Deep Learning for Image Captioning

By
Siddesh Pillai (srp4698@rit.edu)

Advisor : Professor Jeremy Brown (jxbvcs@rit.edu)
Colloquium Advisor : Professor Leon Reznik (lr@cs.rit.edu)

Introduction (1/4)

- Deep Learning is a rapidly growing segment in AI
- Deep Learning is machine perception; ability of a machine to interpret using complex models which help us classify, cluster and predict

Algorithms that mimics the human brain

Experts: Andrew Ng, Yann LeCunn, Geoffrey Hinton, Yoshua Bengio, Li Fei-Fei, Alan Karpathy etc.

Image is taken from google

Introduction (2/4)

- Some context aware applications such as
  - Voice Recognition – Uses a deep Long Short-Term Memory (LSTM) recurrent neural network for acoustic modelling of speech recognition http://googleresearch.blogspot.ch/2015/09/google-voice-search-faster-and-more.html

Image is taken from google

Introduction (3/4)

- Image Search – http://googleresearch.blogspot.com/2013/06/improving-photo-search-step-across.html

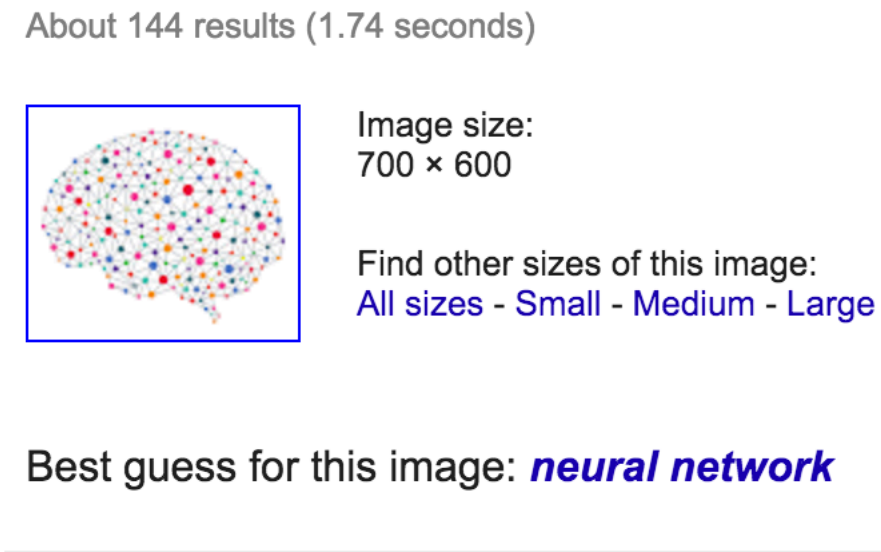


Image is taken from google

Introduction (4/4)

- Some of the areas explored using deep learning are scene detection, emotion detection, smart assistant, facial recognition, self driving vehicles, etc.

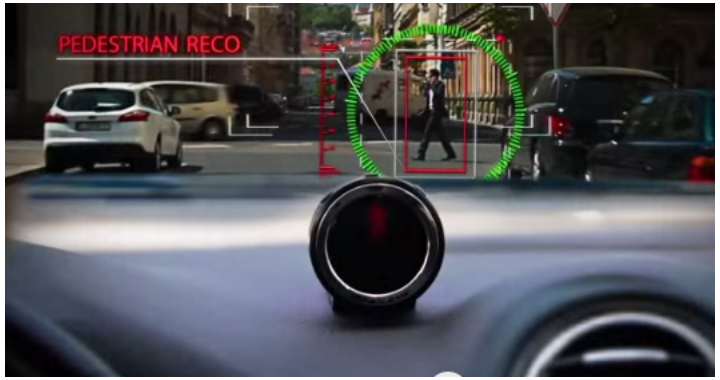
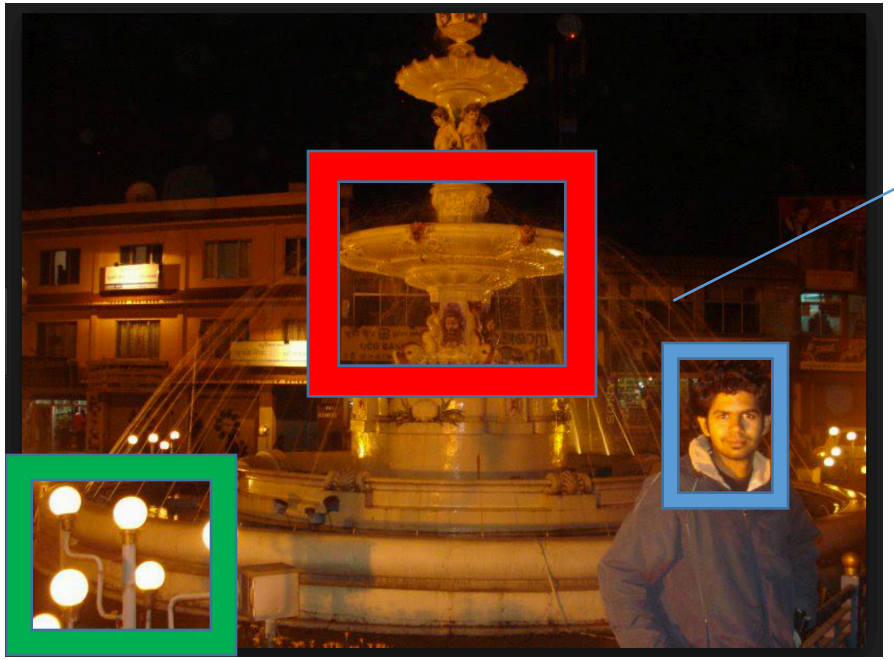


Image source: https://www.google.com/selfdrivingcar/ , http://www.computervisionblog.com/2015/03/mobileyes-quest-to-put-deep-learning.html

The Problem

- The aim is Image Captioning using deep learning techniques
- The general idea is you capture an image and the application will perform the object recognition and provide a semantic description of the scene



lamp, person, fountain (object recognition)

A person is near to lamp and fountain (semantic description)

How is it different to Google Image Search?

- Google Image search – Analyzes the images' distinct features such as lines, points, textures and creates a mathematical model which searches against the billions on images and retrieve similar images, show the best guess and webpages associated with the best guess.
- This experiment is through supervised learning, CNN for object recognition and I will be fine tuning some features at the fully connected layers of the CNN which then fed LSTM recurrent network.

Related Work

- Dense Image Annotations
  - Holistic scene [1] understanding in which the scene type, object and their spatial support of an image is inferred
- Generating Descriptions
  - Developed a log-bilinear model [2] which generates the full description of the images however their model uses a fixed window context
- Neural Network in visual and language domains
  - Convolutional Neural Network [3] [4] has emerged as the most efficient for image classification and recognition

Milestones

- Milestone 1
  - Datasets – Confirm the image data set(Flickr) 40k images (35k training + 3k test + 2k validation)
  - Explore the torch library
  - Train (Supervised learning) the dataset on my machine using convolutional and recurrent neural network and test for image recognition

Milestones

- Milestone 2
  - Derive the sematic alignment of words from the image recognition process using a multimodal recurrent neural network
  - Score the accuracy of the natural language descriptions of the images
- Milestone 3
  - Develop an android app which will provide an interface of this experiment
  - Code freeze
  - Measure the results and efficiency
  - Document the project
  - Design the poster

Expected Results

- Successfully train the images
- Measure the accuracy of Object Recognition
- Score the correctness of the captions generated

References

1. L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 2036–2043. IEEE, 2009
2. R. Kiros, R. S. Zemel, and R. Salakhutdinov. Multimodal neural language models. ICML, 2014.
3. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradientbased learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.
4. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.

Thank you