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

Sources: Canadarm, Canadarm2

Data Science Problem and Motivation

- Access to a vast database of uncaptioned satellite images
  - We aim to caption these images
- Associating an image with a caption makes it accessible
  - Sort images based on content
  - Return queries
  - Evaluate similarity

Sources: Image adapted RSICD optimal dataset

Overall Goal

- Develop a pipeline
  - Given a set of labelled satellite images, develop an end-to-end image captioning system.
- Create a visualization tool
  - Aim to help users interact with model and data

Specific Objectives

- Data preprocessing
  - Transform and store data in a well-defined and reproducible structure
- Model development
  - Extract features from image
  - Generate sentence from features
- Model evaluation
  - Apply n-gram based and semantic similarity based evaluation metrics
- Visualization
  - Generate captions for unseen user uploaded images, and upload results to the database
  - View previously generated image/caption pairs and evaluation scores

Data Description

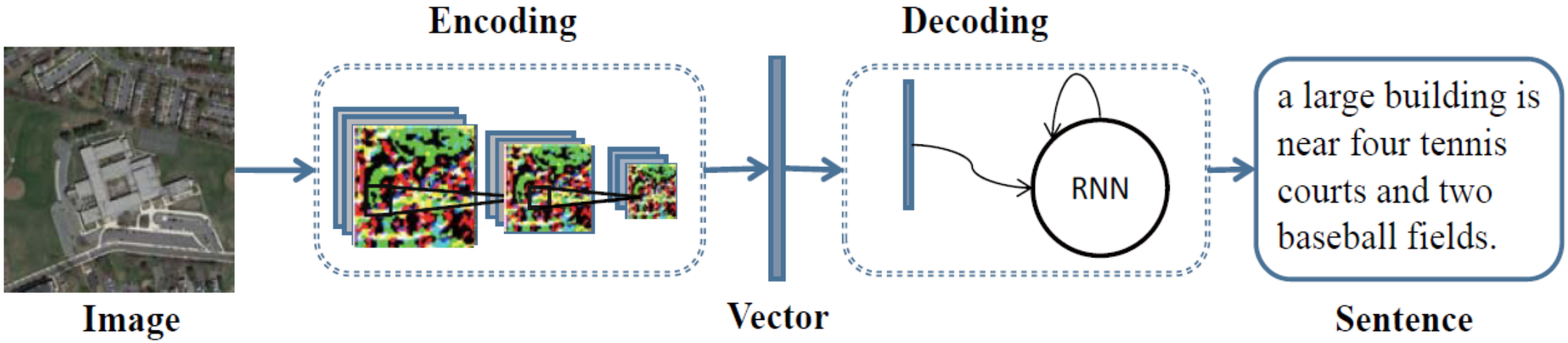
- There are three labeled datasets:
  - UCM_Captions
  - RSICD (Remote Sensing Imaging Captioning Dataset)
  - Sydney_Captions
- Total 13,634 Images
- Train/valid/test on UCM_Captions and RSICD
- Test on Sydney_Captions for generalization ability

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.

Sources: Image adapted from Lu, X. et al. (2018) [1]

Data Science Techniques

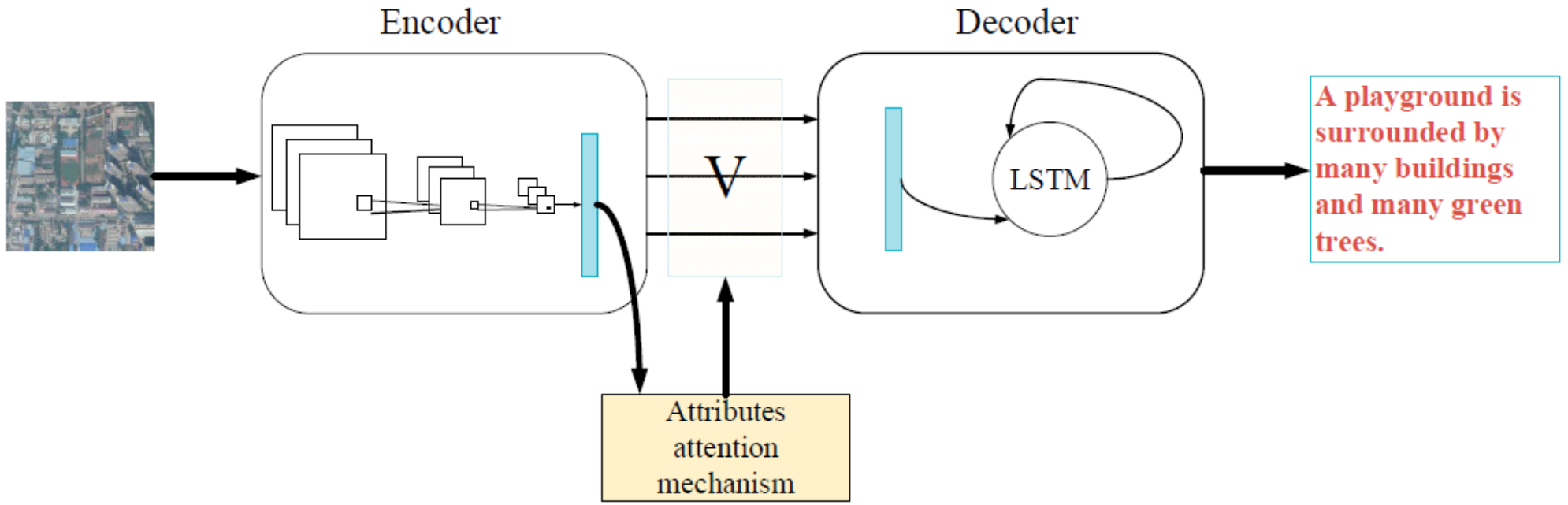
Baseline Model Architecture (CNN + LSTM)



Sources: Image adapted from Lu, X. et al. (2018) [2]

Data Science Techniques

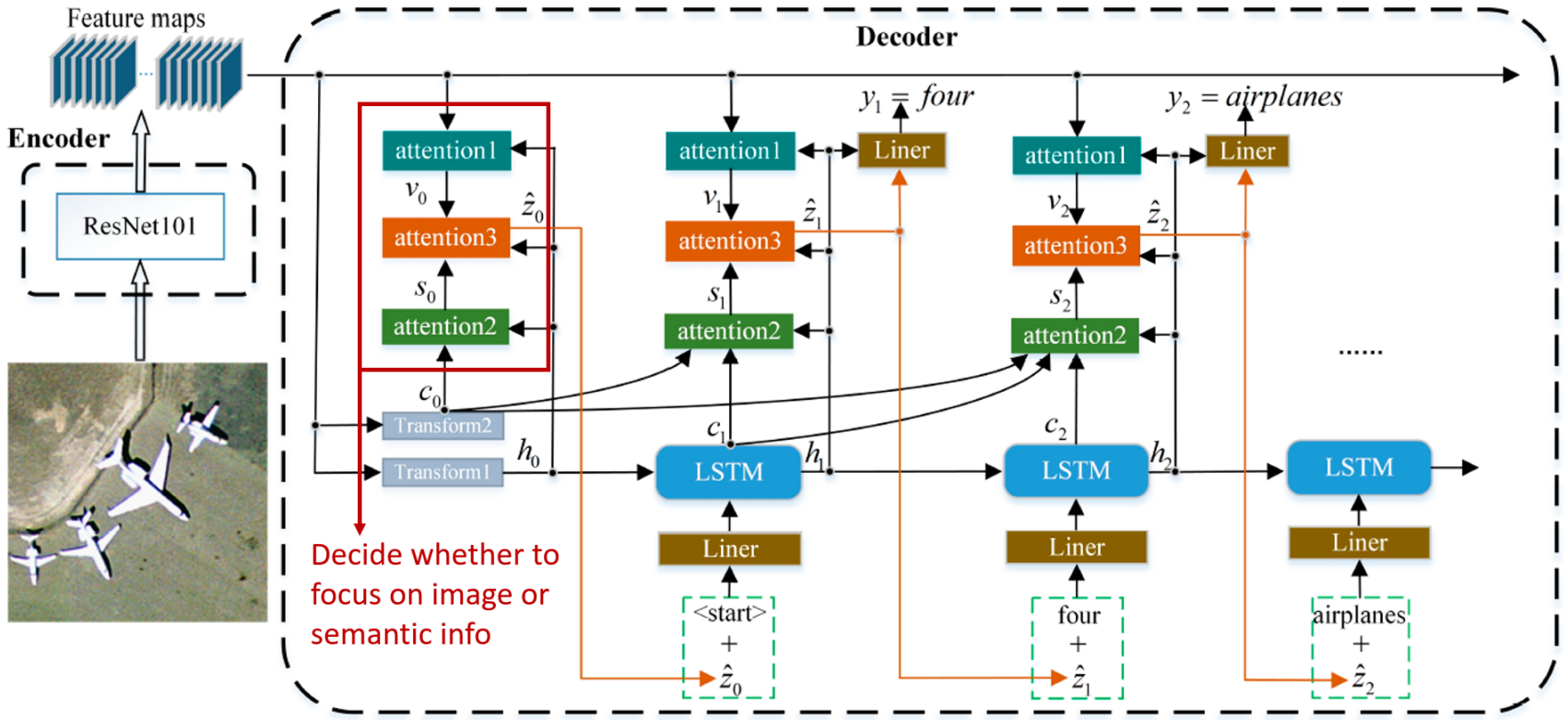
Attention Model Architecture (CNN + Attention + LSTM)



Sources: Image adapted from Zhang, X. et al. (2019) [3]

Data Science Techniques

Multi Attention Model Architecture (CNN + Multi-Attention+ LSTM)



Sources: Image adapted from Li, Y. et al. (2020) [4]

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
  - …

Data Science Techniques: Evaluation Metrics

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

Data Science Techniques: Evaluation Metrics

Semantic-based metrics:

- Universal Sentence Encoder Similarity
- SPICE: semantic scene graph

N-gram based metrics:

- BLEU-1 score = 0.74
- BLEU-2 score = 0.46
- BLEU-3 score = 0
- BLEU-4 score = 0

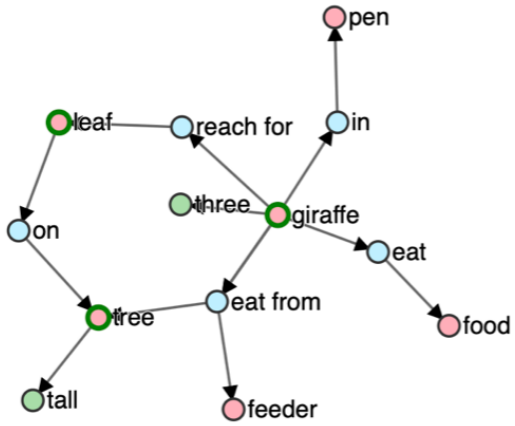
Semantic based metric:

- SPICE score = 0.3

Reference captions

"Three giraffes are eating food from the feeder."
"there are three giraffes embracing in the wild"
"three giraffes are together in their pen and some trees"
"Three giraffes eating from a tall tree together."
"Three giraffes reaching for leaves on a tall tree."

Reference scene graph

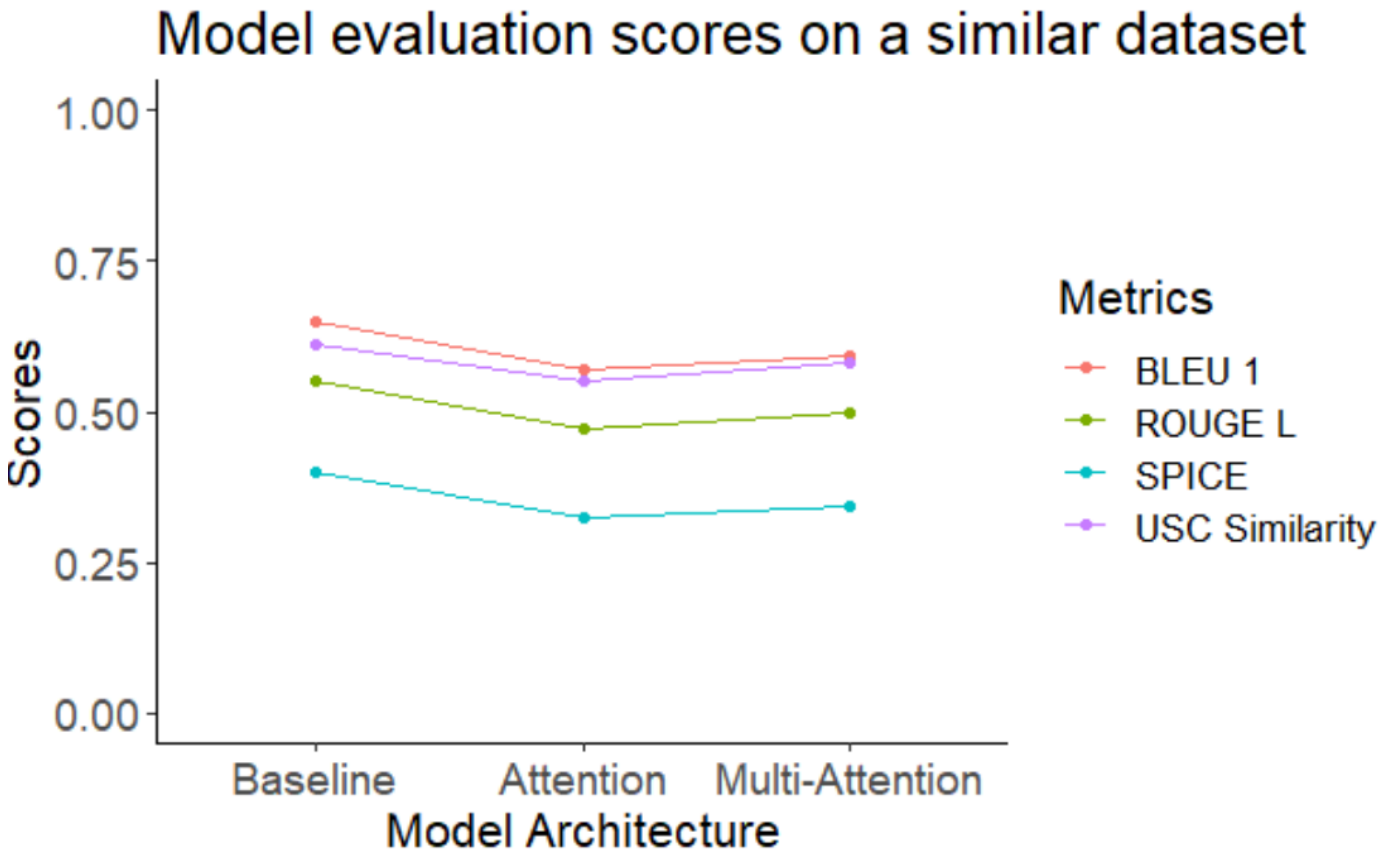


Candidate caption & scene graph

"two giraffes eating leaves from a tree"

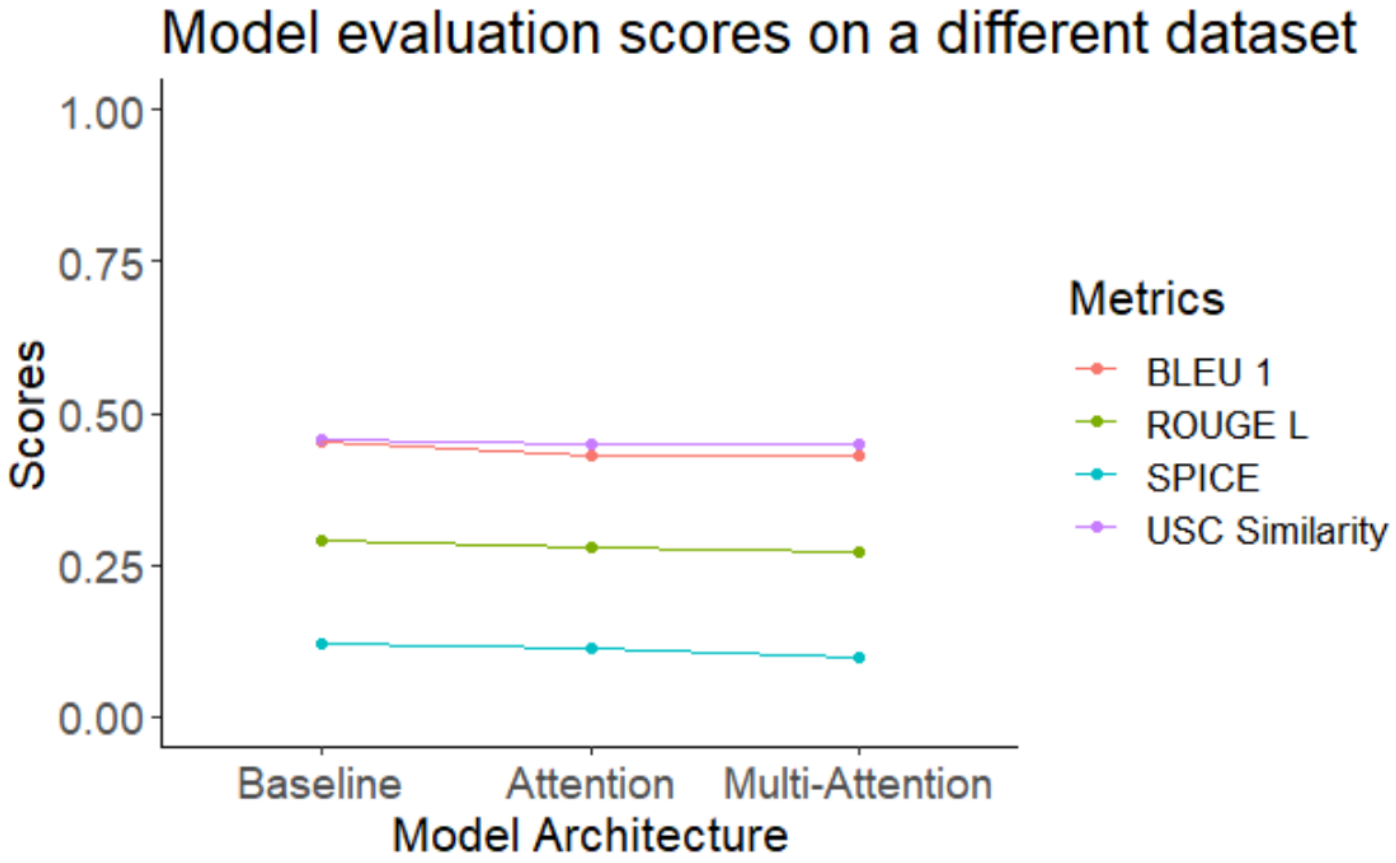
The Baseline Model Shows The Best Performance

- Test data and train data are from the **same** datasets
- BLEU 1 and ROUGE L scores range from 0.5 to 0.8 on the baseline in literatures



Models Show Poor Generalization Capabilities

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



Other considerations

1. CNN models learned from scratch

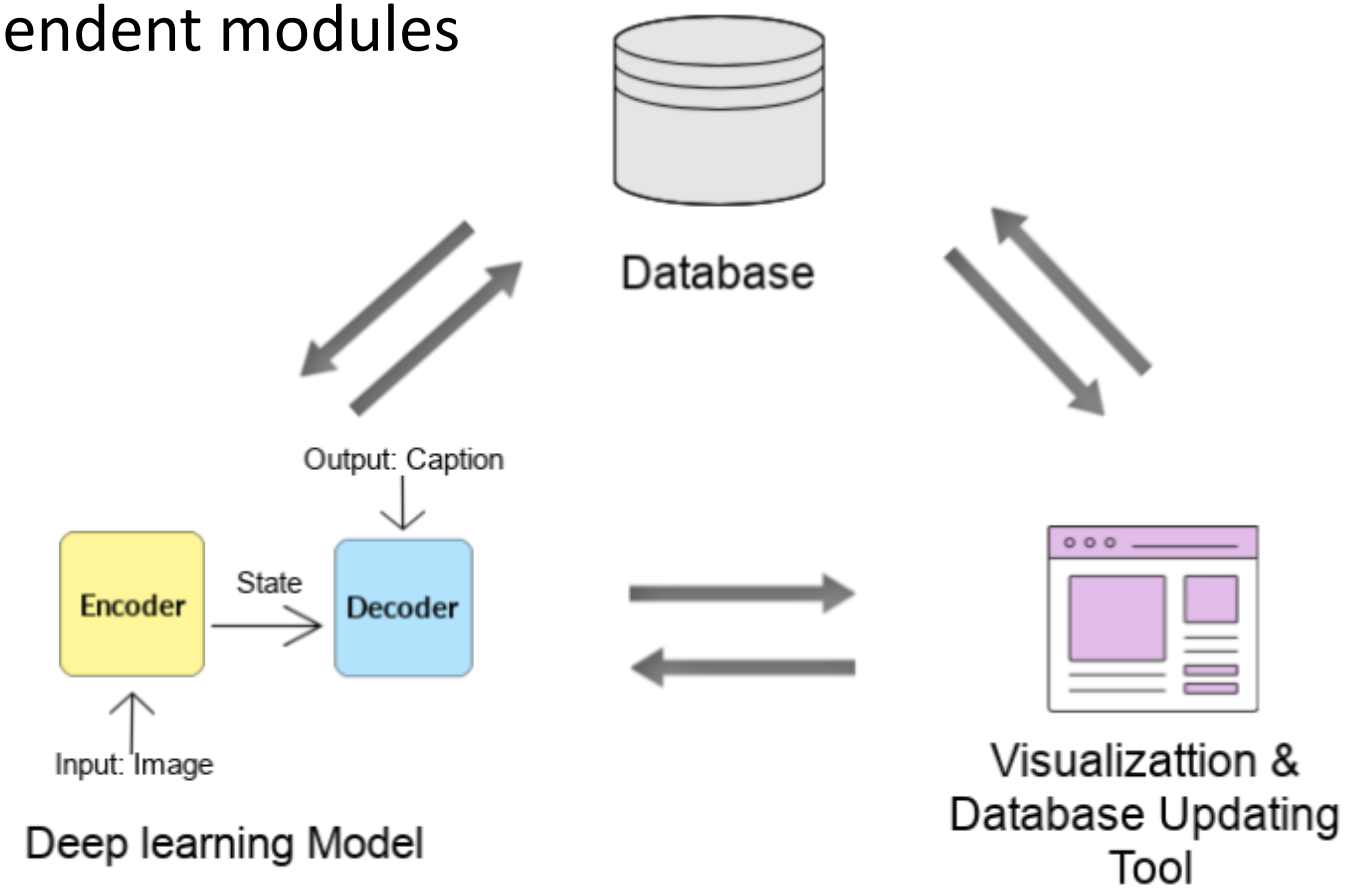
2. Embeddings learned from scratch

Future Improvements

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

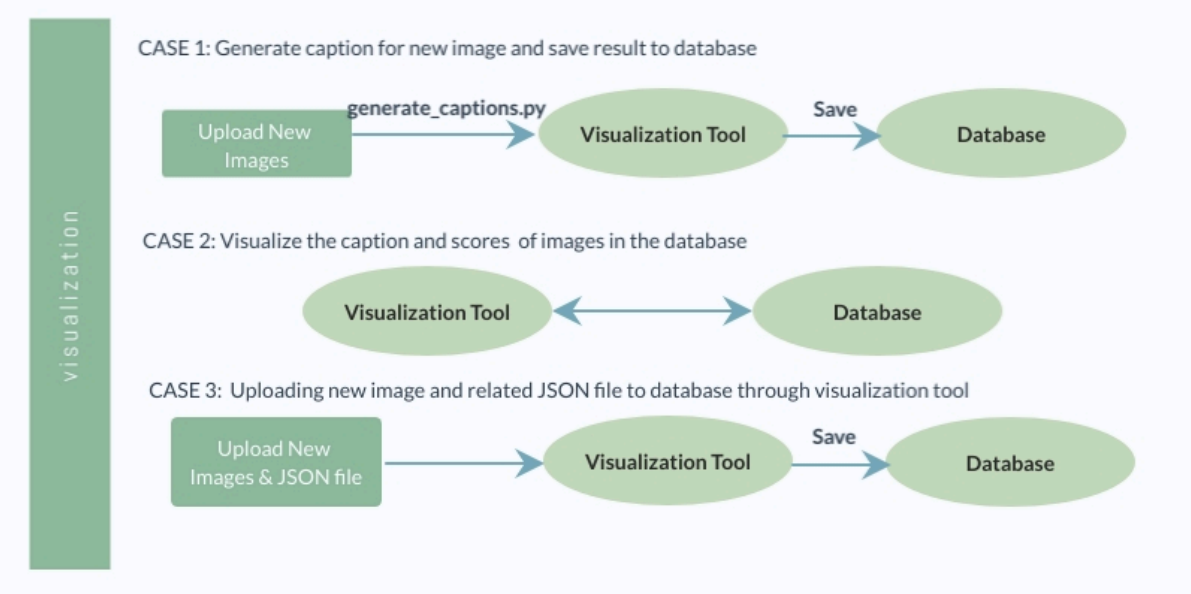
Final Data Product

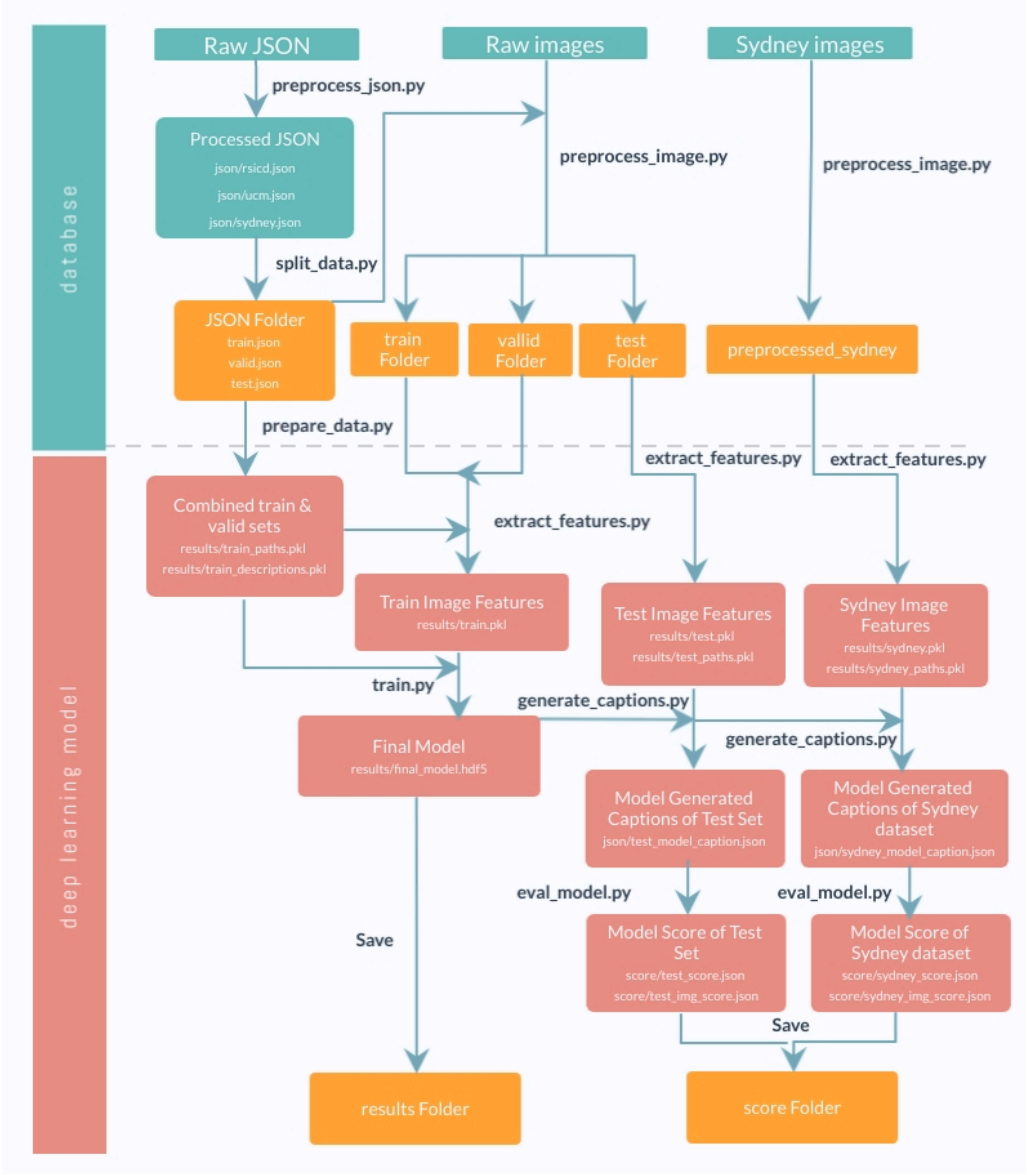
- Complete image captioning pipeline
- 3 independent modules



Product Pipeline

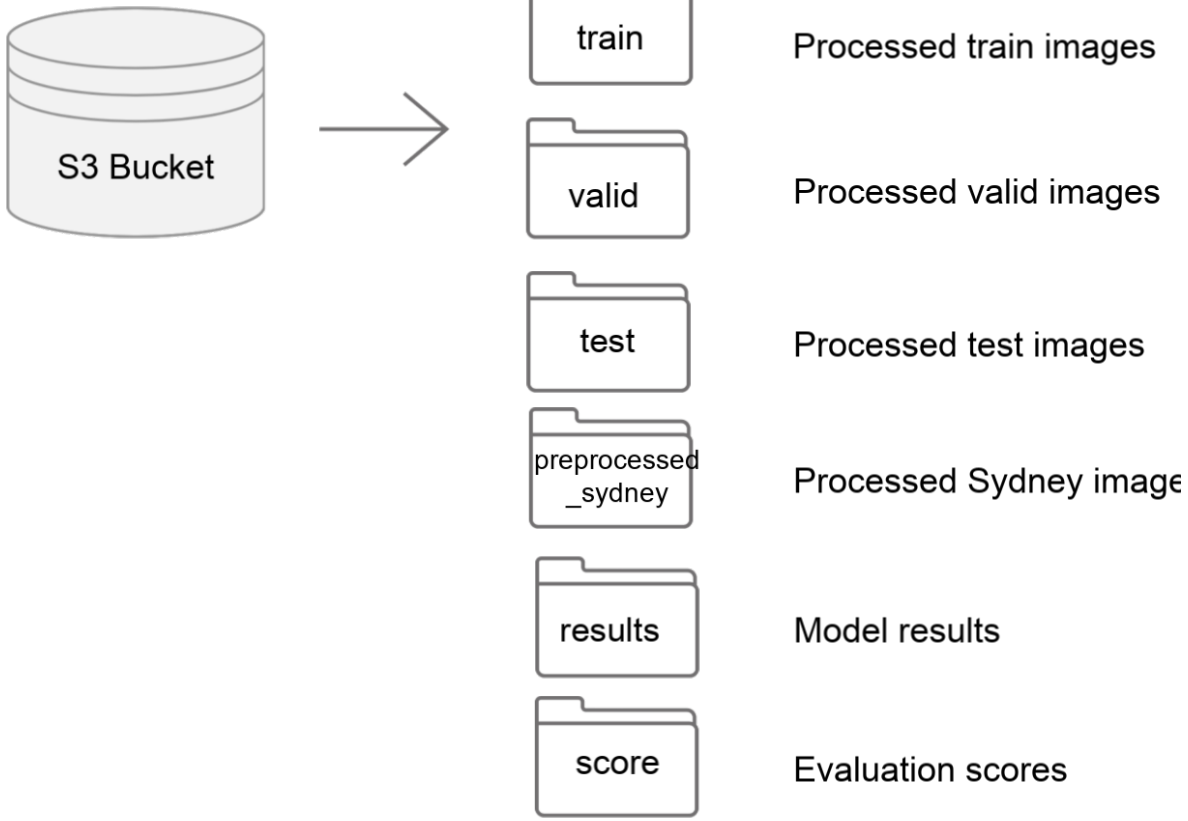
- Database & Model: GNU Make
- Visualization tool: Django





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

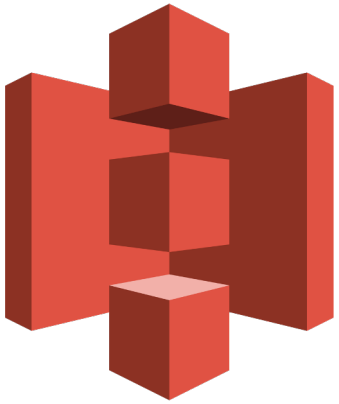
Sources: AWS logo, Pytorch logo

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

Sources: HTML logo, CSS logo, Javascript logo , django logo, S3 logo

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

References

1. B. Qu, X. Li, D. Tao, and X. Lu, “Deep semantic understanding of high resolution remote sensing image,” International Conference on Computer, Information and Telecommunication Systems, pp. 124–128, 2016

2. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring models and data for remote sensing image caption generation. IEEE Trans. Geosci. Remote Sens. 2018, 56, 2183–2195

3. Zhang, X.; Wang, X.; Tang, X.; Zhou, H.; Li, C. Description Generation for Remote Sensing Images Using Attribute Attention Mechanism. Remote Sens. 2019, 11, 612

4. Li, Y.; Fang, S.; Jiao, L.; Liu, R.; Shang, R. A Multi-Level Attention Model for Remote Sensing Image Captions. Remote Sens. 2020, 12, 939

5. SPICE: Semantic Propositional Image Caption Evaluation. Peter Anderson, Basura Fernando, Mark Johnson and Stephen Gould. In *Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, the Netherlands, October 2016*.

Thank you

Questions?