Visual Question Answering

Combining Natural Language Processing and Computer Vision

Pictures are worth 1000 words*

*in natural language

Motivation

Humans convey knowledge most naturally through language and visuals

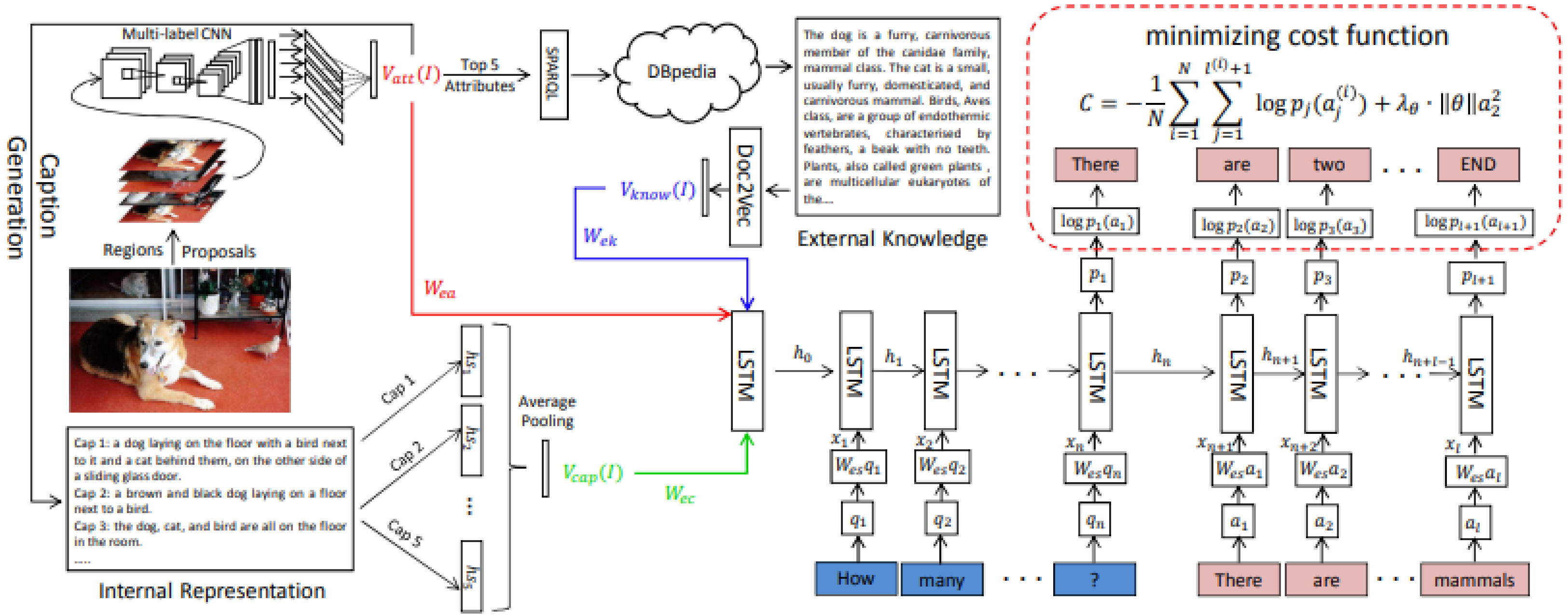
Before Visual Question Answering (VQA), there was a disconnect between the information gleaned from images and the incredible user interface of natural language

After VQA

- We can ask questions about the content of images
- “Is my mole cancerous?”
- “Did this car run a red light?”
- There are so many possible extensions as well

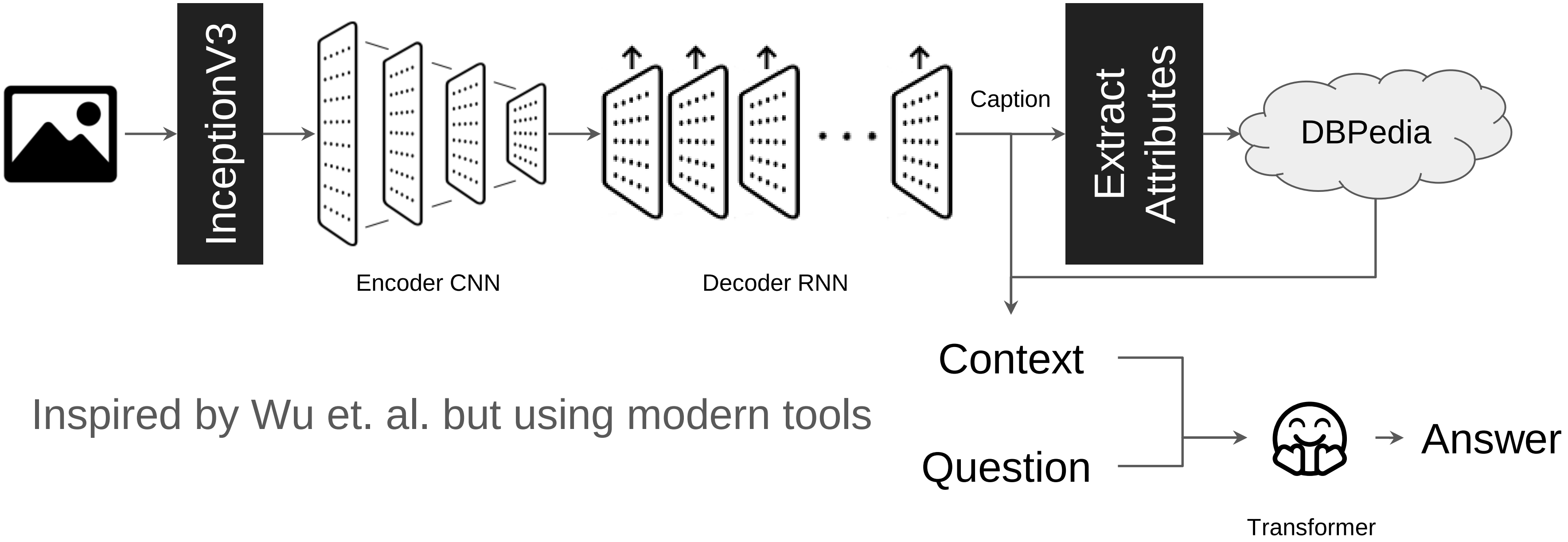
Method

Inspiration Method



Wu et. al. (2016)

Our Method

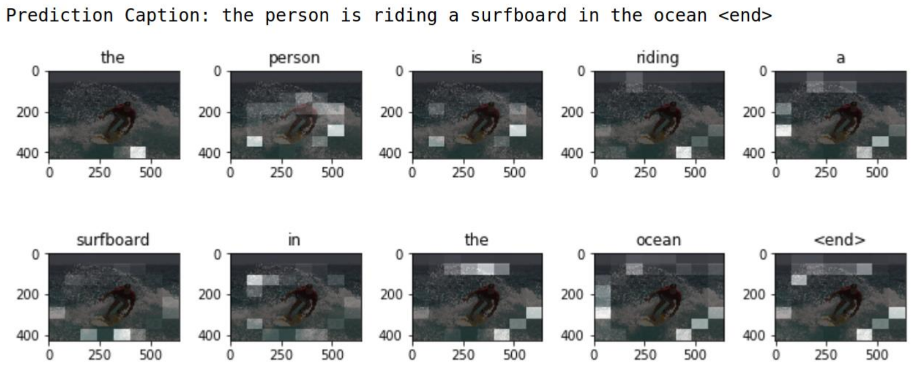


Inspired by Wu et. al. but using modern tools

Generating Captions

Convolutional Neural Network and Recurrent Neural Network using Tensorflow

1. Preprocess the COCO Captioning images using InceptionV3 to extract features
2. Create a vocabulary of all the terms used in the training captions and turn them into vectors
3. Create the model in an encoder-decoder pattern
4. Train the model
5. Evaluate the model



Gaining Attributes from Captions

We use the captions to extract the attributes for which we want to query the knowledge base because we only want the important attributes, which the captions mention.

1. Tokenize the caption
2. Lemmatize
3. Remove stop words
4. Any remaining terms get sent to the KB query

The person is riding a surfboard in the ocean

↓

Ride, surfboard, ocean

Querying for External Knowledge

We query DBpedia, which is Wikipedia as a knowledge base, for the extracted attributes, and concatenate those to the caption to create the context for the question.

- If a term can't be found in DBpedia, it is skipped, but adding more knowledge bases would be a good extension to the project

The dog or domestic dog, (Canis familiaris or Canis lupus familiaris) is a domesticated descendant of the wolf which is characterized by an upturning tail. The dog derived from an ancient, extinct wolf, and the modern grey wolf is the dog's nearest living relative. The dog was the first species to be domesticated, by hunter–gatherers over 15,000 years ago, before the development of agriculture.

Fine-Tuning a 🤗 Transformer for Question Answering

We used the Tensorflow method of fine tuning a 🤗 Transformer

- First, we had to create a context for all of our images using the attribute extraction and querying process
- Then we used the context and question along with the answer to do supervised training on the Transformer

Training Times

Using a Google Colab GPU, it took ~34 minutes to run 20 epochs to train the captioning model on a GPU to get .285 loss on the training set

Fine-tuning the 🤗 Transformer took 80 mins with 10K questions and answers

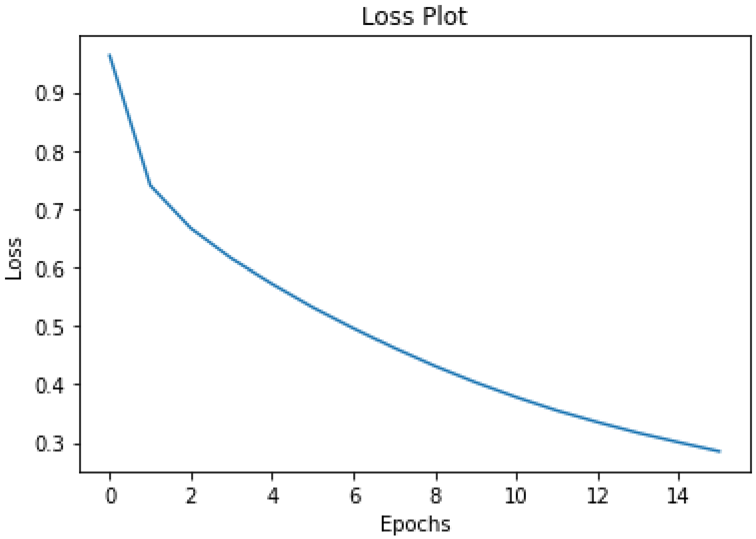
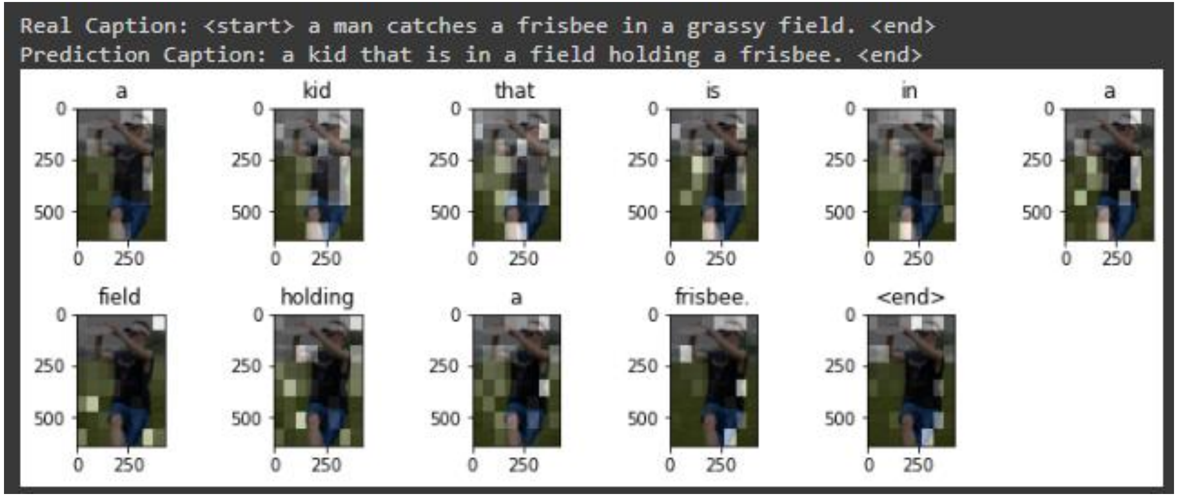
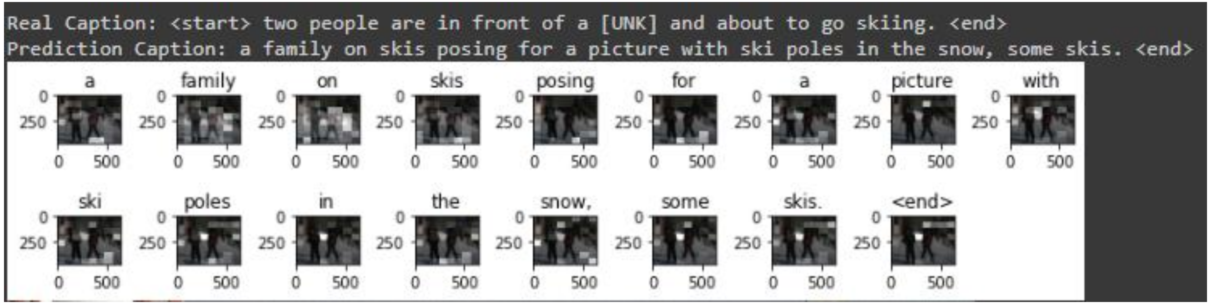
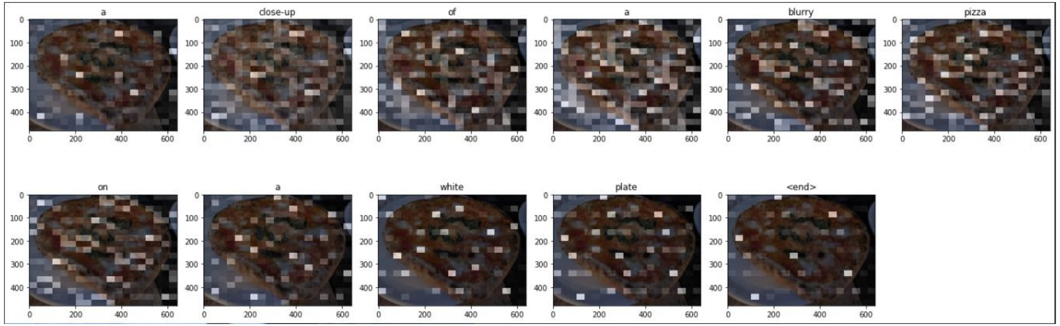


Image Captioning Model output


Real Caption: <start> a man catches a frisbee in a grassy field. <end>
Prediction Caption: a kid that is in a field holding a frisbee. <end>
a kid that is in a
field holding a frisbee. <end>


Real Caption: <start> two people are in front of a [UNK] and about to go skiing. <end>
Prediction Caption: a family on skis posing for a picture with ski poles in the snow, some skis. <end>
a family on skis posing for a picture with
ski poles in the snow, some skis. <end>

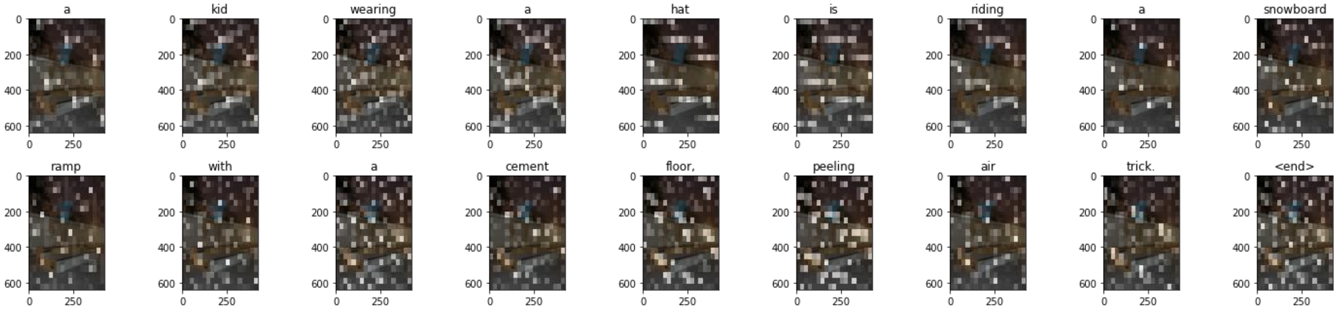
Real Caption: <start> a cheese pizza sitting on a plate <end>
Prediction Caption: a close-up of a blurry pizza on a white plate <end>



Real Caption: <start> two people talking on their cell phones on the bus. <end>
Prediction Caption: children squat up in bus. <end>

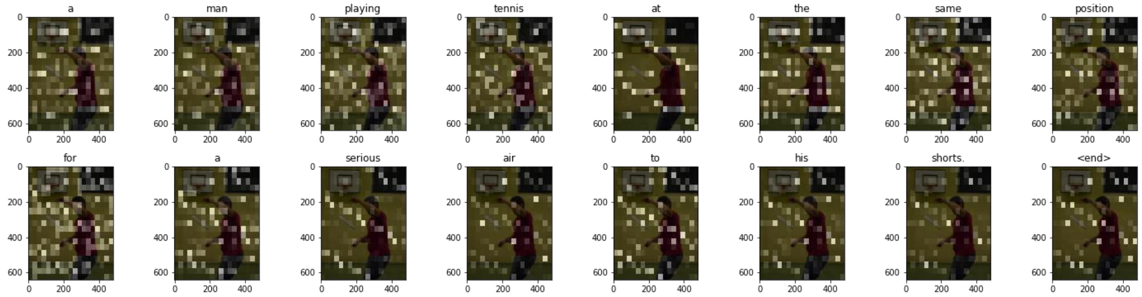
Real Caption: <start> altered photograph of a skateboarder doing a trick at night <end>

Prediction Caption: a kid wearing a hat is riding a snowboard ramp with a cement floor, peeling air trick. <end>



Real Caption: <start> a man is playing with a frisbee at an indoor [UNK] <end>

Prediction Caption: a man playing tennis at the same position for a serious air to his shorts. <end>



VQA Model output

Question What is their near to car?
Possible correct answers {'answer_start': [0, 0, 0, 0], 'text': ['bus', 'sidewalk', 'curb', 'tree']}
Captions generated : ['a', '[UNK]', 'bus', 'is', 'traveling', 'down', 'a', 'tarmac.', '<end>']
Predicted Answer: bus

Question What sport are they playing?
Possible correct answers {'answer_start': [0, 66], 'text': ['snowball fight', 'skiing']}
Captions generated : ['a', 'bunch', 'of', 'people', 'skiing', 'across', 'a', 'hill', 'near', '[UNK]', '<end>']
Predicted Answer: winter

Question What kind of animal is this?
Possible correct answers {'answer_start': [2], 'text': ['cat']}
Captions generated : ['the', 'orange', 'and', 'white', 'cat', 'is', 'sitting', 'near', 'a', '[UNK]', '<end>']
Predicted Answer: cat

Question What are the boys learning in this sport?
Possible correct answers {'answer_start': [0, 0, 0, 0, 0, 0, 21, 0], 'text': ['how to balance ball', 'soccer', 'dribbling', 'ball handling', 'ball control', 'yes', 'kicking', 'patience']}
Captions generated : ['three', 'children', 'playing', 'with', 'a', 'frisbee.', '<end>']
Predicted Answer: frisbee

Question What is the cat playing with?
Possible correct answers {'answer_start': [0, 0, 0, 0, 0, 0, 0, 0], 'text': ['purse', 'cord', 'key chain', 'toy', 'scissors', 'string', 'strap', 'ribbon']}
Captions generated : ['a', 'very', 'cute', 'cat', 'eating', 'on', 'top', 'of', 'a', 'bed', '<end>']
Predicted Answer: bed

Question Where are the people flying the kites?
Possible correct answers {'answer_start': [0, 0, 0], 'text': ['beach', 'on beach', 'in background']}
Captions generated : ['several', 'kites', 'flying', 'on', 'top', 'of', 'the', 'beach', '<end>']
Predicted Answer: several kites flying on top of the beach

Next Steps

Future Improvements

- Adding more knowledge bases to the query to get more data about more attributes
- Training the models more with more diverse datasets
- Increasing embedding layer size and vocabulary size in image captioning to capture some fine level features
    - Consider if the question asks a person’s shirt color, we can’t capture that at the moment

Extensions

- Turn this system into a library that can be trained with more specific datasets for specialized purposes (at the moment, it’s very general)

Thank you for listening!

Any questions?