Bilkent University

Department of Computer Engineering

CS 491 - Senior Design Project

Project Short-name: Deeplay

Project Specifications Report

Team "*Ludens*"

Project Group Members:

F. Serdar Atalay, Ekinsu Bozdağ, Gökcan Değirmenci,

Onur Sönmez, Gökçe Özkan

Supervisor: Prof. Dr. Halil Altay Güvenir

Jury Members: Shervin R. Arashloo, Uğur Güdükbay

Innovation Expert: Mustafa Sakalsız

Submitted at October 15, 2018

Table of Contents

1. Introduction

In the field of computer science, there are many companies, engineers and data scientists that extensively use deep learning. As the amount of necessary and appropriate data to be used during training increases, the performance increases with deep learning algorithms[1]. With the current advancements on machine learning, especially on its subset area of deep learning, intelligent applications become extremely data-hungry. The need for uncompromised quality, actionable data increases everyday. That is to say, data becomes the “new oil” of our era. People label their data to do training and they need more and more data to get the most accurate results for their systems. Labelling data does not require high level education in the field. Thus, when requirements are well understood, any person with adequate hand-eye coordination can label training data. By bringing crowdsourcing and gamification concepts, our project *Deeplay* promises entertainment to people, while making them help those who are interested in or working on deep learning. *Deeplay* provides a game platform in which there will be different categories of games that mainly aims to make users label data.

1.1 Description

As our *Senior Design Project*, we propose an innovative *Data Labeling Platform* that combines gamification, human-in-the-loop and blockchain concepts. It will be a new, innovative approach to bonding humans and machines: letting machines ask us help for the data they need to improve. The platform will get massive amount of data from the clients(businesses) and in return provides high quality training data as a output.

In *Deeplay platform*, there will be one game at the beginning which will require the user to label different data. The game will be available with 2 options: single-player and multi-players. In any case, the reliability of whether the user labels the data correctly is a very important issue for the game. Therefore, in 1 player mode, the user will be sent some already correctly labeled data without their label, to check the correctness of the user's performance. We are planning to use OpenImages[2] dataset as our initial labeled dataset to solve cold-start problems. If s/he can do it correct for many times, the system will accept that user as trustworthy for the next data, which will be the real unlabelled data similar to Google reCaptcha system. In multiplayer mode, different users will be able to play the same game and label the same data competing with each other. This will help our system to check the correctness of the answers by making comparison with other players' labeled pixels.

The data will be supplied from people who want their data to be labelled. Those people will be able to integrate different games they develop to label some specific data into our platform. Other companies, scientists etc. will also be able to send their unlabelled data to the best game category for themselves.

1.2 Constraints

1.2.1 Economic

- Our system will run on several nodes(servers) and use blockchain for transactions which require small amount of fee.

1.2.2 Implementation

- The platform will consists of different parts such as a cross-platform mobile application(Android, iOS), sophisticated backend subsystem and a Web application for clients that want to upload their data and oversee the ongoing labeling processes.
- For distributed data processing and streaming we will look at open-source frameworks such Apache Spark, Hadoop, MapReduce.

- For auto label predictions, we will examine ML frameworks and libraries such as PyTorch, SparkML, Scikit-Learn.
- The user interface of the web application may built with JavaScript and one of its popular UI libraries such as React.js.
- We are planning to use blockchain technology to create smart contracts and track the game usage. It will create two-sided market platform and monetize the efforts of game developers and labeled data demanders.
- We may store the chunks of data in decentralized storage and use IPFS mechnasicm to download and stream files from not just single-HTTP server, from different nodes.

1.2.3 Security

- We need to securely transfer company’s sensitive data to hosted games. Our distributed system will eliminate any third-party services and provide secure network for the companies. It should be impossible for third-party to trigger data transfer.

1.2.4 Project Timeline

- We will do research about decentralized storage, blockchain and how to solve crowdsourcing problems until November 2018.

1.2.5 Distribution

- Since our platform is a two-sided market platform for both labeled data demanders and game developers. All the marketplace including App Store, Google Play Store and web platform we are providing are potential targets.

1.2.6 Ethical Constraints

- Development of the app will abide to code of Ethics framed by the National Society of Professional Engineers [3] .

- User’s private information will not be shared with third parties.Also through authentication, the actual user’s information will be protected.
- Our application should not provide private information without the approval of companies with third parties.

1.2.7 Sustainability Constraints

- Users will be able to rate product and provide feedbacks about product.
- System allows users to report bugs and enhancement requests .

1.3 Professional and Ethical Issues

The platform builds upon the principle of “Privacy by Design” to assure that data we label, process and use are not affected by privacy issues.

Main purpose of our application is collecting raw data from companies and give it to users, for them to process. For instance a company sends landscape images with birds over and here on the image. We then give this image to users to locate the birds on the picture. Difference between raw and processed data is that one of them has requested information on it. During this process there are two main ethical issues, company’s private information and user’s a.k.a player’s.

1.3.1 Private Data of Clients

Our clients may not always want to share the details of the project they work on, so, the raw data that we obtained from the company may hold great importance about their project. That's why we need a secure environment in our application to protect company data. Also, the user’s processed data should be kept safe until we process the data into a usable form for companies and send it.

1.3.2 Sensitive Data of Users

There is also standard user privacy issues. Such as user’s name, mail, age, occupation etc. Users private information will not be shared with third party applications or companies. The private information will be handled with General Data Protection Regulation (GDPR).

2. Requirements

2.1 Functional Requirements

2.1.1 User Functionality Requirements

- Any user or foundation who wants to label their AI/ML training data may use the platform as an **end-to-end pipeline for their “data labeling” needs**. Our system will decentralize the process and distribute the data across a network of nodes.
- Labeled data demanders may select the game they want to label according to the category of the game.
- Labeled data demanders may use the blockchain to create smart contracts to track the game usage and monitor the number of end-users who labels it within an interactive user interface.
- Labeled data demanders may put a threshold for the number of required labels for each data. Clients may also want more accurate results for their needs and put high threshold level for their data.
- Any user or foundation may post a labeling task that will be asked during the game.
- Any user or foundation may host their data labeling games on the platform.
- Multiple users may play the game in real-time multiplayer game session so that the accuracy and reliability of the labeling would increase.

- End-user may play any game and label random data(e.g. Image segmentation) on the platform. Our system will use human computation for crowdsourcing.

2.1.2 System Functionality Requirements

- The system needs to cross-validate crowdsourced data set of responses according to predefined consensus rules and crowdsourcing algorithms.
- The system will monetize the efforts of the game developers according to the value they created.

2.2 Non-functional Requirements

2.2.1 Usability

- The heart of the platform consists different types of addictive games that made by thinking about the end-user first. So it means that the users should be able to use the platform easily and intuitively.
- Plot and play instructions of the games should be easily understandable.

2.2.2 Performance

- The application should be available to its users 24/7/365 with at most 0.5% overall downtime statistics (99.5% availability).
- Filtration and preprocessing steps of the batch labeling tasks should be fast enough to keep up with the deadlines given by the clients.

2.2.3 Extensibility

- The system should support easy integration for the games. Thus, more games may be integrated into the system in order to increase the number of options for the data labeling. For example, in the future, there might be human-labeled sound clips, or other special-kind of sensitive data which needs to be labeled with the help of gamification.

- The architecture and implementation actively caters to future business needs.

2.2.4 Security

- The system should ensure end-to-end security of sensitive data of users and private information of companies.

2.2.5 Scalability

- The system should be scalable enough to handle the growing amount of work and implement the sharding as a way of distributing data across sets of multiple machines.

2.2.6 Robustness

- The system should be robust. Sudden traffic increase should not be an issue to the system.

3. Existing Systems

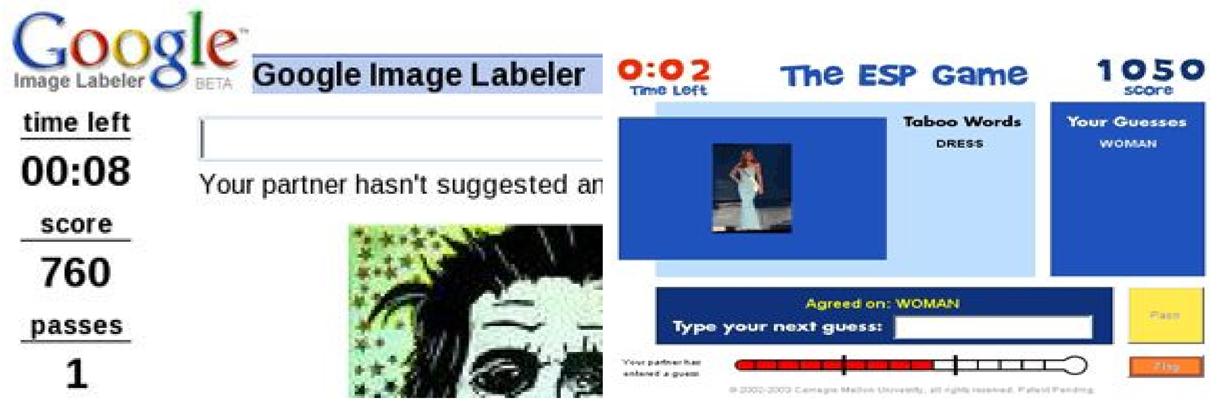
There are many systems that use crowdsourcing for data labeling. The biggest examples include Google's Crowdsource[1] and Amazon's Mechanical Turk[2]. Google's Crowdsource does not offer any fee to the contributors, but the users can level up and open badges as they label more data. The Mechanical Turk, however, offers micro payments for labeling, but they limit the amount of users in order to keep the process and the users verifiable and accountable.

Also, we all know that Google uses reCAPTCHA to collect their labeled data by using the slogan "protects your websites from spam and abuse". Actually, all tech giants like Google, Amazon crowdsource their training data and dominate the marketplace. The ESP Game and Google Image labeler are also the great examples of crowdsourcing.[4] Each gamified app pairs up anonymous partners with each other for a round of images

[1] https://crowdsource.google.com/

[2] https://www.mturk.com/

and label images according to the consensus taken from the users. What they do is actually using human-based computation to solve crowdsourcing problems.



Your partner hasn't suggested an

Another example is hCaptcha[3]; which is a drop-in replacement for reCaptcha that earns website owners money and help companies get their data labeled.

We intent to gamify the labeling process in order to keep users playing and get entertained in a fun way while labeling our data. Briefly, what makes our project innovative and different from the existing similar projets is that it brings game developers and people who demand labelled data in one domain. It does not serve data for specific companies, but any demander will be able to get their data labelled by using the games in our platform. Giant companies do their crowdsourcing own and dominate the market. We are aiming to provide a platform in which every company may do their own crowdsourcing and do not need to integrate the system into their own platforms just like reCaptcha, hCaptcha. Only requirement for labeled data demanders is to use our existing platform. End-users will be able to select different games according to their taste from the platform and while playing the game, they will be labelling data for the demanders.

[3] https://hcaptcha.com/

4. Conclusion

Deep learning technique is getting more popular and used by a broader range of people. This increases the demand and need of labelled data. In an other side, there is an endless demand for games. In our platform, using crowdsourcing and gamification concepts, we bring these two popular fields together. In the platform, there will be one sample game making people label a set of data while playing. Labelled data demanders will be able to send their data into the game. The platform will provide opportunity for game developers to put their game related to data labelling in our platform, which will increase the options for both the players and the data demanders to select the most accurate category or game for their data type. In this platform, we will provide game developers a domain where they can put their labelling based games, help demanders to collect labelled data and provide the users entertainment with game.

5. References

[1] J. Brownlee, "What is Deep Learning?", *Machine Learning Mastery*, 2018. [Online]. Available: https://machinelearningmastery.com/what-is-deep-learning/

[2] "Open Images Dataset V4", *Storage.googleapis.com*, 2018. [Online]. Available: https://storage.googleapis.com/openimages/web/index.html

[3] Code of Ethics National Society of Professional Engineers, nspe,org, 2016. https://www.nspe.org/resources/ethics/code-ethics/.

[4] https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2008-132.pdf