Project Report:

Digital Collection Accessibility Solution for New York Public Library

Team:

Abhijeet Rajesh Sant (ars1125@nyu.edu)
Nikita Girish Amlani (nga261@nyu.edu)
Rohit Saraf (rs6785@nyu.edu)
Shreya (ss13337@nyu.edu)

Supervisor:

Prof. Jean-Claude Franchitti (jcf@cs.nyu.edu)

Project Manager:

Sharon Denning (sharondenning@nypl.org)

1. Introduction

The modern world publishers, content editors and web developers carry a virtuous intent of broadcasting scholarly substances in the form of contextual images and videos. It has been ascertained that domesticating figures and pictures engenders a much higher value of a product or service than using mere texts. Consequently, imagery and recordings became critical for user engagement on the internet. Engaging the audience by clicks, shares, and downloads escalated immensely underestimating one of the most important aspects of accessibility of images, alt-text.

Throughout this whole process of grabbing user attention the most overlooked characteristic; alt text has recently gained spotlight due to the doctrine of developing a heartening brand impression. In the language of web development (Hyper text markup language and Cascading Style Sheets), alt text is a small narration of an image that appears on the webpage. Alternative text also known as alt-text is a short, text narration capturing the essence of a visual object in an electronic medium to improve user experience.

Following image serves as an example of a simple alt-text.

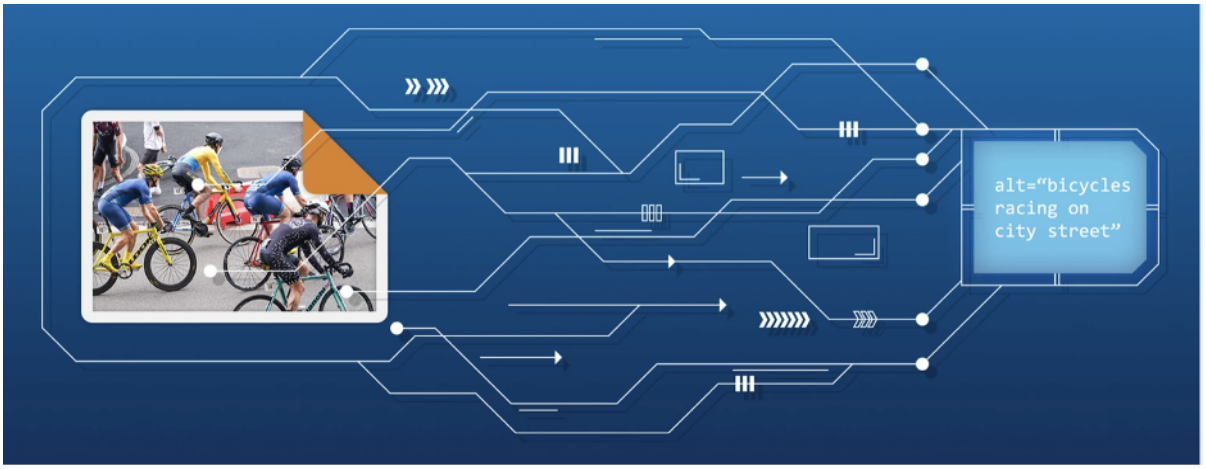


Figure 1.1 Alt-text example

The digital world should be widely available to everyone, including the visually impaired for whom accessibility might become a problem if they fail to comprehend information. It is easy to miss out particulars in digital versions thus appropriately written alt-texts can radically reduce pessimism.

Following are some encouraging points to use alt-text to our advantage.

1. Equalization of opportunity on a digital platform to access objects is requisite and alt text empowers the visual impaired populace to grasp unhindered content thus transforming the platform to be ADA (American Disability Act) compliant.
2. Images tend to take a significant period of time for uploading and downloading and individuals aren't equipped with lightning speed internet. Here alt-text can be distinguished as a contingency resort for broken images or images that don't load easily.
3. Alt-text on images is a contributory cause for Search engine optimizations
4. Alt text which are concise descriptions can allow engine crawlers to catalog images on the web page.
5. Alt text can also boost customer base by transforming simple images into hyperlinked search results and thus companies/individuals can use Google's Search Engine Results page to their advantage.

For a large content warehouse like the New York Public Library, it is demanding to render descriptive text for each object. Addressing this concern computer vision comes to our rescue. Image recognition services and its techniques have boosted our confidence in generating captions even for monochrome & disproportionate pictures.

This project responsibly generates image captions and allows the user to confirm the correctness of monochrome images by using CCN and LSTM techniques and corroborating with new metrics like BLEU score.

Succeeding pointers like approach and implementation, outcome and performance, challenges faced will create an entire premise of the project.

Approach and implementation of the Model

To create the end-to-end solution for the problem statement, a CCN-LSTM model was trained and tested which would provide alt text for the images. An interface was created using the Python Flask framework using javascript, HTML, and CSS in the frontend. The model can be retrained from the interface as more alt texts are generated and saved. This would ensure continuous improvement to the performance of the model and eventually provide perfect alt texts for the images.

To train the model, images were first downloaded from the FSA collection of NYPL using a python script. There were 34700 images that could be accessed using the

NYPL API. Along with the images, the titles associated with them were also downloaded which were then divided into training and test datasets. After observing the titles, it was apparent that the titles were not descriptive of the images. To improve the training data the images were run through detectron_model by Facebook which is pre-trained on millions of images. These generated captions were descriptive of the images.

The next step was to train the model which could generate the alt text for the images. We chose the VGG model for this solution. Keras provides this pre-trained model directly and can be easily imported and trained on the images you have. We used the Keras tools to resize the images to the preferred size of 224 x 224 pixels. We then pre-compute the photo features using the pre-trained model and save them to a file so that we don't have to reload them. We then load these features later and feed them into our model as the interpretation of a given photo in the dataset. These are saved in the features.pkl file.

The text was then cleaned. The titles were converted to lowercase, all punctuations were removed and words with length less than one character were removed. It is important to have a vocabulary that is both expressive and as small as possible. A smaller vocabulary will result in a smaller model that will train faster. This dictionary of image identifiers and descriptions is saved to a new file named descriptions.txt, with one image identifier and description per line.

The model is then defined. It has three layers as described below-

- **Photo Feature Extractor**. This is a 16-layer VGG model pre-trained on the ImageNet dataset. The photos are preprocessed using the VGG model (without the output layer) and going forward this would be used as input.
- **Sequence Processor**. This is a word embedding layer for handling the text input, followed by a Long Short-Term Memory (LSTM) recurrent neural network layer.
- **Decoder** Both the feature extractor and sequence processor output a fixed-length vector. These are merged together and processed by a Dense layer to make a final prediction.

The model is then fit on the training dataset. The model learns fast and quickly overfits the training dataset. For this reason, when the skill of the model on the development dataset improves at the end of an epoch, it will save the whole model to file.

The saved model is evaluated for the skill of its predictions on the holdout test dataset by generating descriptions for all photos in the test dataset and evaluating those predictions with a standard cost function. The actual and predicted

descriptions are collected and evaluated collectively using the corpus BLEU score(The Bilingual Evaluation Understudy Score) summarizes how close the generated text is to the expected text. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0.

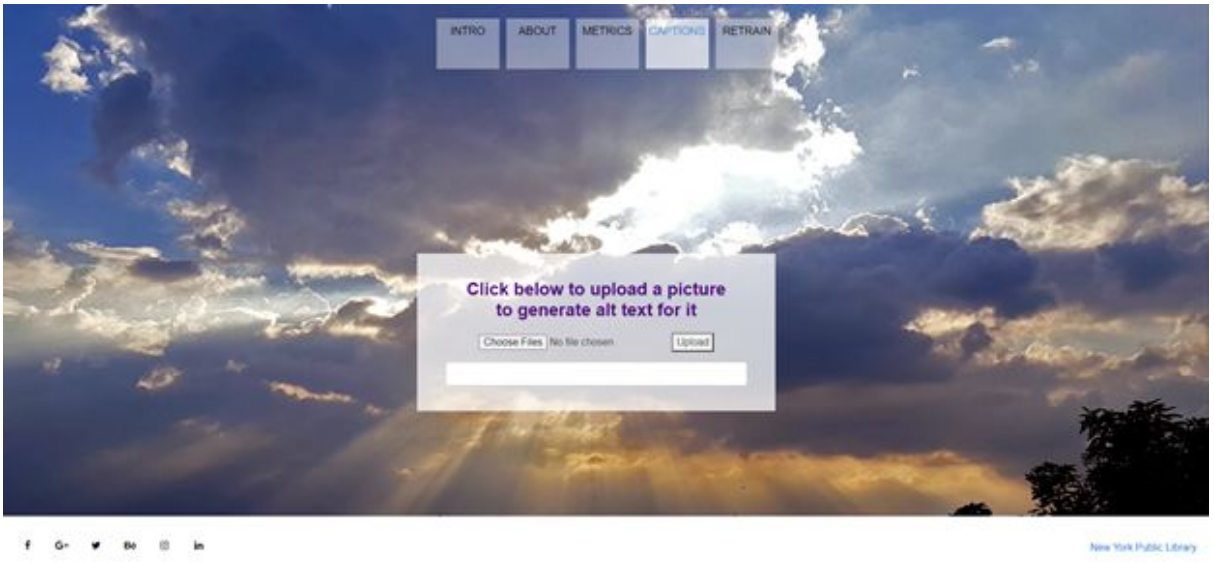
After the model achieved the BLEU score of 0.44 with weight 3 and 0.64 with weight 0, it was integrated into the interface created. The interface has a section which allows users to retrain the model after enough data is collected from their end. This ensures the continuous learning of the model.

Result

The end to end system developed provides the solution of increasing the accessibility of NYPL Digital Collection. Our development is focused with NYPL's FSA (Farm Security Administration) collection. Our application generates ALT Text in an interactive mode or in a batch mode.

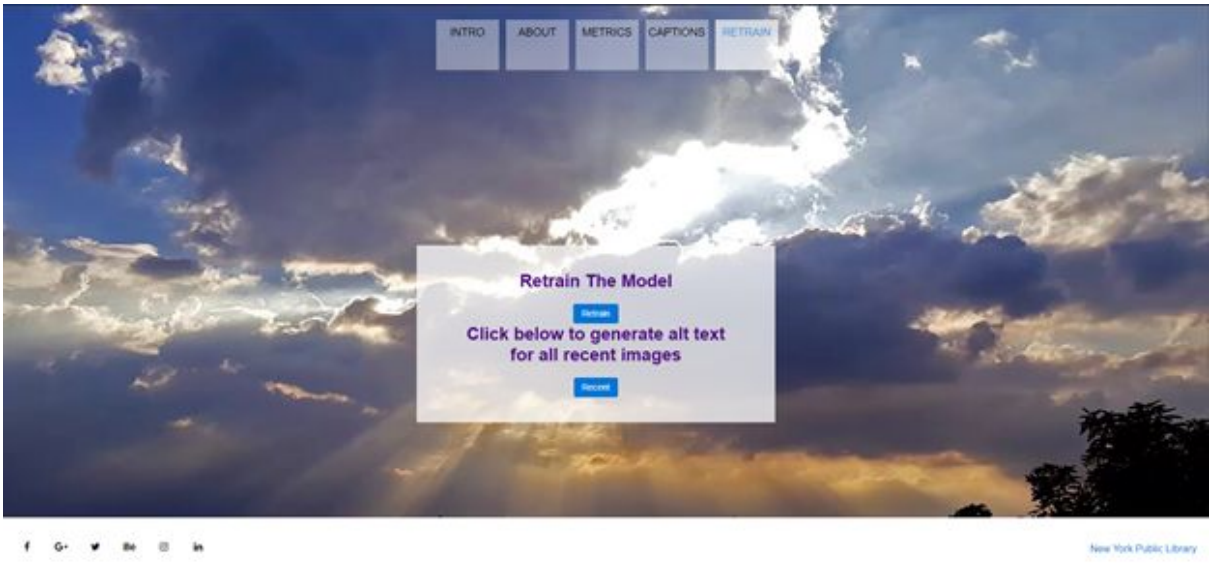
Interactive Mode

In the Captions Tab we can execute the model in an interactive mode. Here the user can generate ALT text to the image they upload. The output text that gets generated is editable so it can also be corrected by the end user and confirmed. The confirmed text then can be used further to re-train the model and hence help to improve vocabulary of the model.



When we try to upload images from the FSA collections, we get the below Alt Texts

In the Retrain tab, we have the option to retrain our model with the changes to the ALT text that got generated in the Captions tab. The retraining of the model takes a significant time hence it is done in the background. This helps to improve the model vocabulary.



Batch Mode

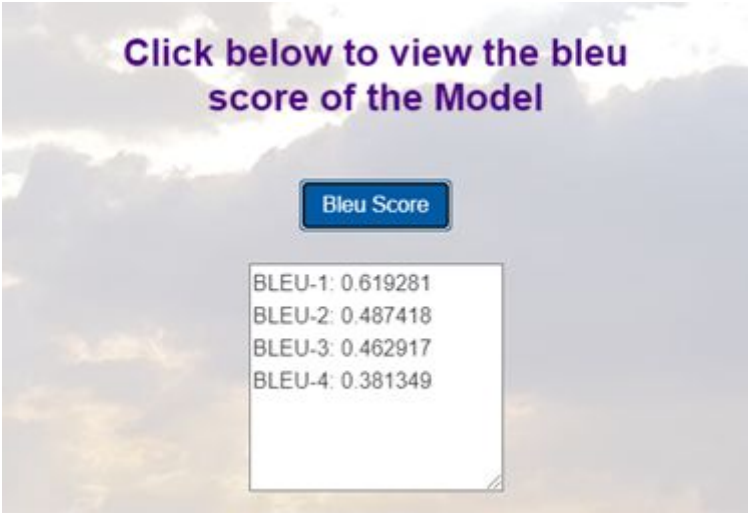
In the batch mode Alt Text generation, the application supports downloading the recently uploaded images to the NYPL FSA Collection and generating their ALT Text in a batch. It works on the idea of downloading the images uploaded after the completion of the latest download. Once the download is complete, it generates the Alt Texts for those images. We can go through the Alt Texts being generated in the local and correct accordingly if needed. These changes can also be further used to retrain the model by selecting the Retrain button.

Evaluate Model

The model keeps improving with each time we train it. The metrics of the model can be evaluated after every retrain we do. We can evaluate the skill of its prediction by generating the Bleu Score of the model. The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0. It is language independent and easy to understand. To calculate the bleu score, we convert the predicted caption and references to unigram/bigrams.

$$\text{modified ngram precision} = \frac{\text{max number of times ngram occurs in referance}}{\text{total number of ngrams in hypothesis}}$$

It is common to report the cumulative BLEU-1 to BLEU-4 scores when describing the skill of a text generation system.



Below is a rough guideline for the interpretation of BLEU score, as provided by google. Hence, the BLEU score generated for the model is acceptable.

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

Challenges Faced

The training data from the NYPL for the FSA collection was not appropriate for training the model. Some of the issues that were present in the dataset can be noted as below:

- There were no alt-texts for images, hence we had to use image titles for creating training data.
- Most of the image titles were just the name of a city, country, or a particular year.
- Half of the images were just the back side of the actual picture and the title was describing the actual front side of the image.

When the model was trained with the above data, the results were not acceptable. To overcome the issue, a bootstrapping model was used to generate alt-text for the images, which was then used as the dataset to train the actual model. The bootstrap model was trained on millions of images, hence it was providing a better starting point by generating alt-texts to train the actual model.

Future Scope

- The model can be further improved by retraining on data that has been verified by the actual users.
- Expand the solution to generate alt-texts for other NYPL collections.
- Implement a feature to detect when the alt-text generated by the model needs human intervention.
- Implement a feature to show the confident score of the alt-text generated by the model.
- Add the exploration search feature on top of the model.