David Kwon Han Lam

Literature Review & Data Description – CKME 136

Introduction

How does a person describe an image? They would usually describe them by highlighting one or more of its features such as color, texture, shape, motion, and location that it possesses.   
  
There is a saying that says a picture is worth a thousand words. Thousands of words may sometimes not be enough to describe an image especially if the image captured is a huge event in a person's life.   
  
Describing an image can be extremely challenging especially if the individual who is being described has not seen this item firsthand.   
  
In games like Pictionary or charades, we tried to give our teammate enough clues to beat the other team in the fastest time. The key is to use the least amount of words or gestures.   
  
Everyone has their own words to describe the same image. Some companies like McDonald's have achieved the goal of getting the general public to describe them the same way. When someone thinks of the golden M or golden arches right way McDonald comes to mind.

My goal from using this data set is to create a classification algorithm that would be able to predict the best word and image combination. By perfecting the algorithm, we can successfully predict what most of the general public uses to describe images.

This type of data is beneficial to help determine which picture best represents a word that can be used to promote a business or individual ad on their webpage, biography, business cards, etc.

Literature Review

Through the process of reviewing works of literature articles I hope to gain a better understanding of a few things such as knowing what are current algorithms in place to study images, techniques used to increase the efficiency of an image analyzing algorithm, and even processes that can refine the dataset before an algorithm is used.

Article 1 - CIDEr: Consensus-Based Image Description Evaluation

In this article, the researchers are trying to evaluate a new algorithm that describes images using human descriptions and compare them with existing algorithms. This study is divided into three parts. First, they use existing models (BLEU, ROUGE, METEOR) to conduct the test. BLEU is a precision-based metric used to compare the difference between a computer-generated sentence and a human-generated sentence. ROUGE is a recall-based metric used to compare the difference between computer-generated summary and human-generated one. METEOR is a combination of precision and recall based metrics. Then they would compare these models with a new model (CIDEr) that was created.  The CIDEr measures the likenesses between a generated sentence compare to one present by humans. Lastly, they would use two datasets (PASCAL-50S and ABSTRACT-50S) to evaluate which model provided the best results. From the conclusion, the researchers note that CIDEr provides the highest accuracy compared to existing models.

Article 2 - Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics (Extended Abstract)

In this article the researchers are trying to evaluate a ranking type framework with a normal framework. To determine the effectiveness of the ranking framework they created a new dataset to perform this task. From the conclusion, the researchers note that ranking type framework mimics more along with human judgement compare to automatic evaluation metrics BLEU and ROUGE.

Article 3 - Simple Image Description Generator Via A Linear Phrase-Based Model

In this article the researchers are attempting to create a simplified model to extract relevant description from a given image. They found that the characteristics used to describe a given image described are usually noun phrases. The interaction between these characteristics is the prepositional phrase and verb phrases. Thus, they would train a model to predict the possible phrases for a given image. The results closely align with the human agreement score. The results provide promise to the researchers and they plan in the future to apply this model to other datasets.

Article 4 - Comparing Automatic Evaluation Measures for Image Description

In this article the researchers are comparing the correlation between human judgement and automatic measures such as BLEU, Smoothed BLEU, TER, ROUGE-SU4, and Meteor. To calculate the correlation the spearman technique has been used. From the results they found that BLEU shows the weakest correlation with human judgement and Meteor shows the highest correlation and all the other measures fall in between. The researchers suggested using Meteor, Smoothed BLEU, or ROUGE-SU4 instead of BLEU because they show a stronger correlation with human judgement.

Dataset

The dataset topic is image descriptions. This dataset consists of a set of images, matched words, and a confidence score for each matched pair. Contributors were shown pictures and words and ask to determine whether if the two pairs are relevant to each other. If they match, they are given a score. The confidence score shows how likely the contributor matched the image with the word.

The image descriptions dataset is download from the website <https://www.figure-eight.com/data-for-everyone/>. The data was created by figure eight (crowd flower).

Approach

## Step 1: <Data Collection>

Download the dataset from the website

The image descriptions data set can be obtained from the website <https://www.figure-eight.com/data-for-everyone/>. Based on the [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/by/4.0/) this data is free to be shared and transformed as long as the author is credited and changes were indicated if made.

This dataset was conducted by the crowdflower on March 30, 2011. They obtained this dataset by asking individuals to partake in a survey. In the survey the researchers provided these individuals with a group of images and asked if a given word matches the explains the imaged well.

In this step I have obtained the raw data which was downloaded from the website and opened with excel. Due to the vagueness and simplicity of the data set I was completely confused about how to move further. Fortunately, I was able to find a glossary of terms that are commonly used in figure eight. The descriptions used in this table are from my interpretation and extraction from the website

## Step 2: <Data Exploration & Feature Selection>

**2.1 Reviewing Variables**

1. *Version 1 – Raw Data*
   * I had selected Jupyter notebook with python coding as my main way to showcase my code, visuals, and results. Before uploading the dataset into Jupyter notebook I attempted to check one of the image links. Upon discovery I noticed that the link was broken and decide to proceed as Jupyter notebook may be able to pull the images.
2. *Version 2 – UTF-8 csv format*
   * When I was attempting to run the CSV document on Jupyter notebook I received an error message UnicodeDecodeError which I believe has some to do with the file format so I converted the file from a CSV to a UTF-8 CSV file through the save as function on excel

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type** |
| Unit Id | Variable used to define each unique an entity in the dataset | Integer |
| Golden | Test question response whether word matches with the phrase   * **True** * **False** | Boolean |
| Unit State | Status of the variable   * **Golden** – test question * **Finalized** – received enough trusted judgments to be considered complete and will no longer collect judgments | String |
| Trusted Judgement | is an answer from a contributor with an accuracy score higher than the [minimum accuracy](https://success.figure-eight.com/hc/en-us/articles/202702975-Guide-To-Test-Question-Settings-Quality-Control-) you set on the settings page. All trusted judgments are included in your results.   * **Range: 3 - 190** | Integer |
| Last Judgement | The time the last judgement has been made   * **Range: Nov 6 - 25** | Datetime |
| Choose One | The selection made by the contributor   * **Yes** * **No** | Boolean |
| Confidence | confidence score describes the level of agreement between multiple contributors (weighted by each contributors’ trust scores), and indicates our “[confidence](https://success.figure-eight.com/hc/en-us/articles/211573026-How-to-Create-Test-Questions-from-High-Confidence-Rows)” in the validity of the aggregated answers for each row of data. The aggregate result is chosen based on the response with the greatest confidence.   * **Range: 0.5243 - 1** | Decimal |
| Choose One Gold | The selection of the best word and picture combination from contributors | Boolean |
| Image | Contains the link where the image is extracted from | URL |
| Tag | the bag of words used in the dataset | string |

**2.2 Data Cleaning**

1. *Version 3 – Dropping Variables*
   * **Golden** variable can be dropped as this is the answer given by the researchers and not the contributors
   * **Unit State** variable can be dropped as the study has already been finished so no need to figure out if additional judgements are needed
   * **Last Judgement** variable can also be dropped from the dataset as the time of last judgement should not affect the accuracy of the decision
   * **Choose One Gold** variable can be dropped because we are not looking for the best pair and word combination, we are trying to mimic whether the contributors will deem the image and phrase matches
2. *Version 4 – Renaming the dataset variables to make it easier to understand*
   * From **Unit Id** to **Unit Id** – no changes made
   * From **Trusted Judgment** to **Trusted Judgment** – no changes made
   * From **Choose One** to **Response** – indicates the response made by a contributor
   * From **Confidence** to **Confidence Score** – added the word score to variable
   * From **Image** to **Image** – no changes made
   * From **Tag** renamed to **Phrases** – when viewing the data set, I noticed that some of the tags were not real words some words were gibberish e.g. ‰ÛÏcarrodecombate‰Û
3. *Version 5 – Converting Variables to a different type*
   * **Unit Id** – no changes made
   * **Trusted Judgment** – no changes made
   * **Response** – converted from Boolean to String
   * **Confidence Score** – no changes made
   * **Image** – the URL links are broken but still useful as it determines the link between the image and phrases - converted from URL to string
   * From **Tag** renamed to **Phrase** – when viewing the data set, I noticed that some of the tags were not real words, but some were gibberish e.g. ‰ÛÏcarrodecombate‰Û

## Step 3: <Feature Selection>

* **Unit Id** – is important variable to keep in the dataset because it the attribute used to identify the different pairs of image and phrases
* **Trusted Judgment** – is not as important as the 4 other variables but may be useful in determining the likeness a contributor will deem the image and pair matches. The higher the number of trusted judgements the increase likelihood of contributor confirming the pair
* **Response** – is important variable to keep in the dataset because contains yes or no responses made by the contributor
* **Confidence Score** – is not as important as the 4 other variables but may be useful in determining the likeness a contributor will deem the image and pair matches. The close the score is to 1 the increased likelihood of contributor confirming the pair
* **Image** – is important variable to keep in the dataset because contains the images used to determine response made by a contributor
* **Phrase** – is important variable to keep in the dataset because contains the phrases used to determine response made by a contributor

## Step 4: <Build Models>

For this step I will be using 3 models to run my revised data set. These models are linear regression, decision tree, and Naïve Bayes. I will be running these models by using python in juypter notebook. The coding for these models can be found online I have including the citation under the work cited section of my report. I will evaluate each model’s effectiveness, efficiency, and stability. I will also be finding additional methods in python to improve them further.

When I attempt to run the dataset, I encounter an error in the python juypter notebook. This occurs because some of the variables that I have kept are categorical and need to be transformed into numerical variables

Model 1 – Linear Regression Model

Model 2 – Decision Tree Model

Model 3 – Naïve Bayes Model

## Step 5: <Summary and Conclusion>

Once the analysis of these models has been completed, we can do a summary and concluded if our initial hypothesis matches the results. Further detail will be provided in the final project.

Work Cited

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4566-4575).

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, *47*, 853-899.

Lebret, R., Pinheiro, P. O., & Collobert, R. (2014). Simple image description generator via a linear phrase-based approach. *arXiv preprint arXiv:1412.8419*.

Elliott, D., & Keller, F. (2014, June). Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 452-457).

“Data For Everyone.” *Figure Eight*, www.figure-eight.com/data-for-everyone/.

“Glossary of Terms.” *Figure Eight Success Center*, success.figure-eight.com/hc/en-us/articles/202703305-Glossary-of-Terms.

“Image Descriptions - Dataset by Crowdflower.” *Data.world*, 21 Nov. 2016, data.world/crowdflower/image-descriptions/workspace/data-dictionary.