

WGU C951  
Task 3  
MACHINE LEARNING PROJECT PROPOSAL  
Derek Brown  
2/15/2023

## Table of Contents

A. Project Overview.....	3
A.1. Organizational Need .....	3
A.2. Context and Background.....	3
A.3. Outside Works Review .....	3
A.4. Solution Summary.....	4
A.5. Machine Learning Benefits.....	5
B. Machine Learning Project Design.....	5
B.1. Scope .....	5
B.2. Goals, Objectives, and Deliverables .....	5
B.3. Standard Methodology .....	6
B.4. Projected Timeline .....	7
B.5. Resources and Costs.....	8
B.6. Evaluation Criteria .....	8
C. Machine Learning Solution Design.....	9
C.1. Hypothesis.....	9
C.2. Selected Algorithm.....	9
C.2.a Algorithm Justification.....	9
C.2.a.i. Algorithm Advantage .....	9
C.2.a.ii. Algorithm Limitation .....	9
C.3. Tools and Environment .....	10
C.4. Performance Measurement.....	10
D. Description of Data Sets .....	10
D.1. Data Source .....	10
D.2. Data Collection Method.....	10
D.2.a.i. Data Collection Method Advantage.....	10
D.2.a.ii. Data Collection Method Limitation .....	10
D.3. Data Normalization.....	10
D.4. Data Security.....	11
References.....	11

## **A. Project Overview**

The project will use machine learning and image recognition to help create appropriate alternative text for images on a webpage. Images that do not seem to have appropriate alternative text will be flagged for manual review.

### **A.1. Organizational Need**

The organization needs to ensure that the images displayed online have appropriate alternative text to meet accessibility guidelines. This means that the alternative text must accurately describe the image for users.

### **A.2. Context and Background**

The organization operates an online clothing store, and as such, the website contains a lot of images of various products. Because of this, it is necessary that the images contain appropriate alternative text to ensure they are accessible for users who rely on screen readers.

### **A.3. Outside Works Review**

Image classification is achievable using a neural network. Menon discusses the possibility in the article *Ultimate Guide To Building Powerful Keras Image Classification Models* (Menon, 2022). The article gives an overview of the entire process of creating an image classification model. It starts by describing a neural network as having three types of layers - an input layer, hidden layers, and an output layer. The input layer and output layer, as the names suggest, take input and return output respectively. The hidden layers are where the input is processed. There can be any number of hidden layers. The article goes on to discussion about selecting a dataset, loading the data, and the creation of a Convolutional Neural Network (CNN). It discusses how to plot the accuracy and how to use the pre-trained VGG16 model and ends with a recap of the discussion. For our purposes, the most relevant part of the article is the part discussing the neural networks, as these will be an integral part of our solution.

Specifically, we will be implementing a CNN which is further elaborated in an article titled *Why are Convolutional Neural Networks good for image classification?* (Mishra, 2019). This article highlights the difference between a CNN and the typical neural network design. Because of the large amount of data needed for an image classification model, a neural network can be very computationally intensive. The CNN reduces the computational power and time needed by reducing the size of the matrices in each of the nodes (neurons). Basically this means that instead of keeping track of each pixel and its relation to all the other pixels, the CNN only

really concerns itself with the pixels immediately surrounding a given pixel. This creates a more efficient way to process the images and doesn't negatively impact accuracy. It has an added benefit of making the neural network more flexible regarding object positioning within the image compared to other neural network models (I.e. the object's position in the image doesn't impact its ability to recognize the object). The article describes the process using some examples to show the process of reducing the size of the matrices using some matrix operations. While the article offers just a brief overview of the process, it is helpful in understanding the core idea about how this type of neural network improves efficiency, and why we would want to implement a CNN in our solution.

An alternative neural network is the Multilayer Perceptron (MLP), which is a type of Artificial Neural Network (ANN). In the article *CNN vs MLP for Image Classification*, the MLP model is briefly compared to the CNN (Dinesh, 2019). Similar to the previously discussed article, this article discusses the amount of data being processed - the MLP "includes too many parameters because it is fully connected. Each node is connected to every other node in next and the previous layer... as a result... overfitting can occur which makes it lose the ability to generalize." The CNN avoids the problem of overfitting by breaking the image into smaller parts and analyzing the parts independently. In addition, the MLP model uses vectors instead of a matrix. A matrix is ideal for images because each pixel is represented as a value in the matrix and their spatial relationship is preserved (I.e. their positioning in the matrix is the same as their positioning in the image). Since the MLP doesn't use a matrix, that means the spatial relationships are lost. The article discusses this in more detail. In summary, the CNN is "spatially invariant" meaning it can learn to recognize an object regardless of its position in the image, whereas MLP struggles when the object is moved to different parts of the image. The conclusion is that while MLP models may work well for other types of data, CNN is the ideal model for image classification, which is exactly what our solution will do.

#### **A.4. Solution Summary**

To accomplish our goal, our solution will build an image classification model as outlined by Menon. To achieve this, the ideal model is the CNN, which is corroborated by both Dinesh and Mishra, primarily because it is capable of preserving spatial data (which is essential for images) and will enable better generalization. This model will be used to generate text for the website's images that can be used as alternative text.

### **A.5. Machine Learning Benefits**

Machine learning is an essential part of automating the process of analyzing web images. Its primary benefit is efficiency - it will make the process much faster than it could be done through manual review. Machine learning will reduce the need for manual review. This will allow reviewers to spend less time on this and more time on other important tasks.

## **B. Machine Learning Project Design**

### **B.1. Scope**

The scope of the project will include:

- The dataset needed to train the model will be collected and prepared.
- A software solution will be developed and trained using the prepared dataset to classify images.
- The solution will scan a webpage and identify images as input items, classify images based on type of clothing depicted, generate a rating for each image comparing the image's actual alternative text to what is expected, and flag images with poor ratings for manual review.
- Documentation will be created describing the collection and preparation of the data used in the dataset.
- Documentation will be created describing the software solution, how to install and use it, and commentary for each major section of code for future maintenance.

The scope will NOT include:

- The solution will not classify items based on color or patterns,
- The solution will not identify text within an image (such as printed text on a shirt).
- The dataset will not include images for categories not explicitly stated.

### **B.2. Goals, Objectives, and Deliverables**

Goals:

- To improve website accessibility by ensuring the images have appropriate alternative text.

Objectives:

- Accuracy - the product must meet the determined threshold for accuracy for classifying the images into the categories.
- Ease of use - the product must be easy to use for end users.
- Proper documentation - the product will include proper documentation for ease of use, maintenance, and troubleshooting.

Deliverables:

- A software product that can be used to analyze images on a website
- Documentation on how to use/install the product

### **B.3. Standard Methodology**

Development will follow the CRISP-DM methodology:

- Business understanding: the project aligns with the business objective of ensuring that the website is usable and accessible for the vast majority of users. What this means in our case is that all the images used for products on the website have appropriate alternative text assigned. To do this, we will need to utilize a dataset of relevant images with descriptive text, and train the program using this data.
- Data understanding: the data needed for this project will include images with accurate text descriptions. The data will be collected from online resources. From there, this data will need to undergo data cleaning and preparation for quality assurance.
- Data preparation: any images that are not relevant to current products on the online store will be excluded. In addition, any images with incomplete data will be removed or will need to be modified (I.e. if the image is missing descriptive text, a description without an image, or a corrupted image file etc.). Each image that remains will then be scaled from their original size to a 64X64 pixel square image format.
- Modeling: for our model, we will use a neural network of 3 layer types (input layer, hidden layer, and output layer). We will train the model to classify five different types of items (shirts, pants, hats, hoodies, and socks) sold in the online store, and will include 300 images from the dataset for each type to train the image. Once trained, the model will be tested against a dataset of 200 images (40 for each type) and will aim to achieve over 90% accuracy for the test data.

- **Evaluation:** the model will be tested against a dataset of test images for accuracy. The initial aim for the model will be at least 90% accuracy to be considered successful. Once this is achieved, it will move to deployment.
- **Deployment:** the model will be available as an internal tool for the organization. It will be available to employees online by logging into the website and navigating to the tool. The model will be monitored and updated periodically to reflect any changes in the store's inventory (I.e. adding or removing a type of product), and will be trained on larger datasets to improve accuracy in the future. A feature will be available for users to report inaccurate results, which will also improve its accuracy.

#### **B.4. Projected Timeline**

March 1 - Project approved

March 8 - Project team meets to plan tasks.

March 13 - Data collection and preparation begins.

March 20 - Data collection and preparation is completed. Training the model begins.

March 22 - Training is completed. Testing begins.

March 27 - Testing is completed. The model is submitted for review.

April 3 - Review is completed. Documentation is reviewed and revised as needed.

April 10 - Documentation is finished.

April 12 - Project is submitted to Project Manager for approval against requirements.

April 17 - The tool is deployed on the website once approved.

#### **Sprint Schedule**

<b>Sprint</b>	<b>Start</b>	<b>End</b>	<b>Tasks</b>
1	March 13	March 20	Data collection/cleaning
2	March 20	March 27	Model training and testing
3	March 27	April 3	Review of model
4	April 3	April 12	Documentation is updated project is submitted for approval
5	April 12	April 17	Tool is deployed

### B.5. Resources and Costs

Resource	Description	Cost
Software Developer 1	Implementing the code and software design for the project. Testing the model and review. Source code documentation.	\$45/hr X 90 hrs = \$4,050
Data Engineer 1	Collecting, cleaning, and preparing training and test data. Model review. Data documentation.	\$50/hr X 55 hrs = \$2,750
Project Manager	Coordinating the project, assigning tasks, reviewing and delivering the final product.	\$60/hr X 50 hrs = \$3,000
Laptop for Software Developer 1	Laptop capable of coding, training, and testing the model.	\$2,000
Laptop for Data Engineer 1	Laptop capable of data collection and cleaning	\$1,000
Laptop for Project Manager	Laptop capable of running project management software	\$1,000
Microsoft Suite License	To create/edit documentation	\$10/Mo X 3 Licenses = \$30
	<b>Total</b>	<b>\$13,830</b>

### B.6. Evaluation Criteria

The criteria below will be used to evaluate and measure the success of the final product:

Objective	Success Criteria
User flag rate	Less than 10% of results being flagged as inaccurate by users
Ease of use	Less than 10% of users complain about difficulty using the tool
Effective Documentation	Less than 10% of users complain about poor documentation



## **C. Machine Learning Solution Design**

### **C.1. Hypothesis**

By utilizing machine learning and implementing a type of neural network, I will be able to create a software solution that is able to accurately classify different types of images that are in use on the organization's website. Once created and trained, the solution's accuracy will be tested by using test images with predetermined descriptors as input, and comparing results with the expected values.

### **C.2. Selected Algorithm**

The selected algorithm for this project is the Convolutional Neural Network (CNN) algorithm. A neural network uses layers of neurons to simulate a neural network similar to that used in a human brain. It consists of three kinds of layers - an input layer, hidden layers, and an output layer. The input and output layers, as their names suggest, take input and produce output respectively. The hidden layers take the input from the input layer, and process the data, passing it between multiple hidden layers before being passed to the output layer, which produces the result (Menon, 2022). A CNN improves this by using filters and some matrix operations performed on the matrices representing the layers to reduce the size of the nodes, which makes it more efficient than some other algorithms (Mishra, 2019).

#### **C.2.a Algorithm Justification**

This algorithm was selected since it offers exactly what the organization needs - image classification with high accuracy. CNN is very useful for image classification and can be trained to be highly accurate. According to Dinesh "CNNs perform better than any other model in such "human-like" tasks like recognizing specific objects in the picture," which makes it an ideal candidate for our task of classifying images from the website (Dinesh, 2019).

##### **C.2.a.i. Algorithm Advantage**

A CNN has the advantage of being able to identify key features in an image without being manually trained or supervised. This means it is able to classify images very effectively once trained, and this ability is very scalable.

##### **C.2.a.ii. Algorithm Limitation**

The greatest limitation to a CNN is the large amount of data needed to achieve the best results. Of course, small datasets can be used to speed things up, but for greatest accuracy, a

large dataset is ideal. As the size of the dataset increases, the computational power and time needed to train the model increases as well.

### **C.3. Tools and Environment**

The solution will be developed on a Windows machine running Windows 10. It will be developed in Python 3.10 using the PyCharm IDE. The python libraries that will be used to create the solution are OpenCV, Tensorflow (Keras), and matplotlib.

### **C.4. Performance Measurement**

Quality will be measured using some test images. These will have predetermined descriptors, and will be used as input. The result will be compared to the descriptor for accuracy. The performance will be measured by how quickly the result is processed.

## **D. Description of Data Sets**

### **D.1. Data Source**

The data will be a subset of the dataset from Kaggle.com's "Fashion Product Images (Small)" dataset (Aggarwal, 2019). The dataset will include only the five categories listed in section B.3 above.

### **D.2. Data Collection Method**

Data will be collected via a direct download from the source.

#### **D.2.a.i. Data Collection Method Advantage**

This data collection method has the advantages of being quick and easy to collect, and it is free.

#### **D.2.a.ii. Data Collection Method Limitation**

The biggest downside to this method is that it will require a lot of data preparation and cleaning before the data is really usable for our purposes.

### **D.3. Quality and Completeness of Data**

To prepare the data, we will need to go through and check the image descriptions for each of the images and keep only the ones that are needed for our goals. The images will also be resized as needed to 64X64 pixels. Images will also be removed if they are missing a description, or if the image is corrupted and unreadable. The dataset will then be divided - 600 images will be set aside for testing (120 for each category), and the rest will be available to use for training.

#### **D.4. Precautions for Sensitive Data**

None of the data used to train the model is sensitive, and no extra precautions will be necessary in that regard. The software solution is to be used only on company owned computers. It should not be copied or transferred to personal devices including flash media, external hard drives, or personal computers. Additionally, this applies to any reports and data generated by the software solution. Any communication between team members regarding the software solution should be relayed using the organization's approved channels, and personal communication methods (personal email, phone, chat, etc.) are prohibited.

#### **References**

- Aggarwal, P., Fashion Product Images (Small) (2019),  
<https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small> accessed on February 13, 2023.
- Dinesh, CNN vs MLP for Image Classification (2019),  
<https://medium.com/analytics-vidhya/cnn-convolutional-neural-network-8d0a292b4498>  
accessed on February 13, 2023.
- Menon K., How To Build Powerful Keras Image Classification Models: Simplilearn (2022),  
<https://www.simplilearn.com/tutorials/deep-learning-tutorial/guide-to-building-powerful-keras-image-classification-models> accessed on February 13, 2023.
- Mishra, P., Why are Convolutional Neural Networks good for image classification? (2019),  
<https://medium.datadriveninvestor.com/why-are-convolutional-neural-networks-good-for-image-classification-146ec6e865e8> accessed on February 13, 2023.