

# 16824 Homework 3

Jiajun Wan  
Andrew ID: jiajunw2

## Task 1: Understanding VQA

- 1.1 Why do you think that it's computationally costly to use all answers as separate classes instead of keeping only the most frequent ones?

We have to use a large number of output neurons, that add computations on final sigmoid/softmax calculation, loss computation, and backward propagation. Keeping only the most frequent ones will largely reduce the amount of calculations.

- 1.2 Complete the `__init__` method of the dataset class for VQA in `vqa_dataset.py`. Specifically, initialize the VQA API and anything you need from that.

```
1 self._vqa = VQA(annotation_file=annotation_json_file_path,  
2                  question_file=question_json_file_path)
```

- 1.3 Implement the `__len__` method of the VQADataset class. Should the size of the dataset be equal to the number of images, questions or the answers?

```
1 def __len__(self):  
2     # total number of questions  
3     return len(self._vqa.qqa)
```

The size of the dataset be equal to the number of questions.

- 1.4 Complete the `__getitem__` method

```
1 q_anno = self._vqa.qa[idx]  # load annotation  
2 q_str = self._vqa.qqa[idx]['question']  # question in str format
```

## Task 2: Building a pipeline for VQA

- 2.1 What should be the output dimension of the trainable linear layer?  
Complete the corresponding part in the `__init__` method of BaselineNet in `models.py`.

```
1 self.classifier = nn.Linear(
2     self.text_encoder.config.hidden_size + 512,
3     n_answers,
4 )
```

It should be `n_answers` (5127)

- 2.2 Implement `compute_vis_feats` that featurizes the image (it can be implemented as an one-liner!).

```
1 def compute_vis_feats(self, image):
2     """Convert image tensors to feature tensors."""
3     return torch.nn.functional.adaptive_avg_pool2d(self.vis_encoder(image), 1).
4        squeeze() # feed to vis_encoder and then mean pool on spatial dims
```

- 2.3 Implement the forward pass of BaselineNet. Make sure to use `compute_vis_feats` and `compute_text_feats`.

```
1 def forward(self, image, question):
2     """Forward pass, image (B, 3, 224, 224), qs list of str."""
3     text_feats = self.compute_text_feats(question)
4     vis_feats = self.compute_vis_feats(image)
5     concat_feats = torch.hstack((text_feats, vis_feats))
6     logits = self.classifier(concat_feats)
7     return logits
```

- 2.4 What is the loss for multi-label classification? (Hint: in HW1 you also tackled a multi-class classification problem)

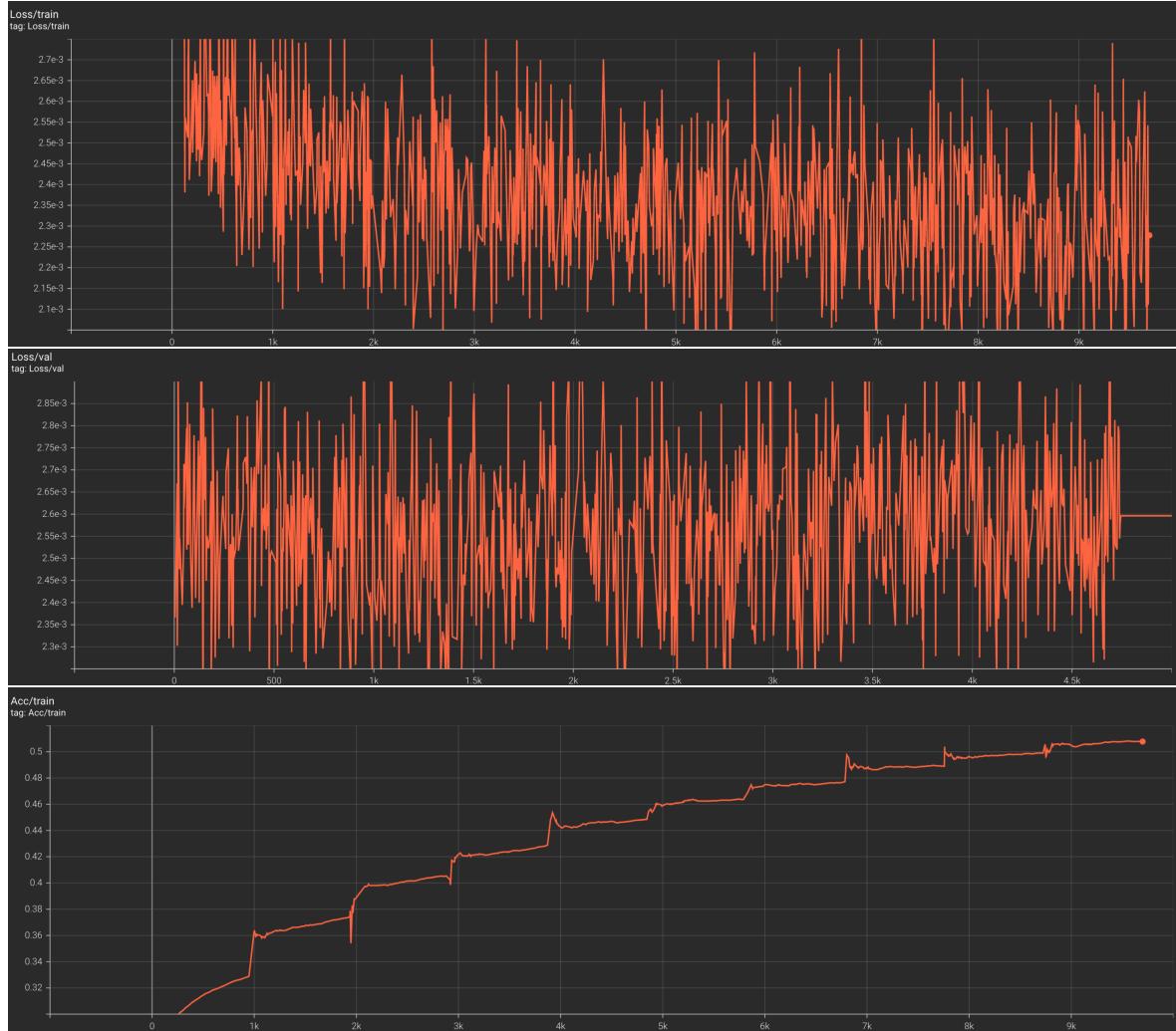
Binary Cross-Entropy Loss

- 2.5 Implement the loss call and the optimization code in the `train_test_loop` method.

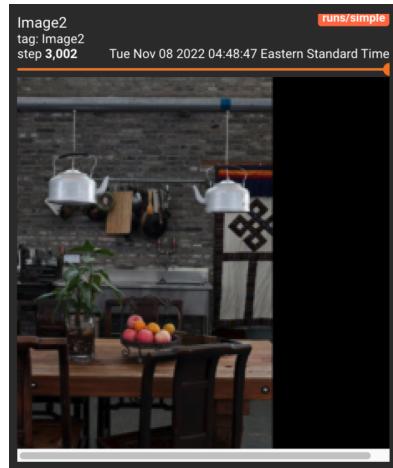
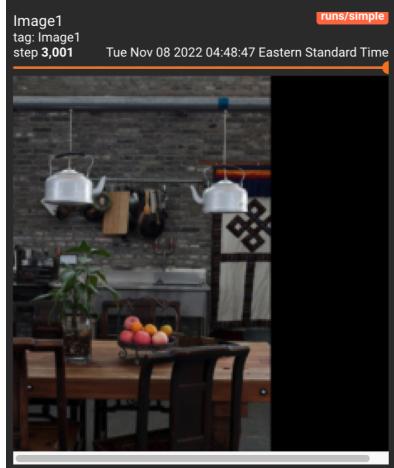
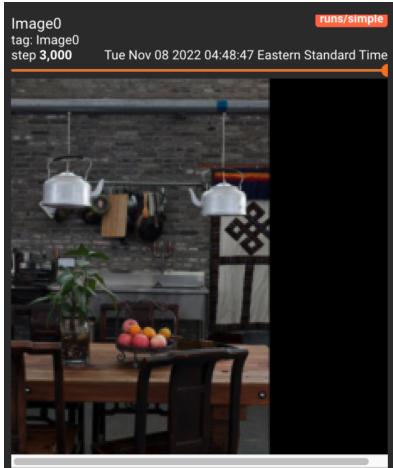
```
1 loss = F.binary_cross_entropy_with_logits(scores, answers)
2
3 # Update
4 if mode == 'train':
5     # optimize loss
6     loss.backward()
7     self.optimizer.step()
8     self.optimizer.zero_grad()
```

**2.6** Complete the `train_test_loop` method to monitor the performance in Tensorboard. Plot the loss and accuracy and include these in your report. Additionally show multiple image-question pairs (at least 3) with the respective answers (predicted and ground-truth).

```
1 # add code to plot the current accuracy
2 self.writer.add_scalar(
3     'Acc/' + mode, n_correct / n_samples,
4     epoch * len(self.data_loaders[mode]) + step
5 )
```







GT Answer0

GT Answer0/text\_summary  
tag: GT Answer0/text\_summary

step 3,000

wood

GT Answer1

GT Answer1/text\_summary  
tag: GT Answer1/text\_summary

step 3,001

yes

GT Answer2

GT Answer2/text\_summary  
tag: GT Answer2/text\_summary

step 3,002

kettle

Predicted Answer0

Predicted Answer0/text\_summary  
tag: Predicted Answer0/text\_summary

step 3,000

Other

Predicted Answer1

Predicted Answer1/text\_summary  
tag: Predicted Answer1/text\_summary

step 3,001

yes

Predicted Answer2

Predicted Answer2/text\_summary  
tag: Predicted Answer2/text\_summary

step 3,002

5  
Other

Question0

Question0/text\_summary  
tag: Question0/text\_summary

step 3,000

What is the table made of?

Question1

Question1/text\_summary  
tag: Question1/text\_summary

step 3,001

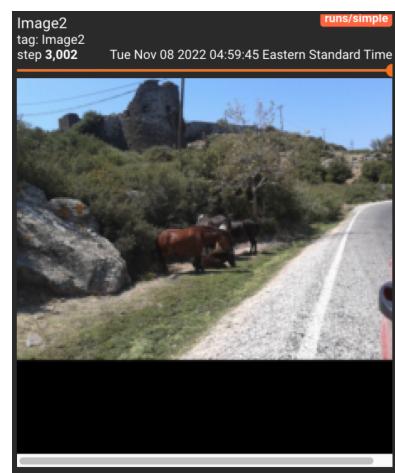
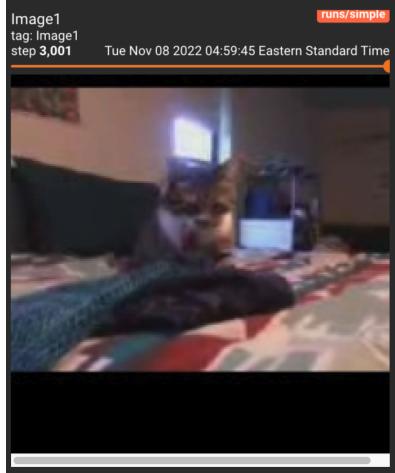
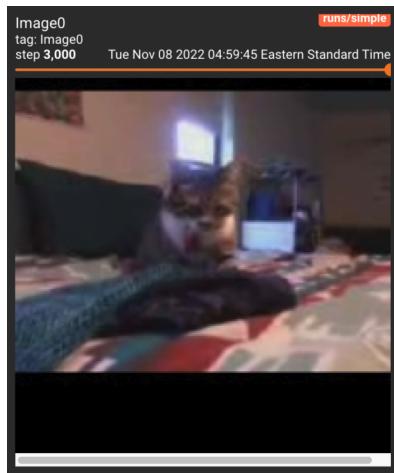
Is the food napping on the table?

Question2

Question2/text\_summary  
tag: Question2/text\_summary

step 3,002

What has been upcycled to make lights?



GT Answer0

GT Answer0/text\_summary  
tag: GT Answer0/text\_summary

step 3,000

camera

GT Answer1

GT Answer1/text\_summary  
tag: GT Answer1/text\_summary

step 3,001

no

GT Answer2

GT Answer2/text\_summary  
tag: GT Answer2/text\_summary

step 3,002

3

Predicted Answer0

Predicted Answer0/text\_summary  
tag: Predicted Answer0/text\_summary

step 3,000

yes

Predicted Answer1

Predicted Answer1/text\_summary  
tag: Predicted Answer1/text\_summary

step 3,001

yes

Predicted Answer2

Predicted Answer2/text\_summary  
tag: Predicted Answer2/text\_summary

step 3,002

2

6

Question0

Question0/text\_summary  
tag: Question0/text\_summary

step 3,000

What is causing the picture to be blurry?

Question1

Question1/text\_summary  
tag: Question1/text\_summary

step 3,001

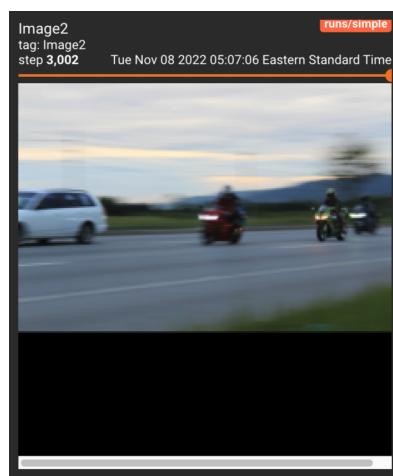
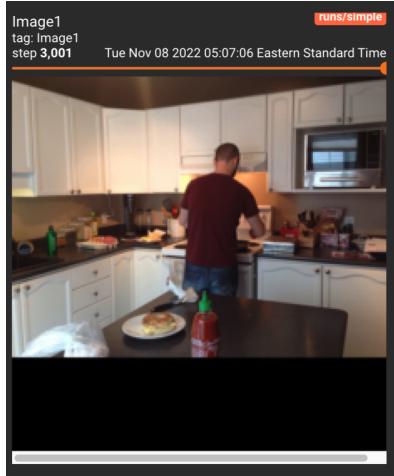
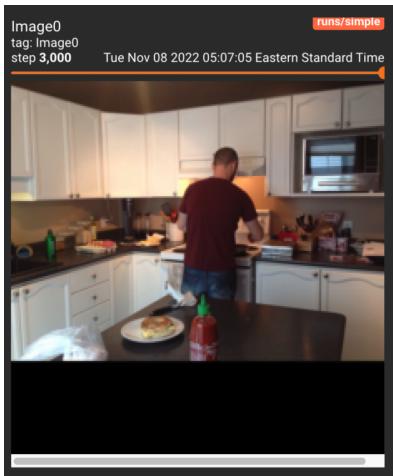
Is it night time?

Question2

Question2/text\_summary  
tag: Question2/text\_summary

step 3,002

How many horses are shown?



GT Answer0

GT Answer0/text\_summary  
tag: GT Answer0/text\_summary

step 3,000

ketchup

GT Answer1 GT Answer0

GT Answer1/text\_summary  
tag: GT Answer1/text\_summary

step 3,001

yes

GT Answer2

GT Answer2/text\_summary  
tag: GT Answer2/text\_summary

step 3,002

yes

Predicted Answer0

Predicted Answer0/text\_summary  
tag: Predicted Answer0/text\_summary

step 3,000

kitchen

Predicted Answer1

Predicted Answer1/text\_summary  
tag: Predicted Answer1/text\_summary

step 3,001

kitchen

Predicted Answer2

Predicted Answer2/text\_summary  
tag: Predicted Answer2/text\_summary

step 3,002 7

2

Question0

Question0/text\_summary  
tag: Question0/text\_summary

step 3,000

What is in the bottle on the kitchen Isle?

Question1

Question1/text\_summary  
tag: Question1/text\_summary

step 3,001

Is the sauce spicy?

Question2

Question2/text\_summary  
tag: Question2/text\_summary

step 3,002

How many headlights are visible?

Final Val Acc: 0.5341694647442228