09-14-2018

Pythia for Vizwiz

Vivek Natarajan*

Tina Jiang*

Meet Shah

Xinlei Chen

Dhruv Batra

Devi Parikh

Marcus Rohrbach

* - indicates equal contribution

facebook Artificial Intelligence Research | A-STAR

facebook Artificial Intelligence Research

Motivation

Two key aspects of the Vizwiz dataset

Motivation

Two key aspects of the Vizwiz dataset

Requires OCR

What does the bottle say?

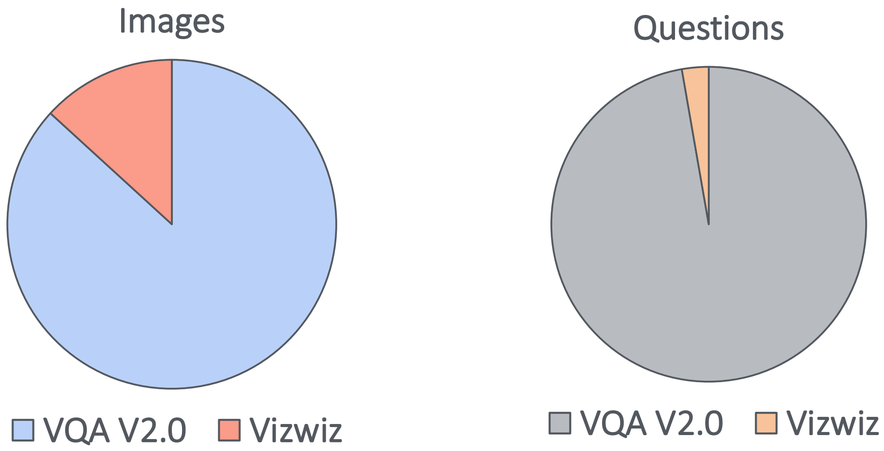
Motivation

Two key aspects of the Vizwiz dataset

Requires OCR

What does the bottle say?

Vizwiz dataset is small



Number of Images - VQA V2.0: 204721, Vizwiz: 31173
Number of Questions - VQA V2.0: 1105904, Vizwiz: 31173

Pythia

Our starting point to the Vizwiz Challenge

Pythia

Our starting point to the Vizwiz Challenge

- Modular framework for VQA research released by the FAIR A-STAR team

Pythia

Our starting point to the Vizwiz Challenge

- Modular framework for VQA research released by the FAIR A-STAR team

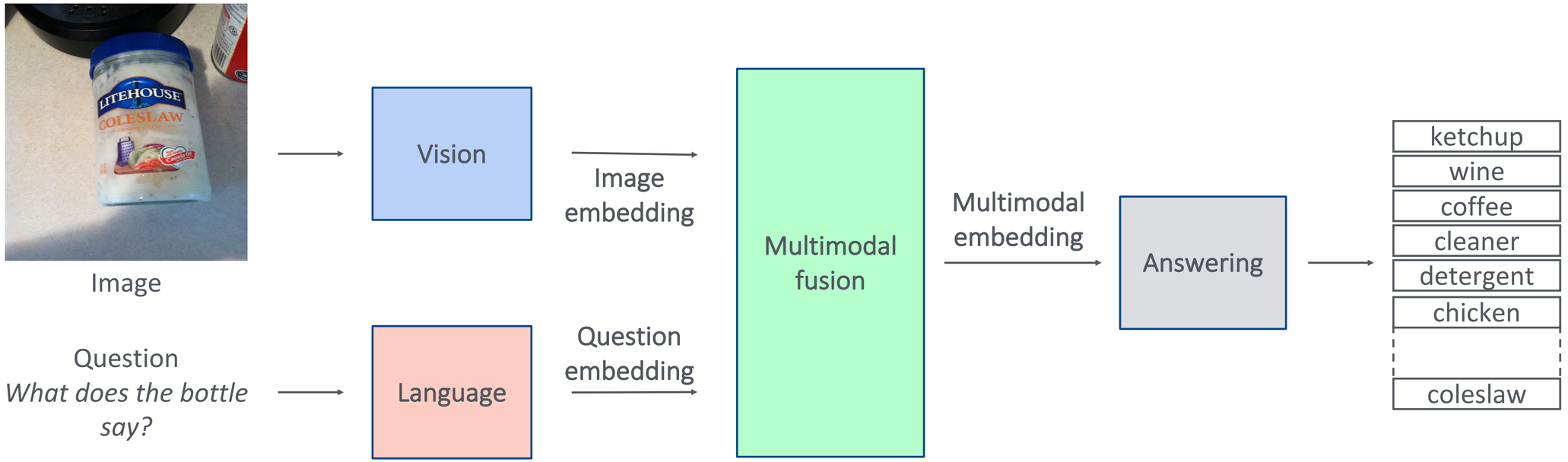
- Pythia v0.1 formed the basis of the winning entry to the VQA Challenge 2018!

Pythia

Our starting point to the Vizwiz Challenge

- Modular framework for VQA research released by the FAIR A-STAR team

- Pythia v0.1 formed the basis of the winning entry to the VQA Challenge 2018!

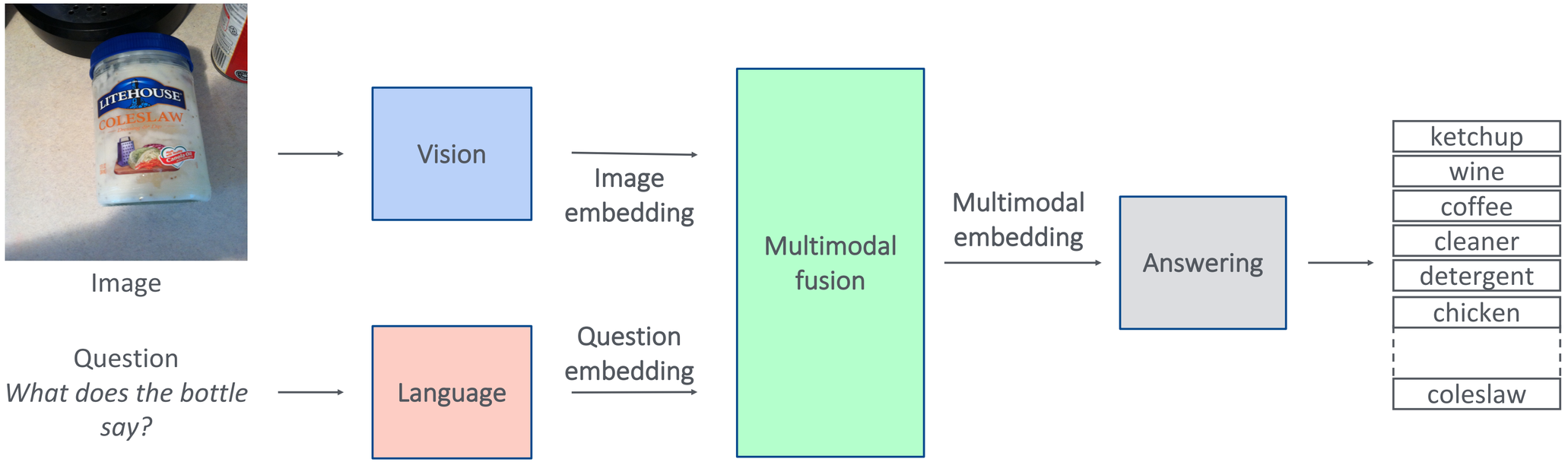


Pythia

Our starting point to the Vizwiz Challenge

- Modular framework for VQA research released by the FAIR A-STAR team

- Pythia v0.1 formed the basis of the winning entry to the VQA Challenge 2018!



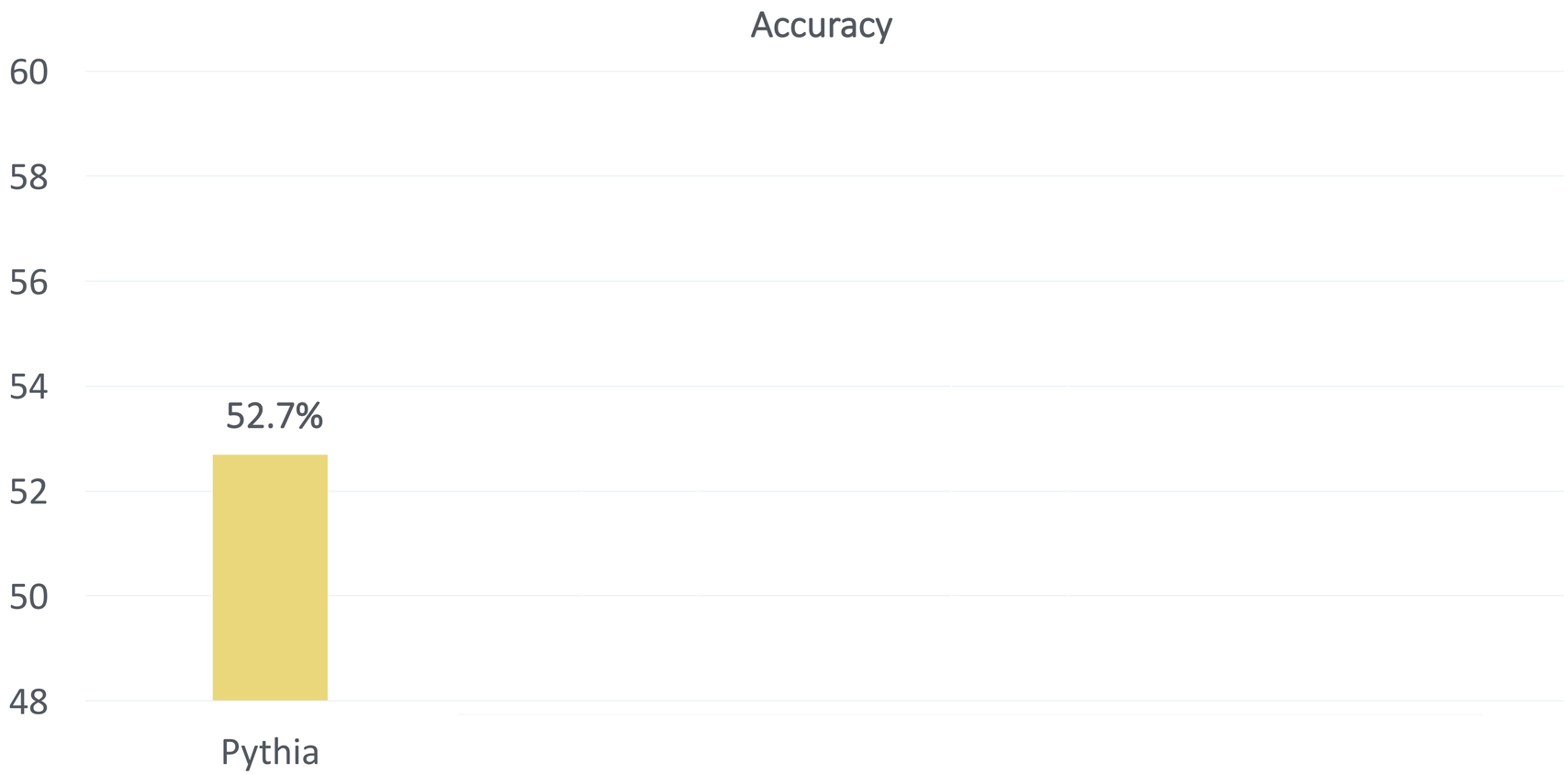
Abstract: https://arxiv.org/pdf/1807.09956.pdf
Code: https://github.com/facebookresearch/pythia

Pythia

test-dev accuracy

Pythia

test-dev accuracy



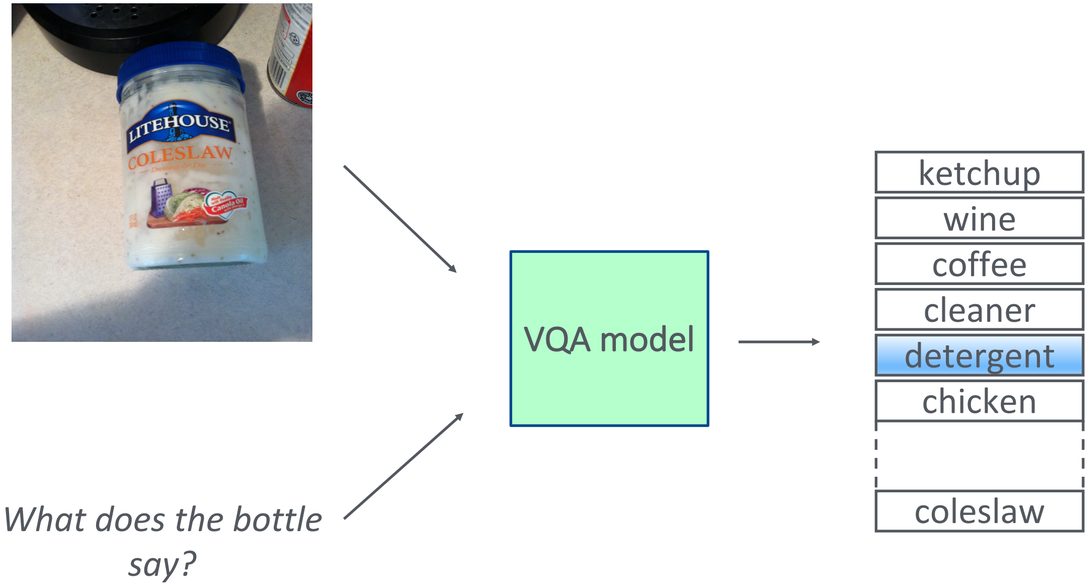
Pythia

Qualitative results

Pythia

Qualitative results

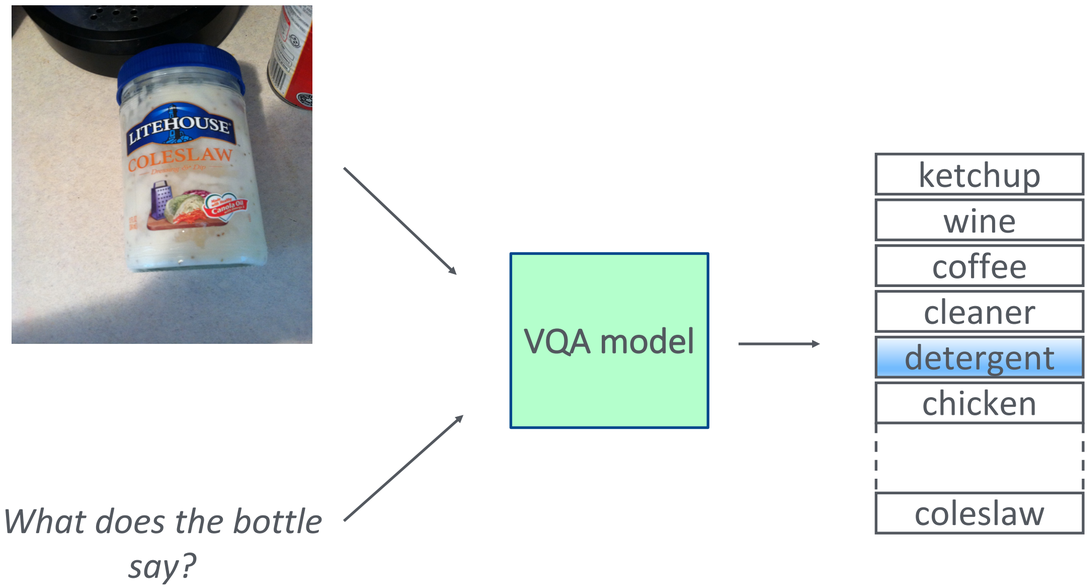
Poor performance on questions that required OCR capabilities



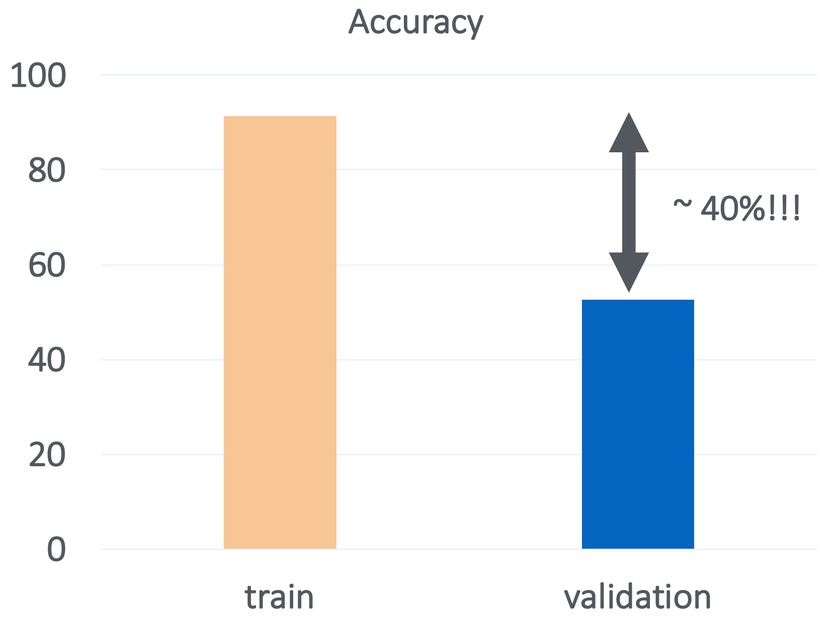
Pythia

Qualitative results

Poor performance on questions that required OCR capabilities



Small dataset, model overfitting



Pythia

Qualitative results

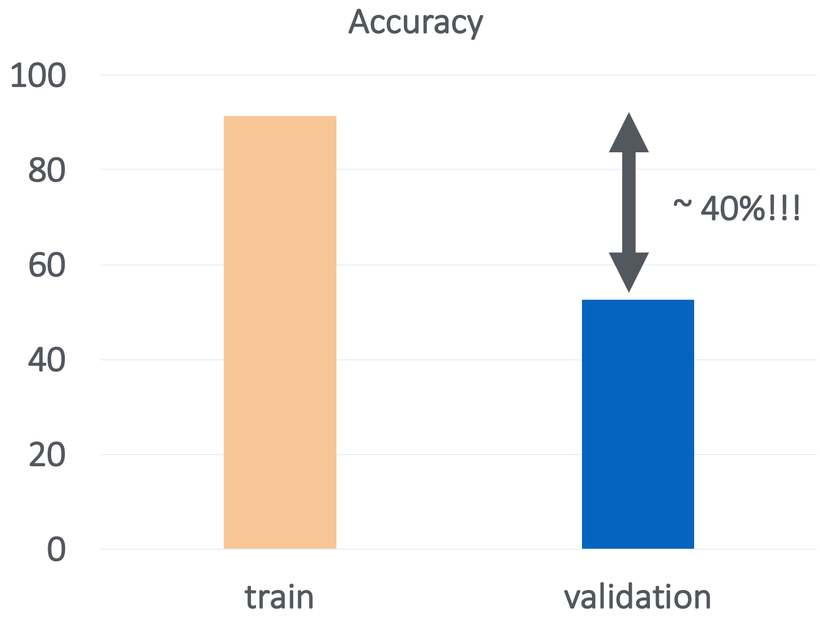
Poor performance on questions that required OCR capabilities

What does the bottle say?

VQA model

ketchup
wine
coffee
cleaner
detergent
chicken
coleslaw

Small dataset, model overfitting



Poor performance on yes/no and number categories which had few examples in the dataset

Pythia

Qualitative results

Incorporate results from OCR into the model

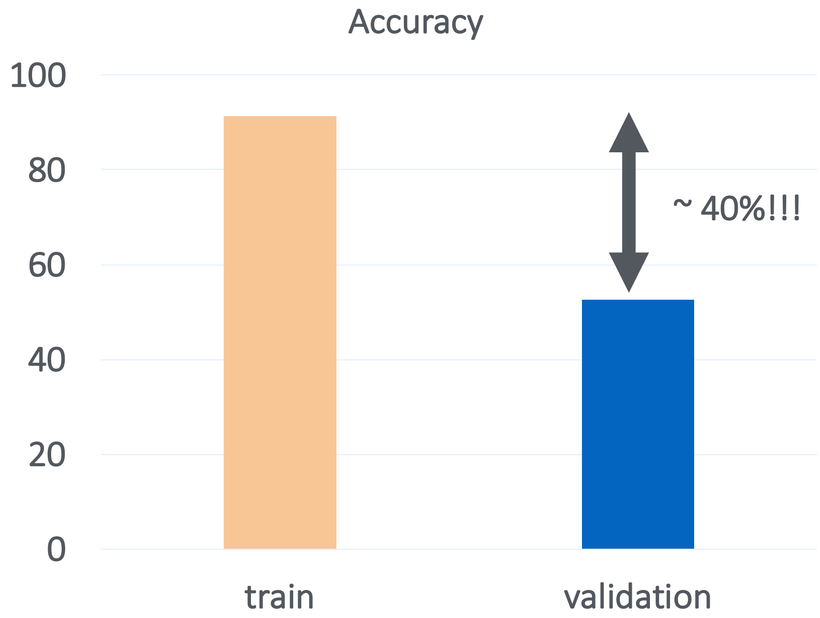
Poor performance on questions that required OCR capabilities

What does the bottle say?

VQA model

ketchup
wine
coffee
cleaner
detergent
chicken
coleslaw

Small dataset, model overfitting



Poor performance on yes/no and number categories which had few examples in the dataset

Pythia

Vision

Language

Multimodal fusion

Answering

Image

Resnet 152

Faster RCNN

Grid convolutional features

Object detection bottom up features

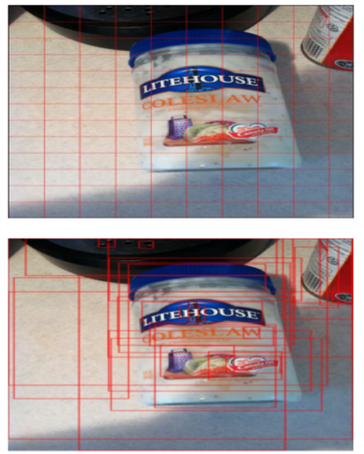


Image embedding

Pythia for Vizwiz VQA

Vision

Language

Multimodal fusion

Answering

Image

Resnet 152

Faster RCNN

Rosetta

Grid convolutional features

Object + text detection bottom up features

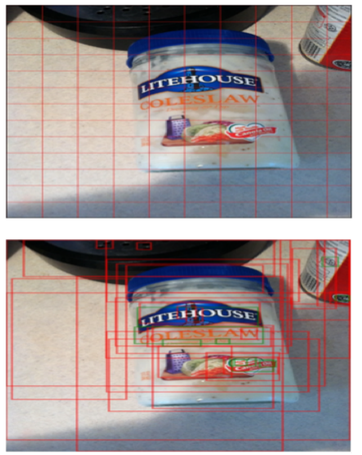
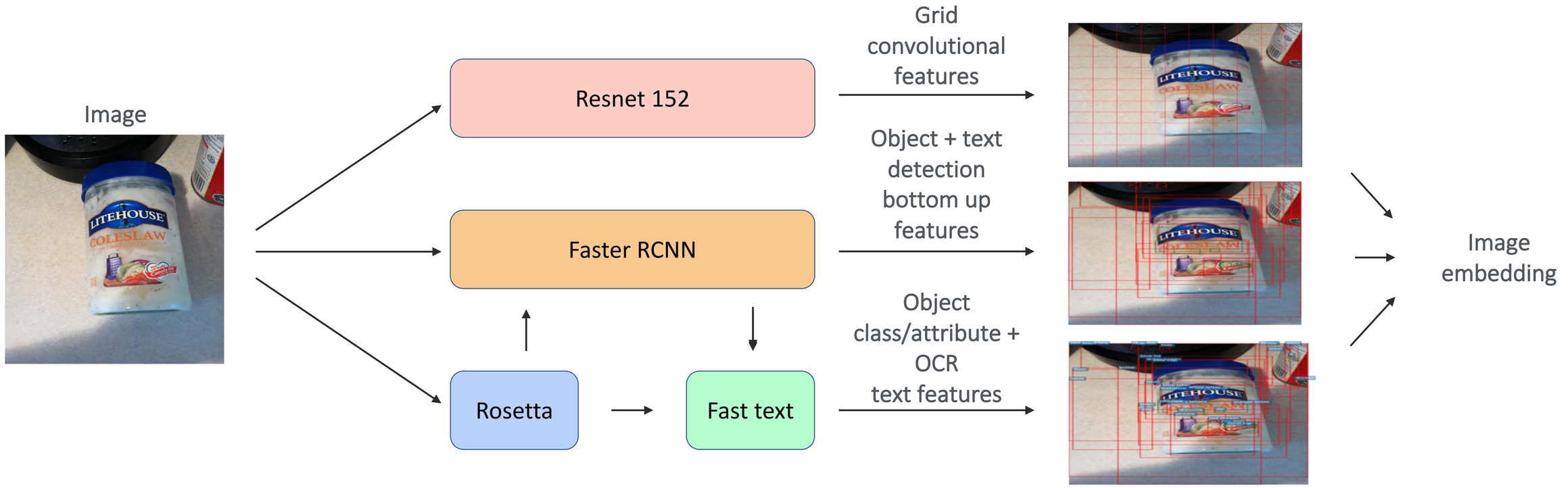


Image embedding

Rosetta: Large scale system for text detection and recognition in images, Borisyuk et al, KDD 2018

Pythia for Vizwiz VQA

Vision | Language | Multimodal fusion | Answering



Fast text: Enriching word vectors with subword information, Bojanowski et al, ACL 2017

Pythia for Vizwiz VQA

Vision

Language

Multimodal fusion

Answering

Image

Resnet 152

Faster RCNN

Rosetta

Fast text

Grid convolutional features

Object + text detection bottom up features

Object class/attribute + OCR text features

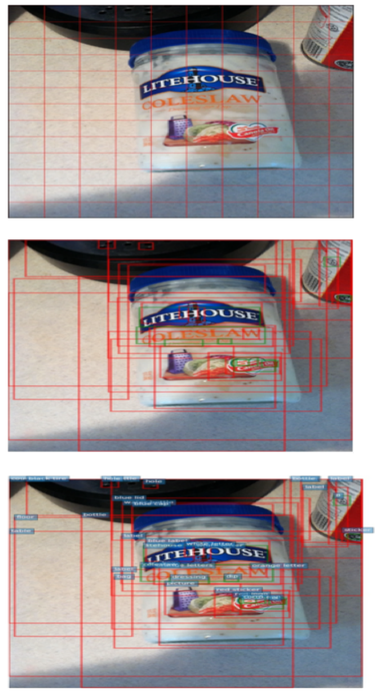


Image embedding

Character level modeling to help address OCR errors

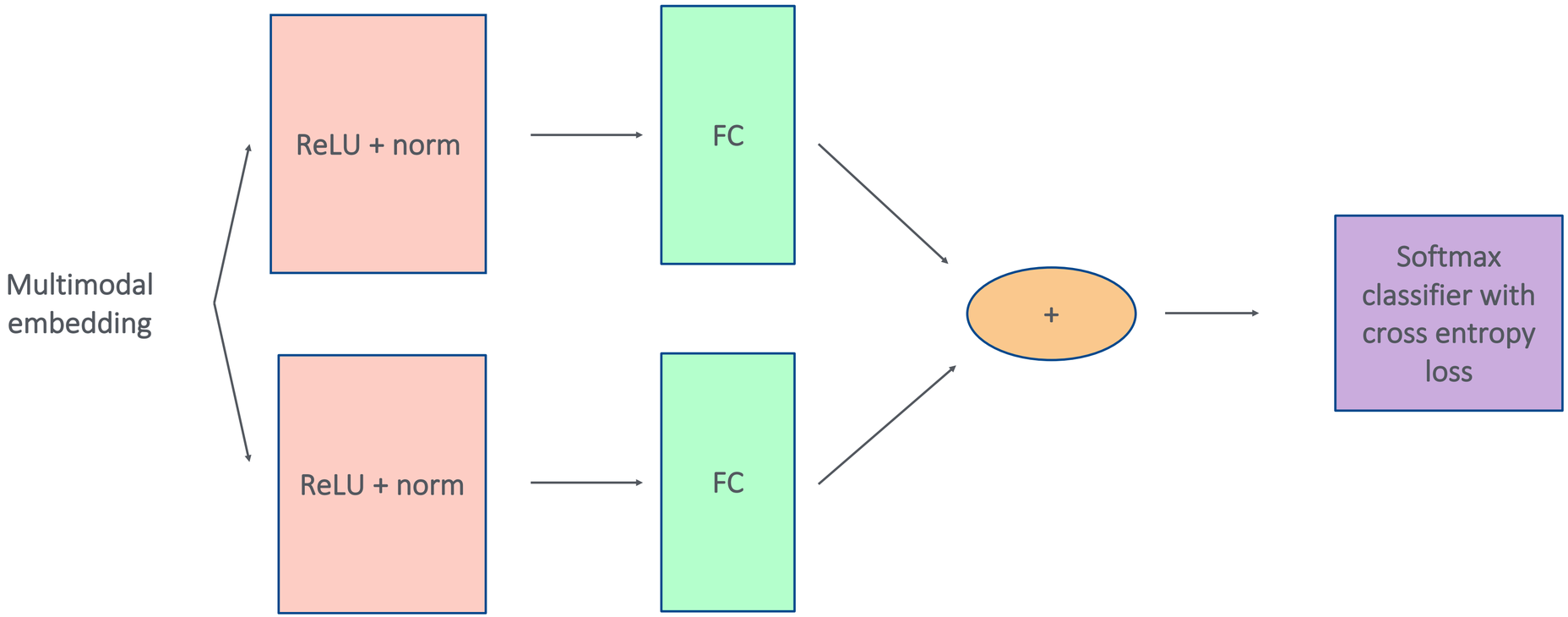
Pythia

Vision

Language

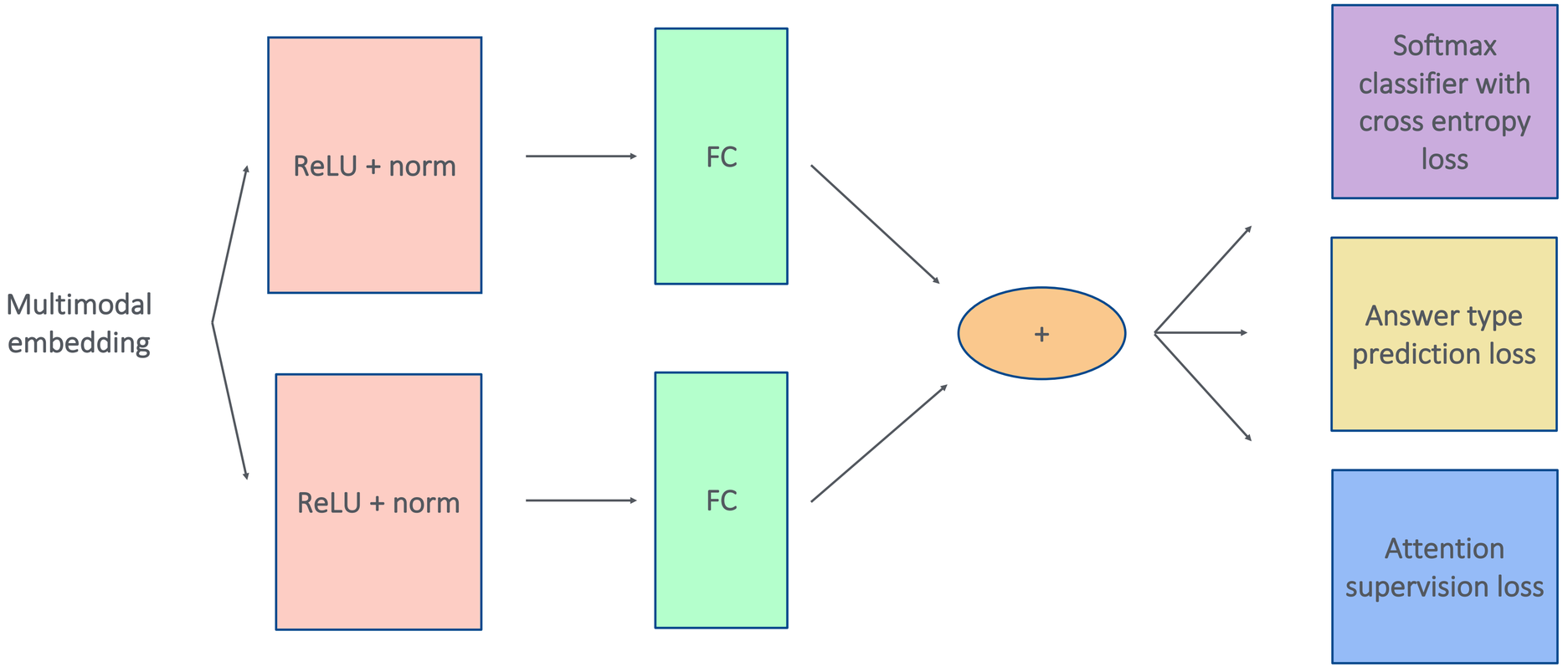
Multimodal fusion

Answering



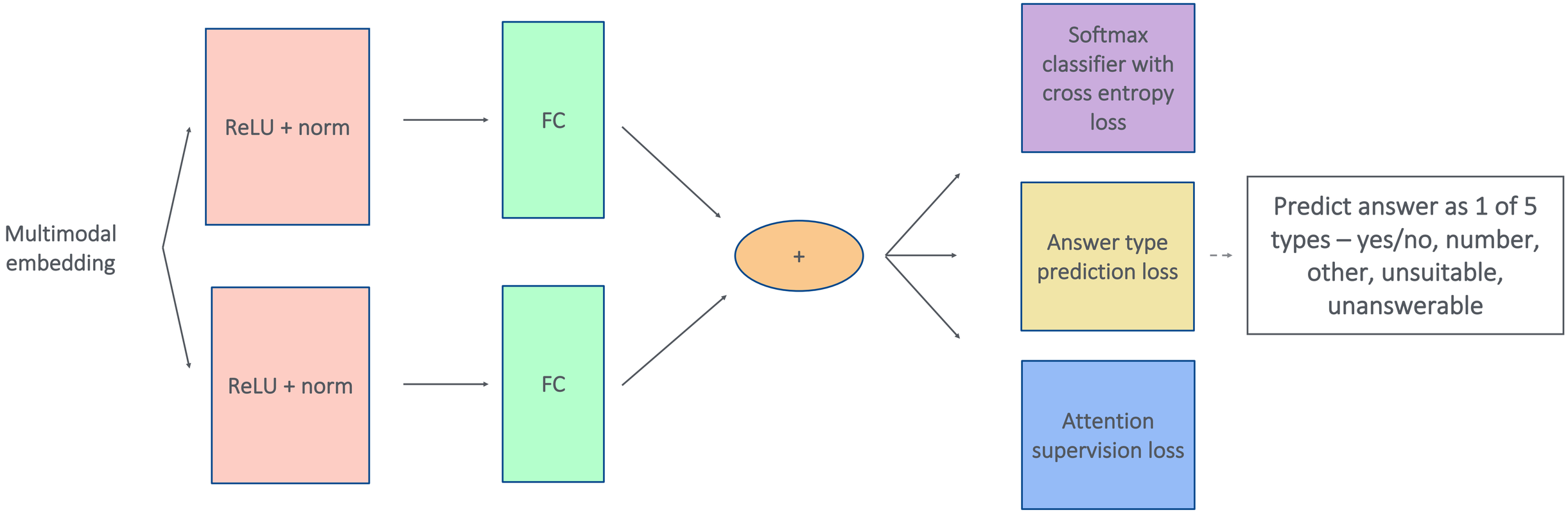
Pythia for Vizwiz VQA

Vision | Language | Multimodal fusion | Answering



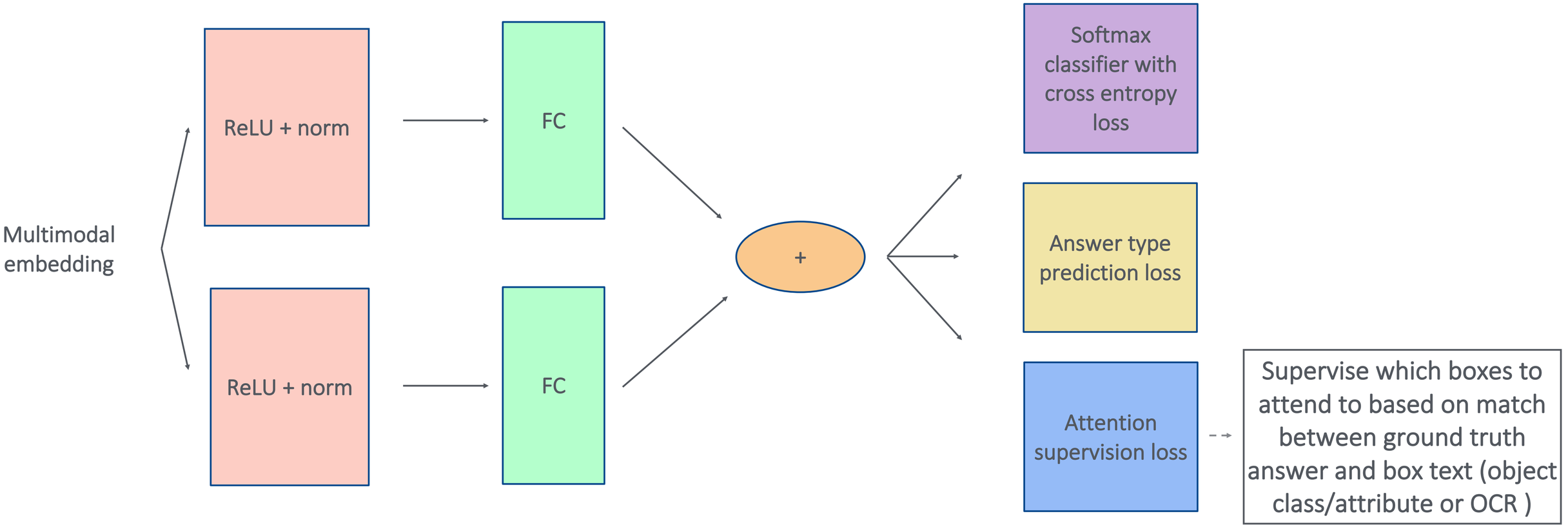
Pythia for Vizwiz VQA

Vision | Language | Multimodal fusion | Answering



Pythia for Vizwiz VQA

Vision | Language | Multimodal fusion | Answering



Pythia for Vizwiz VQA

Vision

Language

Multimodal fusion

Answering

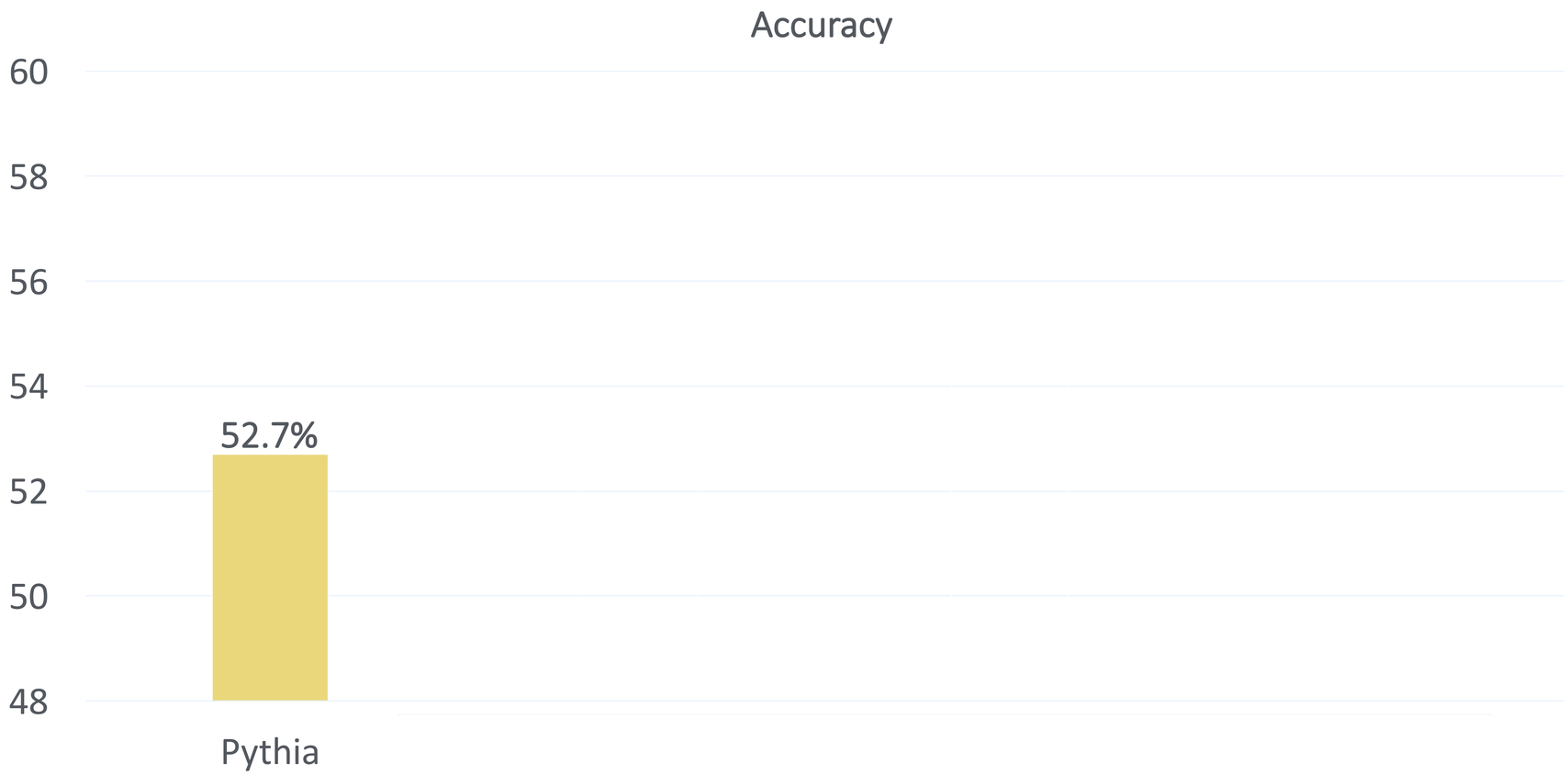
Question
What kind of wine? Thanks.
Ground truth answer
Moscato

Attention supervision loss

Supervise which boxes to attend to based on match between ground truth answer and box text (object class/attribute or OCR)

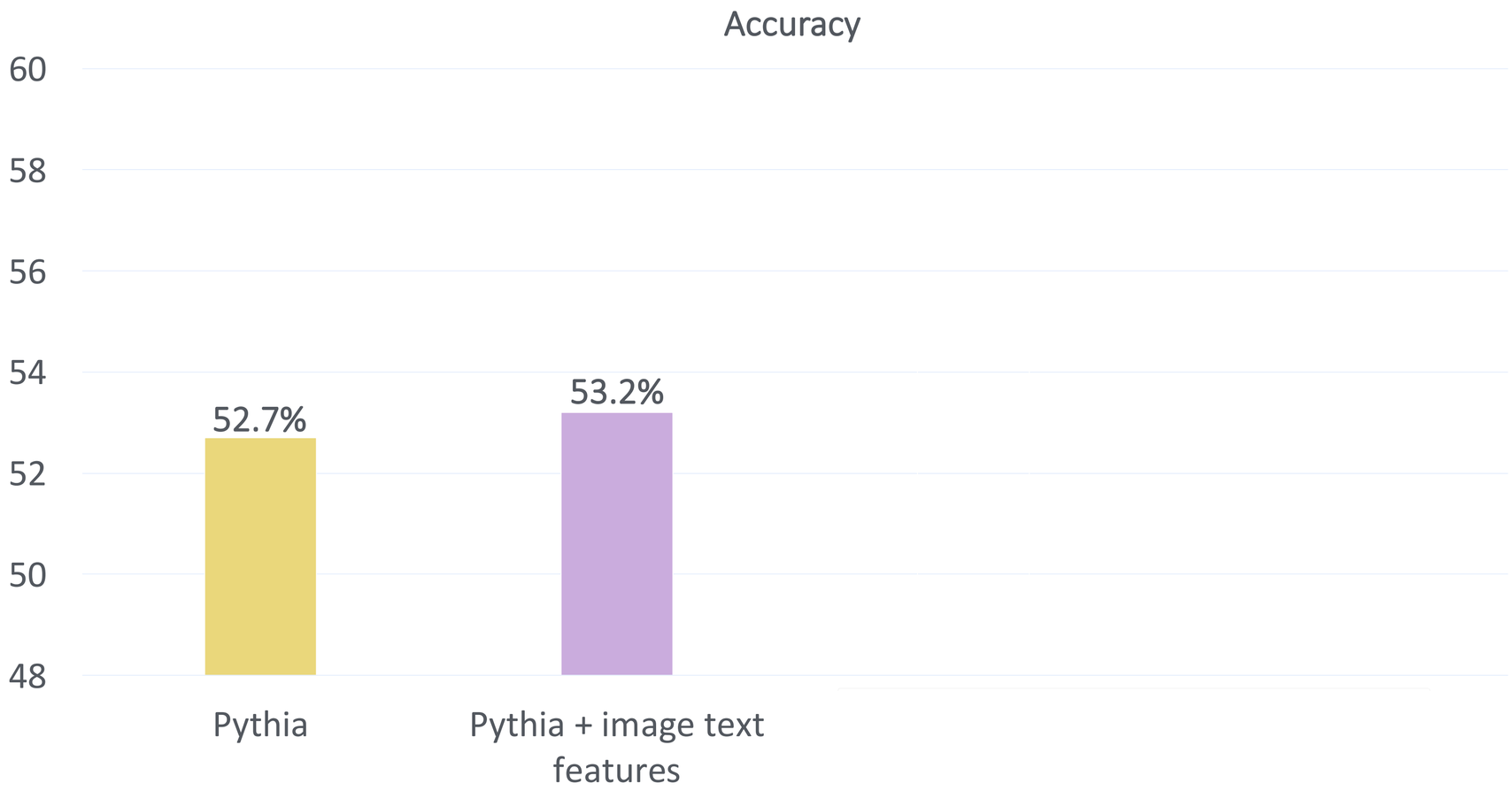
Pythia for Vizwiz VQA

test-dev accuracy



Pythia for Vizwiz VQA

test-dev accuracy



Pythia for Vizwiz VQA

Qualitative results

What does the bottle say?

Image
+
Question

VQA model

ketchup
wine
coffee
cleaner
detergent
chicken
coleslaw

Pythia for Vizwiz VQA

Qualitative results

What does the bottle say?

Image
+
Question

VQA model

ketchup
wine
coffee
cleaner
detergent
chicken
coleslaw

Pythia for Vizwiz VQA

Qualitative results

What does the bottle say?

Image
+
Question

VQA model

ketchup
wine
coffee
cleaner
detergent
chicken
coleslaw

Pythia

Qualitative results

Incorporate results from OCR into the model

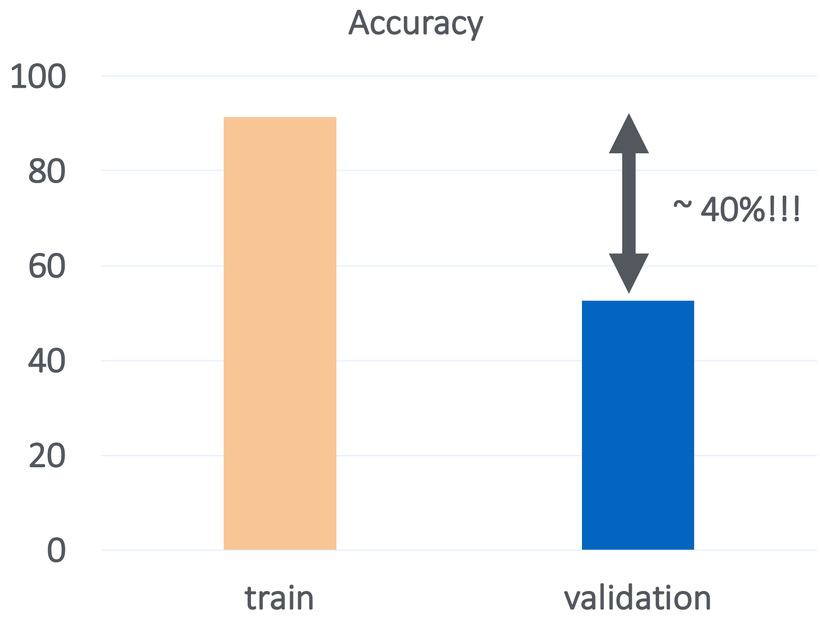
Poor performance on questions that required OCR capabilities

What does the bottle say?

VQA model

ketchup
wine
coffee
cleaner
detergent
chicken
coleslaw

Small dataset, model overfitting



Poor performance on yes/no and number categories which had few examples in the dataset

Pythia

Qualitative results

Incorporate results from OCR into the model

Poor performance on questions that required OCR capabilities

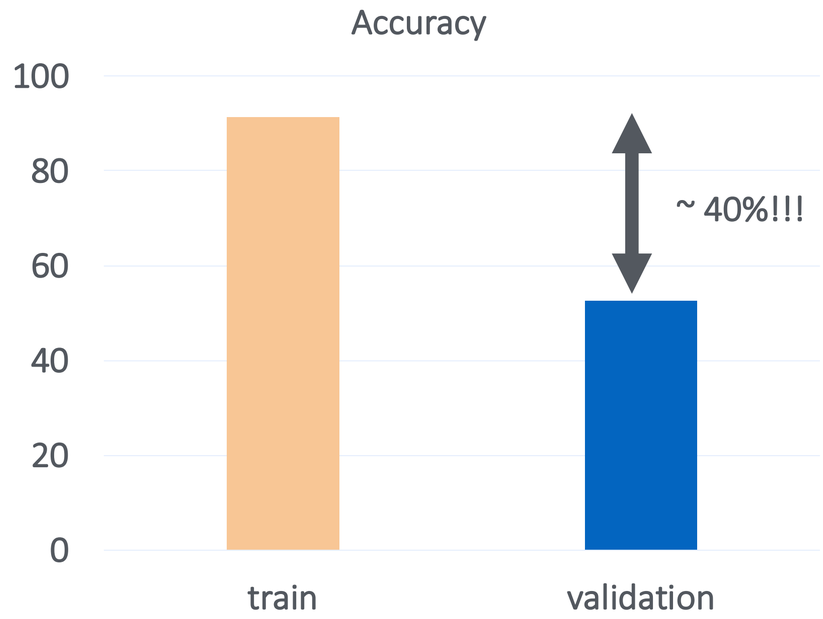
What does the bottle say?

VQA model

ketchup
wine
coffee
cleaner
detergent
chicken
coleslaw

VQA pretraining

Small dataset, model overfitting



Poor performance on yes/no and number categories which had few examples in the dataset

Pythia for Vizwiz VQA

Pretraining on VQA dataset to initialize model parameters

Pythia for Vizwiz VQA

Pretraining on VQA dataset to initialize model parameters

- Vision ✅

Pythia for Vizwiz VQA

Pretraining on VQA dataset to initialize model parameters

- Vision ✅
- Language ✅

Pythia for Vizwiz VQA

Pretraining on VQA dataset to initialize model parameters

- Vision ✅
- Language ✅
- Multimodal fusion ✅

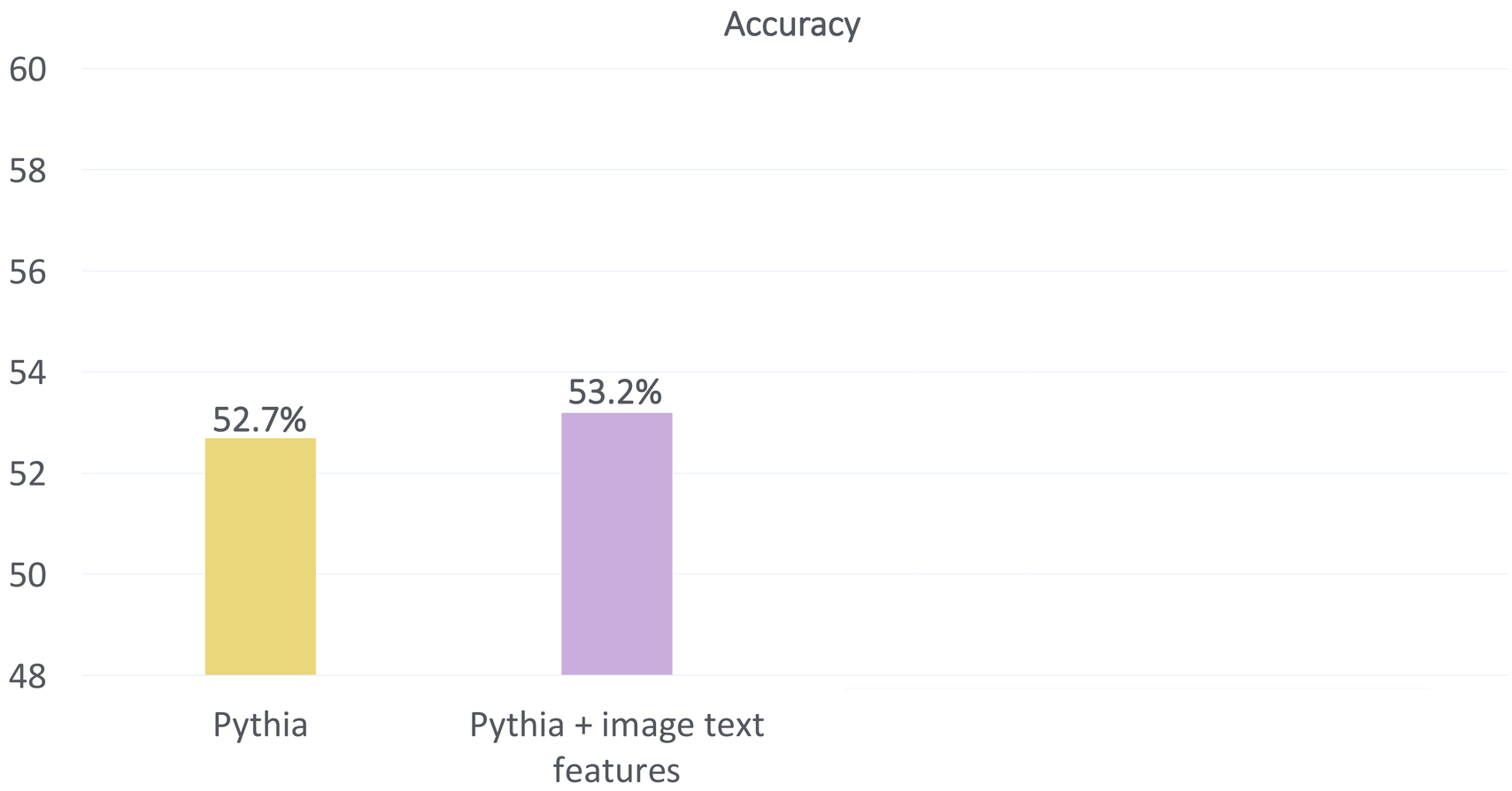
Pythia for Vizwiz VQA

Pretraining on VQA dataset to initialize model parameters

- Vision ✅
- Language ✅
- Multimodal fusion ✅
- Answering 💔
  - The answer spaces are different

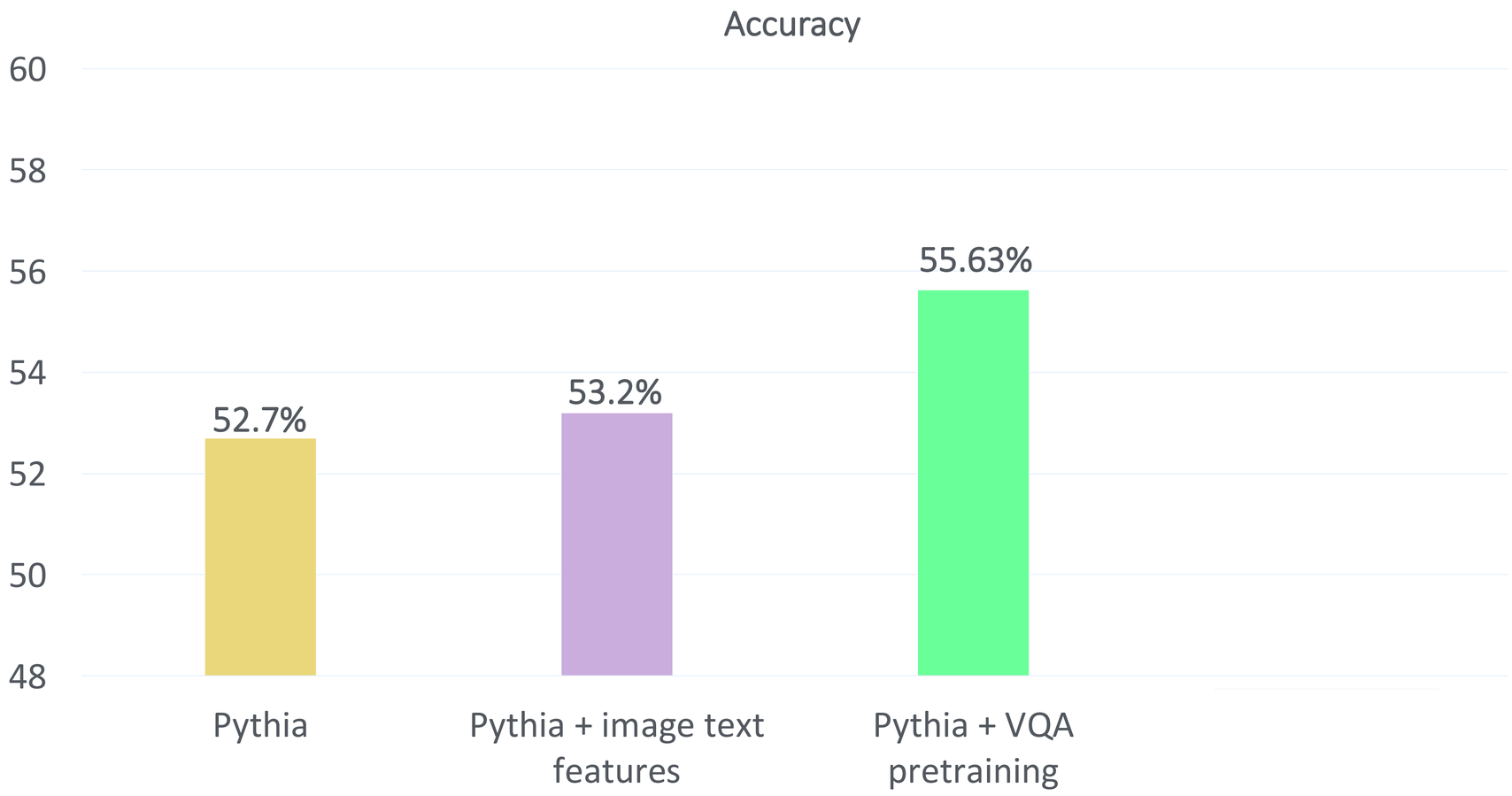
Pythia for Vizwiz VQA

test-dev accuracy



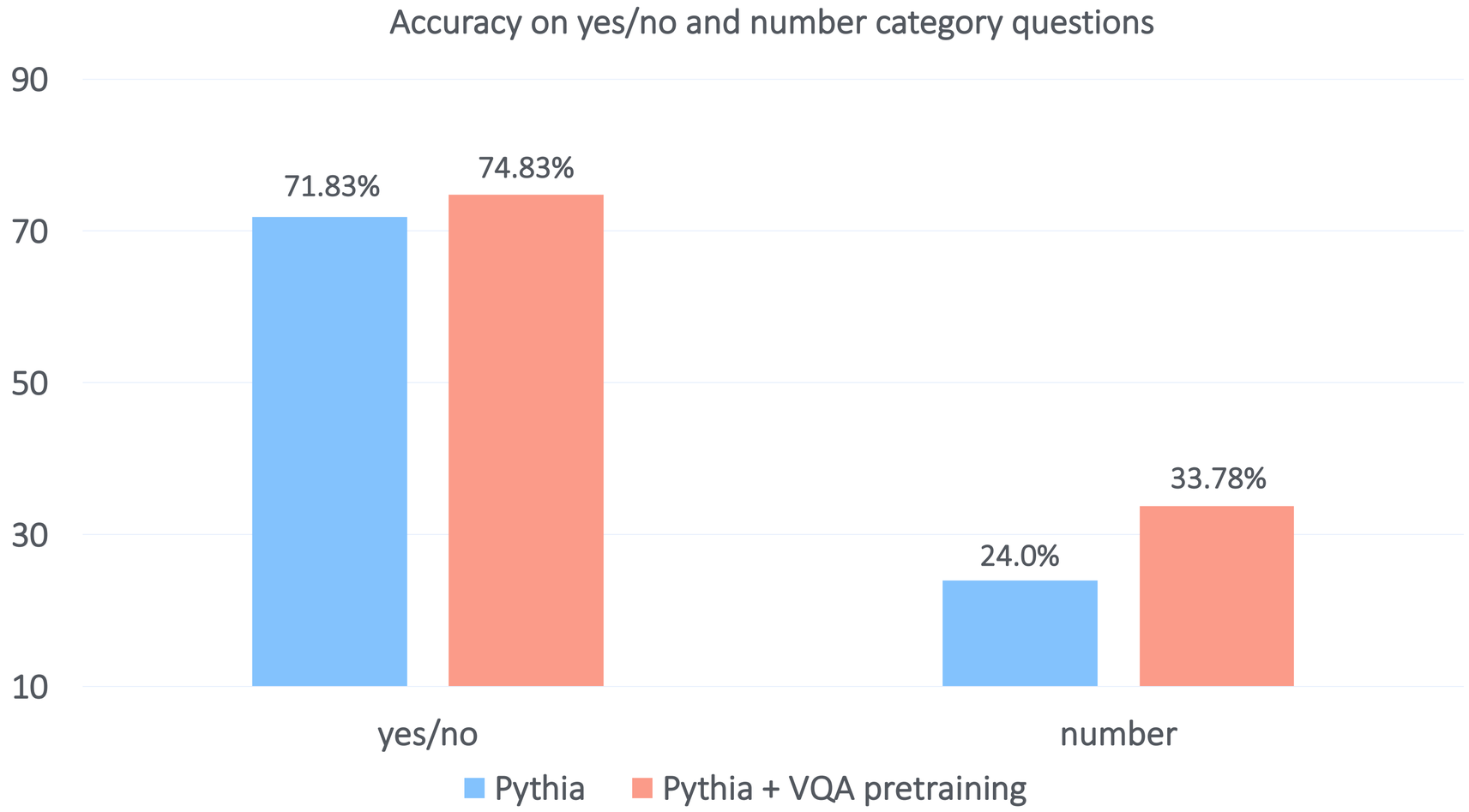
Pythia for Vizwiz VQA

test-dev accuracy



Pythia for Vizwiz VQA

Performance on yes/no and number category questions



Pythia for Vizwiz VQA

Recipes used in VQA Challenge 2018

- Diversified model ensemble

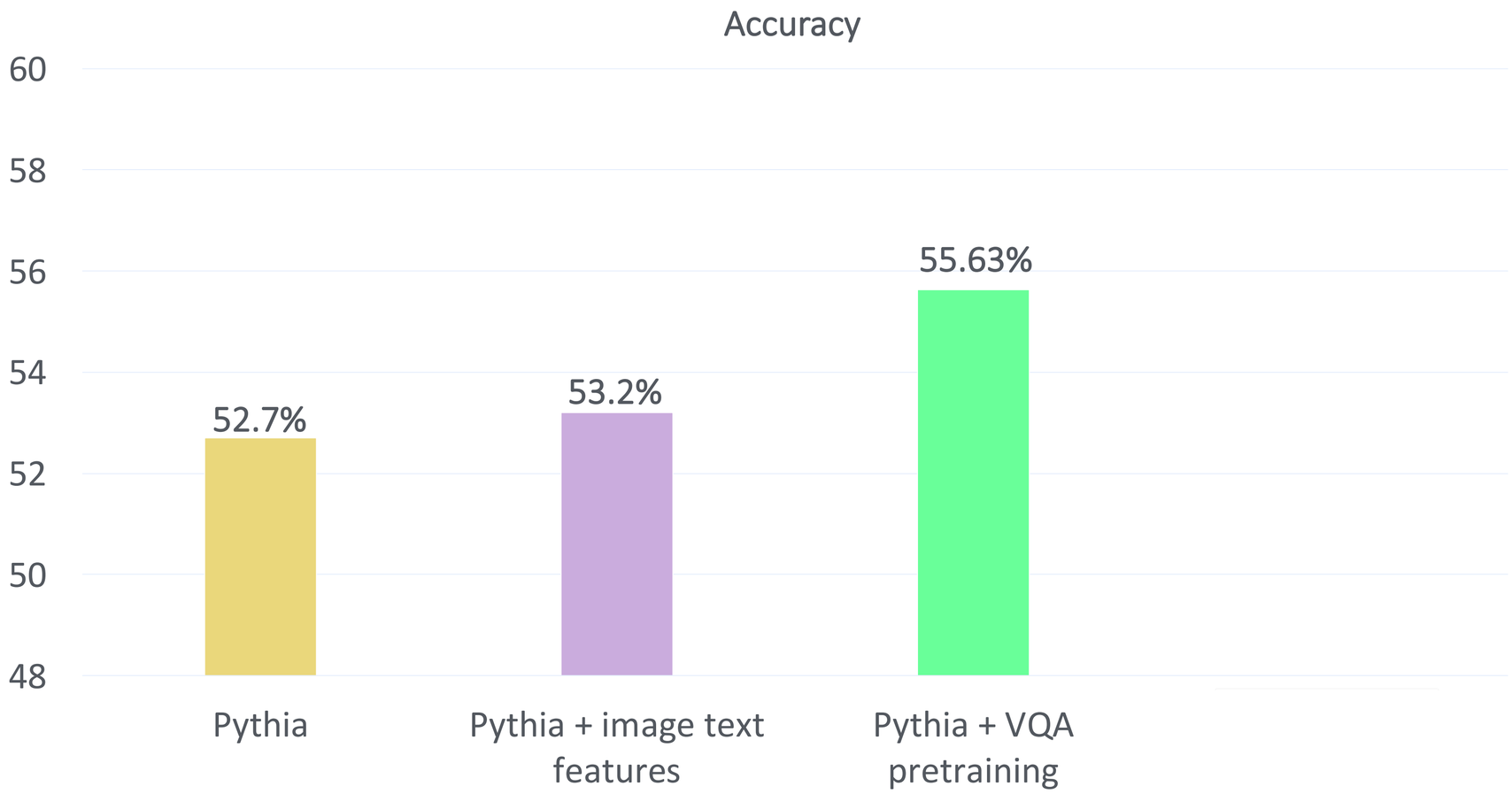
Pythia for Vizwiz VQA

Recipes used in Vizwiz Challenge 2018

- Diversified model ensemble ✅

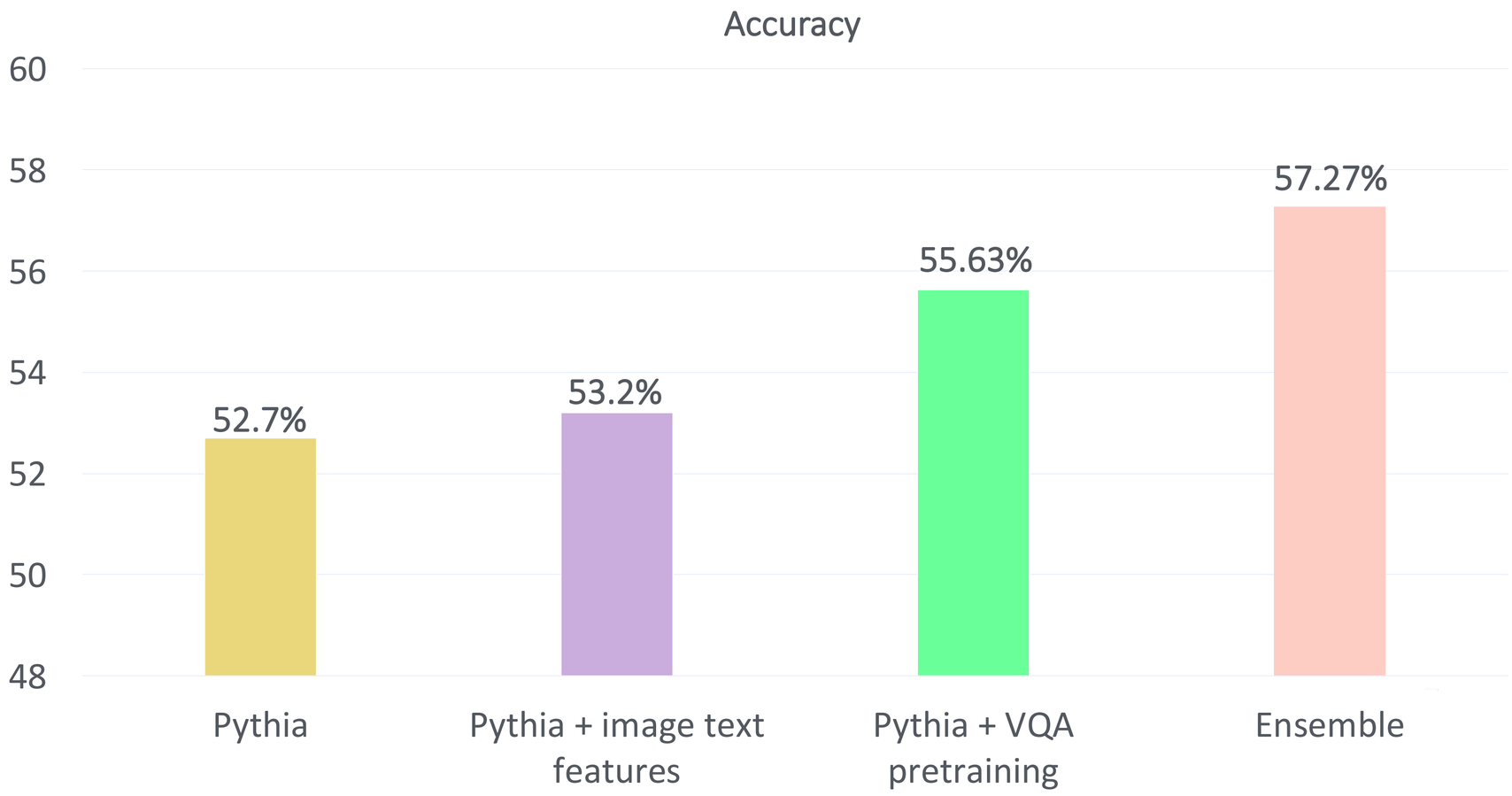
Pythia for Vizwiz VQA

test-dev accuracy



Pythia for Vizwiz VQA

test-dev accuracy



Pythia for Vizwiz Answerability

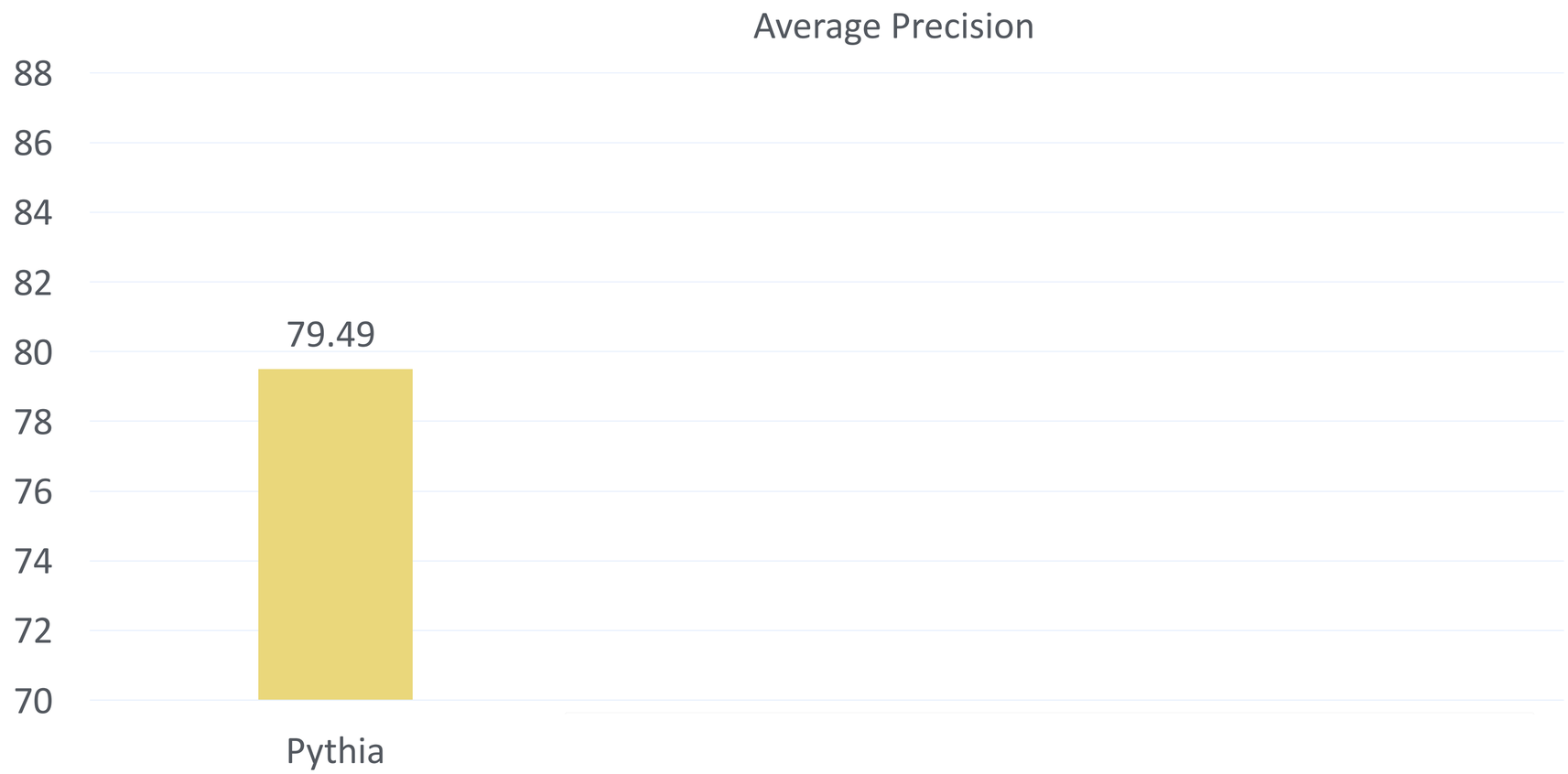
Details

- Same model architecture as for the VQA track except answering classifier module

- Answering classifier predicts one of 3 classes – unanswerable, unsuitable or answerable

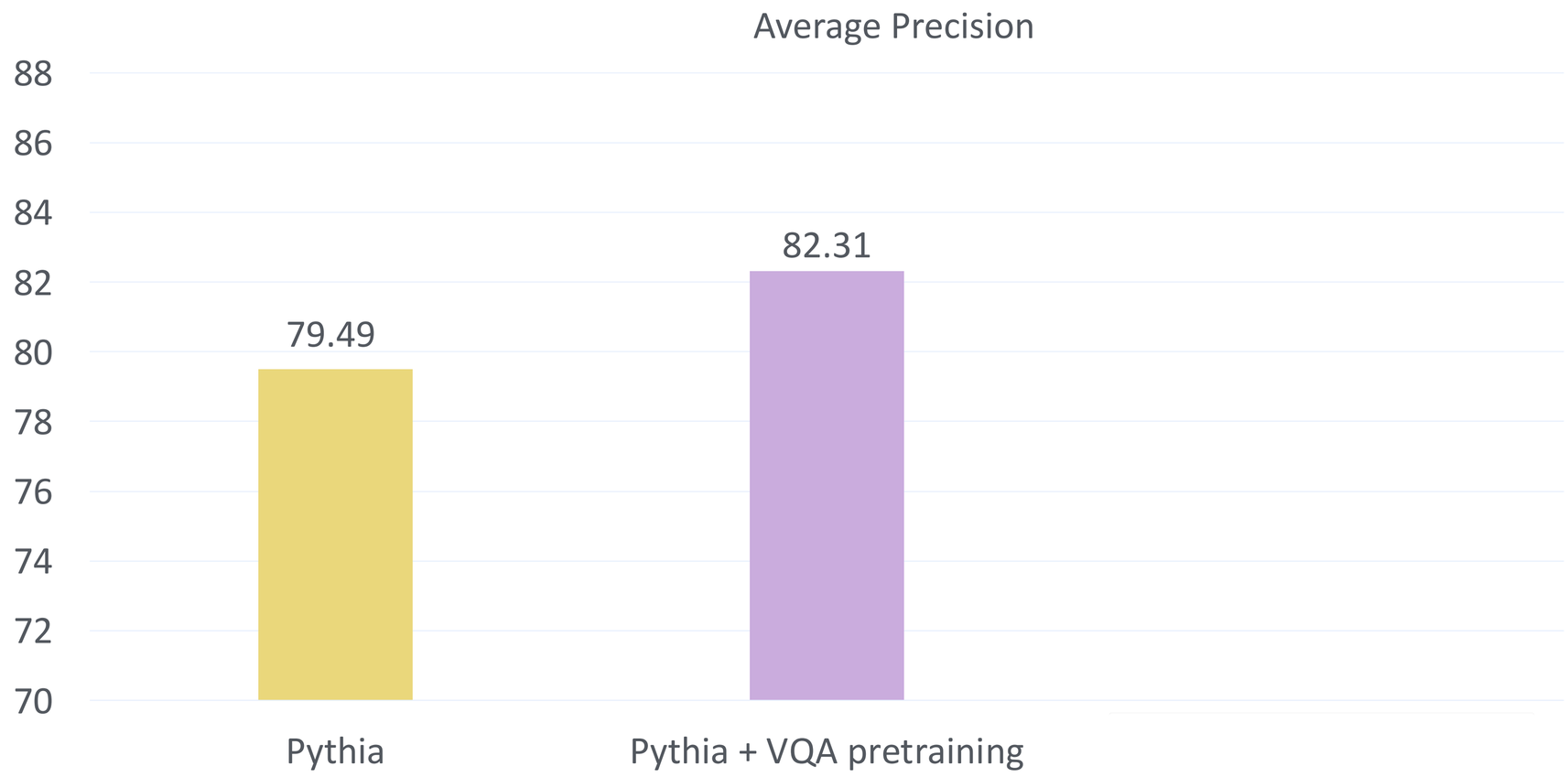
Pythia for Vizwiz Answerability

test-dev average precision



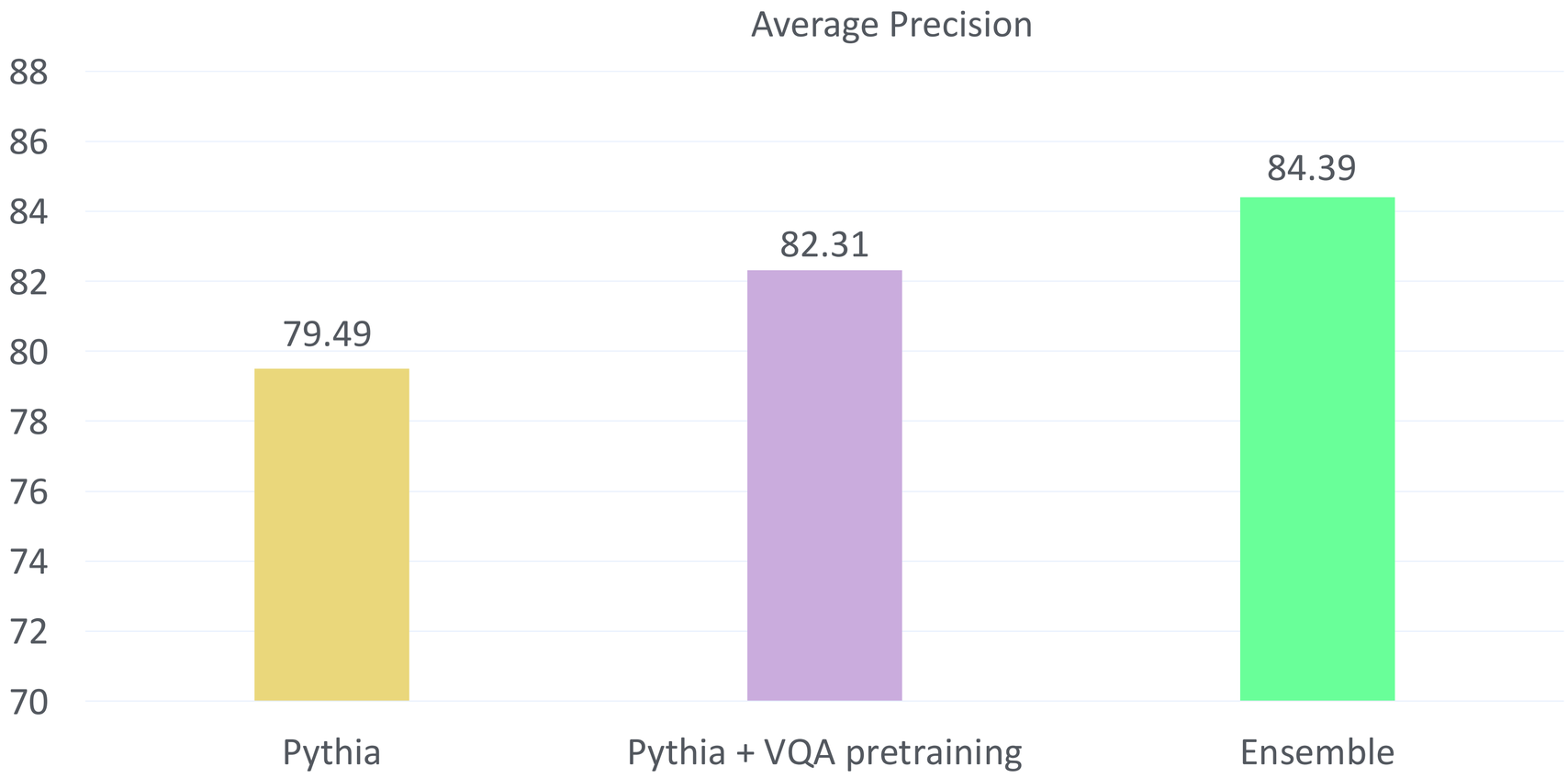
Pythia for Vizwiz Answerability

test-dev average precision



Pythia for Vizwiz Answerability

test-dev average precision

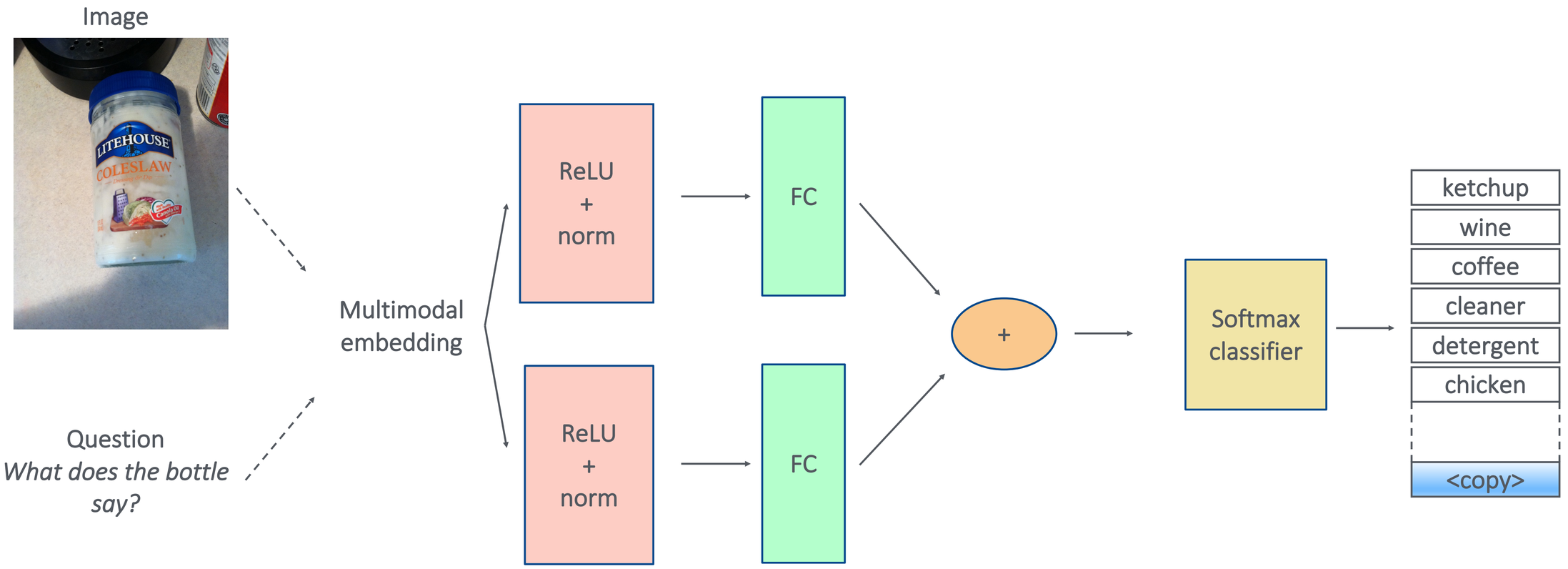


Observations

- Dataset a tad too small to train more complex models
    - Prone to overfitting and high variance in runs
- VQA accuracy not ideal as it penalizes long tail answers even if only one or two tokens differ with ground truth
- Human performance on the validation dataset is quite poor – 57.49% while the best possible performance using the most common answer is 93.16%

Ongoing research

Dynamic answer generation with copy mechanism



Ongoing research

Dynamic answer generation with copy mechanism

Image

Question
What does the bottle say?

Model

<copy>

Use attention on bounding box text

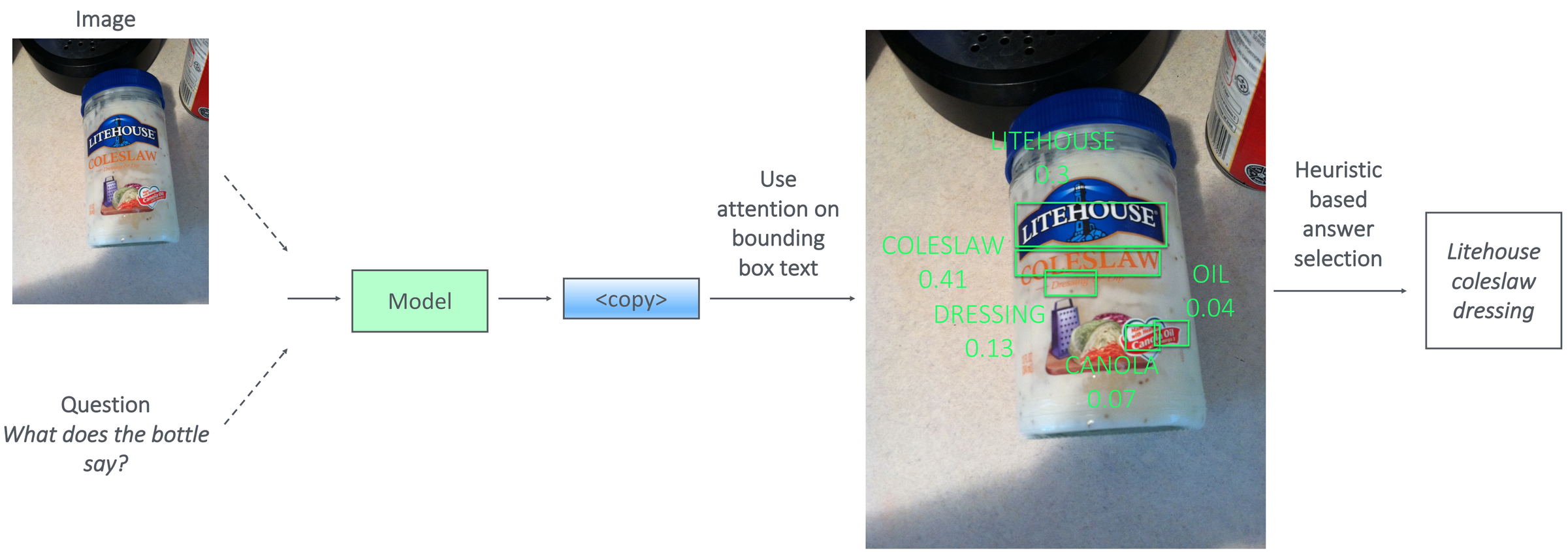
Heuristic based answer selection

Litehouse coleslaw dressing

Ongoing research

Dynamic answer generation with copy mechanism

Accuracy: 54.21% with no VQA pretraining
(Baseline: 53.2%)



Pythia for Vizwiz VQA

Recipes used in VQA Challenge 2018

- Diversified model ensemble

- Data augmentation

Pythia for Vizwiz VQA

Recipes used in Vizwiz Challenge 2018

- Diversified model ensemble

- Data augmentation

Question rephrasings:
What product is this? -> What does the label say?

Pythia for Vizwiz VQA

Recipes used in Vizwiz Challenge 2018

- Diversified model ensemble ✅
- Data augmentation 💔

Pythia for Vizwiz VQA

Recipes used in Vizwiz Challenge 2018

- Diversified model ensemble ✅
- Data augmentation 💔

More work needed to ensure distribution of augmented data matches that of the Vizwiz dataset distribution

Conclusions

- Pythia is awesome!
- OCR text features and VQA pretraining – the two key ingredients of our challenge entry
- However open questions remain on reading text to answer questions, incorporating external knowledge sources and generating tail answers