Visual Question Answering

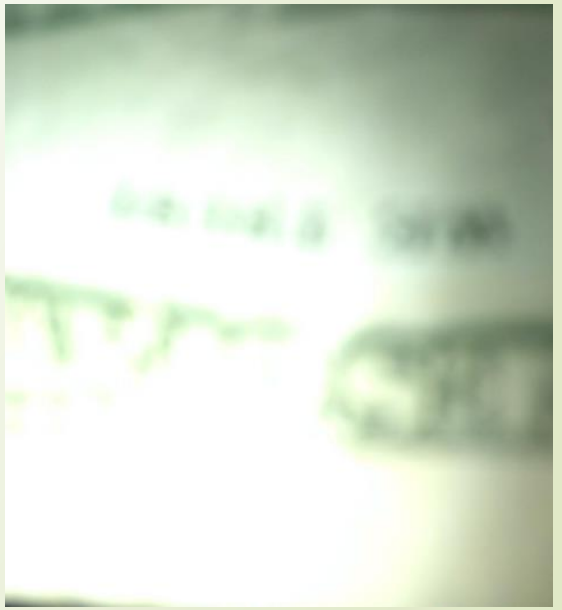
with image captions

Motivation

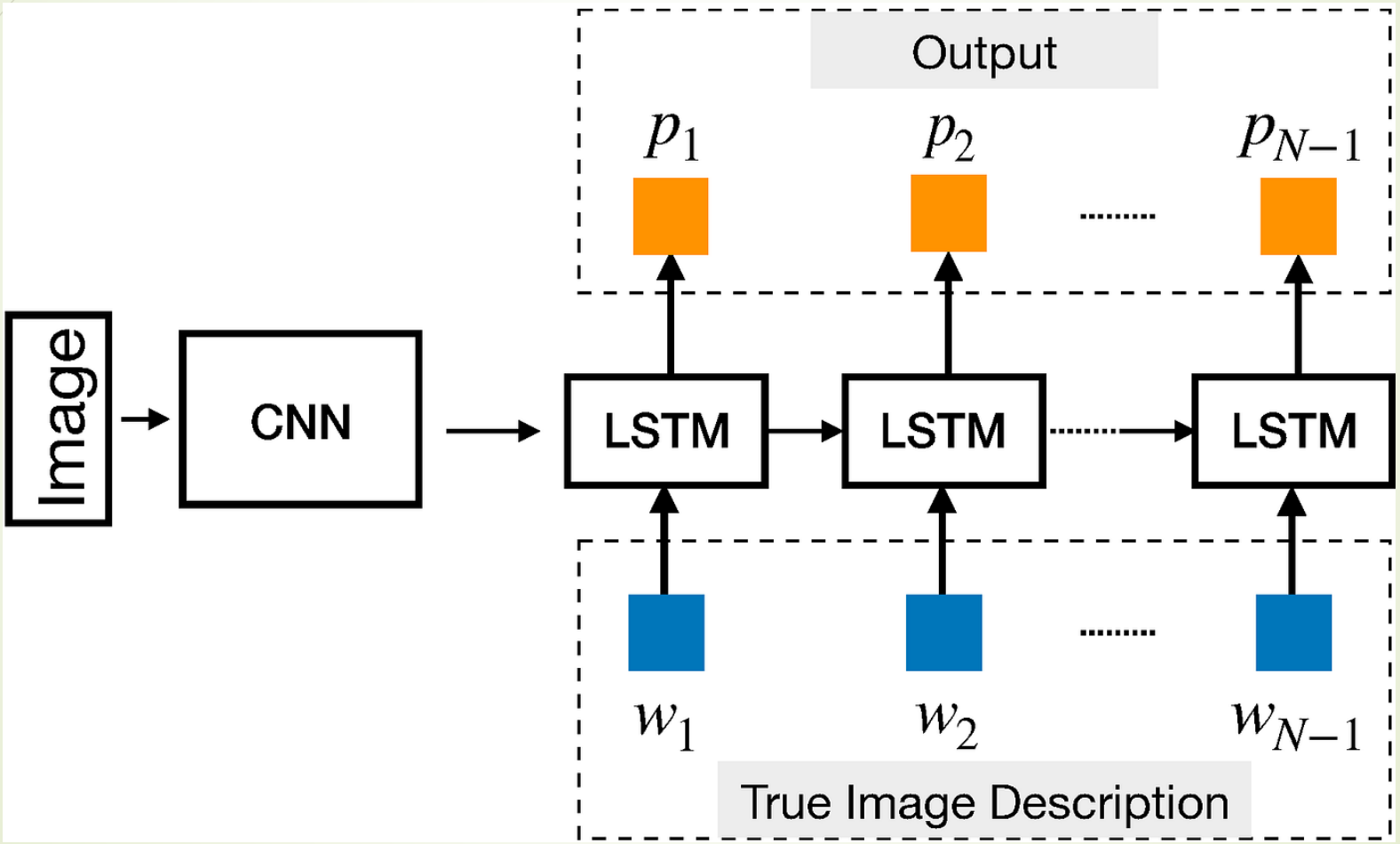
Develop a computationally cheap model for visually impaired people.

Datasets

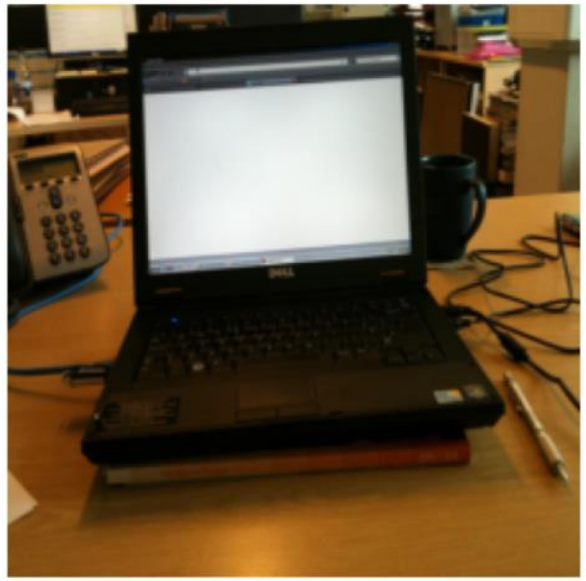
- MSCOCO Captions
- VizWiz



Modeling Image Captions



Results



Generated Caption: [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK]

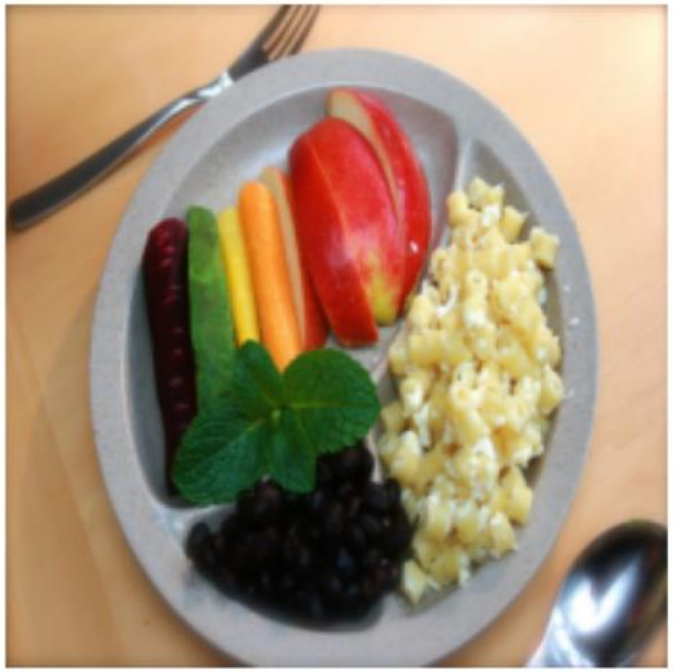
Generated Caption: a a a a a a a a a a

Most prob

Generated Caption: a plate of with with a a a a .

Generated Caption: a man of a a a a a a .

K-prop



Generated Caption: a plate of of with

Generated Caption: a man riding a a

Most prob

K-prop

Generated Caption: a man with a a a a a a .

Generated Caption: an bathroom is with on

Generated Caption: a man of a a a a a a .

Generated Caption: a group of a in a the in a

VQA Model

- Generate image captions with vit-gpt2-image-captioning transformer
- Vision Transformer fine-tuned on ImageNet(base-sized model)
- Bert Tokenizer (questions, captions)
- Bert Model -> embeddings for questions and captions
- GPT Tokenizer (answers)
- Fusing questions, captions and image features
- GPT for generation of open-ended answers

Results

```
Validating:   1%|█                                                                    | 1/103 [00:00<00:36,  2.82it/s]
['., and and::', '., of:::', '...,::', ',, and and and and', '., of of::', '., and and::', '...\n\n\n', '.,,,\nA', '. of and and::', '.,,,\n\n', '\n,,,\n\n', '., and and and 2', '.-\n\n\n\n', '.,\n\n\n\n', '.:,\n\n\n', ',, and::,']
Validating:   2%|██                                                                   | 2/103 [00:00<00:31,  3.19it/s]
['.,,,\n\n\n\n\n\n\n', ',,,:::::::\n', '.i- A A A A A\n\n\n', '.,,\n\n\n\n\n\n\n\n', '.,,,..\n\n\n\n\n\n', ',,\n\n\n,,,,,,.', ',,:::::::\n\n', ',. and\n::...\n', ', the and and and::::::', '.l and and:::::::', '.,-:: - -\n\n\n\n', '...\n\n\n\n\n\n\n\n', '..,,,,,,,,.\n', '., 2::::::\n\n', ',.,,\n\n\n\n\n\n\n', '.,,,,,,,\n\n\n\n']
Validating:   3%|██                                                                   | 3/103 [00:00<00:28,  3.55it/s]
[',l and and and and and', '..,::::', '., of,,,,', '.,:::::', '.,:::::', '.. from from from::', '.,,\n\n\n\n', ',,,\n\n\n\n', ".'s and and:::", '., and and::', '.l and::::', '.. and::::', ',. and and and and and', '.l. A A\n\n', ',,,,\n\n\n', '.l and::::']
Validating:   4%|███                                                                  | 4/103 [00:01<00:27,  3.59it/s]
['.. and and and and::', ',l.\n\n\n\n\n', ', of and and and and and and', '...\n\n\n\n\n', '.,......', ',,,:::""', '. to and and and and and and', '.,,,,\n\n\n', '.,,,,,,,\n', ', and and and,,,,', '.,\n\n\n\n\n\n', ',us and and and and::', ',., a::::', '.,,\n\n\n\n\n\n', '.:,,,,,,,', '.,\n\n\n\n\n\n\n']
Validating:   5%|███                                                                  | 5/103 [00:01<00:28,  3.49it/s]
['.,\n\n', ',, and and', '.,.\n', '.,. A', ', of and:', '.,.\n', '., and:', '.,,,', ',,,\n', '.. and and', '.. and,', '.,, and', '.,,\n', '...:', ',...', '., A\n']
Validating:   6%|████                                                                 | 6/103 [00:01<00:27,  3.49it/s]
[',,\n', '., and', ',,,', '.l.', '.,,', '...', '., and', ',,,', ',.,', ',, and', '. and and', '..,', '.l.', '...', '., and', '.. the']
Validating:   7%|█████                                                                | 7/103 [00:01<00:25,  3.72it/s]
['.. 2 and and and:', ',, and and:::', '., and and:::', '.,,,,\n\n', '.., and and the the', '.,., A A A', '., and and and::', ',,, and and and,', ', of and and and and and', '., and::::', '...\n\n\n\n', '.,,,\n\n\n', ',, and and a a a', '., the::::', '......\n', '..\n\n\n\n\n']
Validating:   8%|█████                                                                | 8/103 [00:02<00:25,  3.74it/s]
['.,,, A', '.-,\n\n', '.on:::', '.,,,,', '.,,::', '.: of..', '\n,,\n\n', '.,,\n\n', '.. and and:', ',, and and and', '.,,,,', '., and and and', ',....', ',,:::', '.,...', '.....']
Validating:   9%|██████                                                               | 9/103 [00:02<00:26,  3.56it/s]
['.,,::::::\n\n', '., and and and and and and: - -', '.,\n\n\n\n\n\n\n\n\n', '.,,,,,\n\n\n\n\n', '......\n\n\n\n\n', ',,,,\n\n\n\n\n\n\n', '.l.,\n\n\n\n
```

Challenges and Future work

- No good results
- Hard to answer open-ended questions without multi-classification

- Using only GPT and exclude BERT
- Trying this model on less noisy data
- Maybe trying to improve the image captioning model

References

- https://arxiv.org/pdf/1505.00468
- https://arxiv.org/pdf/1603.02814v2
- https://arxiv.org/pdf/2311.00308

Thank you. Questions?