Caption This - Group BKKST

Advanced Practical Data Science, MLOps

AC295

Team Members: Al-Muataz Khalil, Ed Bayes,
Stephen Knapp, Matthew Stewart, Shih-Yi Tseng

Outline

- Project Scope
- Project Workflow
- Process Flow
- Data
- Models

Problem Definition

The World Health Organization (WHO) estimated that 314 million people have visual impairment across the world, including 269 million who have low vision, and 45 million who are blind (Ono et al 2010). Many people with visual impairments rely on screen readers in order to access the internet through audio, and thus depend on image captions (Yesilada et al 2004). Therefore, accessibility, as well as automatic indexing and other goals, make accurate image captioning an important priority (Hossain et al 2018).

Proposed Solution

We will design, build, and deploy at-scale an application which receives an image of an everyday activity which then assigns a caption of the image contents, based on state-of-art computer vision and natural language models. Additionally, the app will provide a visualization of the image components reflected in the caption through saliency maps.

Project Scope

Proof Of Concept (POC)

- Set up CI/CD pipeline
- Store Flickr8k and COCO datasets in GCP bucket
- Conduct image feature extraction and EDA
- Test baseline model (efficientB0 net)
- Verify models predict labels for unseen photos
- Visualize image components through saliency maps

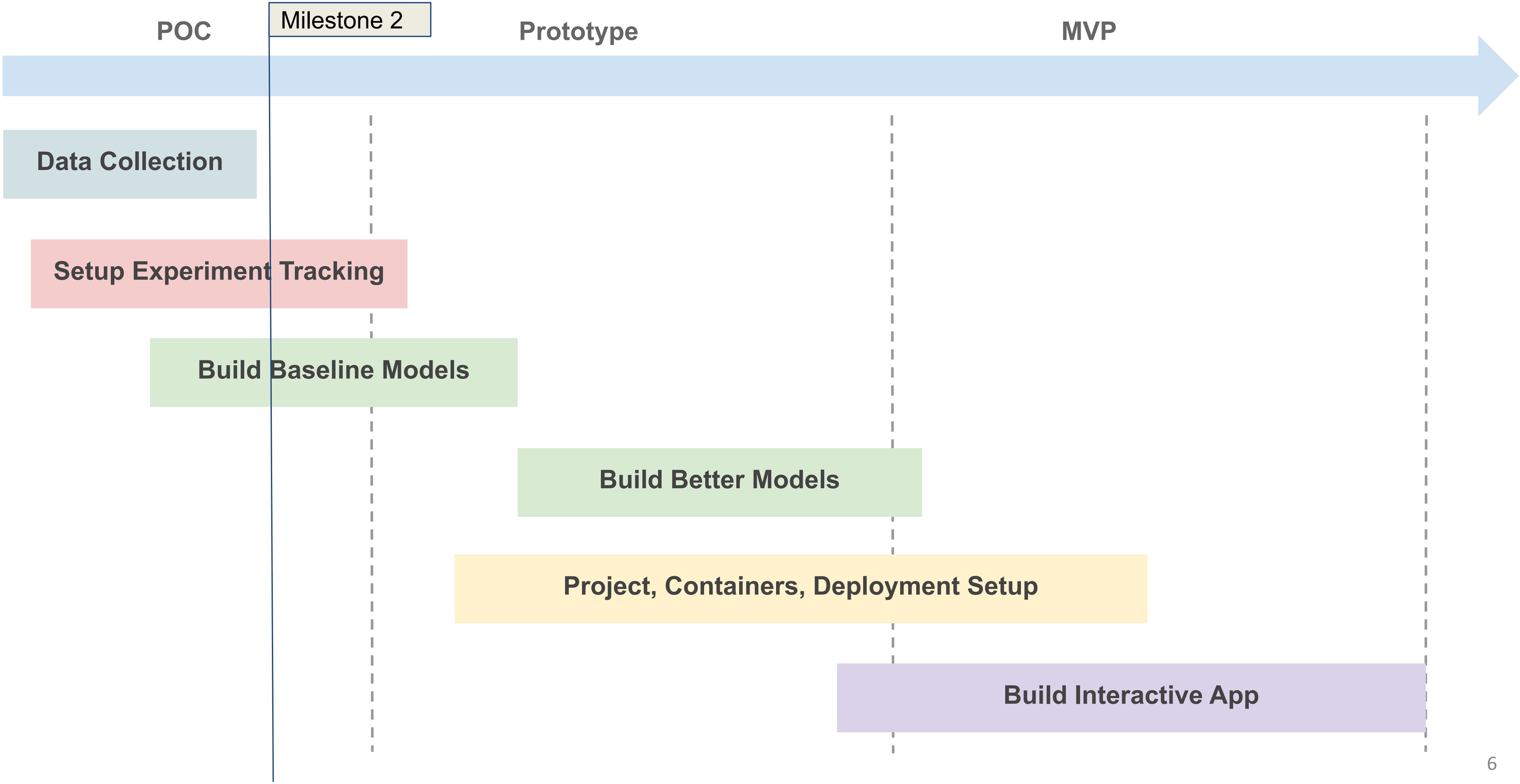
Prototype

- Create ‘looks like’ mockup of UX using figma
- Deploy one model to Fast API to service model predictions as an API

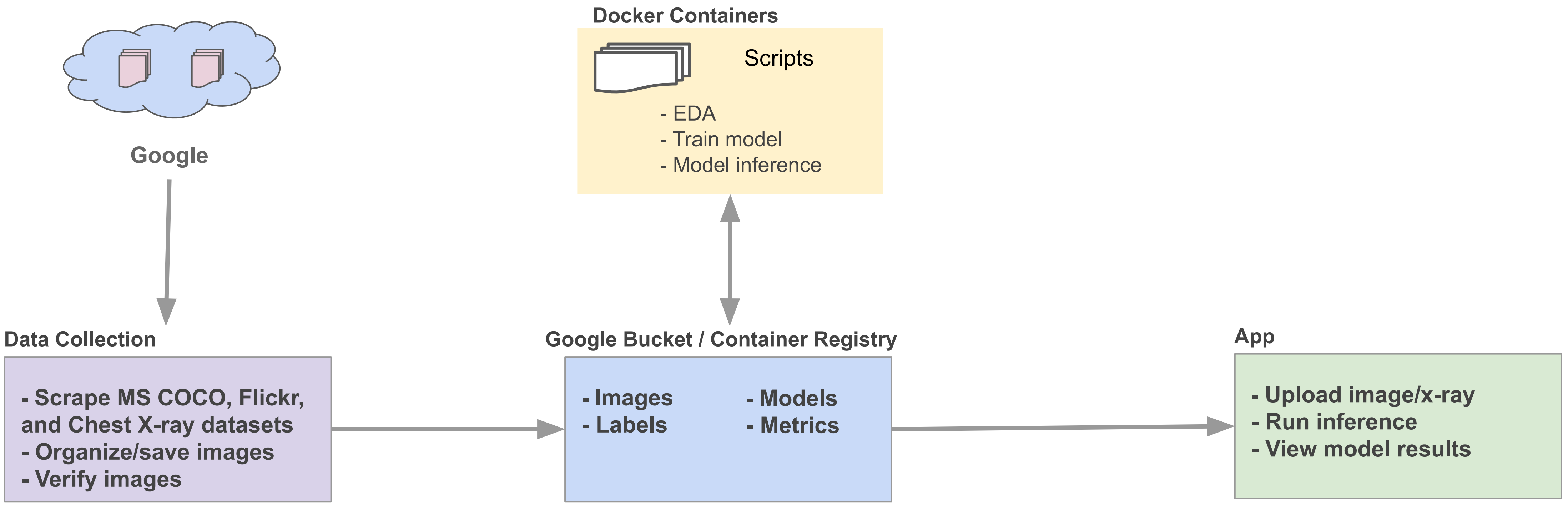
Minimum Viable Product (MVP)

- Create App that labels unseen photos
- API Server for uploading images and predicting using best model

Project Workflow



Process Flow



Google

Docker Containers

Scripts

- EDA
- Train model
- Model inference

Data Collection

- Scrape MS COCO, Flickr, and Chest X-ray datasets
- Organize/save images
- Verify images

Google Bucket / Container Registry

- Images
- Labels
- Models
- Metrics

App

- Upload image/x-ray
- Run inference
- View model results

Data

- Public Google bucket (link here) containing MS COCO and Flickr 8K datasets used during this project.
- **Flickr 8K:** 8,091 images from one of six categories, each with 5 corresponding image captions.
- **MS COCO (2014):** 164K images split into training (83K), validation (41K) and test (41K) sets.
- Both datasets are standardized datasets used for benchmarking and released under CC0 license (public domain).
- **(Tentative) Chest X-Rays Images and Reports:** 1,000 chest x-rays and XML medical reports from the Indiana University hospital network (also released under CC-BY-NC-ND 4.0 license).

Data

A brown dog in two black collars running through a grassy field .

a small brown and black dog lying down in a furry rug .

A man on the street standing by his bicycle .

Friends and family dance on a beach by their vehicles .

A man feels on top of the world on top of a large rock formation .

A dog leaps into the air to catch a ball in its mouth .

A dog leaps into the air to catch a ball in its mouth .

Some children watching fish in a pool .

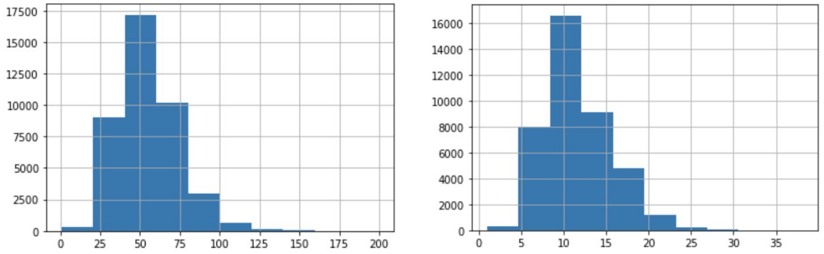
Two gray dogs jump at each other over the tall grass .

Data Details

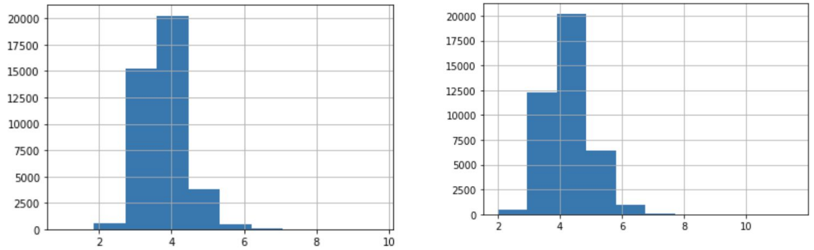
- Total number of images: 8091
- Label counts: 40455
- Images details:
    - Width range: 164-500 px
    - Height range: 127-500 px
    - Memory: 4113.97 MB

Caption Data Analysis

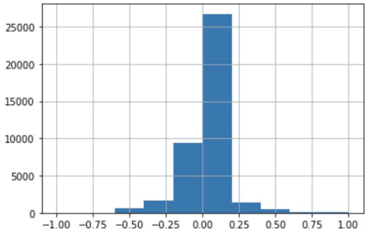
Captions are generally between 25 and 100 characters and 5 to 20 words



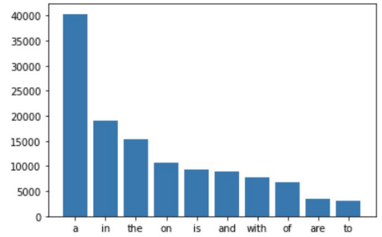
with an average word length of 3 to 4 characters (with and without stopwords)



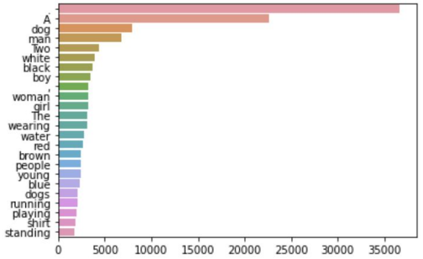
the majority of the sentiment polarity scores cluster around zero meaning most are pretty neutral



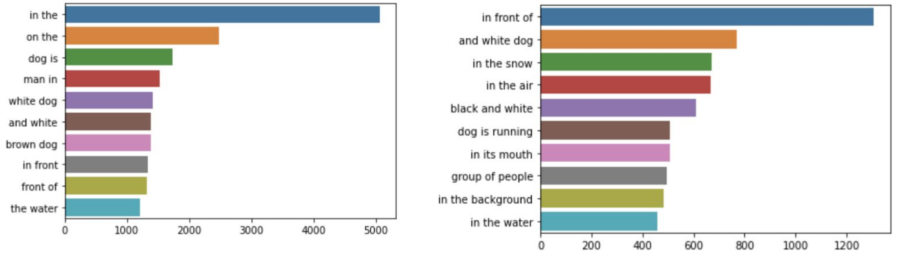
Breaking down captions into individual words top stop words are



and plotting lemmatized words removing stopwords shows the top words are 'dog' and 'man'

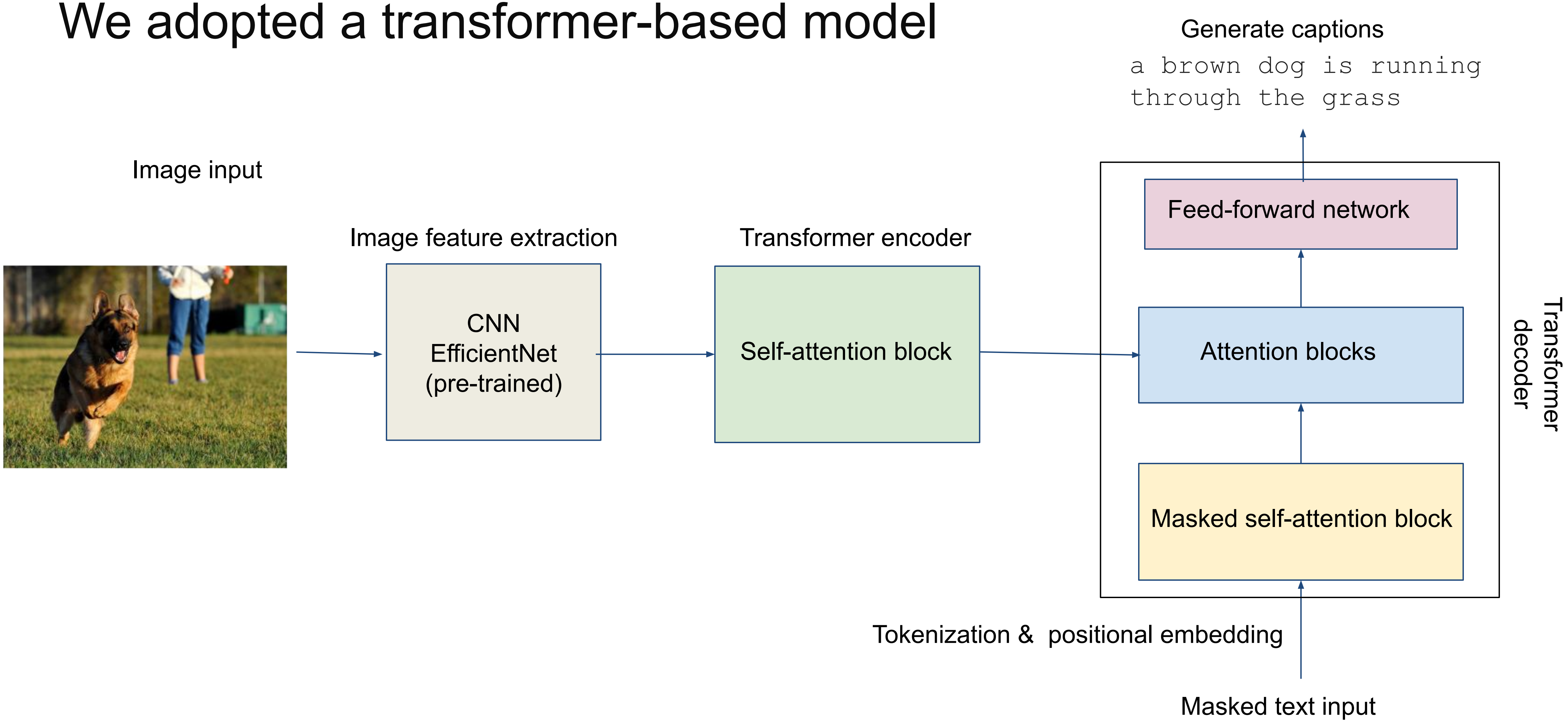


using ngram exploration, we can look at which pairs of words come up the most, and which trigrams



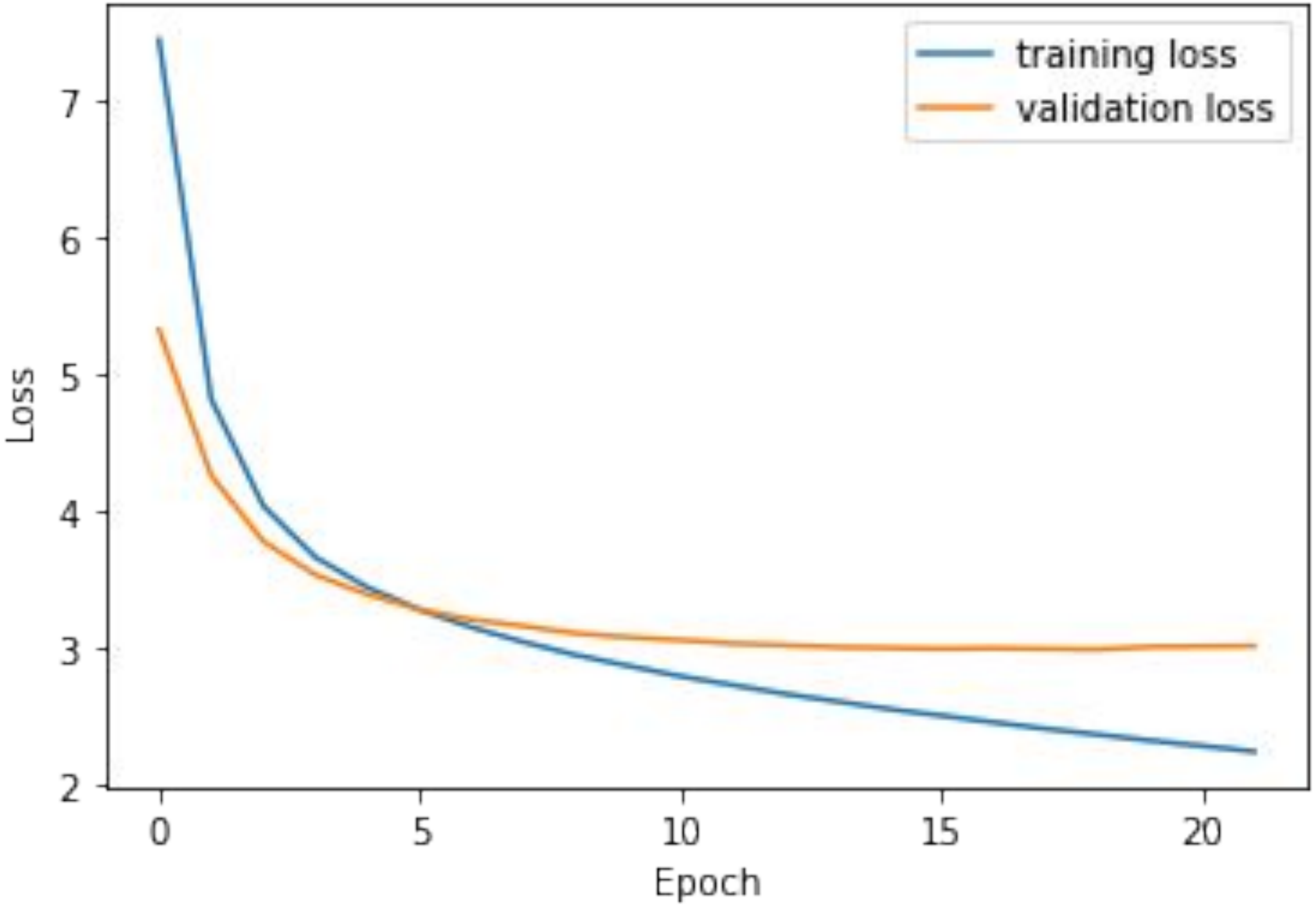
Baseline Model

We adopted a transformer-based model



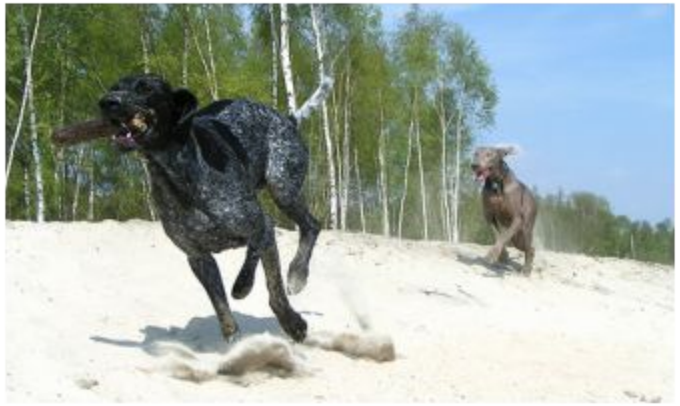
Models - Training Progress

Trained on 72% of Flickr8k data, validated on 18% (10% saved as test data)

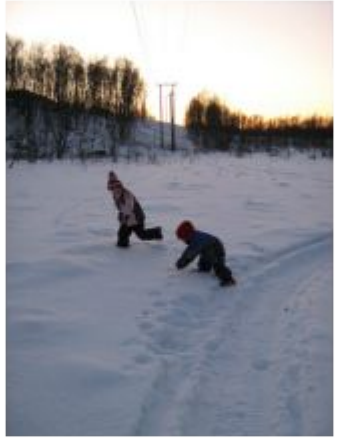


Test Results

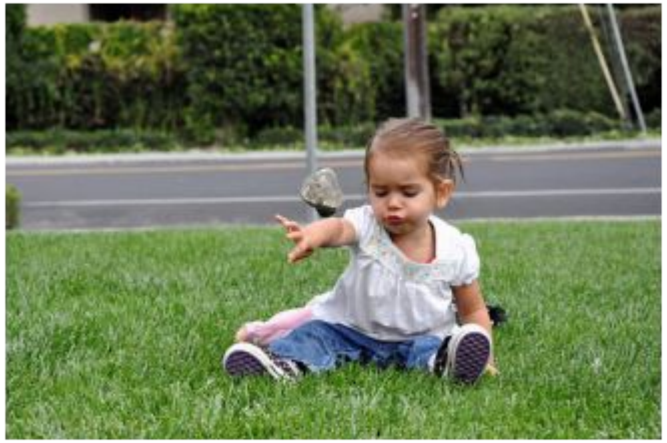
Generate captions on test images



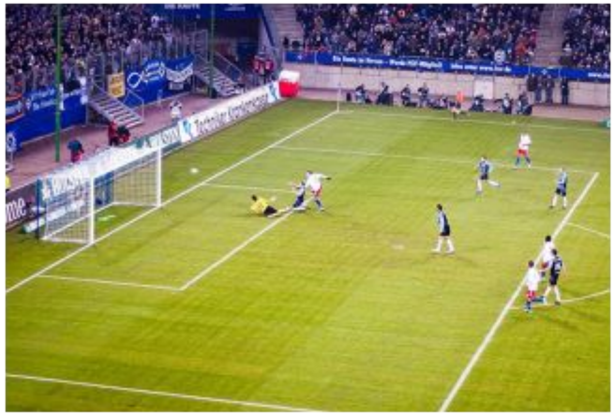
a black dog is running on the beach



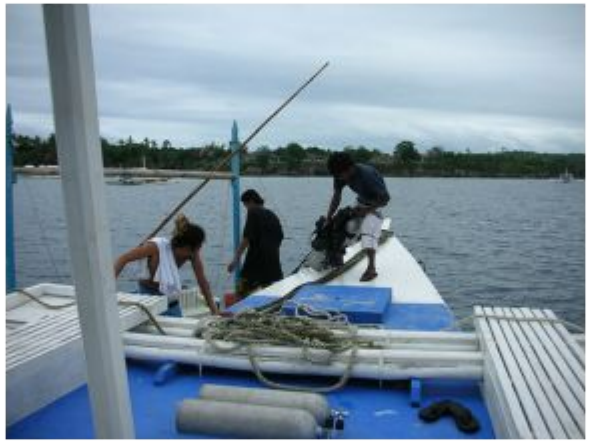
two children playing in the snow



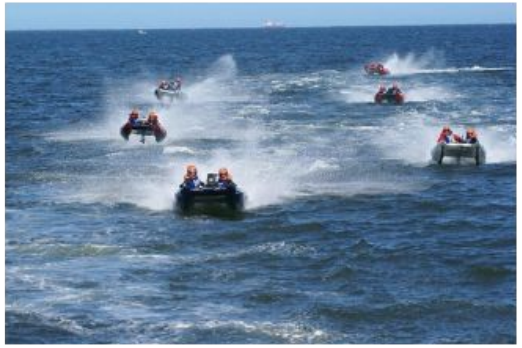
a little boy is playing with a toy on the grass



a soccer player is running on the field



a man is sitting on a wooden dock near a lake



a red boat is being ridden over a wave

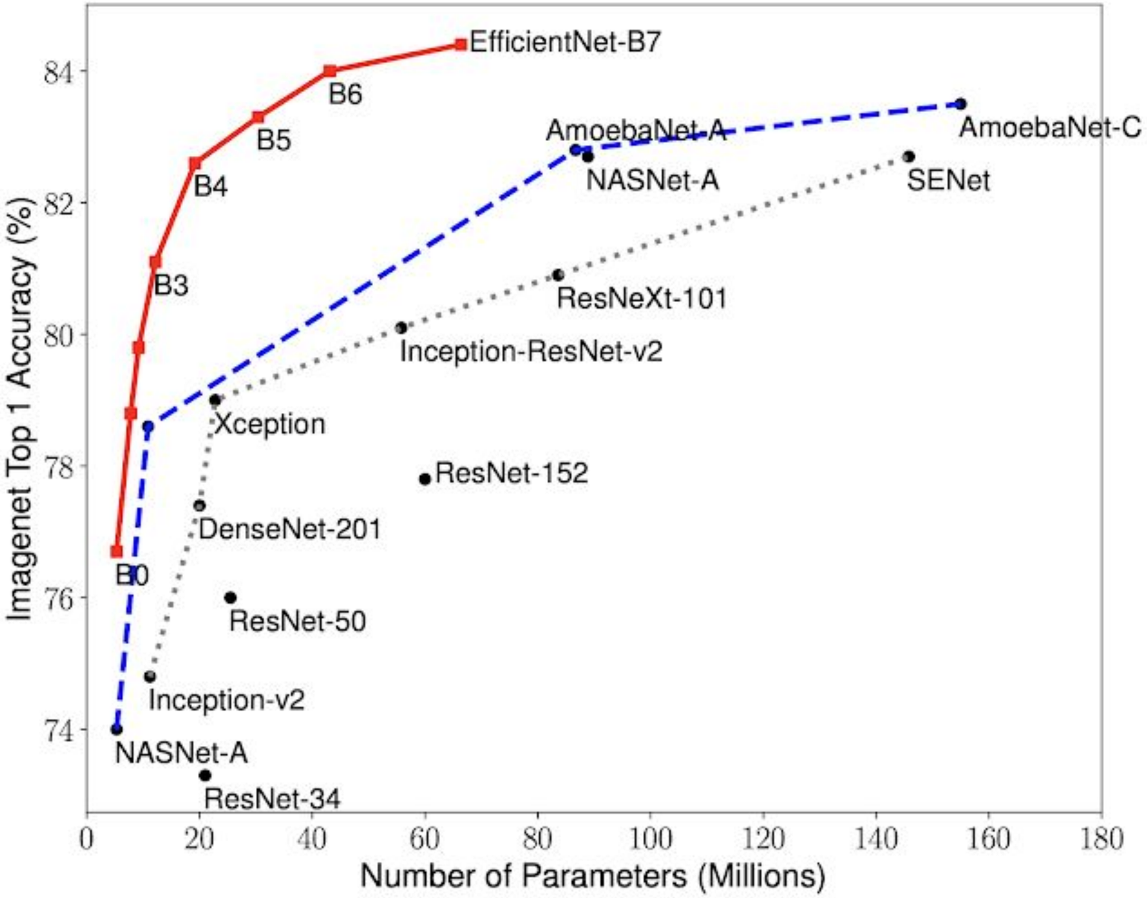
Future Improvements

Image Processing and Object Detector

- EfficientNet B4 to B7, EfficientDet-B7
- Using pretained models on larger datasets
- More data (ImageNet-21k, Flickr8k, MSCOCO)

Image Captioning Layers

- Pretrained natural language decoders (GPT2)



Supporting Notebooks in Repo

- EDA: https://github.com/skgithub14/AC215_KKST/blob/main/notebooks/EDA.ipynb
- Baseline Model: https://github.com/skgithub14/AC215_KKST/blob/main/notebooks/Image_captioning_with_RNN_SYT.ipynb
- Combined: https://github.com/skgithub14/AC215_KKST/blob/main/submissions/milestone2_KKST/Milestone2_EDA_with_baseline_models.ipynb