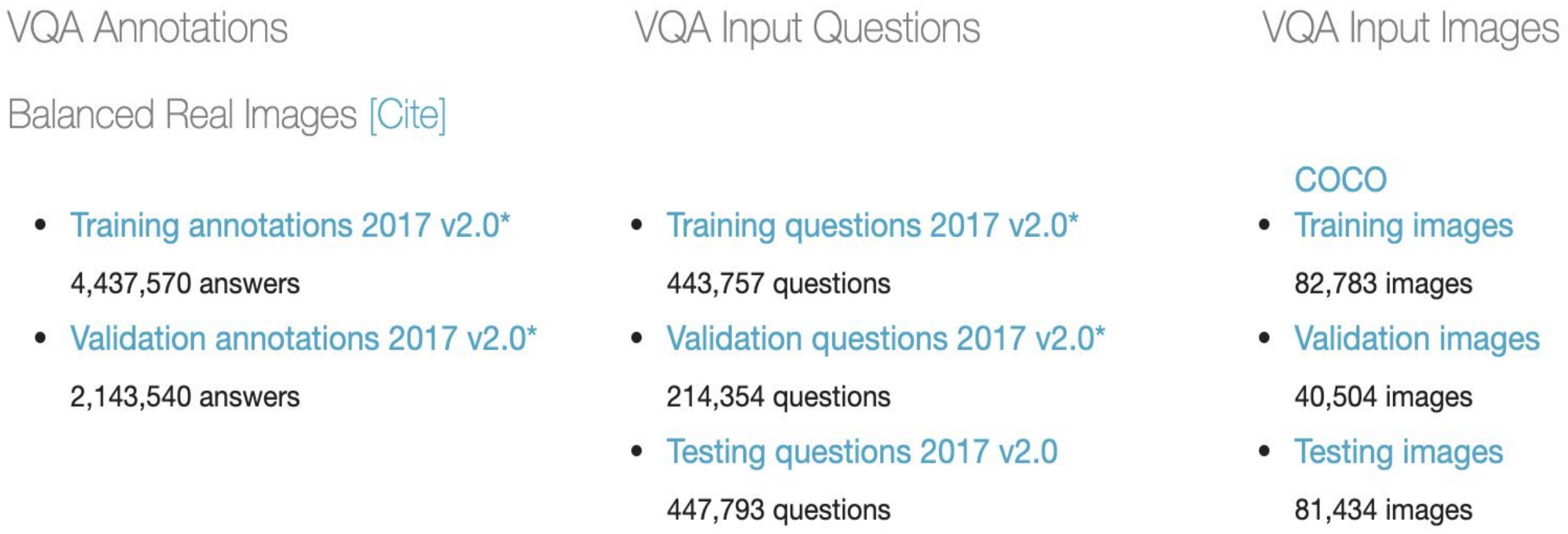
Sprint 2

VQA for Visually Impaired

What have we done:

- Found open-source datasets:
  - VQA v2
  - VizWiz
- Found open-source models
  - Pythia [https://arxiv.org/abs/1807.09956]
  - UNITER [https://arxiv.org/pdf/1909.11740]
  - A Strong Baseline for VQA [https://arxiv.org/pdf/1704.03162]
  - Original VQA Model [https://arxiv.org/pdf/1505.00468]
  - VinVL [https://arxiv.org/abs/2101.00529]
- Successfully trained 3 models
  - A Strong Baseline for VQA
  - Original VQA Model
  - VinVL

Data sets: VQA v2



VQA Annotations

Balanced Real Images [Cite]

- Training annotations 2017 v2.0*

  4,437,570 answers
- Validation annotations 2017 v2.0*

  2,143,540 answers

VQA Input Questions

- Training questions 2017 v2.0*

  443,757 questions
- Validation questions 2017 v2.0*

  214,354 questions
- Testing questions 2017 v2.0

  447,793 questions

VQA Input Images

COCO

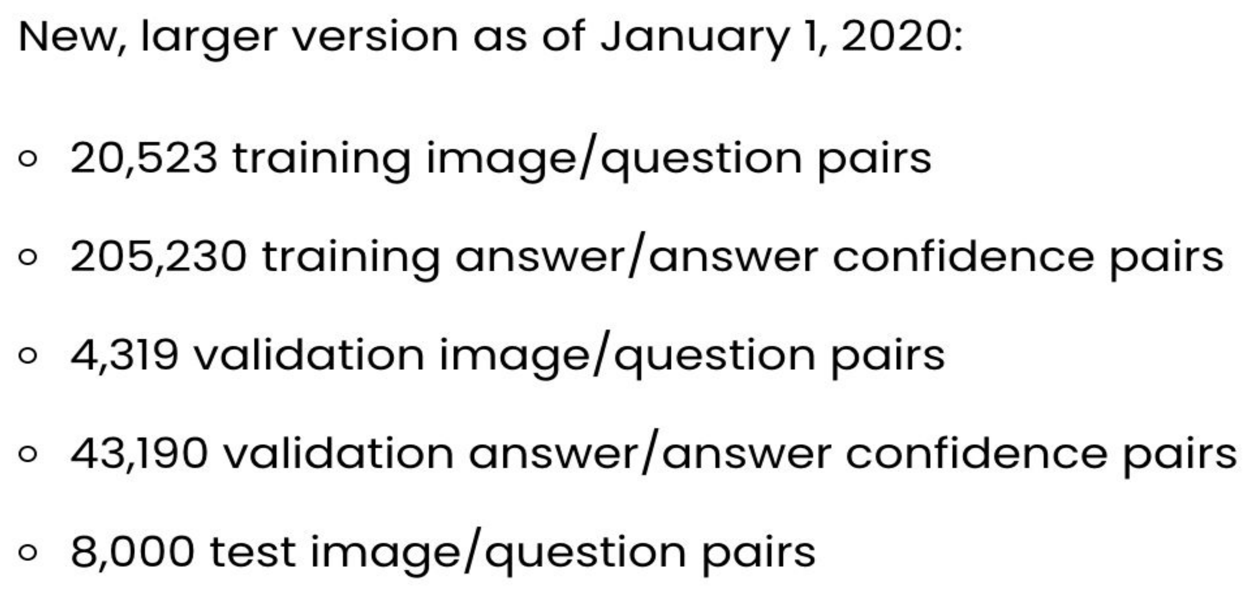
- Training images

  82,783 images
- Validation images

  40,504 images
- Testing images

  81,434 images

Data set: VizWiz



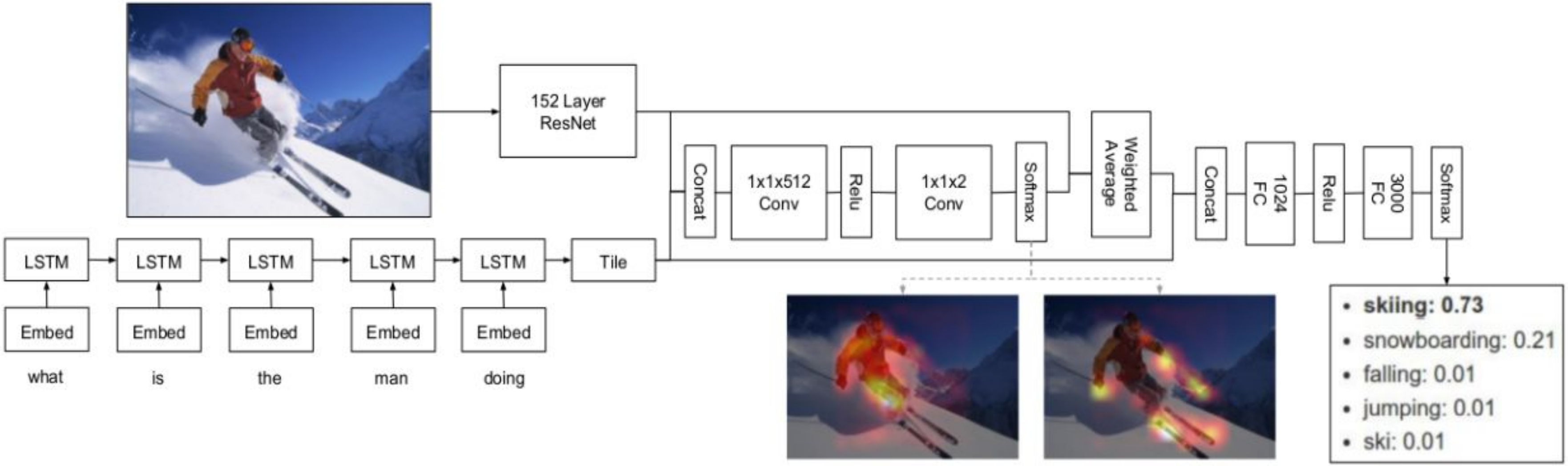
New, larger version as of January 1, 2020:

- 20,523 training image/question pairs
- 205,230 training answer/answer confidence pairs
- 4,319 validation image/question pairs
- 43,190 validation answer/answer confidence pairs
- 8,000 test image/question pairs

Model: A Strong Baseline For Visual Question Answering

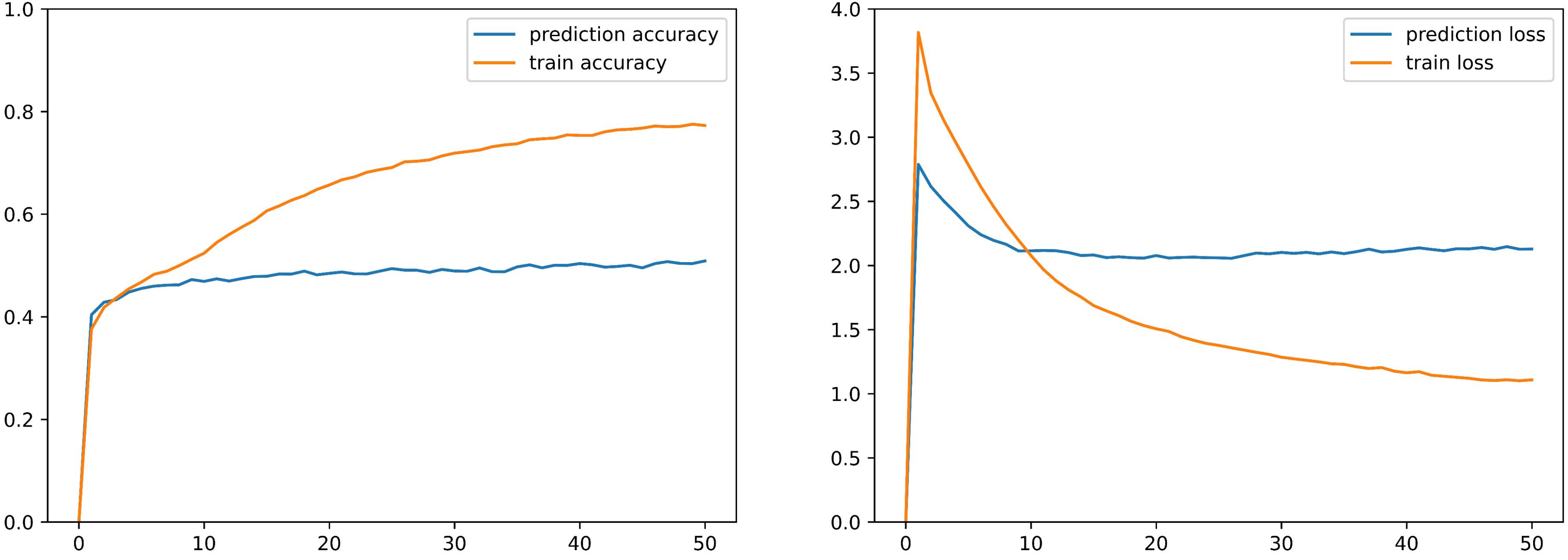
- Visual feature are extracted using a pretrained (on ImageNet) ResNet-152.
- Input Questions are tokenized, embedded and encoded with an LSTM.
- Image features and encoded questions are combined and used to compute multiple attention maps over image features.
- The attended image features and the encoded questions are concatenated and finally fed to a 2-layer classifier that outputs probabilities over the answers (classes).

Model: A Strong Baseline For Visual Question Answering



- **skiing: 0.73**
- snowboarding: 0.21
- falling: 0.01
- jumping: 0.01
- ski: 0.01

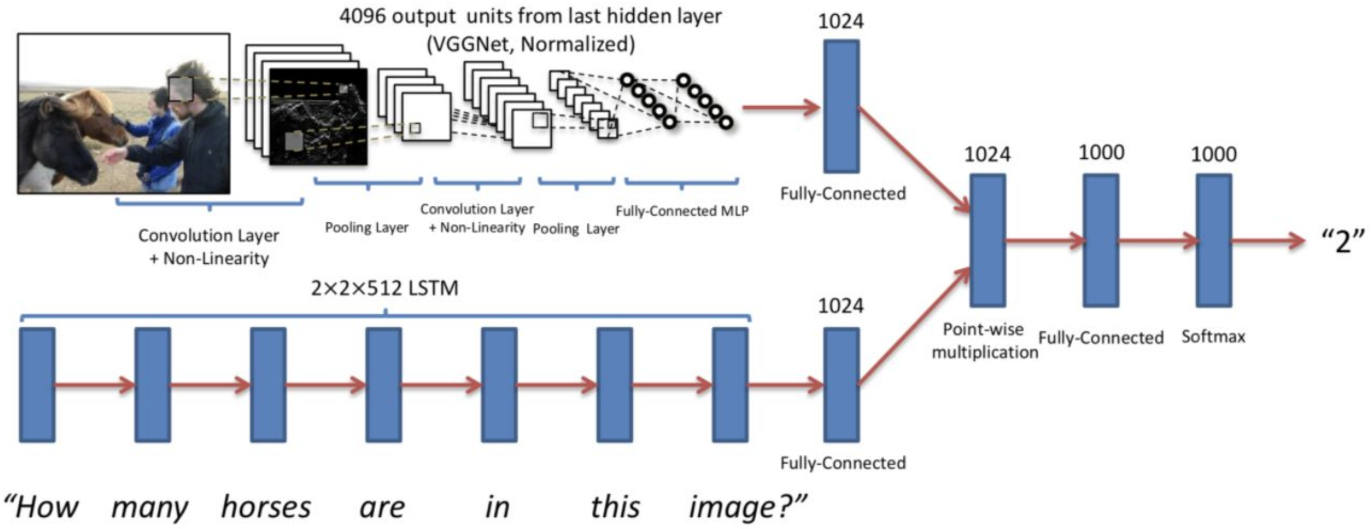
A Strong Baseline For Visual Question Answering : Results



Model: Original VQA model

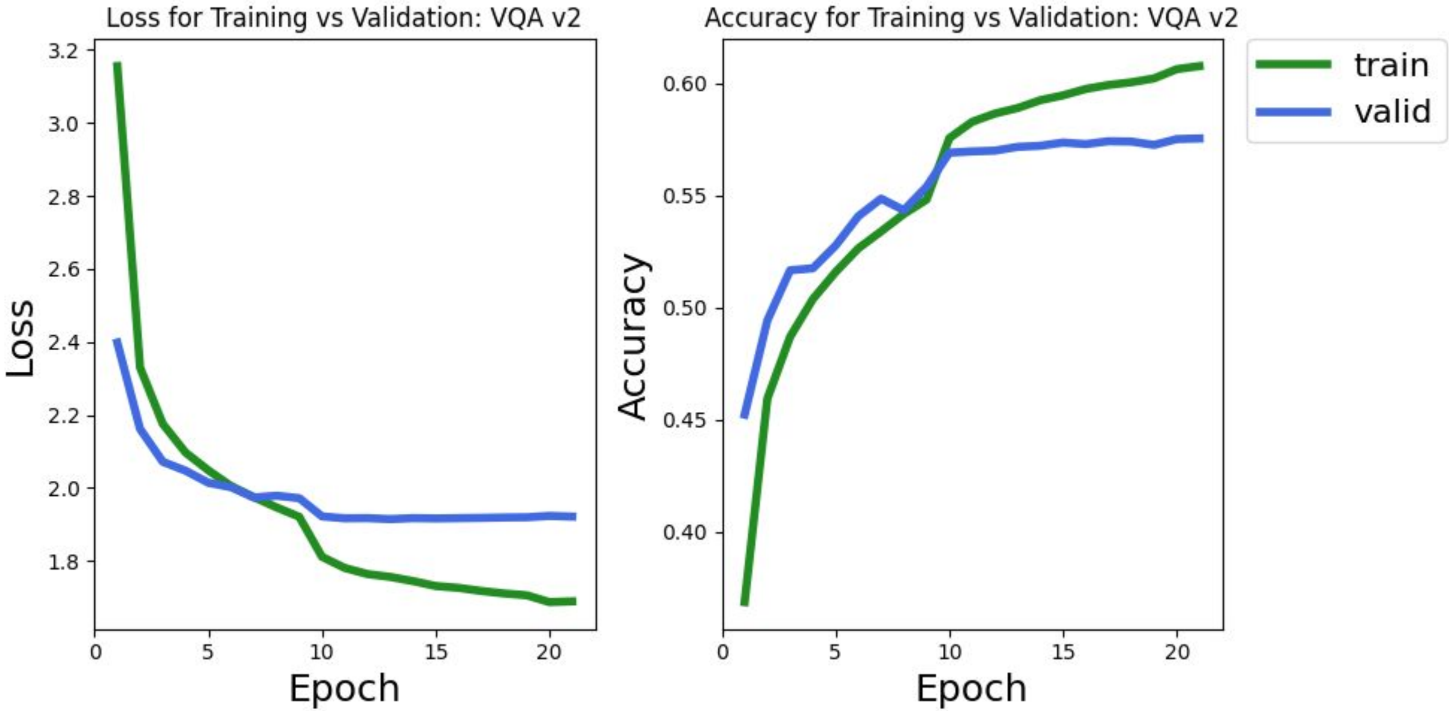
- To encode the image: The last hidden layer of a VGGNet was used, followed by an L2 normalization
- To encode the question: Deeper LSTM with two hidden layer is used to get a 2048-dim embedding of the question
- Question Images Fusion: Element wise multiplication
- Output: Fully connected layer followed by softmax to obtain a distribution over answers

Model: Original VQA model

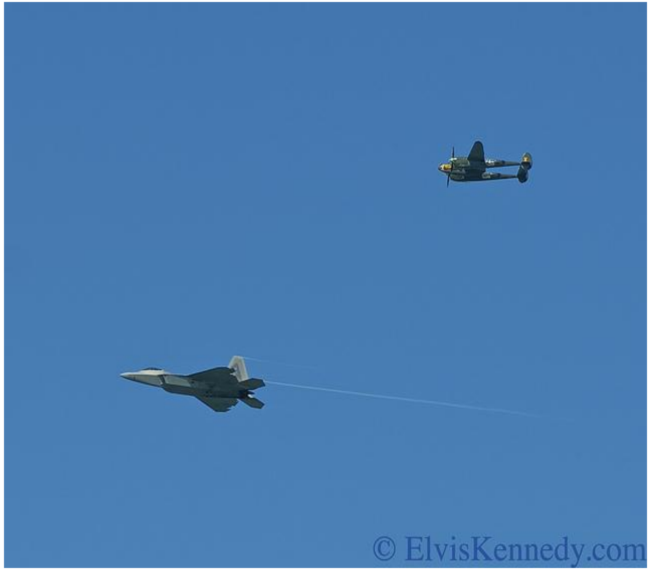


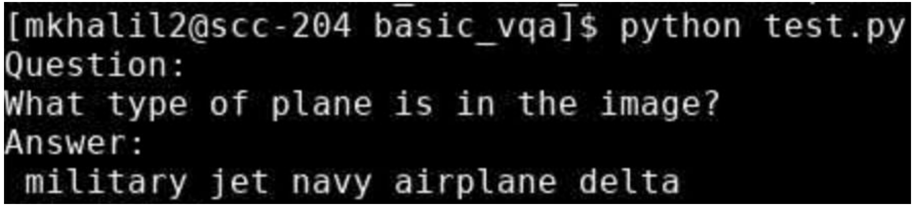
https://arxiv.org/pdf/1505.00468.pdf

Original VQA: Results



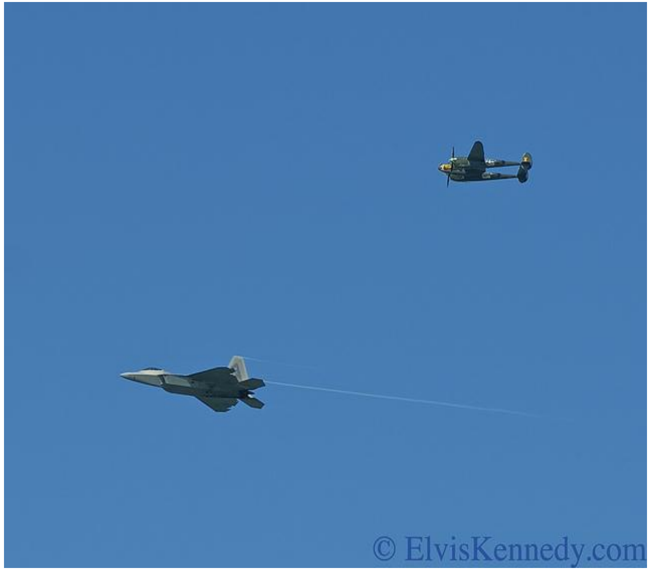
Original VQA: Test

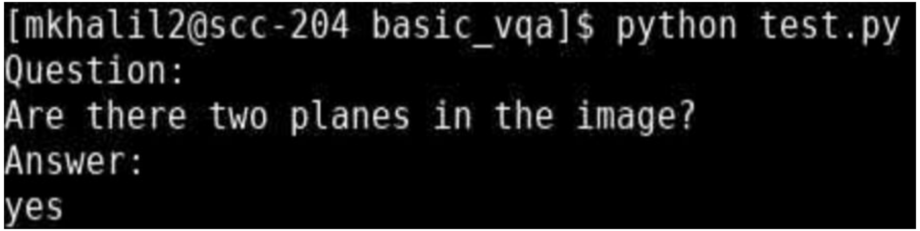




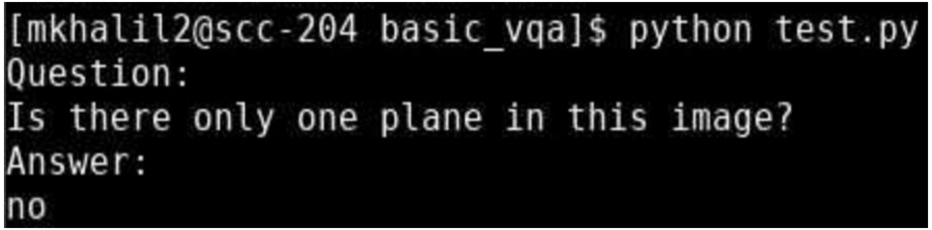
```
[mkhalil2@scc-204 basic_vqa]$ python test.py
Question:
What type of plane is in the image?
Answer:
 military jet navy airplane delta
```

Original VQA: Test



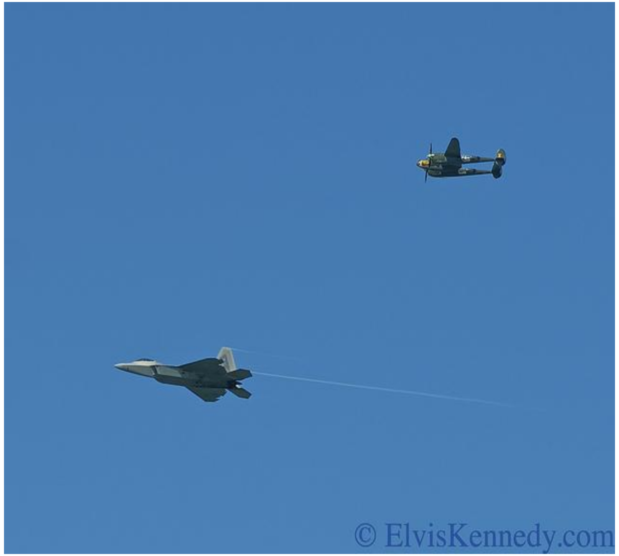


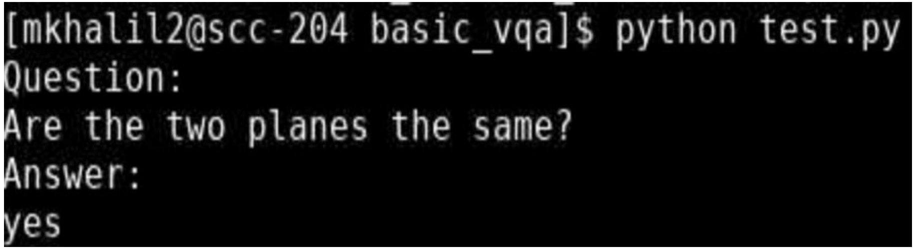
```
[mkhalil2@scc-204 basic_vqa]$ python test.py
Question:
Are there two planes in the image?
Answer:
yes
```



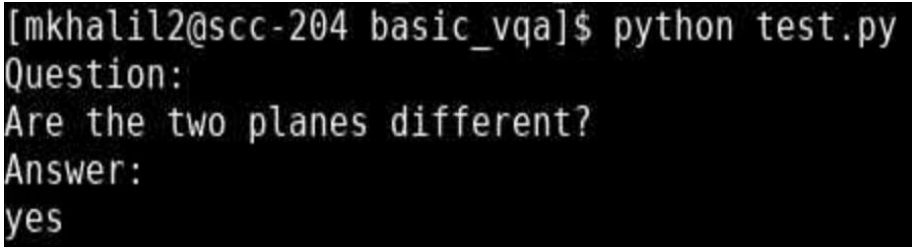
```
[mkhalil2@scc-204 basic_vqa]$ python test.py
Question:
Is there only one plane in this image?
Answer:
no
```

Original VQA: Test





```
[mkhalil2@scc-204 basic_vqa]$ python test.py
Question:
Are the two planes the same?
Answer:
yes
```

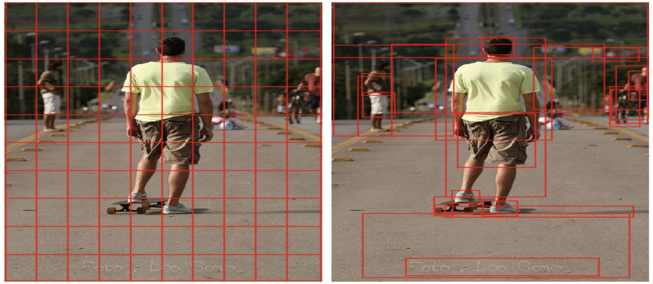


```
[mkhalil2@scc-204 basic_vqa]$ python test.py
Question:
Are the two planes different?
Answer:
yes
```

State of the Art Models:

Pythia Model: Pythia v0.1 https://arxiv.org/pdf/1707.07998.pdf

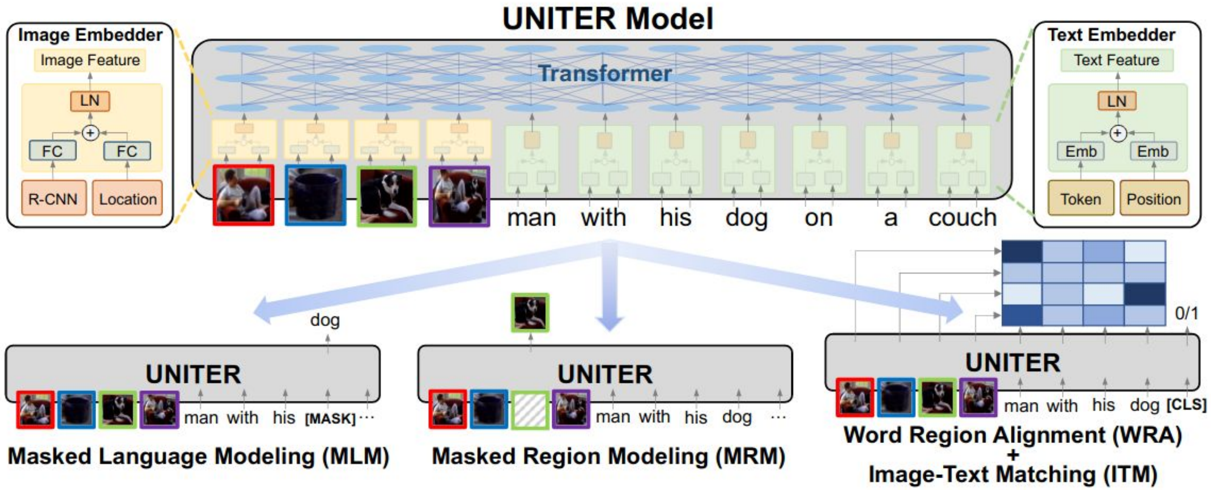
- Winning entry for the 2018 VQA Challenge
- Bottom up top down model: enables attention to be calculated at the level of objects and other salient image regions.
- Bottom up, identifies feature vectors, while top down provides feature weighting.
- Re-implementation of the bottom up top down model with changes in :
  - Image features fine tuning
  - Learning rate schedule
  - Data augmentation



State of the Art Models:

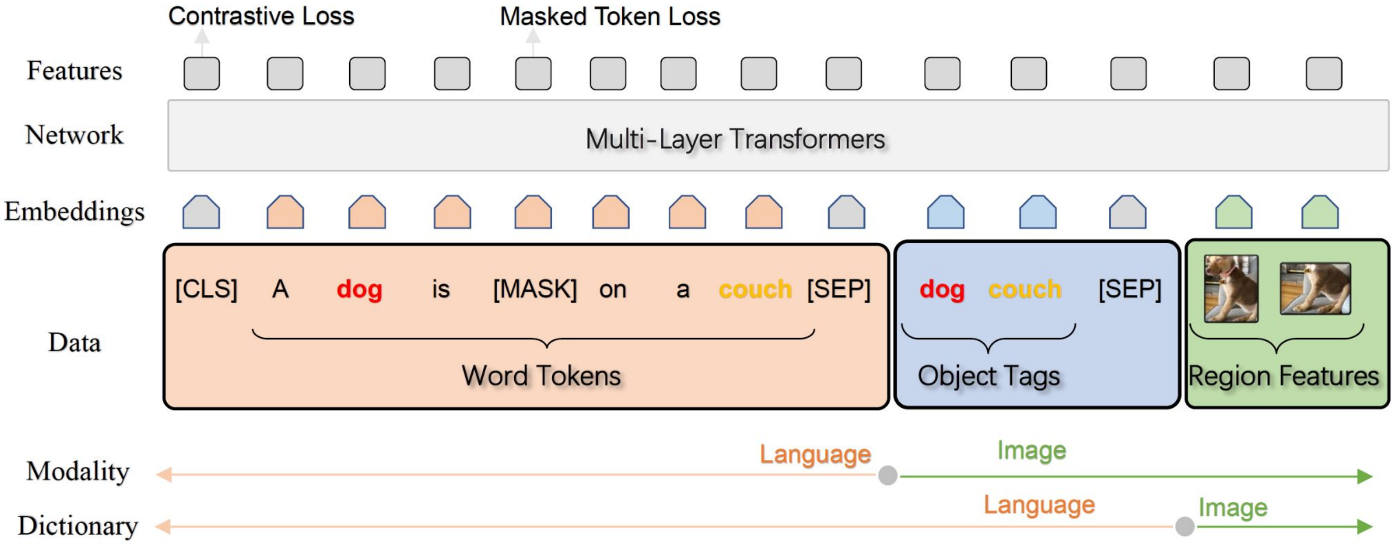
UNITER

- UNiversal Image-TExt Representation Learning
- Transformer-based model using pre-training techniques from NLP



VinVL

- Transformer based V+L
- Improved on previous cutting edge V+L model Oscar [https://arxiv.org/abs/2004.06165]
- Currently ranked 1st on VQA v2.0



Sprint 3 Goals:

- Keep trying to train state of the art models
- Design a test procedure to perform consistent testing across models