SCENE DESCRIPTION GENERATION

Using Deep Learning Methods

Group 12

Abirami Dhayalan - Arjun Sridhar - Batul Petiwala

Scene Description

- For the image we textually describe its:
  - Visual content
  - Objects in the image
  - Interaction between objects

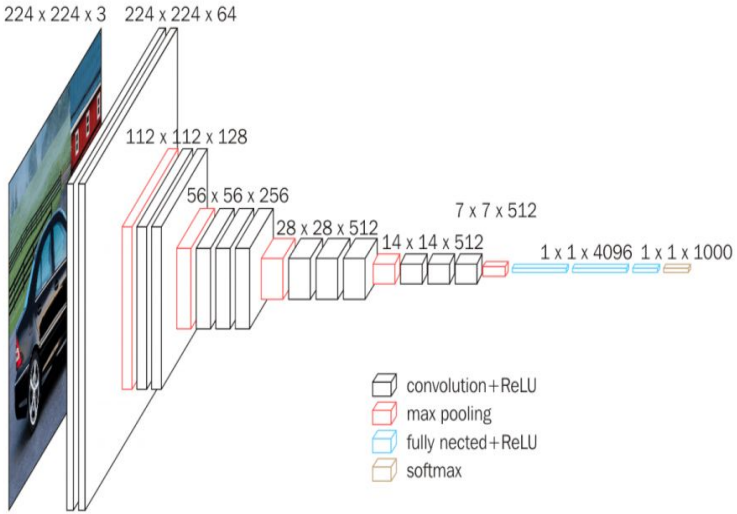
- Very important in helping to replicate human perception task
  - Image caption methods are not new but they can be computationally expensive

- Our approach is to try and use a combination of existing state of the art methods

Person riding motorcycle on raceway

Related Work

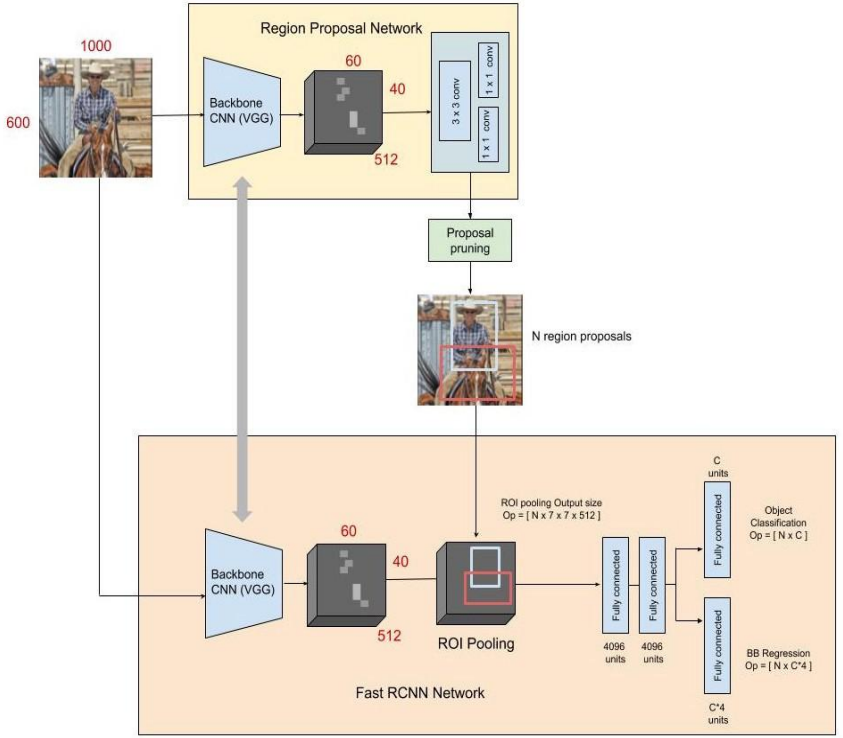
- Deep Learning for Image Classification
  - Train network to classify images into class
  - Convolutional Neural Networks (CNN)
  - ResNet-50
  - VGG16



[1]

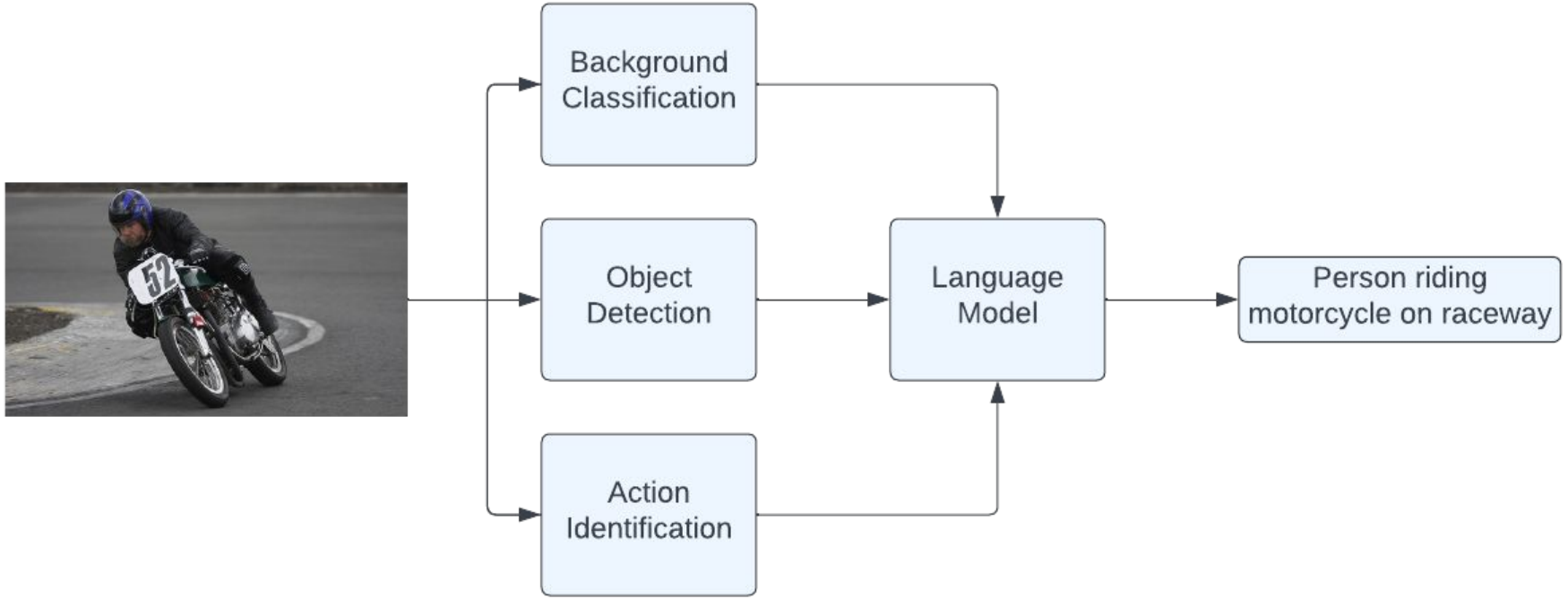
Related Work

- Object Detection
  - Faster-RCNN
  - Extension of Fast R-CNN
  - Uses Region Proposal Network (RPN)
  - Fast R-CNN + RPN = Faster R-CNN



[2]

Approach


Background Classification
Object Detection
Action Identification
Language Model
Person riding motorcycle on raceway

Background Classification

- Pretrained CNN Model on Places365 dataset
  - ResNet50 architecture
  - 85% top-5 accuracy
- Dataset has over 10 million images and more than 400 unique scenes
- Authors have provided PyTorch open-source models to use: https://github.com/CSAILVision/places365 [3]

Image from dataset with label food_court[3]

Object Detection

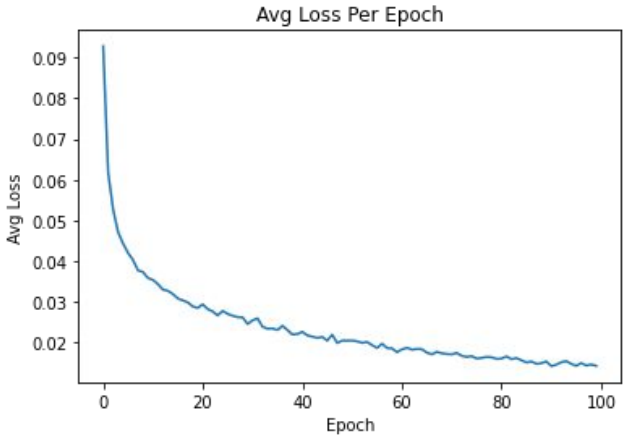
- PyTorch Pretrained Faster R-CNN model
  - ResNet50 Architecture

- Detected objects are passed on to the Language Model

Person and Surfboard detected in the image

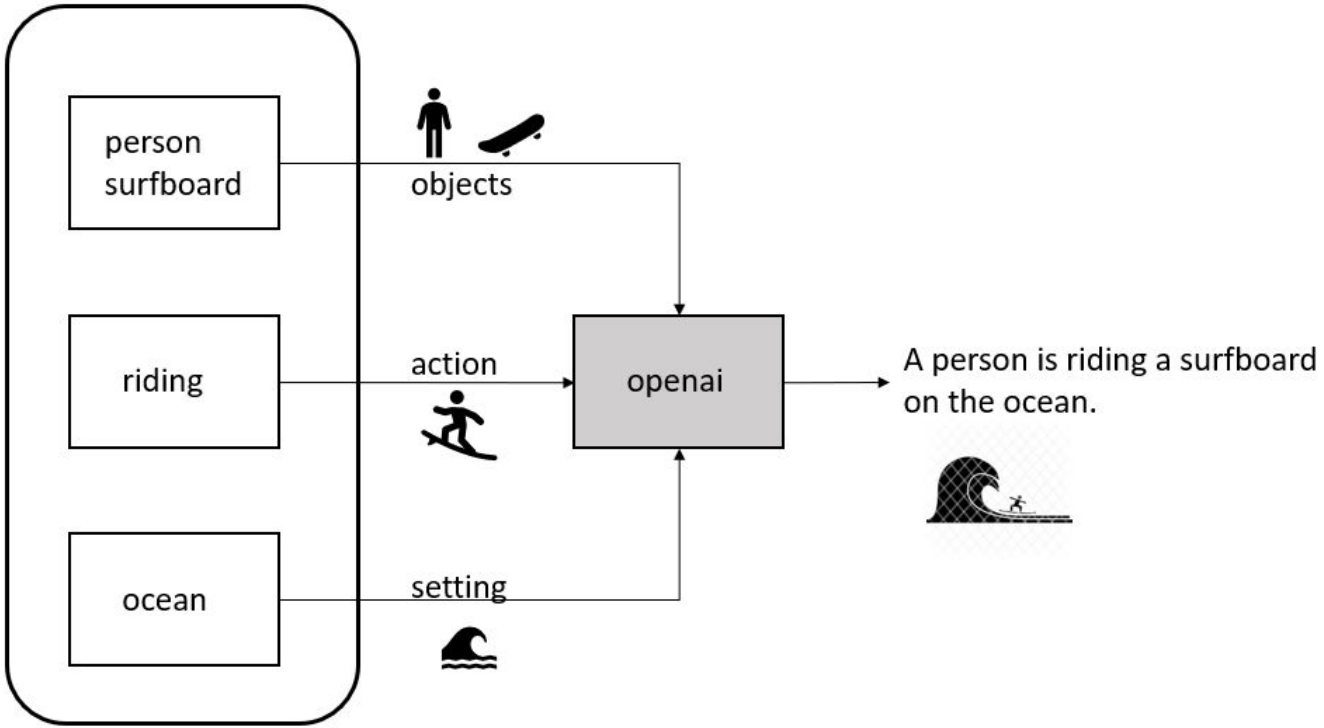
Action Identification

- Trained model on Stanford 40 Actions dataset
  - 62% accurate
- Dataset has 4,000 training images, over 5,000 test and 40 actions [4]
  - Around 200 images per action [4]
- ResNet50 architecture
  - Would like to compare with other architectures such as VGG16



100 epochs, 32 batch size, 0.001 learning rate

Language Model

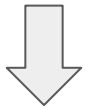


Openai GPT 3

- Generative Pretrained Transformer 3 [5]

- Uses Deep Learning to produce human text [5]

Person fishing river

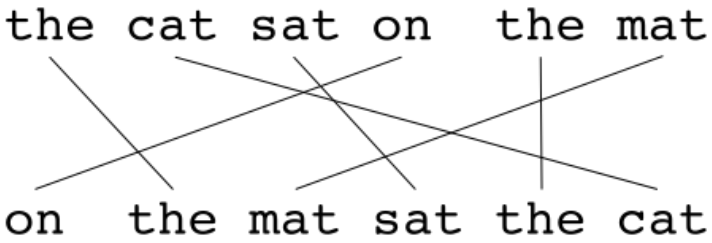


A person fishing in a river

Evaluation

METEOR

- Metric for the evaluation of machine translation output [6]
- Higher score => closer to reference caption [6]



[7]

BLEU

- BiLingual Evaluation Understudy [6]
- Similarity between the model translated & reference captions [6]

Demo

- Flickr8k Dataset: 8,000 photos and up to 5 captions for each photo [8]

Caption in the Dataset: Person riding their bicycle on the street with a backpack on

Demo

Action Predicted: **Riding**

Scene Predicted:

1. residential_neighborhood
2. parking_lot
3. motel
4. gas_station
5. street

Caption Generated:
A person riding a bicycle with a backpack in a residential neighborhood.

Ground truth:
Person riding their bicycle on the street with a backpack on

Meteor Score:
0.74

BLEU Score:
0.67

References

- [1] https://neurohive.io/wp-content/uploads/2018/11/vgg16-1-e1542731207177.png
- [2]https://towardsdatascience.com/faster-r-cnn-for-object-detection-a-technical-summary-474c5b857b46
- [3] Places: A 10 million Image Database for Scene Recognition B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017
- [4] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. Internation Conference on Computer Vision (ICCV), Barcelona, Spain. November 6-13, 2011.
- [5] https://arxiv.org/pdf/2005.14165.pdf
- [6]https://medium.com/explorations-in-language-and-learning/metrics-for-nlg-evaluation-c89b6a781054
- [7] https://upload.wikimedia.org/wikipedia/commons/2/27/METEOR-alignment-a.png
- [8] Hodosh, Micah, Peter Young, and Julia Hockenmaier. ”Framing image description as a ranking task: Data, models and evaluation metrics.” Journal of Artificial Intelligence Research 47 (2013): 853-899

Thank you!