
VQA – Visual Question Answering

Team# 1 | #5 Jada, Santosh | #11 Mehta, Avni | #14 Muppala, Raji | #19 Patel, Harshil

---

Project Preproposal

- **Project Goal and Objectives**
  - **Motivation**

The motivation for undertaking VQA project was primarily our interest in doing a challenging project in multi-discipline Artificial Intelligence. In this present tech era, with AI research and computer vision on demand, we want to work on images to train the machine, see its analyzing capabilities and improve its accuracy.

  - **Significance / Uniqueness**

Several projects have worked on visual question answering; however, unlike our application, they are fairly restricted with the questions from a fixed set. Our application involves open-ended, free-form questions and answers provided by humans. Additionally, our application will have the capability to transform the question from speech-to-text so that it could be used by visually impaired.

  - **Objective**

The goal of this project is to implement the visual question answering system with our machine giving it a vision to understand the image and providing it visual common sense to analyze. Then it would answer open ended free-form questions of natural language like English. We shall train the deep learning model with thousands of images. Then we extract the important features of these images using Convolutional Networks and make the model identify actions and objects. With this analyzing capability we shall ask the machine, free form questions in natural language and now the machine should have a RNN/LSTM to understand text and answer it. This system could help visually impaired people to understand an image.

  - **System Features**

This application will include a web application that enables the user to either select an image from the available images or upload an image. The user can then ask a question about the selected image. The main purpose of this application would be to provide an accurate natural language answer in real-time.

- **Related Work**
  - **Projects done by others**

There have been several papers that are working on tasks of image tagging like [3, 4], image captioning [5, 6] and text-based Q&A [7, 8]. A team from MIT has worked on 'Simple Baseline for Visual Question Answering' [9].

- **Bibliography**

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. arXiv preprint arXiv:1505.00468, 2015.

[2] http://visualqa.csail.mit.edu/

[3] J. Deng, A. C. Berg, and L. Fei-Fei. Hierarchical Semantic Indexing for Large Scale Image Retrieval. In CVPR, 2011

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classifi- cation with Deep Convolutional Neural Networks. In NIPS, 2012

[5] G. Kulkarni, V. Premraj, S. L. Sagnik Dhar and, Y. Choi, A. C. Berg, and T. L. Berg. Baby Talk: Understanding and Generating Simple Image Descriptions. In CVPR, 2011.

[6] A. Farhadi, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every Picture Tells a Story: Generating Sentences for Images. In ECCV, 2010

[7] A. Fader, L. Zettlemoyer, and O. Etzioni. Open Question Answering over Curated and Extracted Knowledge Bases. In International Conference on Knowledge Discovery and Data Mining, 2014

[8] A. Fader, L. Zettlemoyer, and O. Etzioni. Paraphrase-Driven Learning for Open Question Answering. In ACL, 2013

[9] B. Zhou, Y. Tian, S. Suhkbaatar, A. Szlam, R. Fergus. Simple Baseline for Visual Question Answering. arXiv:1512.02167

Project Plan

❖ **Prioritized Features**

1. We would perform the task of training the machine with a small data set of 10 to 15k images.
2. Firstly, we would train the model on a question which would have yes/no answers.
3. Then we would try for the answers that contains number.
4. Finally, we would try for multiple-choice and open free-form questions.

❖ **Datasets**

We will be using MS COCO dataset obtained from http://mscoco.org/dataset/#download
The following table describes the data:

	Training Set	**Validation Set**	**Testing Set**
Images	82,783	40,504	81,434
Questions	2443,757	214,354	447,793
Answers	4,437,570	2,143,540	-

❖ **Technologies**

The following are the technologies and programming languages which we would use for the project:

- Python (programming language)
- Tensor flow (for deep learning models)
- Pyspark (for spark)
- HTML,CSS (front end)
- Google Cloud Service

❖ **Schedule for 3 Increments**

The following are the tasks we would like to complete with every increment:

1. Increment 1:
   - Find a dataset (labeled/unlabeled data)
   - Preprocess and then set up the environment for development
2. Increment 2:
   - Test already available model.
   - Change the parameters, models and try to get better accuracy.
3. Increment 3:
   - Development of web interface for using the trained model with test data.
   - Service Implementation with testing.

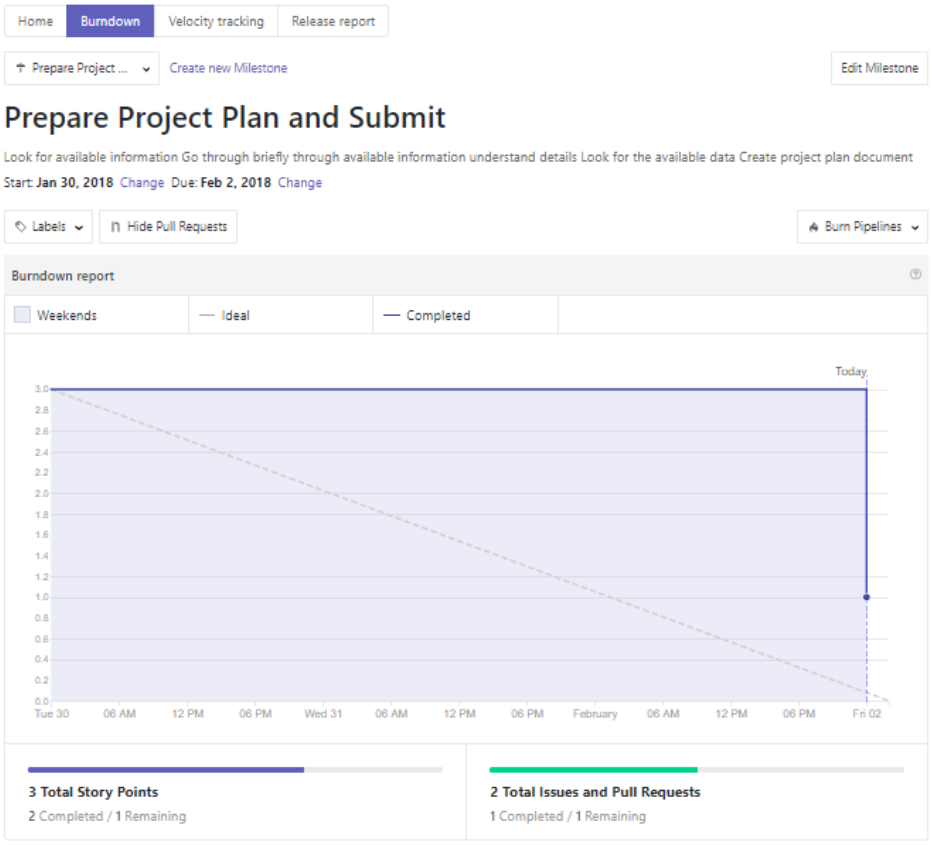
❖ **Project Timelines, Members, Task Responsibilities**

- **Timelines:** Increment 1: 2/23/2018, Increment 2: 3/19/2018, Increment 3: 4/23/2018, Final Project Video/PPT: 4/30/2018, Project Demo: 5/3/2018, Final Project Package: 5/7/2018
- **Team Members:** Santosh Jada, Harshil Patel, (Trinadha) Raji Muppala, Avni Mehta
- **Task Responsibilities:** All team members will have different contributions to the project (25% each)

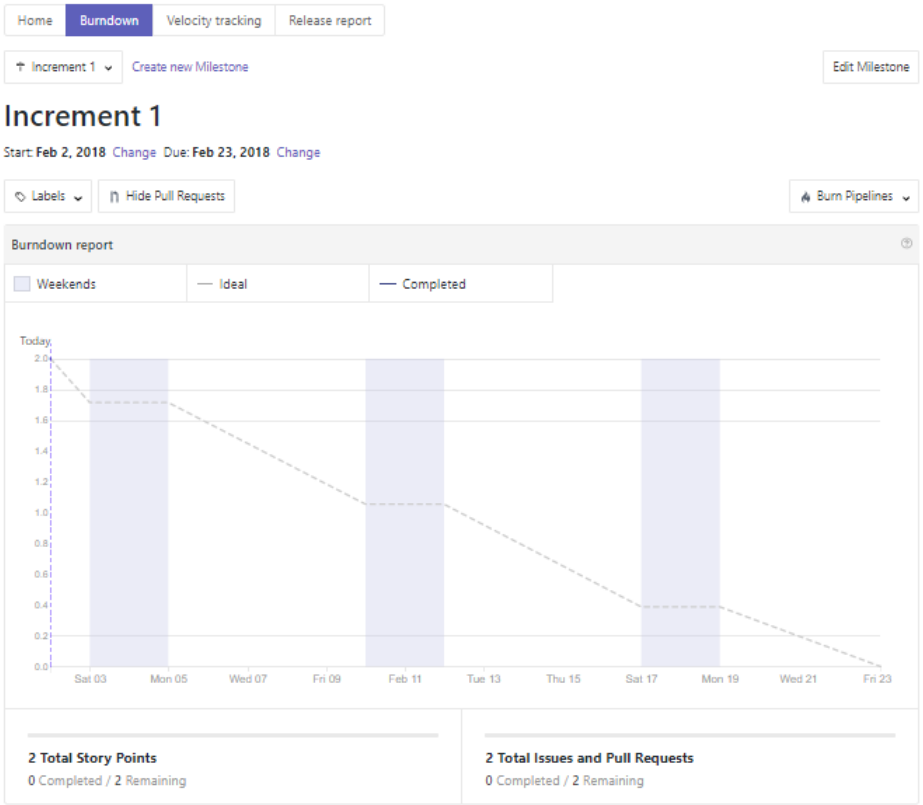
❖ **Burndown Chart**

All the milestones and issues are created in ZenHub for this project.
The following are the screenshots for few of them:

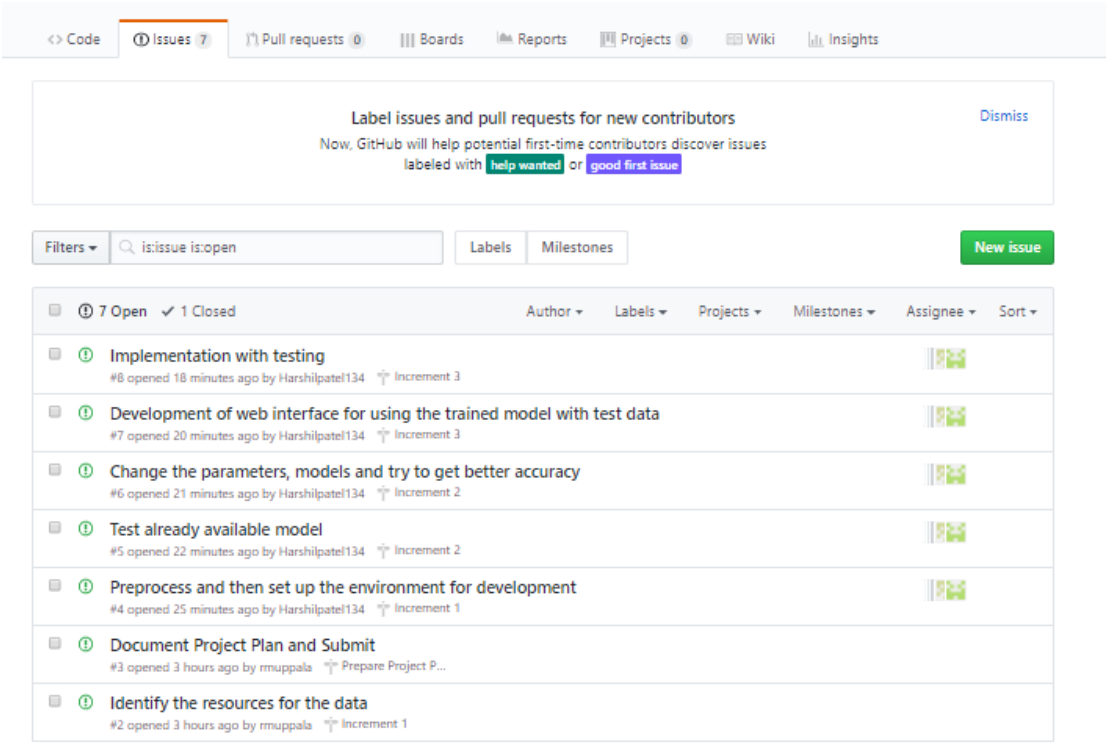
Burndown chart for Project Proposal-



Burndown chart for Increment 1 -



Issues-



Milestones-