**Motivation**

- **Vision** is the most important sensory stimulus. 3.4 million people in the US and 285 million worldwide are deprived of this gift.
- Dealing with simple day-to-day tasks becomes an ordeal for these individuals and they are also plagued with safety concerns.
- Powered by Deep Learning, our system takes in an image of a scene and generates a rich, semantic description in the form of speech, to give the visually impaired a sense of their surroundings.

**Data**

- Used the **MS COCO 2014** dataset. It contains images of objects from 80 classes.
- Pascal has only 20 categories while COCO spans over 80. This helps the system generalize better.
- ImageNet is too big a dataset for our application.

**COCO** also is a standard dataset for object detection, segmentation and captioning of images. COCO has bounding boxes along with the image class.

**Model**

- Input:
  - Images with bounding boxes around objects + 5 Captions
- Output:
  - Best sentence/speech describing the input image

**Architectures**

- **VGG**: Simple - uses only 3x3 convolutions but number of parameters is extremely high. (Owing to the FC layers).
- **ResNet**: Add rectangle to tackle vanishing gradient problem and can train very deep NN. The number of parameters are significantly less (~12.4 million).
- **CNN**: 
- **RNN**: 
- **TTS**:

**Results**

- **Show and tell**: A generative model based on deep recurrent architecture that generates natural sentences.
- **Top-Down**: Uses faster R-CNN for bottom-up attention and uses task specific context for the top down mechanism to predict an attention distribution on image regions.
- **Attention2**: Is a self critical sequence training (reinforcement) which uses its own test time inference algorithm to normalize the rewards it experiences.

**Discussion**

- The model seems to reflect the bias in the training dataset. For example, whenever the network sees a woman, it correlates it with a ‘woman holding a phone’, and an umbrella corresponds to ‘rain’, buildings are most often predicted as ‘clock towers’.
- Started with Neuraltalk2 GitHub repository which was in Lua, ran on Caffe. Migrated to a PyTorch implementation as this is more widely used.
- Fixed a lot of bugs in the ImageCaptioning.pytorch GitHub repository and switched to a CNN fine-tuneable version.
- Gradient clipping, optimization algorithm, learning rate (decay) and many such hyper-parameters were varied, but the repository already had carefully tested optimal values.

**Future Scope**

- This model can be ported to a mobile platform as an application for generating auditory descriptions for the visually impaired.
- Building a new dataset by appending vocal description of objects, we can build a potential end-to-end system for this application.

**References**