Multilingual & Color Focused Image Captioning For Visually Impaired Using Deep Learning Techniques

Abdul Sohail Ahmed , Poojan Gagrani , Sourab Rajendra Saklecha , Yuti Ashwin Khamker

Department of Applied Data Science, San Jose State University

DATA 270: Data Analytics Processes

Submitted to : Dr. Eduardo Chan

November 03, 2023

Data Collection

Data Collection Plan

Data Collection is the process of collecting data systematically from various sources. It is one of the most critical steps in the project as it serves as the base from where the whole data analysis begins. It is the building block to make informed decisions, solve relevant problems, evaluate performance, identify trends, and perform a risk assessment and management. The quality of the data collected plays a vital role in determining the accuracy of the machine-learning models. Models created from high-quality data help to make the most accurate decision.

For the task of multilingual and color-focused image captioning two sources were selected to go ahead with MS-COCO (Microsoft Common Objects in Context) and Flickr30k. These datasets are considered benchmark datasets which are mostly used in projects that perform image captioning. The sources and datasets are as follows:

1. MS-COCO dataset created by Microsoft Research, consists of a diverse set of images gathered from various platforms along with human-generated annotations (Lin et al., 2014).
2. The Flickr30k dataset provided by Young et al. (2014), contains images with their associated captions extracted from the website called Flickr.

The MS-COCO dataset is one of the most influential datasets that is used in the field of computer vision. It is used for image captioning, detection of objects, segmentation of semantics, and detection of relationships between images and objects. It was created by Microsoft research by performing extensive web scraping on various platforms to gather a diverse set of images to ensure the comprehensiveness of data. Collected data of 2017 went through different selection processes to filter unnecessary images like the one with poor quality and less content to select 163,957 images that were rich in quality and represented relevant information. The selected images were provided to the team of human annotators to create useful and relevant annotations. This process includes the creation of bounding boxes around the images, identification of the object labels and showing them with the images, and providing detailed textual content which is the information about the image (caption). There are 80 types of object categories

included in the MS-COCO dataset like animals, people, household items, etc. Each image in the dataset has five different descriptive captions which makes it useful for image captioning. The 2017 dataset was split into train, test, and validation sets with 118,287; 40,670; and 5,000 instances which made the creation of the model much faster, ensuring faster hyperparameter tuning and reducing the computational cost. The reason for choosing this dataset was its coverage of diverse images with annotations of high quality. The images had multiple captions with each image properly categorized into object classes which make the dataset very easy to work around with all required pieces of information. Due to the high quality of images, the color characteristics can also be easily extracted from them. FiftyOne is the user interface or tool that allows faster and more efficient access to the MS-COCO data repository. It enables the creation of subsets from the original dataset or allows filtering of data to select data of a particular object class.

Table 1 illustrates the data collection plan for MS-COCO dataset. It provides brief information about the dataset, what will be done once the data is collected, and metadata information. This dataset is considered as the benchmark dataset as it is widely used because of it's high quality annotations and great variation in the object classes of the images.It is used in the project for training different machine learning models so that they can generate the highly descriptive, color-focused captions. For training models dataset of total size 18 GB is used so that models can get better hang of the images with their object classes and their corresponding captions so that when they were validated against 1 GB dataset they generate semantically correct with high quality color emphasized descriptions (captions) of the images.

Table 1

Data Collection Plan for MS-COCO Dataset

Data Collection Plan

Group number: 3 | Date: 10/01/2023

Project leader: Abdul Sohail Ahmed | Project title: Multilingual Image Captioning For Visually Impaired

Description of the data collected

The Common Objects in Context dataset, or COCO, is a valuable tool for tasks like segmentation, object detection, and caption generation. It has a number of noteworthy features, such as superpixel-based material segmentation, accurate object delineation, and the ability to recognize objects in relation to their surroundings. COCO has over 330,000 photos, most with annotations which contains over 1.5 million unique object instances across 80 different object categories and 91 material categories. The variety of the dataset is increased by the five insightful captions that support every picture.

What will be done with the data once it has been collected?

After obtaining the data, We preprocess the data and carry out any necessary data augmentation.We then use it to train and assess the deep learning models for our computer vision tasks, including segmentation, captioning, and object detection.

Key Variables - A summary of the chosen input variables (Y's) and/or output variables (X's)

		1	2	3
What?	**Variable title**	Images	Captions	Multilingual Color Focused Captions
	Input (X) or output (Y) variable?	X	X	Y
	Data type	Image	String	String
	Collection method	To facilitate easy use of the dataset, COCO has partnered with the team behind the open-source tool FiftyOne. It is a tool that helps in visualization and access to COCO data resources and serves as an evaluation tool for model analysis on COCO		
Historical data	Historical data exist?	No	No	No
Who?	**Data collector**	Poojan Gagrani		
	Operational definition exist?	No	No	No
	Data collector trained?	Yes	Yes	Yes
	Resources available for data collector?	Yes	Yes	Yes
When?	**Start date**	9/25/2023	9/25/2023	9/25/2023
	Due date	10/3/2023	10/3/2023	10/3/2023
	Duration (in days)	8	8	8

Additional Comments

Another way of using the data from the dataset , MSCOCO also offers an API. The COCO API assists in loading, parsing, and visualizing of annotations in COCO. The API supports multiple annotation formats . For additional details see: CocoApi.m, coco.py, and CocoApi.lua for Matlab, Python, and Lua code, respectively, and also the Python API demo.

The Flickr30k dataset is a popular resource in computer vision, particularly for tasks related to generating descriptions for images which are publicly available. This dataset comprises approximately 31,000 photos sourced from Flickr which is a photo-sharing platform. The images were selected according to predetermined standards, such as having a variety of themes, a wide range of visual content, and high-quality photos. For generating corresponding captions for these images, the researchers

employed human annotators to generate relevant captions in natural language form. For each image in Flickr30k, five unique captions were curated giving various explanations. As our project focuses on generating relevant captions in multiple languages that are color-focused, the reason for choosing the Flickr30k dataset is because of well-established evaluation metrics, this dataset is used as a benchmark that is commonly used to assess the performance of deep learning models for image captioning. Also, it has different captions for the same image which allows for describing the variation of colors. Data size is about 9 GB.

Table 2 shows the data collection plan laid out for the Flickr30k dataset. There are a total of six inputs and one output. The image itself would be one of the inputs along with five different captions for each of the images. A multilingual, color focused caption string would be the output. All these captions would be strings of characters. The way of collecting the data has been briefly described in the Data Collection Plan (DCP). We do not have any historical data for this dataset. The primary data collector has also been mentioned in the DCP, and this data collector was also trained on the usage of this dataset. The start date and end date of the data collection process has been defined in lines with the gantt chart.

Table 2

Data Collection Plan for Flickr30k Dataset

Data Collection Plan

Group number: 3 | Date: 10/01/2023

Project leader: Abdul Sohail Ahmed | Project title: Multilingual Image Captioning For Visually Impaired

Description of the data collection

The Flickr30k dataset is a popular resource in computer vision, particularly for tasks related to generating descriptions for images which are publicly available. This dataset comprises approximately 31,000 photos sourced from Flickr which is a photo-sharing platform. The images were selected according to predetermined standards, such as having a variety of themes, a wide range of visual content, and high-quality photos.

What will be done with the data once it has been collected?

The Flickr30k dataset contains data that can be used for training and assessing machine learning models for tasks like language generation and picture captioning. This will improve the capabilities and performance of computer vision and natural language understanding systems. The data must be divided into two sets after pre-processing: a training set and a testing set.

Key Variables - A summary of the chosen input variables (Y's) and/or output variables (X's)

		1	2					3
What?	**Variable title**	Images	Caption 1	Caption 2	Caption 3	Caption 4	Caption 5	Multilingual Color Focused Captions
	Input (X) or output (Y) variable?	X	X	X	X	X	X	Y
	Data type	Image	String	String	String	String	String	String
	Collection method	Data was first requested from the team that manages Flick30k (managed by the University of Illinois). After reviewing the request, we were granted access to the dataset via a link. This link contained the images and the five corresponding captions						
Historical data	Historical data exist?	No	No	No	No	No	No	No
Who?	**Data collector**	Yuti Khamker						
	Operational definition exist?	No	No	No	No	No	No	No
	Data collector trained?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Resources available for data collector?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
When?	**Start date**	9/25/2023	9/25/2023	9/25/2023	9/25/2023	9/25/2023	9/25/2023	9/25/2023
	Due date	10/3/2023	10/3/2023	10/3/2023	10/3/2023	10/3/2023	10/3/2023	10/3/2023
	Duration (in days)	8	8	8	8	8	8	8

Additional Comments

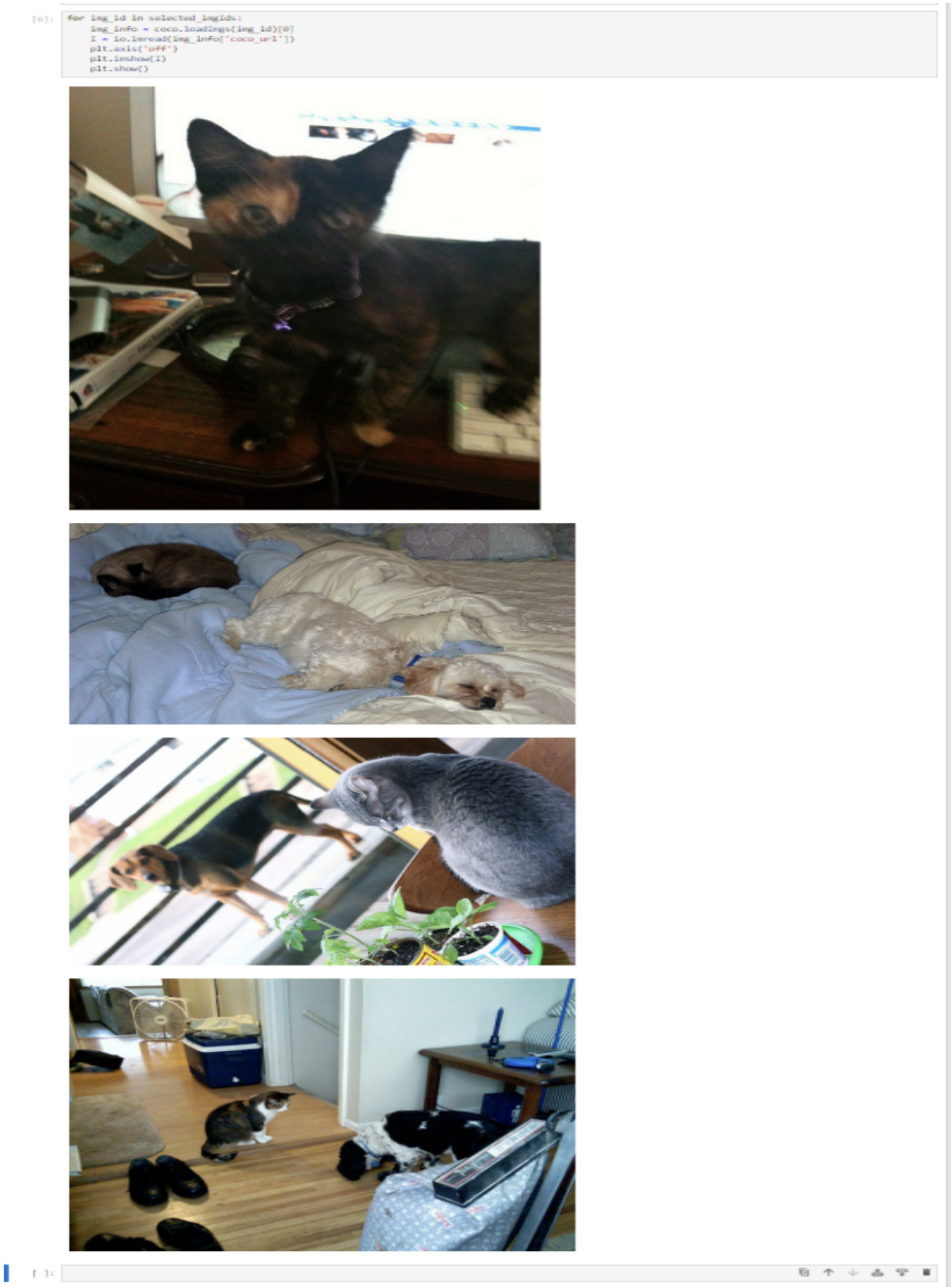
This dataset is about 9 Gigabytes in size. Flick30k augments the 158k captions from Flickr30k with 244k coreference chains, linking mentions of the same entities across different captions for the same image, and associating them with 276k manually annotated bounding boxes.

Dataset Snapshots

The figure 1 illustrates the snapshots of the images obtained from the COCO dataset. These were collected using the API built on python and hosted by MS-COCO.

Figure 1

Dataset sample extracted from the COCO API



The figure 2 provided below are the snapshots of the images obtained from the Flickr30k dataset. Data was first requested from the team that manages Flickr30k (managed by the University of Illinois). After reviewing the request, we were granted access to the dataset via a link. This link contained the images and the five corresponding captions for each of the images.

Figure 2

Sample Images extracted from the Flickr30k dataset

Synthetic Data

Using synthetic images for our image datasets that are to be used for image captioning can help improve the model's robustness and diversity by introducing novel scenes and objects not present in real-world data. Synthetic images can also be carefully designed to address specific challenges, such as low-light conditions or rare scenarios, enhancing the model's ability to generate accurate and informative captions in a variety of situations.

Another useful strategy that contributes to dataset diversification is image augmentation, which involves applying transformations such as flips, rotations, and color adjustments to improve the model's capacity for generalization and to produce captions for different image variations.

Figure 3

Displaying the original image of "cat" extracted from COCO API.

```
[333]: I = io.imread(img['coco_url'])
       plt.axis('off')
       plt.imshow(I)
       plt.show()
```

Figure 4

Identifying and Extracting objects from the original image

```
[340]: plt.imshow(overlay_image_rgb)
       plt.axis('off')
       plt.show()
```

Figure 5

Inverting colors by highlighting objects and backgrounds from the original image

Object Highlight

```
[341]: highlight_color = (0, 255, 0, 255)  # Green color with full opacity
       highlight_image = Image.new('RGBA', (img['width'], img['height']), highlight_color)

[342]: highlighted_object = Image.composite(highlight_image, image, mask)

[343]: background_color = (0, 0, 255, 255)  # Blue solid color
       solid_background = Image.new('RGBA', (img['width'], img['height']), background_color)

[344]: final_image = Image.composite(highlighted_object, solid_background, mask)

[345]: final_image_rgb = final_image.convert('RGB')

[346]: plt.imshow(final_image_rgb)
       plt.axis('off')
       plt.show()
```

Figure 6

Blurring background in the original image

```
[350]: plt.imshow(object_image_rgb)
       plt.axis('off')
       plt.show()
```

Figure 7

Cropping the original image to keep focus on the main object

```
[352]: plt.imshow(cropped_image)
       plt.axis('off')
       plt.show()
```

Figure 8

Rotating the original image by 45 degrees

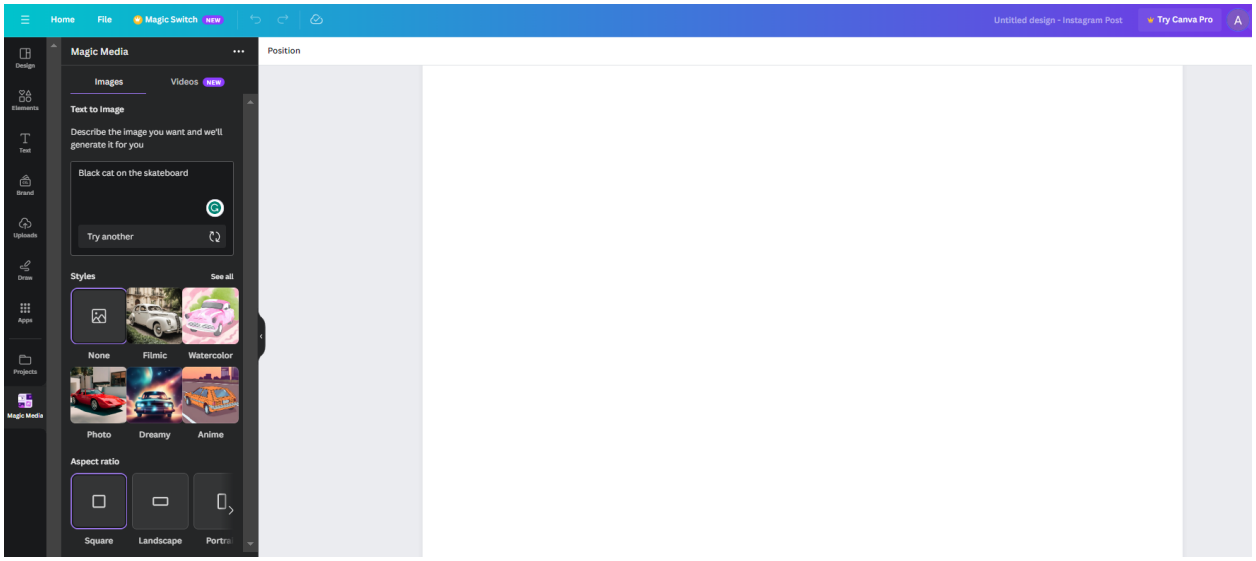
```
[353]: angle = 45
       rotated_image = image.rotate(angle, expand=True)

[354]: plt.imshow(rotated_image)
       plt.axis('off')
       plt.show()
```

AI is a powerful technology that is revolutionizing the current world. There are some of the online available AI tools such as Deep AI, AI Art Generator by Hotpot.ai, and Magic Media by Canva that can be used to generate images from the text which is the captions given as an input. Since, no actual images are used from the dataset and the images are generated from text descriptions, they can be considered as synthetic data. These tools can generate large amounts of synthetic data which is very close to actual data through the process of text to image synthesis which can be used to perform operations such as data augmentation, prevention of privacy by hiding sensitive information, detection of anomaly etc.

In Figure 9 the prompt “Black cat on the skateboard” was given as an input to the Magic Media, there were some of the modifications that could be made on the generated image like selecting the styles among none, filmic, watercolor, photo, dreamy, and anime that will be followed while generating image or the aspect ratio whether it would be placed in landscape, portrait, or square.

Figure 9

Prompt to generate synthetic image from text in Canva

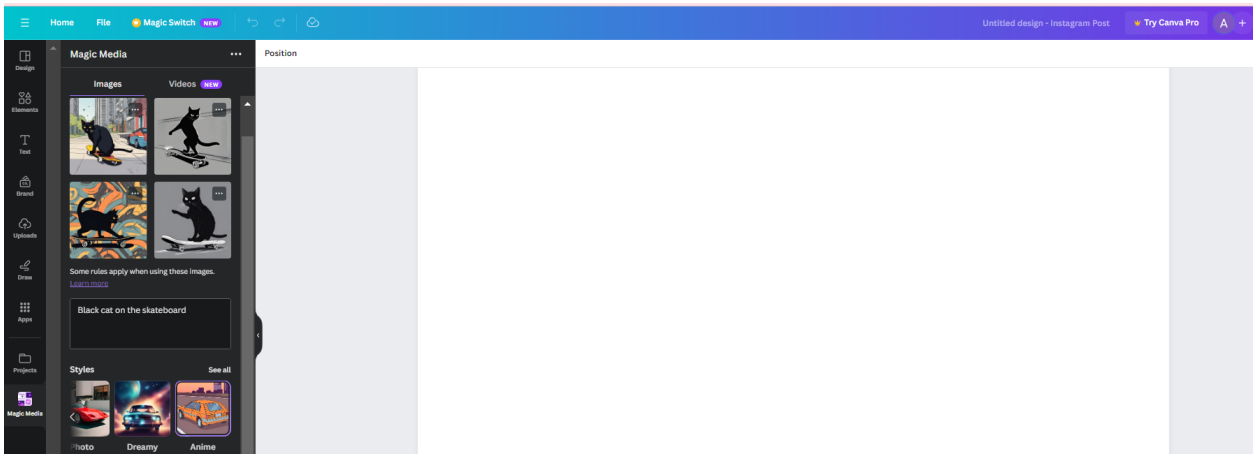


It is quite evident in figure 10 that prompt was able to run successfully and different images with black cat on the skateboard were created. There were different options provided in the output so that the user can select the image that suits the need. Since, the images are artificially generated in this scenario so

they can be considered as synthetic data. These images can be used for the purpose of data augmentation if required.

Figure 10

Generated images in Canva



Among the different sets of images the image in figure 11 is selected because it can be used to extract some meaningful information like having the object car present in the background which can also be detected and the image can be placed in the "cars" object category. Another thing to note is that this image can be used to create captions with color characteristics. For example, this image can generate a caption as "black cat on yellow skateboard with red and blue car in the background".

Figure 11

Selected image from set of generated images

References

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *In Proceedings of 2014 European Conference on Computer Vision*, 740–755. Springer. https://doi.org/10.48550/arXiv.1405.0312

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, *2*(1), 67–78. https://doi.org/10.1162/tacl_a_00166