.

**Post Conditions**

**Success end condition:**

Successfully registered the user.

**Failure end condition:**

User does not get registered.

**Minimal Guarantee:**

* Only through valid details the user gets registered.
* Two users can’t register with same username.

**Trigger:**

Unauthorized user opens the app.

**Main Success Scenario**

1. Open the app.
2. If not logged in,
   * 1. Hit Create Account button.
     2. Enter details.
     3. Hit Register.

otherwise, Logout the current user and follow step 2 above.

**Frequency:**

Once,unless another user wants to create account.

**Use Case: Image upload by the user.**

**Id:**  UC- <UC-003>

**Description:**

User selects a particular Image from the Phone Gallery or Clicks the image through Camera.

**Level:** User Goal

**Primary Actor:**

Application User

**Pre-Conditions:**

User must be logged in.

**Post Conditions**

**Success end condition:**

Selected a valid Image file which gets uploaded successfully.

**Failure end condition:**

Selected an invalid file due to which file will not get uploaded.

**Minimal Guarantee:**

The file will only get uploaded if it’s valid.

**Trigger**

User starts the Image Captioning process by clicking the Image Captioning button.

**Main Success Scenario**

1. Open the app.
2. Click Image Captioning Button.
3. Upload the Image successfully.

**Frequency:**

About 10 times per hour.

**Use Case: Speech Recognition.**

**Id:**  UC- <UC-004>

**Description:**

Speech Recognition is an extension to the overall application and part of digital assistant.As the User speaks, his speech gets recognized and it can be utilized for google search and other actions.

**Level:** Sub-Function

**Primary Actor:**

Application User

**Pre-Conditions:**

* User must be logged in.
* Speech Button is selected.

**Post Conditions:**

**Success end condition:**

User speech is recognized.

**Failure end condition:**

* Speech Recognizer does not receive any speech input.
* Speech Recognizer is unable to decode the input due to language limitation.

**Minimal Guarantee:**

The user will be notified of the error.

**Trigger:**

User starts the Speech Recognition process by clicking the Speech Recognition button.

**Main Success Scenario**

1. Open the app.
2. Click Speech Recognition Button.
3. The user Speaks.
4. The speech gets recognized.

**Frequency:**

Once a day.

**Use Case: Caption Receipt.**

**Id:**  UC- <UC-005>

**Description:**

User receives the description of Image uploaded by him.The user can get description in the form of speech or text form.

**Level:** User Goal

**Primary Actor:**

Application User

**Pre-Conditions:**

* Image must be uploaded.
* Captioning algorithm applied.

**Post Conditions:**

**Success end condition:**

The image is described to the user successfully.

**Failure end condition:**

The system is unable to describe image correctly.

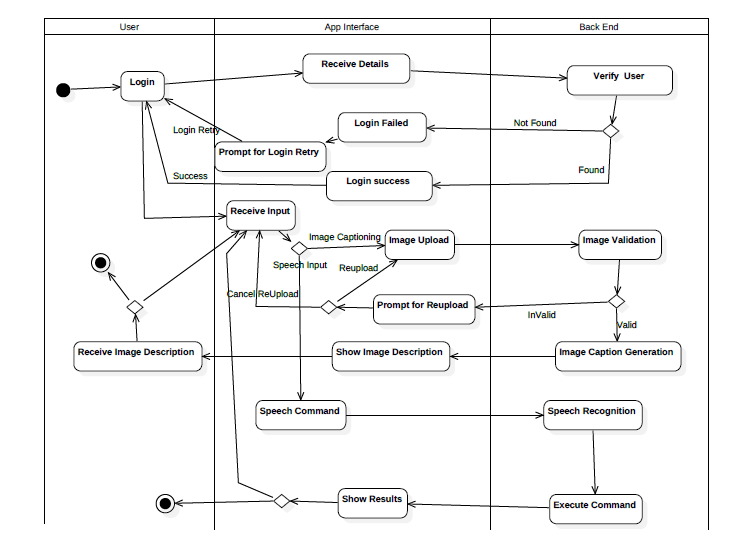
**Trigger:**

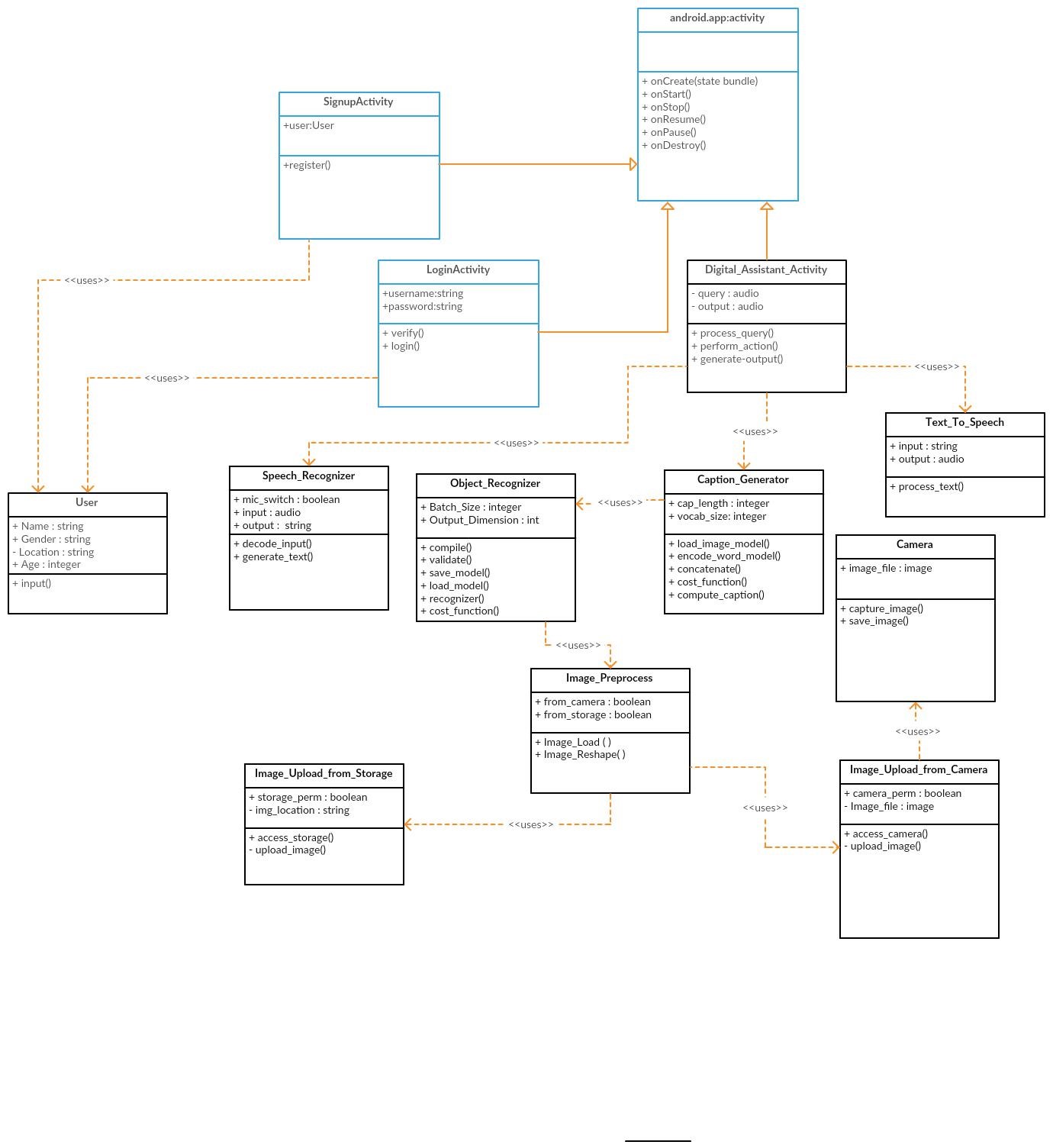
Image upload by the user.

**Main Success Scenario:**

1. Open the app.
2. Click Image Captioning Button.
3. Upload the Image.
4. Image caption is generated.

**Activity Diagram:**

****

Class Diagram 

# Software Requirements Specification:

# Introduction

## Purpose

* Apps in modern world are unamenable towards specially abled humans who have a hard time interacting with them.
* To Design an app, that can describe images in a meaningful way in the form of speech.
* An app that takes input in the form of voice and returns result to the user in the form of voice.

## Project Scope

The goal is to design an android application which covers all the functions of image description and provides an interface of Digital assistant to the user. A Digital assistant help the user to provide answer to his questions which would be given in speech form as a command.

By using Deep learning techniques and Natural language processing project performes:

1. Image Captioning: Recognising Different types of objects in an image and creating a meaningful sentence that describes that image to visually impaired persons.
2. Text to speech conversion.
3. Speech to text conversion and Identifying result for users query.

## References

[https://developer.android.com](https://developer.android.com/)

[mscoco.org/dataset/](http://mscoco.org/dataset/)

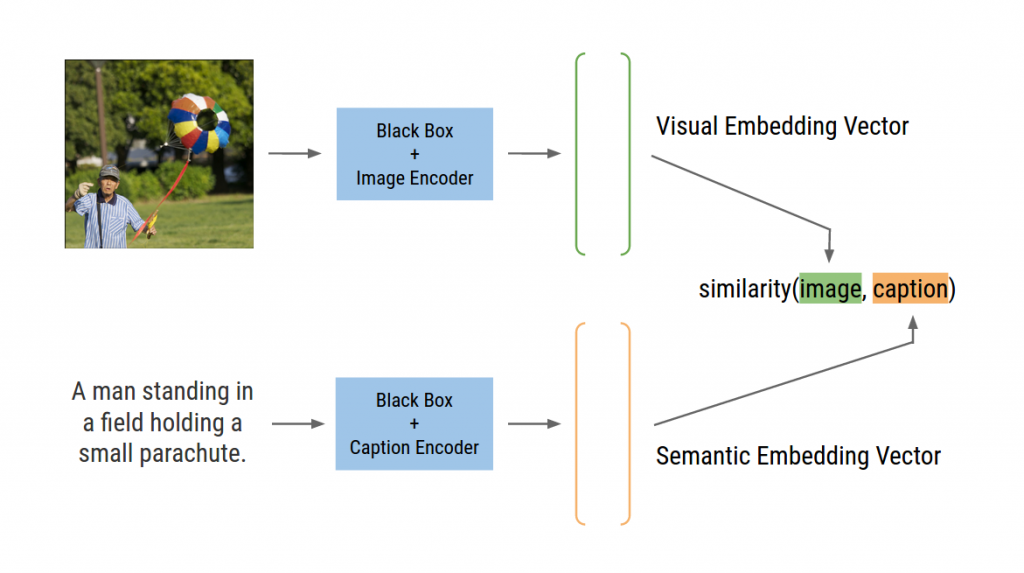
<http://cs.stanford.edu/people/karpathy/deepimagesent/flickr8k.zip>

# **Overall Description**

## Product Perspective

Our project named ”**AISH**” is a self contained project which aims at recognizing the objects in the image and then describing image complete in meaningful way. This project can act as vision for the visually impaired people, as it can identify nearby objects through the camera and give the output in audio form. The app provides a highly interactive platform for the specially abled people. The app implements:

* an image encoder — a linear transformation from the 4096 dimensional image feature vector to a 300 dimensional embedding space
* a caption encoder — a recurrent neural network which takes word vectors as input at each time step, accumulates their collective meaning, and outputs a single semantic embedding by the end of the sentence.
* A cost function which involved computed a similarity metric, which happens to be cosine similarity, between image and caption embeddings



system diagram of the image encoder and caption encoder working to map the data in a visual-semantic embedding space

## Product Features

* Captioning the given image.
* Recognizing input in the form of speech.
* The interface is in the form of a digital assistant which serves the user requests.
* Output generation in the form of speech.

## User Classes and Characteristics

The goal is to design a multi-utility virtual assistant such that any user without any restriction can use it for simplifying their simple daily works and provide them applications for entertainment just by their voice and a simple click. The app specifically targets Visually Impaired people. Brief knowledge regarding Smart phone is required as the app receives input in the form of speech. English language is used for interaction.

## Operating Environment

* Mobile devices with RAM greater than 1GB.
* Android platform
* Sufficient storage space

## Design and Implementation Constraints

* The software has to be integrated onto the user smart phone that in turn has an extremely limited support for machine learning APIs.
* Response time for captioning the given image needs to be reasonable.
* The interface is to be built keeping in mind that it can be easily operated by visually impaired people.

# **System Features**

## User registration<Priority:1>

Description

* Authenticate and Login user to the system.The input can be voice or text.
* New users should be able to register to the system.
* Registered user should be able to change his password if he forgets his password.
* Registered user should be able to update his profile.

Stimulus/Response Sequence

opens the app.

Speak command for register.

Enters the details.

Speak for sign up.

## Speech Recognition<Priority:1>

Speech Recognition is an extension to the overall application and part of digital assistant. As the User speaks, his speech gets recognized and it is utilized as a command for the action.

* When app is opened the speech recognizer will be running in the background for user inputs.

## Image Captioning<Priority:1>

User receives the description of Image uploaded by them. The user can get description in the form of speech or text form.

Stimulus/Response Sequence

* Open the app.
* Click Image Captioning Button or speak the command.
* Click or Upload the Image.
* Image caption is generated.

**Text to speech<Priority:1>**

The text generated after image captioning is described to the user through speech.

**Digital assistant** <Priority:1>

* All functionality is encapsulated inside digital assistant.
* It accepts all commands from the user and generates captions of the images.
* Social media Posts: It grabs the images from the user account and generates captions. The social media sites can be facebook and instagram.
* Additional Features<Priority:2>:

Weather : Reading weather forecast.

News: Reading the recent news for the user.

# **External Interface Requirements**

1. **USER INTERFACES:**

**Login Activity:**

The user interacts with this activity for authentication which is required for storing the user information and captionized content on the server so that the user can access them later.

Login activity requires username and password field.

If the login fails the user is notified by an error message and the error speech.

S**ignup Activity:**

User registers through this activity. The user details like name, email and phone number are requested.User is notified if there is already another user with same username or any other kind of data validation error.

**Tabbed Activity:**

**Tab 1: Captioning**

In this tab the interface for capturing and uploading the image is placed. The user gives voice command for capturing the image and uploading it for captioning or he can press the button and follow the procedure..

The output is a text describing the image in the textview. The text will then be read for the user. In the background the captioning algorithm is implemented.

The tab has relative layout.

**Tab 2: Tools:**

Tools tab contains additional features like -

* loading the images from the user's social media account on facebook or instagram for captioning.
* Reading news or weather report.
* Sending suggestions

The tab has linear list view layout for listing the features.

**Tab 3: Profile**

Profile has textviews and edittext for changing the personal information.

There is a button for logout.

**Social media fragment:**

it has a listview which lists all the images downloaded from social media account with there captions and the items has a button for text to speech conversion for individual images.

The command can also be received through voice for reading the captions.

1. **HARDWARE INTERFACES:**

* Smartphone running Android OS (user)
* MySQL server on Linux
* GET and POST method for communication between Android and MySQL Server.

1. **SOFTWARE INTERFACES:**

**Operating System:** Android

**Language**: Java

**Database:**MySQL database.

**Libraries:**

Keras and Tensorflow for deep learning models.

Retrofit library for communicating with MySQL database.

CloudRail for api integration for multiple social media sites.

# Other Nonfunctional Requirements

## Performance Requirements

**Performance** - The software is designed for the smart phone and cannot run from a standalone desktop PC. The software will support simultaneous user access only if there are multiple terminals. Only voice information will be handled by the software. Amount of information to be handled can vary from user to user.

**Usability** – The software has a simple GUI and easy to use. It has been designed in such a way that visually impaired people can easily use it with minimal problems. The voice commands are particularly helpful for them.

**Reliability** – The reliability of the software entirely depends upon the availability of the server. As long as the server is available, the software will always work without a problem.

**Security:** Credentials of the user are encrypted and application is accessible only to authenticated users.

**Manageability:** Once the image captioning algorithm is devised, no frequent changes will be required. Easily manageable

Appendix A: Glossary

Definitions, abbreviations and acronyms

Table 1 gives explanation of the most commonly used terms in this SRS document.

**Table 1: Definitions for most commonly used terms**

|  |  |  |
| --- | --- | --- |
| **S. no.** | **Term** | **Definition** |
| 1. | Virtual Assistant | An intelligent personal assistant is a software agent that can perform tasks or services for an individual. These tasks or services are based on user input, location awareness, and the ability to access information from a variety of online sources (such as weather or traffic conditions, news, stock prices, user schedules, retail prices and so on.) |
| 2. | Natural language processing | Natural language processing is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. |
| 3. | Socket | Sockets provide the communication mechanism between two computers using TCP. A client program creates a socket on its end of the communication and attempts to connect that socket to a server. When the connection is made, the server creates a socket object on its end of the communication. |
| 4. | Training Set | A training set is a set of data used to discover potentially predictive relationships. A test set is a set of data used to assess the strength and utility of a predictive relationship. Test and training sets are used in intelligent systems, machine learning, genetic programming and statistics. |

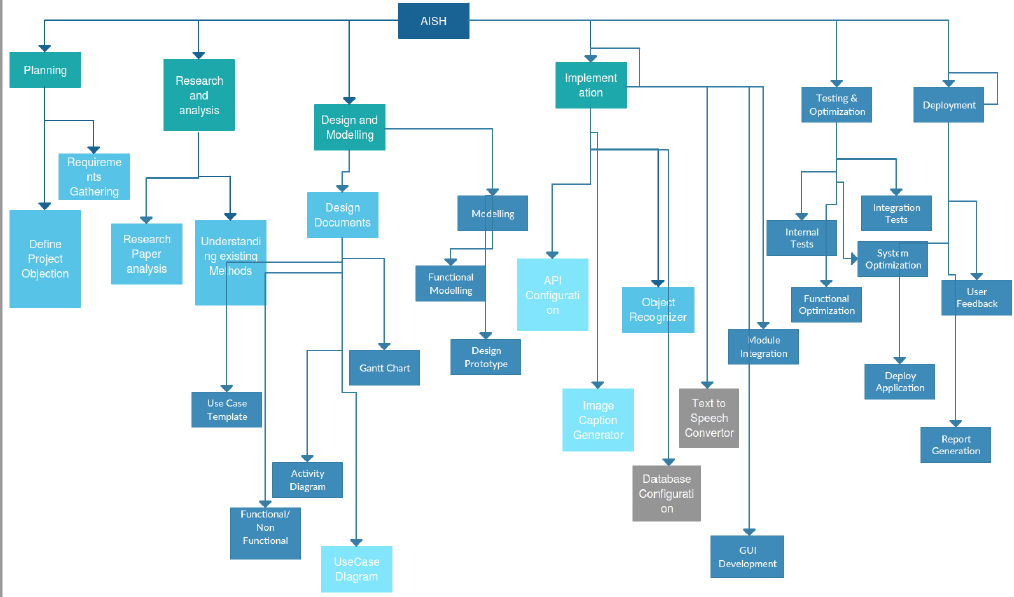
## Abbreviations

Table 2 gives the full form of most commonly used mnemonics in this SRS document.

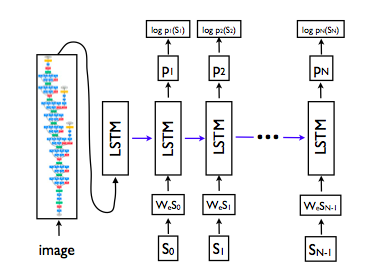
**Table 2: Full form for most commonly used mnemonics**

|  |  |  |
| --- | --- | --- |
| **S no.** | **Mnemonics** | **Definition** |
| 1. | NLP | Natural Language Processing |
| 2. | API | Application Processing Interface |
| 3. | SDK | Software Development Kit |
| 4. | JVM | Java Virtual Machine |
| 5. | NLTK | Natural Language Tool Kit |

**WBS:**

****

**Section 4: Design Specifications**



Flowchart of the proposed system

**References**

1. COLLOBERT, R., WESTON, J., BOTTOU, L., KARLEN, M., KAVUKCUOGLU, K.,

AND KUKSA, P. Natural language processing (almost) from scratch. The Journal of Machine Learning Research 12 (2011), 2493–2537.

1. KARPATHY. The Unreasonable Effectiveness of Recurrent Neural Networks. [http:](http://karpathy.github.io/2015/05/21/rnn-effectiveness/) [//karpathy.github.io/2015/05/21/rnn-effectiveness/](http://karpathy.github.io/2015/05/21/rnn-effectiveness/).
2. KARPATHY, A., AND FEI-FEI, L. Deep visual-semantic alignments for generating image descriptions. arXiv preprint arXiv:1412.2306 (2014).
3. KIROS, R., SALAKHUTDINOV, R., AND ZEMEL, R. Multimodal neural language models. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (2014), T. Jebara and E. P. Xing, Eds., JMLR Workshop and Conference Proceedings, pp. 595–603.
4. SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014).