Team: Aditya Kismatrao, Manas Mishra, Pratik Pandey, Yash Srivastava

Project Proposal: Image Captioning

Problem: We, as a race, are in an era where social media has completely enveloped our day-to-day lives. Major players in this sector deal with digital data in the size of billions. To connect with the world and tell their stories, people upload pictures with captions. They come with catchy phrases to draw people to their profiles. But, what if we could automate this captioning process; what if we could take an image, study it using an algorithm and generate relevant headings and captions? This would not only enhance the way we use social media, but it would transform the way we look at images in general. Our project in deep learning aims at helping with this and serving as a template for various sectors of society to benefit from its applications.

Motivation: Image captioning can not only enhance the way social media is used, it can also help in improving lives of communities, especially the visually impaired. Developing applications for them would help them listen to images. In addition to that, image captioning can also play a vital role in the healthcare industry where it can support medical image reading and finding anomalies in them to help physicians better. All-in-all, image captioning could revolutionize the way we look at images.

Challenges: One of the challenges is image semantics richness where the images used for testing must be semantically related to those used for training the model. Inconsistent objects during training and testing is another hurdle in which the existing training process relies heavily on the selection of data sets. Cross-language text description of images is also an issue and how to implement cross language text description of images is a key problem and a research difficulty in image captioning.

Techniques used: We are planning on using the Flickr 8K dataset for our project. Considering that we have two diverse sets of features, images and captions (text), we would be using a combination of architectures that can encompass the features present in the data. The techniques used are:

1. Convolutional Neural Network (CNN): Using pre trained-ImageNet CNN model to obtain image feature vectors.
2. Word Vectors: Build index dictionary to encode word to integers using GloVe word2vec model.
3. LSTM: Since we are approaching it as a supervised learning problem, we would be using partials captions with image vectors to train the LSTM model to get target captions.

Expected Outcomes: The basic objective of image captioning is to describe an image in text format and hence that would be the major outcome of the project. The resultant captions would cover major image properties like objects, activities, counts and colours.

Theoretical and practical implications: Image Captioning has been widely used in different forms ranging from image-indexing for visually-impaired individuals to group images on platforms like Google and Facebook.The idea can be extended to video-captioning, audio-video annotation, and the popular Visual-Question Answering (VQA) challenge which becomes more user-interactive as we ask questions regarding the input image and receive answers in various formats.