Image Descriptions

David Kwon-Han Lam

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 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 life.

Describing an image can be extremely challenging especially if the individual who is being described to 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 has achieve the goal of getting the general public to describe them the same way. When someone think of the golden M or golden arches right way McDonald comes to mind.

# Introduction

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 majority 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 card etc.

# Literature Review

Through the process of reviewing literatures article I hope to gain a better understandings of a few things such as knowing what are current algorithms in place to study images, techniques used to increase efficiency of 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 describe images using human descriptions and compare it with existing algorithms. This study is divided into three parts. First, they use existing models (BLEU, ROUGE, METEOR) to conduct the test. BLEU is precision-based metric used to compare the difference between a computer-generated sentence and human generated sentence. ROUGE is a recall-based metric used to compare the difference between computer generated summary and human generated one. METEOR is 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 compare 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 normal framework. In order 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 use to describe a given image described are usually noun phrase. The interaction between these characteristics are 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 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 judgment and automatically 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 judgment and Meteor shows the highest correlation and all the other measures fall in between. The researchers suggested to use Meteor, Smoothed BLEU, or ROUGE-SU4 instead of BLEU because they show stronger correlation with human judgement.

# Dataset

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

explain the details. Descriptive statistics of the attributes and datasets can also be provided here.

# Approach

Create a block diagram for the steps of your approach to clearly provide an overview. For example, if you first scrapped twitter, second applied NLP techniques to extract keywords, third labelled the tweets as positive and negative using a set of keywords, and fourth build a classifier, then you should create a box for each of the steps with arrows connecting one step to the next one. A sample block diagram is shown below.

## 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 author is credited and changes were indicated if made.

This dataset was conducted by 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.

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

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type** |
| Unit Id | variable used to define each unique entity in the dataset | Integer |
| Golden | Test question response whether word matches with 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 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 answer 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 image is extract from | url |
| Tag | the bag of words used in the dataset | string |

## Step 2: <Data Cleaning and Exploratory Analysis>

|  |  |
| --- | --- |
| **Before** | **After** |
| Unit Id | No changes were made as each variable has a unique unit id |
| First Response |  |
| Unit State | Status of the variable   * **Golden** – test question * **Finalized** – received enough trusted judgments to be considered complete and will no longer collect judgments |
| 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** |
| Last Judgement | The time last judgement has been made   * **Range: Nov 6 - 25** |
| Choose One | The selection made by the contributor   * **Yes** * **No** |
| Confidence | Kept the variable the same |
| Best Combinations | when looking at the word groups in the filter section of the excel I was able to find only a few yes response  e.g. from the “a” word group only 2 / 13 response were yes |
| Image | Kept the variable name but noticed the links for this dataset is broken changed it to a string variable by removing the shortening the length of the variable e.g. http://clic.cimec.unitn.it/~elia.bruni/crowdflower/mirflickr/100k-images/im923434.jpg to im923434 |
| Phrase | Since this variable contain the real words and non words I have changed tag to phrase |

## Step 3: <Ranking>

Rank the words in groups to

## Step 4: <Exploratory Data Analysis>

Analyze the data to see what results occurred

## Step 5: <Summary and Conclusion>

Once analysis is completed, we can do summary and concluded if our initial hypothesis matches the end results.

# Challenges

# Results

Explain your results here. Consider that you need to communicate your results to executives in an organization. For example:

1. Insert tables and/or charts showing the results
2. Write description of the tables and charts, such that they show the usefulness for an organization
3. Identify the evaluation measures, such as accuracy, precision, recall, etc.

# Conclusions

Give a short summary (one to two paragraphs) of your analysis and conclude the discussion by defining the usefulness of your analysis.

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).

<https://www.figure-eight.com/data-for-everyone/>

<https://success.figure-eight.com/hc/en-us/articles/202703305-Glossary-of-Terms>

<https://data.world/crowdflower/image-descriptions/workspace/data-dictionary>