AutoCap

Image Captioning in AC-215

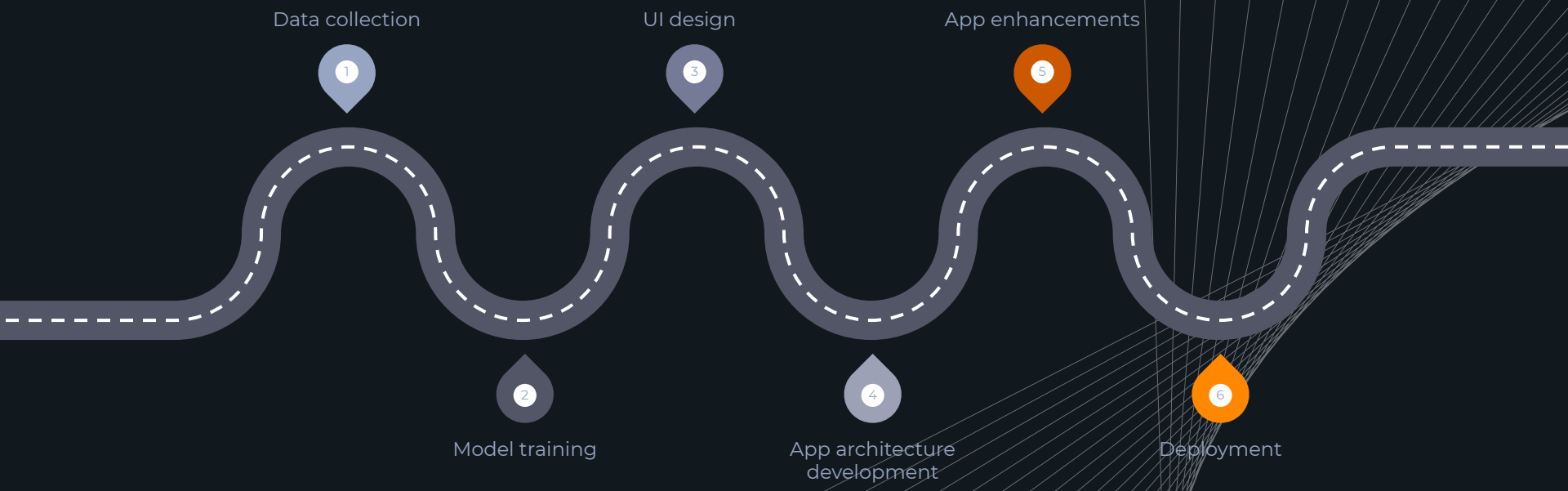
Problem Definition

The **visually impaired** rely on screen readers to access the internet through audio. Additionally, photo album **indexing**, automatic **social media** captioning, and image **filtering** are just some of the many applications that make accurate automatic image captioning an important priority.

Proposed Solution

The objective of this model is to take an image as an input and produce a sequence of text that serves as the caption of the image. We can achieve this with a deep learning, **encoder-decoder model**, where the image is encoded into features that are then decoded into a sequence of text.

AutoCap Roadmap



Data Exploration

MS-COCO Dataset

330k images

5+ captions per image

Training subset

118,287 RGB images

591,753 captions

53,953 unique words

Sample images of MS-COCO train dataset

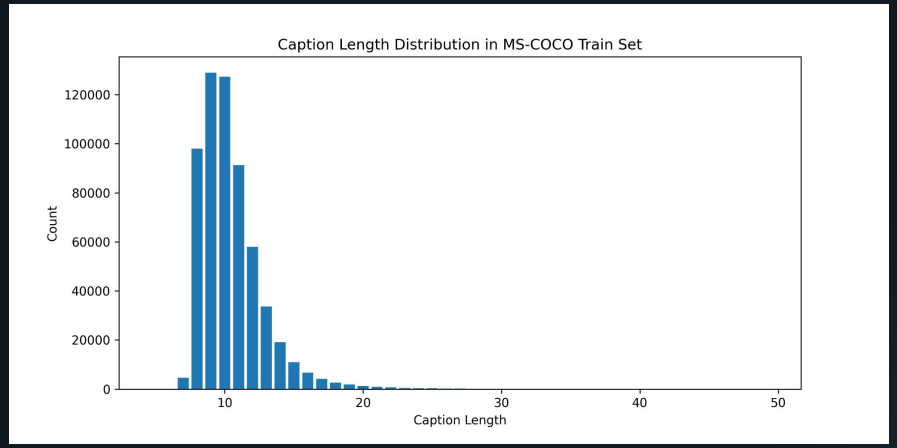
<start> A man and woman standing in front of a bar. <end>
<start> Man and woman standing outside a bar featuring happy hour! <end>
<start> A couple poses under a banner of a turkey. <end>
<start> People are standing next to each other under a turkey sign. <end>
<start> A newly married couple standing next to each other. <end>

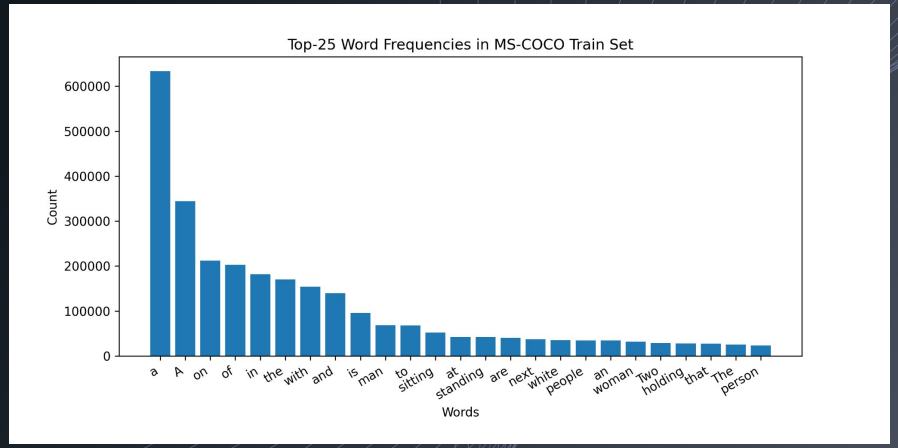
<start> A person in an office holding an Apple cell phone. <end>
<start> A person holding an older version of an ipod. <end>
<start> Small apple cell phone image from the back of a silver case. <end>
<start> A man holding a silver colored iPhone cell phone. <end>
<start> The back of an iphone in front of a destk. <end>

<start> A couple of young men playing a game of frisbee. <end>
<start> Two men in a field by a parking lot play frisbee <end>
<start> two people playing frisbee in a park next to cars <end>
<start> Two people playing tennis together in a field. <end>
<start> Some people playing with a disc in a big grassy field. <end>

<start> A plain, empty kitchen wit a light turned on <end>
<start> A clean kitchen filled with kitchen appliances like a refrigerator and a stove. <end>
<start> A long,clean and spacious kitchen with microwave and other gazette. <end>
<start> A clean kitchen filled with kitchen furniture and accessories. <end>
<start> A brown kitchen with a light on in the middle of it. <end>

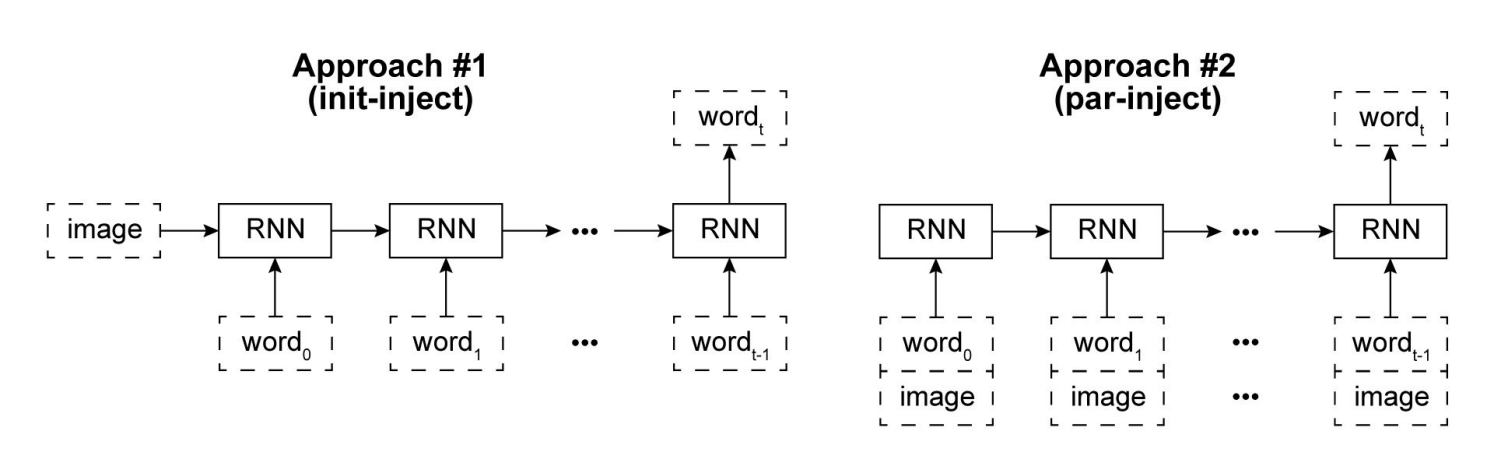
Data Exploration





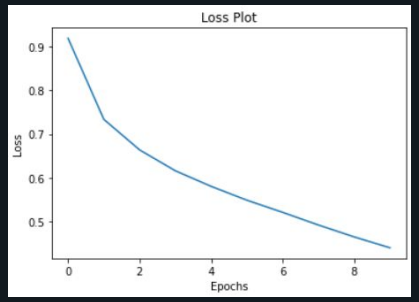
Baseline Modeling

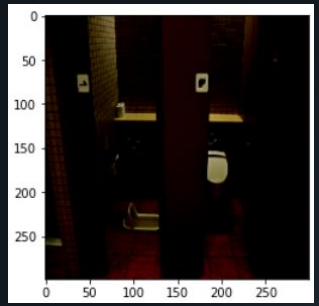
- Used convolutional neural network, InceptionV3, pretrained on the ImageNet dataset to extract high-level features from the images
- Pass features to RNN to predict the next word for each caption



Baseline Modeling

Approach #1

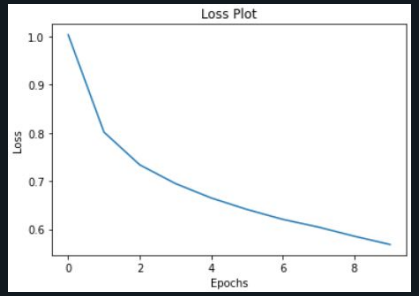


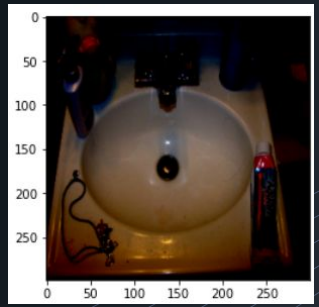


Actual: a bathroom with two different styles of toilets

Predicted: a bathroom with marble tile floor and a shower curtain

Approach #2



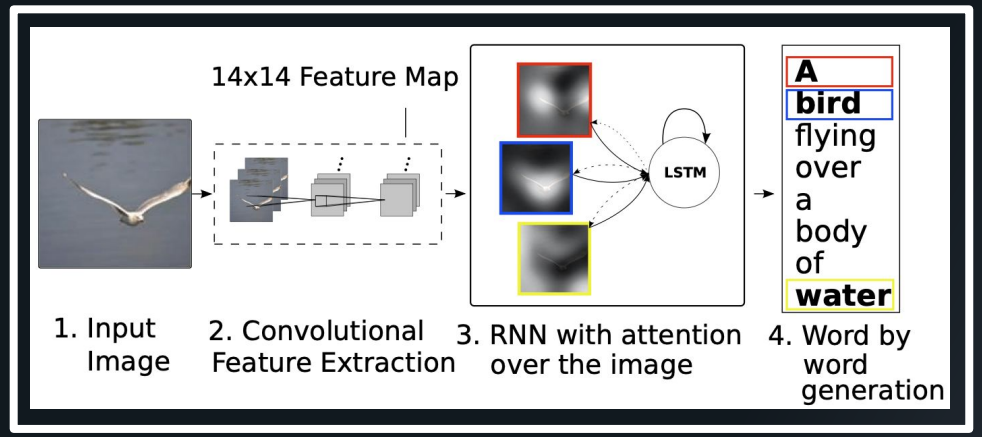


Actual: a white sink that has a necklace a rubber <unk> toothpaste and some beauty items lying around them

Predicted: a bathroom with a sink and a sink

Note: Due to limited time & computational resources, we only look at training loss. In the future we should include validation data to better assess model performance.

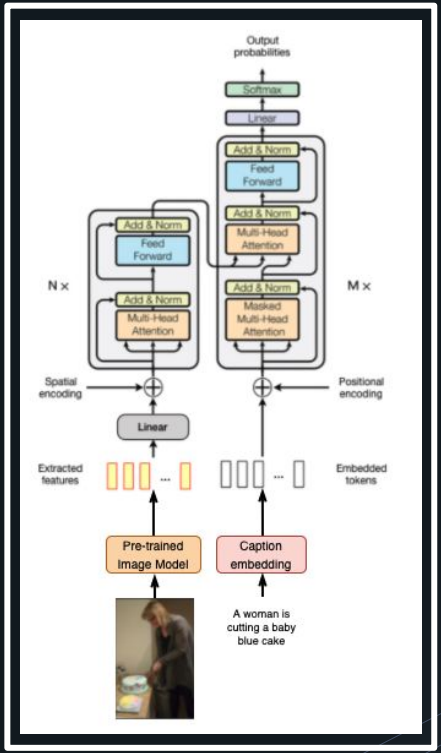
RNN-Attention



Actual: A tan and brown clock tower with sky in the background

Predicted: a clock tower next to a brick building

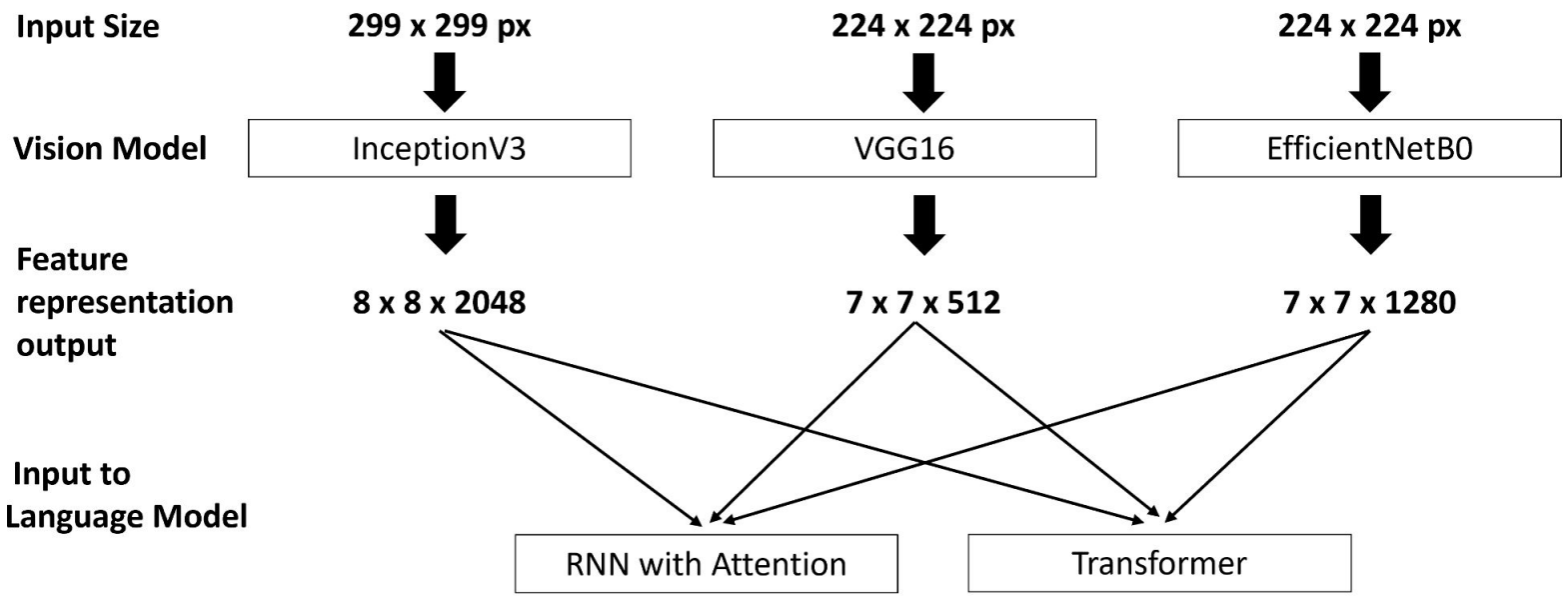
Transformer



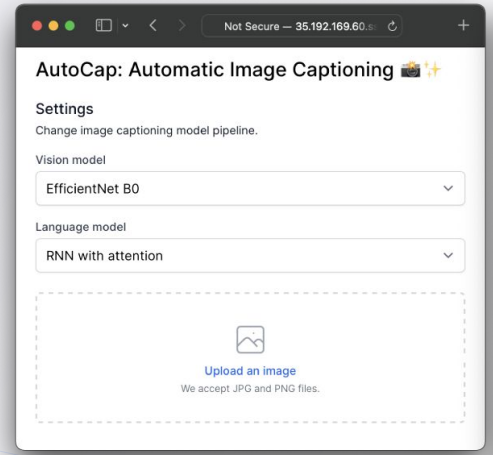
Actual: living room with a sectional couch and a coffee table

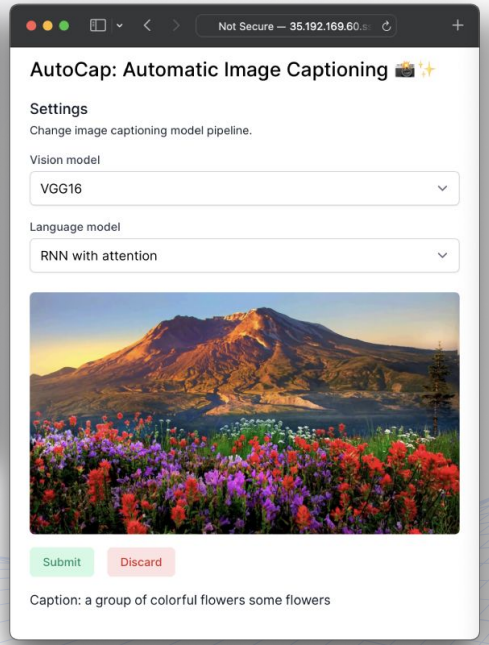
Predicted: a living room with a couch and a coffee table

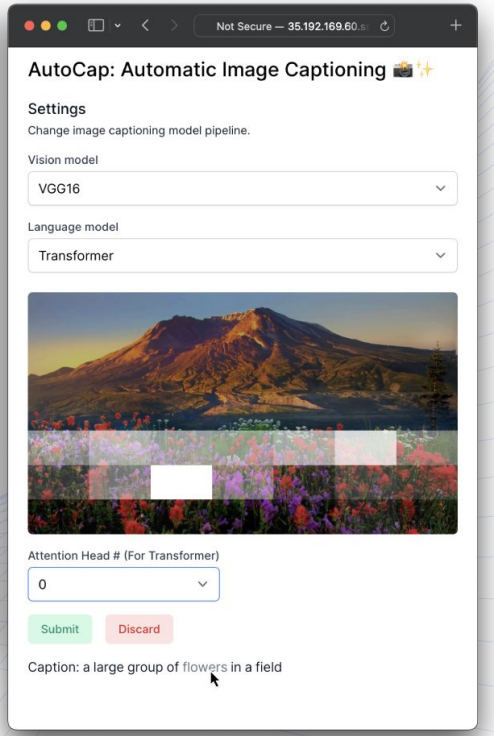
Final Model Stack



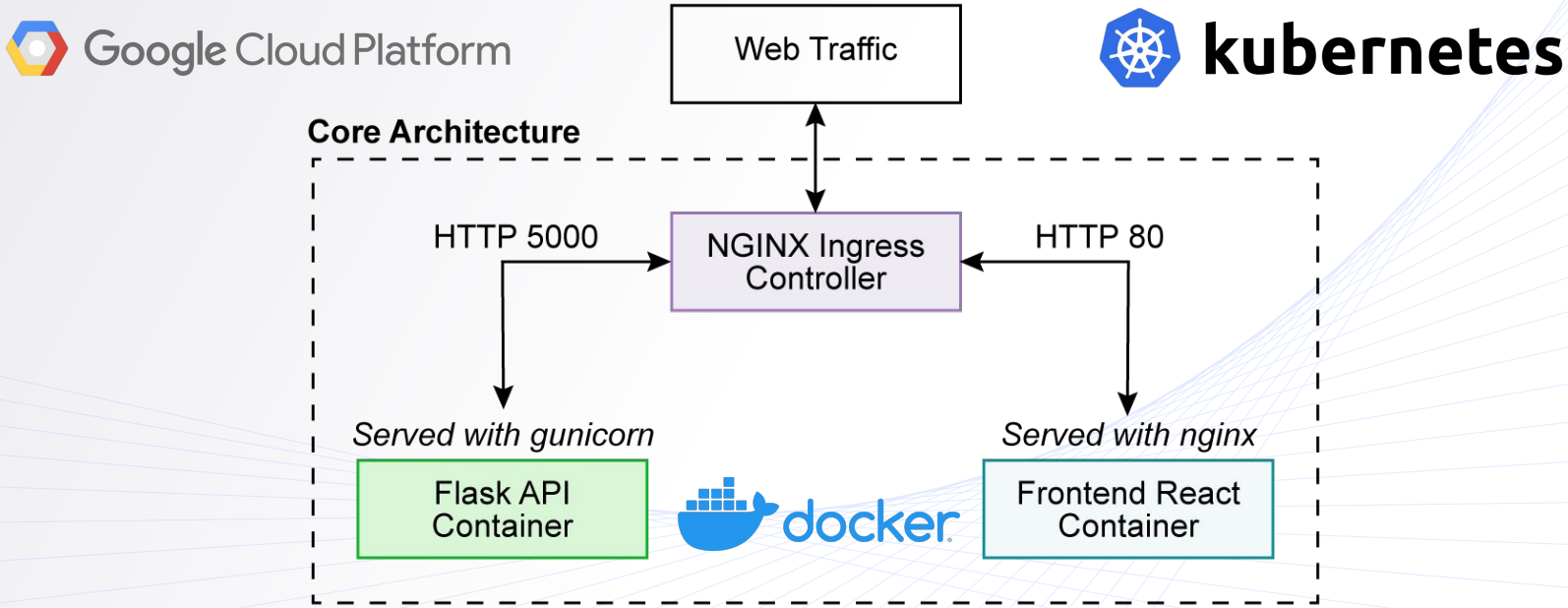
Web App


Not Secure — 35.192.169.60
AutoCap: Automatic Image Captioning
Settings
Change image captioning model pipeline.
Vision model
EfficientNet B0
Language model
RNN with attention
Upload an image
We accept JPG and PNG files.


Not Secure — 35.192.169.60
AutoCap: Automatic Image Captioning
Settings
Change image captioning model pipeline.
Vision model
VGG16
Language model
RNN with attention
Submit
Discard
Caption: a group of colorful flowers some flowers


Not Secure — 35.192.169.60
AutoCap: Automatic Image Captioning
Settings
Change image captioning model pipeline.
Vision model
VGG16
Language model
Transformer
Attention Head # (For Transformer)
0
Submit
Discard
Caption: a large group of flowers in a field

Deployment


Google Cloud Platform
Web Traffic
kubernetes
Core Architecture
HTTP 5000
NGINX Ingress Controller
HTTP 80
Served with gunicorn
Served with nginx
Flask API Container
docker
Frontend React Container

Chrome Browser Extension

The AutoCap Chrome extension interfaces with the backend server API to iterate through all images without an `alt` attribute on a webpage, upload these images, and generate image captions. Returned captions are added to the corresponding `img` tag.

Future Improvements

- Leverage pre-trained **GLoVe embeddings** to decrease model training time and improve performance
- Add additional **language models**
- Create user **feedback mechanism** to capture caption corrections; update model weekly based on updates
- Create a similar **browser extension for Safari**
- Improve **security** functionality
- Develop image tagging for **offensive content**

The Team

Kamran Ahmed

Harvard University
Program in Neuroscience

Luke Sagers

Harvard University
Computational Health Informatics

Brendan O'Leary

Harvard University
School of Engineering

James Parker

Harvard Business School
Harvard School of Engineering