INNOVATION
IN SPACE
AND DEFENCE

IMAGE CAPTIONING OF EARTH OBSERVATION IMAGERY

MDS-MDA JOINT CAPSTONE PROJECT

Dora Qian, Fanli Zhou, James Huang, Mike Chen

MDA

- A Canadian aerospace company
  - Developed Canadarm and Canadarm2
- Access to a vast database of satellite images
  - These images are uncaptioned

Motivation and Goals

- Associate an image with a caption makes it accessible
  - Sort images based on content
  - Return queries
  - Evaluate similarity
- Develop a pipeline
  - Generate a model from raw data
- Create a visualization tool
  - Aim to help users interact with model and data

Objectives

- Data Processing
  - Transform and store data in a well defined and reproducible structure
- Creating the Model
  - Extract features from image
  - Generate sentence from features
- Evaluating the Model
  - N-gram based and semantic based metrics
- Visualization Tool
  - Caption and upload new images to database
  - View previously generated image-caption pairs and evaluation scores

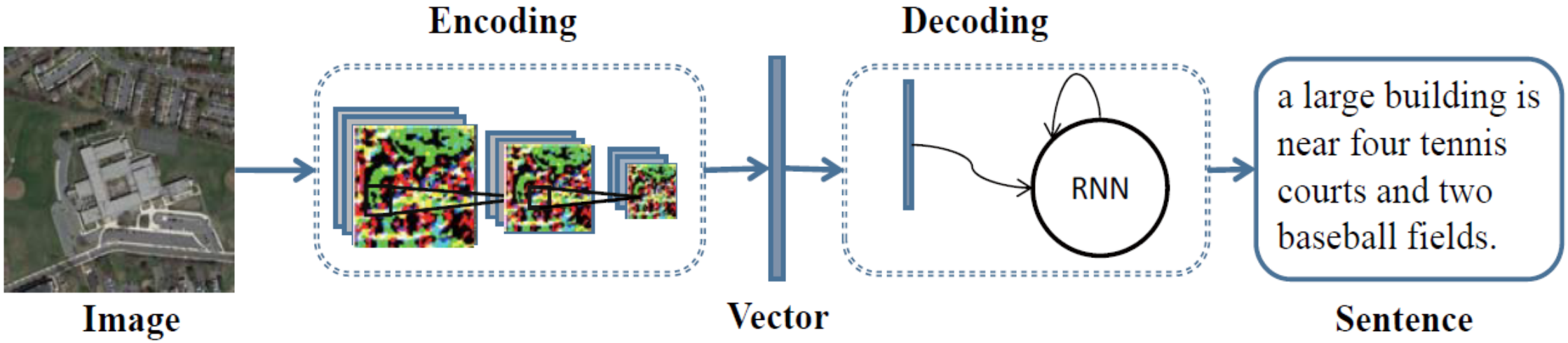
Data Description

- There are three labeled datasets:
  - UCM_Captions
  - RSICD (Remote Sensing Imaging Captioning Dataset)
  - Sydney_Captions
- Total 13,634 Images
- Train/valid/test on UCM + RSICD
- Test on Sydney for generalization ability

1. Four planes are stopped on the open space between the parking lot.
2. Four white planes are between two white buildings.
3. Some cars and two buildings are near four planes.
4. Four planes are parked next to two buildings on an airport.
5. Four white planes are between two white buildings.

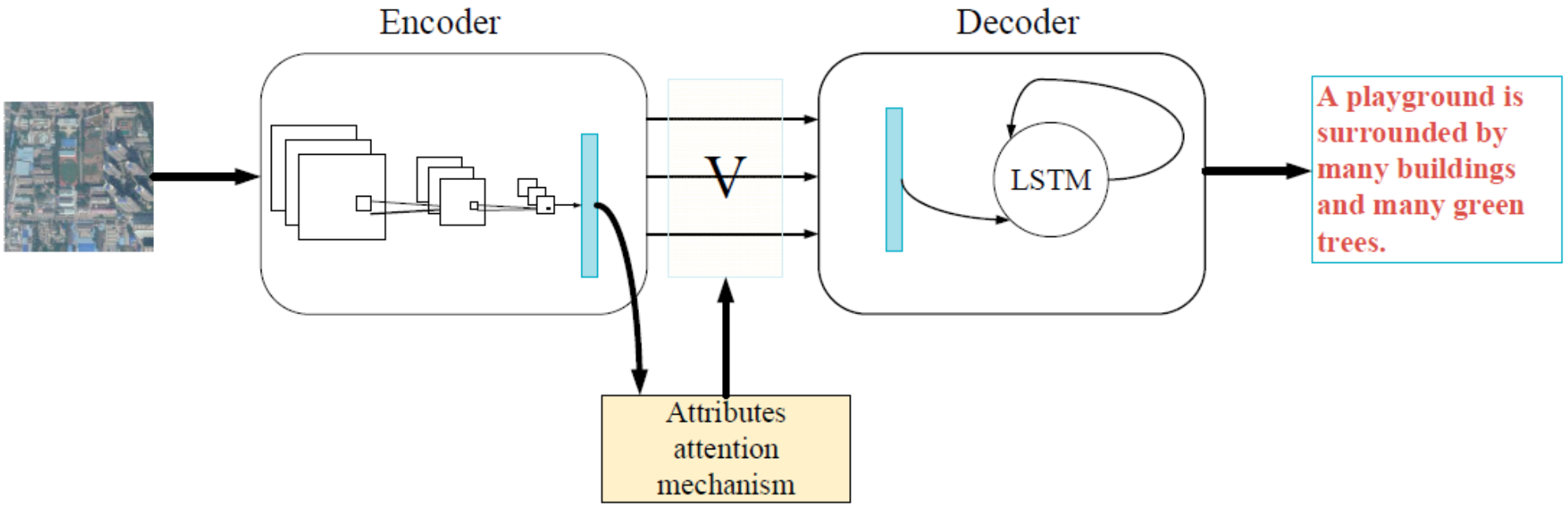
Data Science Techniques

Baseline Model Structure (RNN + LSTM)



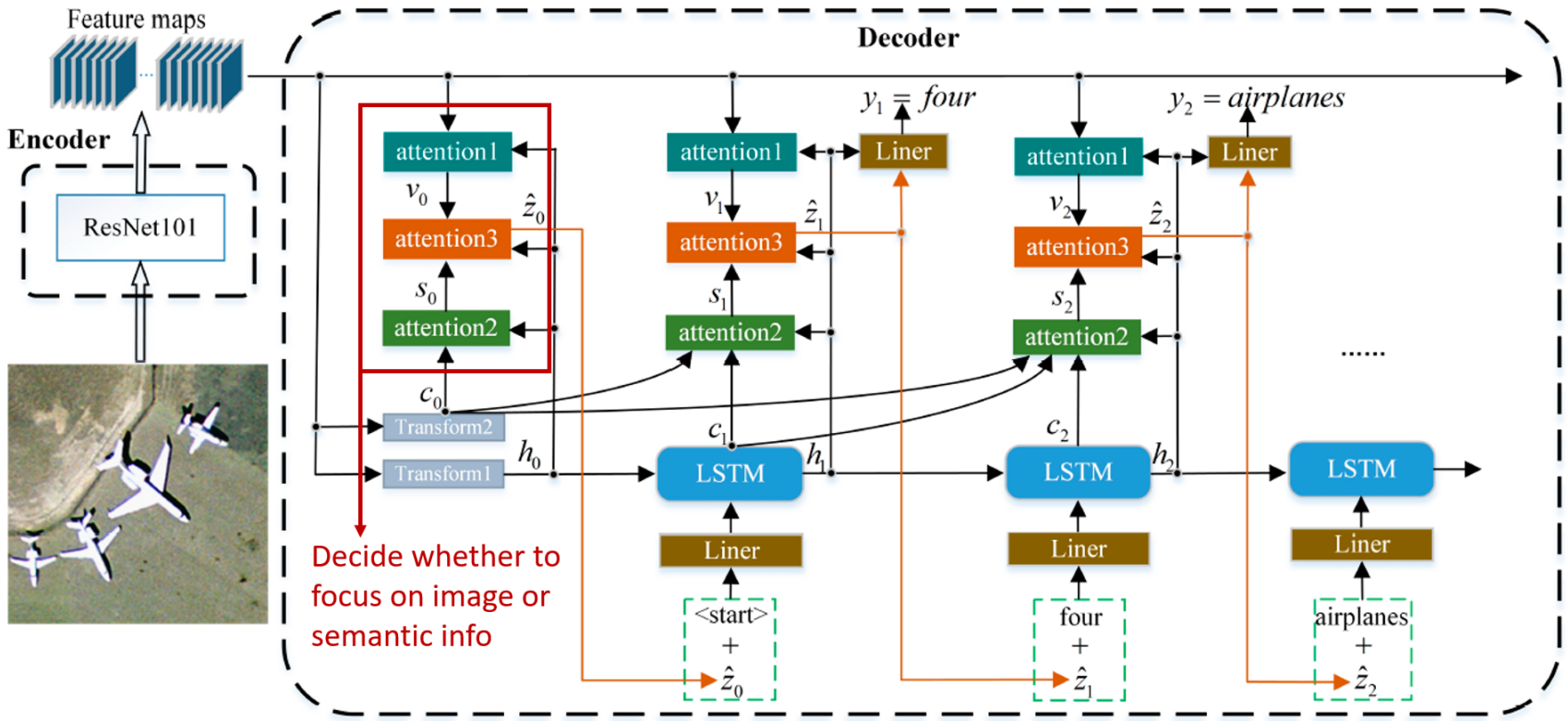
Data Science Techniques

Attention Model Structure (RNN + Attention + LSTM)



Data Science Techniques

Multi Attention Model Structure (RNN + Multi-Attention+ LSTM)



Data Science Techniques: Transfer Learning

1. Pre-trained CNN:
   - InceptionV3
   - Vgg16
   - …
2. Pre-trained embeddings weights :
   - GloVe (200d)
   - Wikipedia2Vec (500d)

- Pros:
  - Good performance
  - Simple to incorporate
  - Reduces training time
  - …
- Cons:
  - Performance depends on task similarity
  - …

Transfer Learning vs. Learning from scratch

1. Pre-trained CNN model performed better than CNN model learned from scratch

2. Pre-trained embeddings weights performed better than embeddings weights learned from scratch

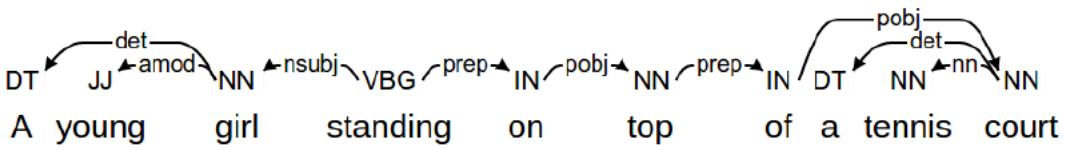
3. We decided to use pre-trained CNN model and embeddings weights

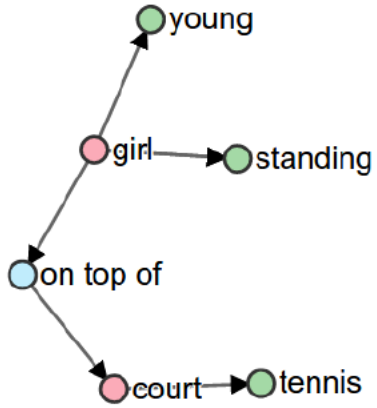
Data Science Techniques: Evaluation Metrics

- Total 9 evaluation metrics
- N-gram based metrics
  - Bleu 1-4
  - Rouge L
  - Meteror
  - CIDEr
  - Commonly used in the community and research papers

Data Science Techniques: : Evaluation Metrics

- Semantic-based metrics:
  - SPICE: semantic scene graph
  - Universal Sentence Encoder Similarity





Evaluation scores from the best model of each structure

Test data and train data are from the **same** datasets

	n-gram Comparison Evaluation							Semantic Evaluation	
MODEL	Bleu 1	Bleu 2	Bleu 3	Bleu 4	METEOR	ROUGE L	CIDEr	SPICE	USC Similarity
Baseline	0.648	0.523	0.440	0.381	0.300	0.553	2.125	0.400	0.612
Attention	0.572	0.435	0.351	0.294	0.256	0.473	1.540	0.324	0.550
Multi-Attention	0.593	0.463	0.340	0.321	0.271	0.498	1.738	0.345	0.583

Evaluation scores from the best model of each structure

Testing Model Generalization

Test data and train data are from **different** datasets

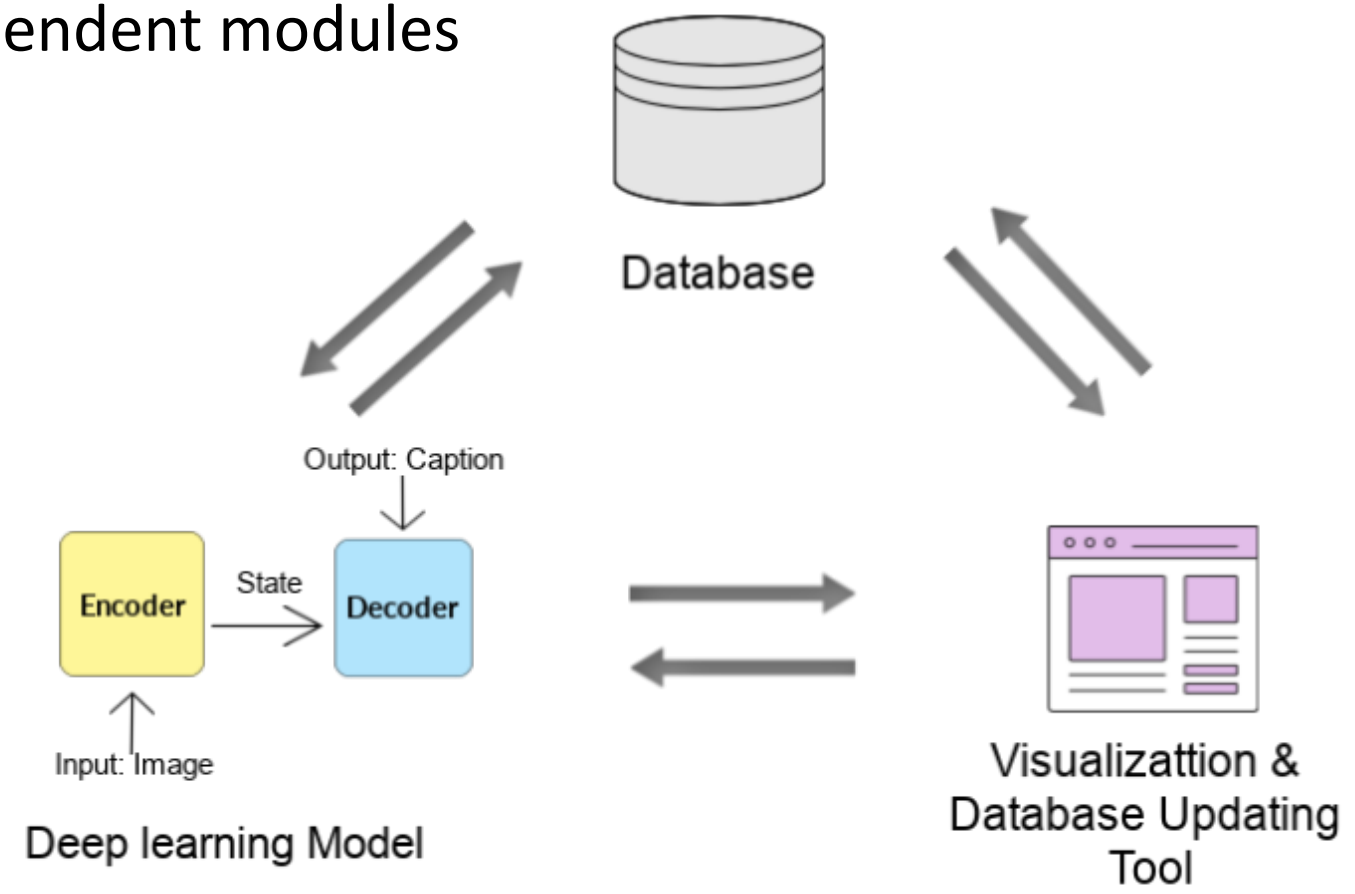
	n-gram Comparison Evaluation							Semantic Evaluation	
MODEL	Bleu 1	Bleu 2	Bleu 3	Bleu 4	METEOR	ROUGE L	CIDEr	SPICE	USC Similarity
Baseline	0.453	0.220	0.117	0.0717	0.145	0.290	0.210	0.119	0.458
Attention	0.431	0.209	0.108	0.0687	0.140	0.280	0.146	0.114	0.449
Multi-Attention	0.431	0.194	0.0778	0.0343	0.133	0.270	0.144	0.0965	0.450

Future Improvements

1. Optimizing hyperparameters
2. Finetuning the pre-trained CNN
3. Extracting features from different convolutional layers
4. Improving attention structures

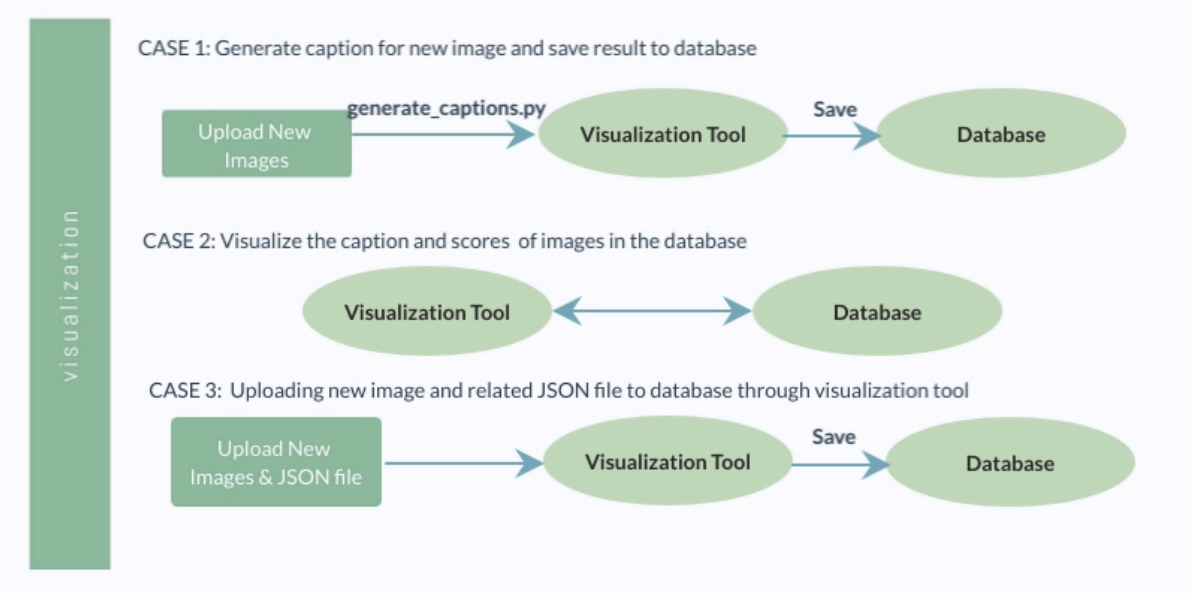
Final Data Product

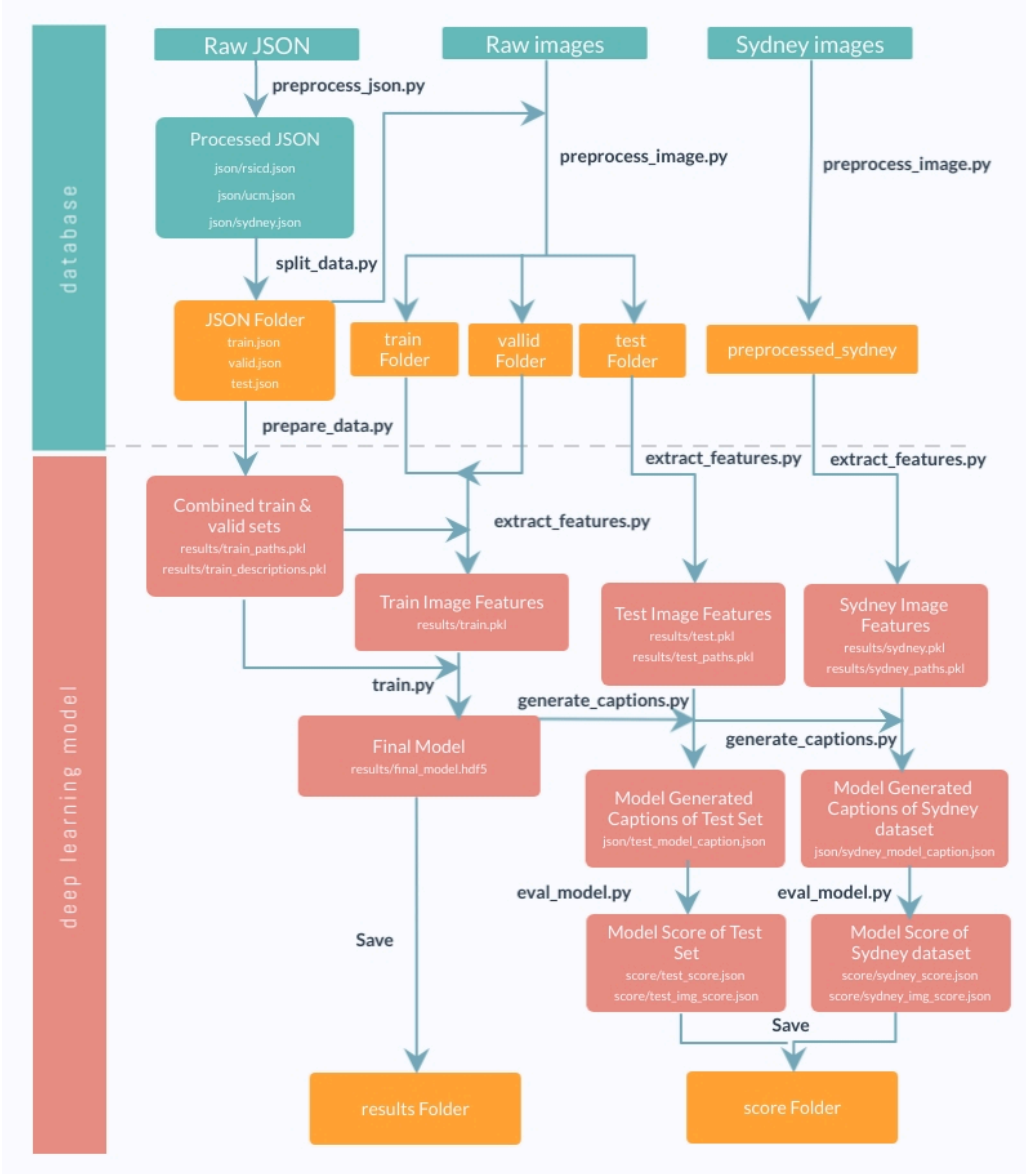
- Complete image captioning pipeline
- 3 independent modules



Product Pipeline

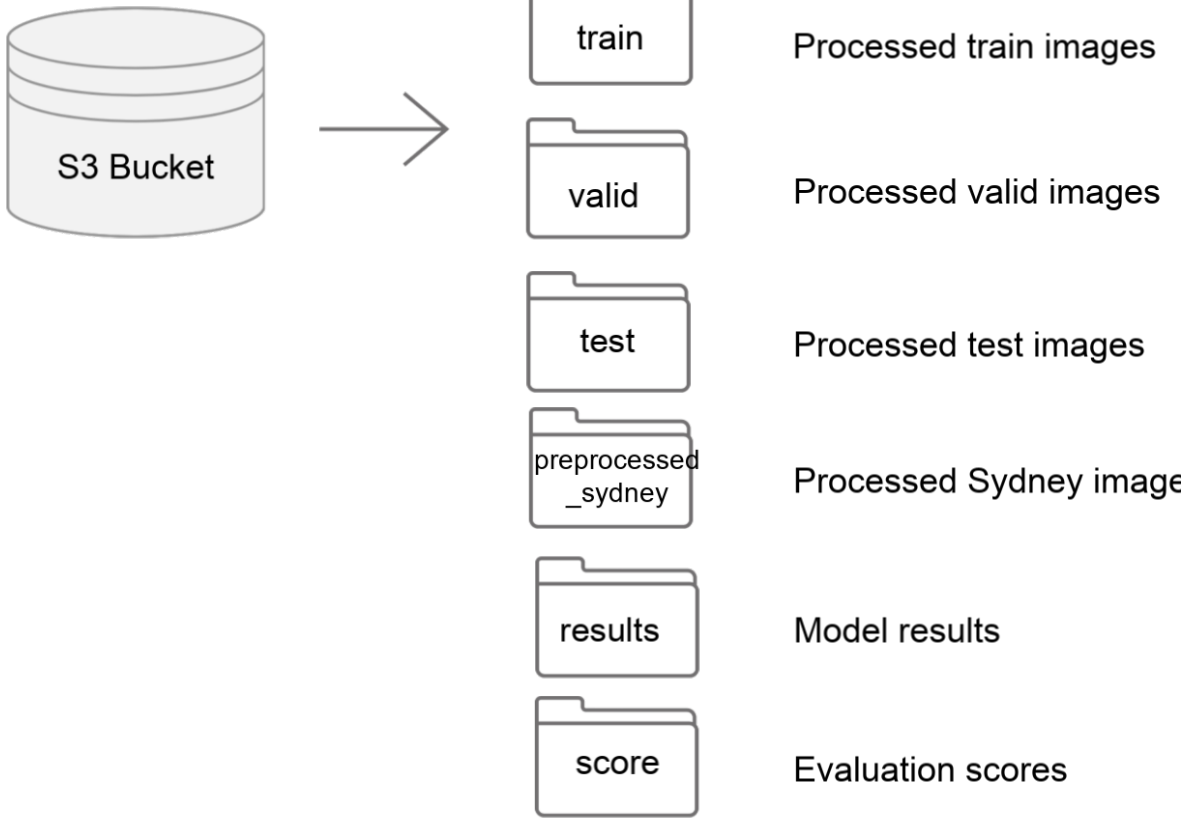
- Database & Model: GNU Make
- Visualization tool: Django





Final Data Product: Database

- AWS S3 bucket
- Advantages:
  - Integrate well with AWS instance
  - Scalability
  - Easy to use



Final Data Product: Deep Learning Model

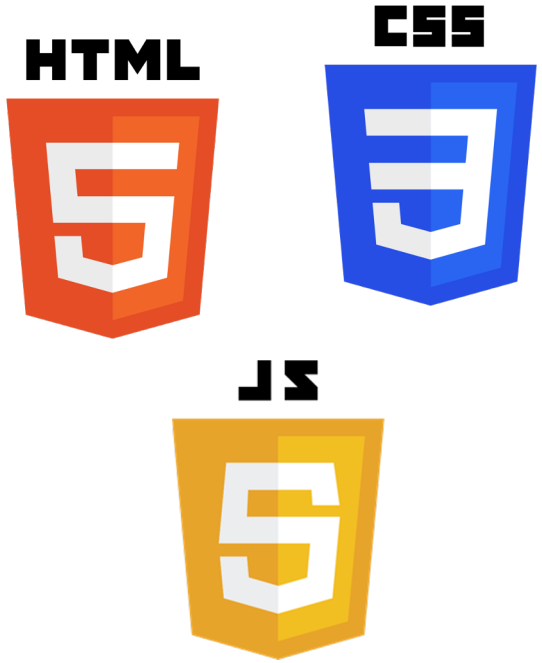
- AWS EC2 P3 instance
- Final model: Baseline model with VGG 16 & Glove Embedding
  - Train
  - Generate Caption
  - Evaluate
- Save trained model, model results and score back to database

Final Data Product: Visualization Tool

Client's Needs

- A visualization tool capable of:
  - Generate caption for new uploaded images
  - Allow users to submit their own captions for the image
  - Showcase the evaluation metrics
  - Allow users to upload multiple images with a json caption file

Final Data Product: Visualization Tool

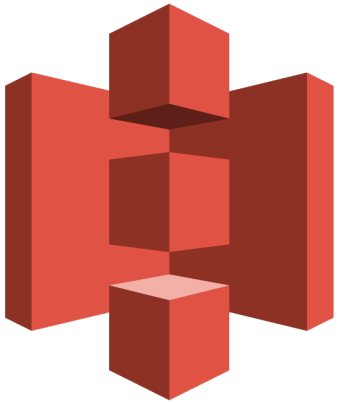


Front End

HTML, CSS, JavaScript

Back End

Django - Python based web framework



Database

AWS S3: To store all the images and caption

json file

Visualization Tool Showcasing

- Demo for showcasing
- Insert screenshot for submission later

Conclusion

- We have been successful in creating a functional data pipeline and visualization tool
  - Our goals were met and all the features we aimed to create are available
  - Performance of the model is fair
- We hope that MDA can iterate and improve upon our work

Limitations

- Relatively small dataset
  - Poor performance of CNN feature extraction models trained from scratch
  - Using pre-trained model trained on ImageNet type images
- Attention model did not perform as expected
  - The captions were not significantly improved and was at times worse
- Short project time
  - Lacked time to add or fine tune more layers

Recommendations

- Explore training a CNN feature extraction model from scratch
  - Use much larger captioned satellite image datasets found online
- Fix or refine attention model
  - A well implemented attention model should yield better results
- Add or fine tune model layers
- More comprehensive cross-dataset performance evaluation

Thank you

Questions?