Task 3

MACHINE LEARNING PROJECT PROPOSAL

Aly Anderson

July 24th, 2021

1. **Project Overview**

**A1. Organizational Need**

A multi-media company, known as ESTELLA, has created an IOS application for users who want to be able to browse, upload, or save a variety of images for their own personal use (inspiring ideas, showing off new styles, things of that nature). Though the app’s functions are satisfactory, the marketing team suggested an idea that would improve the application and make it simpler for customers to use. The feature suggested is to have all images categorized by a certain subject, such as interior design or fashion. This will allow users to look up something by category instead of just browsing through random images to find something they will enjoy. This will require an image recognition algorithm that can automatically detect an object or face and be able to sort it into different groups. For this reason, this proposal intends to assist ESTELLA’s issue by not only providing a solution, but also a project plan in which the new feature can be implemented easily and efficiently so that customers can enjoy their time using the application.

**A2. Project Context and Background**

This project will assist ESTELLA in its needs to enhance user-performance for its application. Its input will collect images that are sent to the organization’s database, and it will output those images containing tags to indicate that it has been categorized into a certain subject. Because the application’s layout was originally just photos that were inserted into the database, this arrangement will allow users to look at specific images depending on the category they were interested in. This will also help users have an easier time navigating through the images and keeping their interest stimulated compared to when scrolling through random images. This project was written with firm belief that it will increase not only its user database due to this convenient new feature, but it also will save the company time comparing to hand-picking each image and putting it into a different category (or letting the users choose the category).

**A3, A3a. Review of Background Works+ Relation to Project Development**

Of course, an important step with Machine Learning is reviewing what others have done to solve problems like ESTELLA’s. One student—whose name prefers not to be shared—wrote on the HBS Digital Initiative that a popular company called Yelp uses a “*deep learning-power image analysis which identifies color, texture and shape of objects in user submitted photographs with 83% accuracy and uses the identifying traits to sort them into categories*” (HBS, para. 3). This allows Yelp to have the ability to sort their images by food, restaurant, menu, etc. Because ESTELLA also needs assistance with sorting images, using a deep learning-power image analysis, just as Yelp does, would benefit their needs greatly. Another example is Pinterest, a company that also works with AI closely, uses another form of the deep learning algorithm. According to James Le, writer of the Towards Data Science, he notes that Pinterest uses the deep learning algorithm that "*can not only determine the subject of an image, it can also identify visual patterns and match them to other photos*” (Le, para. 7). The customers of ESTELLA are supposed to be browsing through a plethora of images. Even if there is one category they want search, it is likely that those users will want to see images that are like the group they chose. Thus, Pinterest’s usage of a VS (Visual Similarity) deep learning algorithm would be useful for ESTELLA’s purposes. A final solution is known as image classification. Image classification uses pixels. While, deep learning uses objects, so that is the main difference between them. A company that uses image classification is a popular company called Google. According to the staff of Google’s developer website, Google uses image classification to “*define a set of target classes (objects to identify in images) and train a model to recognize them using labeled example photo*” (Staff, para. 2). Again, ESTELLA could also use this because it gives the same end-result of being able to target photos and put them into different labels. Whether image analysis, visual similarity, or image classification is used, it all can help ESTELLA to accomplish the goal of allowing users to be able to search and browse through photos that are categorized into different subjects—rather than scrolling through random photos until they find something that they were looking for. Reviewing what other companies have done to accomplish this will allow ESTELLA to grow as a company, improve its capabilities, and increase its user count.

**A4. Summary of the Machine Learning Solution**

The need that ESTELLA has is that the organization wants to create a feature in their application that allows users to search or browse photos in categorized sections instead of random photos that are in the system’s gallery. This proposal suggests using a Machine Learning type known as Image Classification which will define the class to identify the desired object in the image by means of pixels. The categories and the “rules” defined for each category must be created first, but once that has been done and the AI knows how to put the images into different groups, it will save a lot of time for the company. By using image classification, the company will not only be able to categorize their images, but every image in that certain class will be seen by the user. Thus, this accomplishes ESTELLA’s goal of categorizing the images in their database.

**A5. Benefits of the Machine Learning Solution**

There are many benefits to machine learning—or specifically, image classification. One benefit is that image classification will save ESTELLA a lot of time. For example, because the company has many images in its database, having to sort through each image and putting every single one into a different group would be time consuming and expensive. Not to mention, having to deal with more users adding images to the database. With image classification, this function can be done automatically, and large quantities of images can be stored within seconds depending on how fast the machine is. This saves ESTELLA a ton of time and money. Another benefit is that image classification is extremely popular in the machine learning scene, so there are many tools—such as Pytorch, Tensorflow, and things of that nature—that can make the process efficient and simple. Of course, these tools will be explained further in the proposal; however, the existence of these tools once again saves the company time because there is no need to build a *second* application to make the AI function automatically. Also, there is no need to hire additional help so that the machine can be built. This, again, saves the company money. As shown above, image classification is useful, popular, and can save companies with a heavy media database a lot of time. This is exactly what ESTELLA needs.

1. **Machine Learning Project Design**

**B1. Scope of the Project**

The scope of the project is defined as follows:

* **IN SCOPE:**
  + For the project to be created, a usage of third-party tools will be needed and in the scope of the project. The third-party tool used will be discussed below.
  + Training engineers within the company to use the third-party tools mention above. Some funding will need to be allocated for the necessary members of the IT team who are not familiar with the environment needed to create the application feature.
  + Administrative access to modify and retrieve data from the company’s database will be needed for the implementation to work.
* **OUT OF SCOPE:**
  + This project will not include the designing of the application’s user interface. As it is only meant to organize, the new design for the application once the feature has been implemented is out of the scope of this project.

**B2. Goals, Objectives, Deliverables**

* Goals:
  + Automatically organize images into different categories.
  + Allow users to scroll through a variety of images by category type.
* Objectives
  + Add the new feature within the 2021 FISCAL year.
  + Improve user base quantity and increase profits by 15%.
* Deliverables
  + Training company engineers on how to use the third-party tool that will be used to create and implement the new feature into the application.
  + Designing the new feature to have the ability to allow users to look through similar images (based on the category they choose) with relative ease.

**B3. Standard Methodology**

There are many standard methodologies used in the industry today. The methodology that will be used for this project is known as SEMMA. It contains five steps which make up its name: Sample, Explore, Modify, Model, and Access. This will be implemented in the project as shown below:

* **Sample:** As the name suggests, a “sample” or subset of dataset will be needed to be selected and used. The purpose of this step is to identify the initial variables needed and how it will affect the plan. The dependent variable is what will be tested, so this will be represented by the subset of gallery images in the dataset, while the independent variable is the variable that is used by the testers to change the dependent variable. This will be the AI system as mentioned above.
* **Explore:** Again, this name explains the process itself. In this early stage, the variables discussed in the sample stage will be used to examine the relationships and influences between the variables. Therefore, the testers of the project will be able to implement the AI feature and “explore” how the features will influence the dependent variable—the subset of images in the gallery.
* **Modify:** While learning about the outcomes in the explore stage, the tester should be taking notes of the result. In other words, once the data is cleaned and prepped from the exploration and sample stages, the lessons learned from those stages are taken note of so that the necessary modifications and refinement can be made. So once the testing of the feature with the samples have been completed, the notes taken from that stage will be written down so that the data given can be improved.
* **Model:** All the information from the modify stage is taken, so that the data is cleaned, and the lessons learned have been given. In the model stage, the application of the note written in the modify stage are applied here. Thus, the mistakes and bugs noted from the modify stage are now going to be fixed in this stage so that the AI feature can be relatively bug-free.
* **Access:** In this final stage, the feature will now be fully tested. However, this test will be different from the ones mentioned in the explore stage. In this testing, it holds an evaluation of the reliability and value it has. Therefore, once the new AI feature has been implemented, this evaluation will be tested by a select few customers (both new-coming and frequent ones) to be evaluated. Depending on this evaluation, the AI feature will either be pushed to the application with the whole database or will be sent back to the modify stage to add (or remove) certain items within the feature.

**B4. Timeline and Milestones**

* **August 1st-August 2nd, 2021** – Once the proposal has been accepted, the sample stage will be started. Since this stage is relatively quick, there does not need to be a long period for it to be completed.
* **August 3rd-August 6th, 2021** – The presentation of technical proof will be presented. While gathering the samples, there will need to be a presentation to explain the process. A presentation also does not need a long period.
* **August 7th-August 19th, 2021** – This is where the project will be sent to the explore and modify stage. Once the feature has been examined and notes have been taken, it will then be set to delivered to the public. This will take relatively longer than the previous dates.
* **August 20th-August 31st, 2021** – The project should be in the model stage where it has been examined and the notes taken from the modify stage will be implemented. A small window will be also given so that the project can enter the access stage to fully evaluate before releasing to the public.
* **September 1st, 2021** – The project has been tested and has been sent to the fully database so that now the feature has been implemented there. The public using the application will now be able to make use of this feature.

**B5. Resources and Costs**

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Project Team. | This will amount to the hours for the project team who will oversee the testing of features and assign the evaluations. They are also the ones who will write up the presentations and present it to the shareholders. The project team is being paid $45 per hour and are expected to work 40 hours a week. This will be for the timeline of 4.5 weeks. There are currently 6 members in the project team and no expectation of new hires. | $48, 600 for 6 members working 40 hours a week for 4.5 weeks. |
| IT Engineers. | The IT team will lead the creation of the feature. There are currently 3 members of the team who are familiar with environment that will be used for creating the feature. The other 2 members will have to take training courses. The 3 with the knowledge are paid $60 an hour and are expected to work 40 hours a week. This will be for the timeline of 4.5 weeks. The remaining 2 who have no knowledge are paid $40 an hour since their training courses are being paid by the company. | $32, 400 for the 3 members who have knowledge of the environment. And $14, 400 for the 2 members who do not have the knowledge. |
| Training course for feature creation environment | There are current members of the IT team who are not familiar with using the chosen environment for making the new feature (TensorFlow). Because of this, timely training courses will be needed to catch the other team members up to speed. | $8,000 for a boot camp on the environment for two members. |
| Online service—TensorFlow and Jupyter Notebook | This is an online and open-source service. | $0 |
|  | **Total Amount:** | $88, 400 |

**B6. Criteria for Successful Execution of Project**

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| Ease of use | One very important criterion that will be used to evaluate the success of the project is how easy it is to use said project. This will be judged by comparing use of previous version of the application to the newer one. Various selected customers will be used in assistance to gauge ease of use. |
| User base increase | Another objective that was listed previously was the growth of the user base. That objective will also be used in evaluation. The measure of success is whether it will meet the objective of growing by 15% or not. If the user base growth meets the percentage or higher, it will be considered a success. |
| Algorithm Efficiency | A final evaluation that will be measured is the algorithm’s efficiency. ESTELLA’s goal is to make the algorithm as efficient as possible with the minimal amount of cost. This will require a measurement using the Big O to determine the efficiency. The company is trying to avoid any expensive algorithm, such as one that runs by linear time (O(n)) or higher. |

1. **Machine Learning Solution Design**

**C1. Hypothesis**

The issue that ESTELLA is facing is that the arrangement of their images in the application are disorganized and random. This creates difficulty for their customers and prevents an ease of use, which forces them to scroll through a variety of images that are unrelated to what they are looking for to find what they desire. Consequently, it creates a decrease in the user-base. However, the solution for this is to create an AI machine that will reorganize all the images within the company’s database and automatically put them into different categories depending on the picture. Therefore, the hypothesis is that if the images are organized, then the relevant pictures will be easier to find, thus increasing user-base due to ease of use.

**C2. Analytical Method**

The identified machine learning algorithm to use for image classification is an algorithm called the KNN or K-Nearest Neighbor. This algorithm will categorize objects based on the training that it has had. In the training, this algorithm uses vectors and labels that are assigned to it and will save the qualities of the “test” image. It will then use those labels and vectors when looking at the next image to see if it is “near” or the characteristics are similar enough to the test image. This is how the images in ESTELLA’s database will be organized.

**C2A. Justification of Algorithm Selection**

There are many reasons why KNN would be a good algorithm to use. One of those advantages is that KNN is extremely easy to implement. The logic used is like human logic—just like python is one of the easiest programming languages to read—KNN is simple and can handle classes so that multiple categories can be used for the image classification. However, no algorithm has only advantages. Therefore are a few disadvantages to this algorithm. One disadvantage is that because the KNN is so simple, this means that the quality of data must be greater because it will not be as accurate if the quality of data is not good. So, ESTELLA must make sure that its data is cleaned and that the quality is good for highest accuracy. Overall, though the KNN does have its flaws, its advantages are what will benefit the company more, which is why this proposal is using KNN.

**C3. Tools and Environments of Solution**

As mentioned previously, there a few third-party tools that will be used. First, the programming language, Python, will be needed to write the code in those third-party tools. ESTELLA will also need to be familiar with the application known as Juypter to load, write, and edit the data. The open-source framework, TensorFlow, will be needed in order fully implement the algorithm. Finally, the libraries NumPy and Matplotlib will be used help make the process go faster, especially when coding the algorithm, less code will be needed due to the process already being automated by the libraries. For example, there will be no need to self-create a scientific computation for the vectors that KNN will use since the library NumPy already has one built-in. Finally, because the software for this can either use a Windows, Mac, or Linux, this proposal suggests ESTELLA’s using the company’s laptops, which use the Windows operating system. These are the tools and systems that will be needed for the project.

**C4. Measuring Performance**

Performance will be measured using a system known as “Confusion Matrix.” Confusion Matrix will help the testers to see how many false positives/negatives and true positives/negatives there are. For example, if there are two images—one of a dog and the other of a woman posing in clothes—it will be put into different classes, animals, and fashion. A false positive would be if the image recognizes the woman as an animal and puts it into the animal section, whereas a false negative would be if the dog was not recognized as an animal and put into the fashion section. The true positive/negative would be putting each image into its proper category. Basing it on this metric will assist in measuring the accuracy of the quality and performance of the algorithm. The goal is to minimize as many false results as possible to increase accuracy. This is how performance for the project will be measured.

1. **Description of Data Set(s)**

**D1. Source of Data**

As mentioned quite a few times in the previous paragraphs, the data (the images/media) is being retrieved from the company database. Using SQL as the desired language, this database was designed by database engineers of ESTELLA with the intention of holding all the information of the company, such as photographs, videos, customer information, employee information, etc. The database is organized in a way that separates all this data so that unrelated details will not be stored in the same place. For example, the customer information (addresses, usernames, etc.) would not be in the same storage as the employee information (payment, status, etc.). This will make it simple for the project to extract all information from the database where all the images, media, and videos are put. Therefore, this project’s source of data will come from ESTELLA’s database.

**D2. Data Collection Method**

The method that will be used to extract and collect the data is called “Incremental Extraction”. It is opposite of a method called “Full Extraction” which finds the source of the data and extracts it completely. On the other hand, incremental extraction will only extract information from a certain time or event in the past. This time or event will need to be defined (which can be done so in SQL, as that is what the database was written in), so it will have to manually written in. This is the method that will be used to collect the data.

**D2A. Advantages and Limitations**

An advantage using this method is that it gives more assurance that the data is accurate, compared to the full extraction method. For example, this method requires a specific event to be written, so data that was flagged for removal or is not an image will not be added if this is specified early on. Accuracy in this stage will be much better. However, a disadvantage for this is that the extraction/collection will take a considerable amount of time because the extraction will sift through data that was specified to be taken. If the database were small, this would not be an issue, but because ESTELLA’s database contains all the company’s information, it will need to look through to see what its target is and is not. Overall, though this method has its limitations, this proposal chose this method because the pros outweigh the cons and will overall be very beneficial for ESTELLA.

**D3. Qualities and Completeness of Data**

After the collection/extraction of data, it will need to be prepared for loading into the machine learning algorithm—The KNN algorithm—to be used. Accuracy is the key to the proposal; thus, data cleansing is what will be used to prepare the data to be ready. The use of a third-party tool that ESTELLA already sponsors—known as Power BI (an analytical tool)—will be used as a helper tool. This data cleansing will involve looking through dirty data to see if anything needs to be addressed. For example, images that are duplicate (perhaps a user posted the same image twice) will be noticed by Power BI because the tool uses a Primary Key (a key assigns a value a unique ID in a database) which assists in distinctiveness of images/records. Since a duplicate image with have the same key assigned to it, Power BI will be able to identify this right away and remove it. Removal of duplicates from the data is the most important step for this project because the KNN algorithm only knows its most similar item. Thus, if there are duplicates, the KNN will go to the next duplicate and the next one and so on. This will waste of lot of unnecessary time for the algorithm and the user looking through images. Once again, this issue will be resolved by means of data-cleansing and the use of Power BI. After it has been cleansed, it will be ready for the AI to go through it by means of TensorFlow and sort through it. This is how the process will be for preparing the data.

**D4. Precautions for Sensitive Data**

Because this project is important for the development of ESTELLA’s new application feature, sensitive data is likely to arise and should be handled in a proper manner. This section intends to explain how sensitive information will be used and handled during the process. First, sensitive data must be identified, and this sensitive data is the user information. Each media that is uploaded by a user contains a fragment of sensitive user information (precise location, for example) which cannot be seen by the public eye but is still seen within the database. If it is seen within the database, employees have access, and this can cause privacy issues. To prevent issues like this, the solution would be to mask this information so that it cannot be read and deciphered by human logic. Cryptography is suggested and a recommended one would be to use a strong hash known as the SHA-256. This encryptor will store sensitive fields of data with a hash key that is unreadable to the human brain. Since the sensitive user data cannot be removed completely, masking it behind cryptography is the next best step and thus is recommended for this project.

1. **Citations**

HBS. November 13, 2018. *Machine Learning at Yelp*. < https://digital.hbs.edu/platform-rctom/submission/machine-learning-at-yelp/>

Le, James. January 23, 2018. *Pinterest’s Visual Lens: How computer vision explores your taste.*< https://towardsdatascience.com/pinterests-visual-lens-how-computer-vision-explores-your-taste-5470f87502ad>

Staff. June 21, 2021. *ML Practicum: Image Classification*. <https://developers.google.com/machine-learning/practica/image-classification>