Deep Learning for Image Captioning

SIDDESH PILLAI (srp4698@rit.edu)
ADVISOR: PROF. JEREMY BROWN
ROCHESTER INSTITUTE OF TECHNOLOGY

MOTIVATION

- Build an image captioning application which narrates the scene and can help partially blind people understand the scene

Caption: A man wearing sunglasses playing a silver guitar

Figure 1: Motivation: Natural Language description of an image

OBJECTIVES

- Develop a model which employs deep learning techniques in order to identify objects in an image and generate descriptions of image regions
- Create an application which interacts with the model in real time and provides description in text and voice
- Evaluate the performance of the model on the basis of how resistant it is to negative effects of scale and color augmentation

APPROACH

- Data was collated from Flickr30k and MSCOCO datasets
- The data consists of over 70k instances of images with annotated captions

@ Two golden brown horses pull a sleigh driven by a woman in a blue coat .
@ Two horses pull a carriage driven by a woman over snow covered ground .
@ Two horses pulling a sled steered by a smiling blond woman .
@ Two draft horses pull a cart through the snow .
@ Two horses are pulling a woman in a cart .

Figure 2: Sample Data: Image with annotated captions

APPROACH

- Used VGGNet[3] and ResNet[2] pre-trained models for image detection
- First step is to create a model to align sentence snippets to visual regions using multi-modal embedding

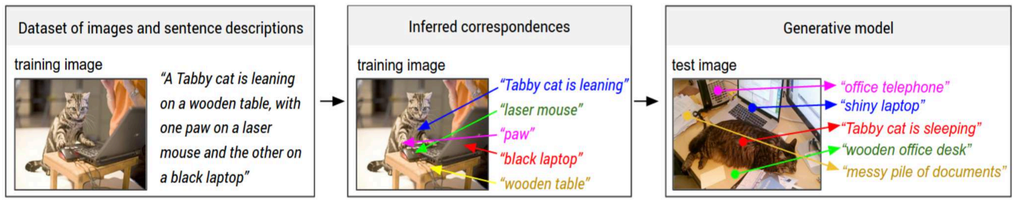

Figure 3: Overview of the model creation[1]

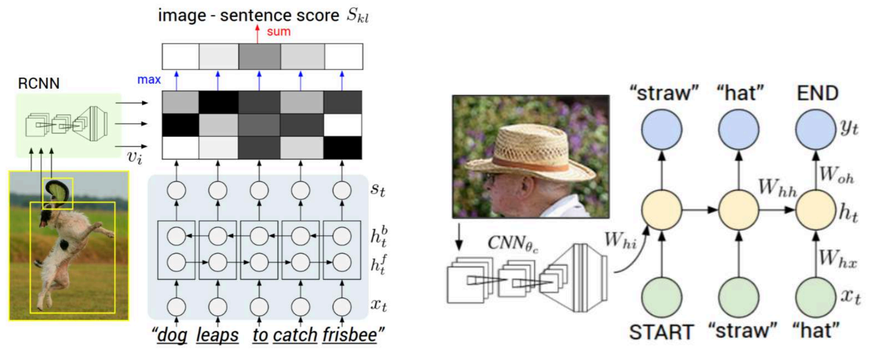

Figure 4: (Left) Evaluating the image sentence score[1] & (Right) Multi-model RNN generative model[1]

Figure 5: Images Augmented with Filters to evaluate resistance to negative effects (L-R) Original Image, Complement, Bright Light, Low Light, Gaussian filter images

- Scale Augmentation is done for high, medium and low resolution
- Computed the semantic similarity of the captions generated from the images with filters and resolutions
- Developed a multi-threaded server using socket i/o which receives an image as input from the client app and processes on the model and returns the captions
- Client app is built on Android platform

RESULTS

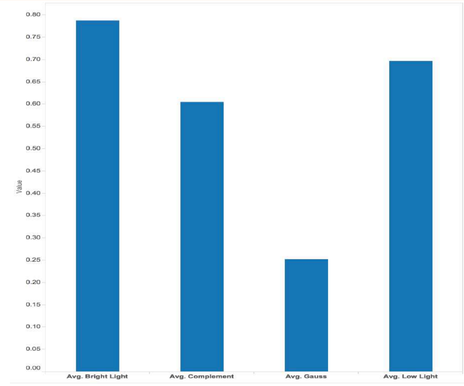

Figure 6: Similarity scores comparison of images with different filters

- Model is trained using Torch and Neural Talk[1]
- ResNet[2] generates better CIDEr score than VGGNet[3]
- Semantic Similarity were computed based on cosine distances
- Total 500,000 images were used to test performance of the model for color and spatial augmentation

CONCLUSION & FUTURE WORK

- Classifier on image meta-data didn't work out well since many images were studio images with fixed image properties
- Bright Light Filter outperforms than the others in terms of semantic similarity
- Incorporating LSTM techniques will produce better results
- Ignore peripheral vision to generate captions based on line of sight

REFERENCES

1. Andrej Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions, 2012

2. Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition CoRR, abs/1512.03385 2015

3. Simonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition CoRR abs/1409.1556 2014