**INTRODUCTION**

In this www world, every day in our life, all have experienced with the huge number of images in a real world which are self-interpret by the individual human being by using their wisdom. Human are naturally programmed to convert the natural scene in to text but it is the complex task for the machine as they are not much efficient like human. Still, human generated captions are considered better as machine need human intervention and programmed accordingly for the better result. Due to the recent development in deep learningbased techniques, computers are capable to handle the challenges of image captioning like detection of object, attribute and their relationship, image feature extraction and generating syntactic and semantic image caption [1]. With the advancement of AI, so many new ideas have revolutionized in the areas of image processing and it has transformed the world in a surprising way. The image captioning Approach (Fig. 1) has wider application in the real world as it provides the better platform for human computer interaction. Due to the emerging application in image processing, image captioning becomes the topic of interest for the academician and researchers. By seeing the Fig. 2, picture someone guess that two dogs are playing with toy and someone might say two dogs hauling in floating toy from the ocean or two dogs run through the water with rope in their mouths, so all of these captions are appropriate to describe this picture. Our brain is so much trained and advanced that it can describe a picture almost accurate but same was not the case with machines. Hence, the main aim of the image captioning is first identified the different objects and their relationship present in the image using deep learning-based technique, generating the textual description using the natural language processing and evaluate the performance of the natural language-based description using different performance matrices. Object detection and segmentation are the part of the computer vision and done with the help of popular CNN and DNN and generating image description (Fig. 3) are the part of natural language processing which is done by RNN and LSTM. CNN works for understanding the objects of the image or scene and provide the answers the various questions about the objects in image like what, where, how, etc.