# Project proposal for Image Captioning

# Domain Background:

Image captioning is the process by which names or description is automatically assigns to digital image.

Objective is to capture objects contained in an image, and also express how these objects relate to each other as well as their attributes and the activities they are involved in

This kind of system could eventually help

* visually impaired people understand pictures,
* Provide alternate text for images in parts of the world where mobile connections are slow, and make it easier for everyone to search on Google for images.
* Semantic image search
* Bering visual intelligence to chat board
* Help visually impaired people to see world around them

The web is filled with billions of images, helping to entertain and inform the world on a countless variety of subjects. However, much of that visual information is not accessible to those with visual impairments, or with slow internet speeds that prohibit the loading of images.

Problem Statement:

Image captioning involves the task of generating natural language descriptions of visual content with the use of datasets comprising of image-caption pairs

Image captioning is to understand objects in images, relate to one another and translating it all into natural-sounding language.

Accurately describing a complex scene requires a deeper representation of what’s going on in the scene, capturing how the various objects relate to one another and translating it all into natural-sounding language.

Accurate image captioning is a challenging task that requires advancing the state of the art of both computer vision and natural language processing.

# Dataset and input

1. Conceptual Captions Dataset
   1. It is a new dataset consisting of ~3.3M images annotated with captions.
   2. Conceptual Caption images and their raw descriptions are harvested from the web, and therefore represent a wider variety of styles.
   3. More precisely, the raw descriptions are harvested from the Alt-text HTML attribute associated with web images.
   4. To arrive at the current version of the captions, we have developed an automatic pipeline that extracts, filters, and transforms candidate image/caption pairs, with the goal of achieving a balance of cleanliness, informativeness, fluency, and learnability of the resulting captions.
2. MS-COCO – COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features: 330K images (>200K labeled)Object segmentation
   1. 330K images (>200K labeled)
   2. 1.5 million object instances
   3. 80 object categories
   4. 91 stuff categories
   5. 5 captions per image
3. Flickr8K dataset
   1. comprised of more than 8,000 photos and up to 5 captions for each photo.
   2. build models on your workstation using a CPU.
4. Flickr30k
5. SBU
6. Adobe\_MIT FiveK, contains random images from Intagram

# Solution Statements:

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing.

There are 2 leading approaches:

One stream takes end to end , encoder decoder framework adopted from machine translation.

Use CNN to extract high level images features and fed them into LSTM for caption generation. Thi went one step further by introducing attention mechanism.

Second stream applies compositional framework, it divides captions generation into several parts: word detector by a CNN, caption candidates generation by a maximum entropy model and sentence re-ranking by deep multimodel semantic model.

The main component include

1. Vision model using deep residual network – deep residual network(ResNets)
2. Language and semantic ranking model
   1. Use of LSTM / GRU for caption generation
   2. use MELM + DMSM, MELM(maximum entropy language model) together with a deep multimodel similarity model(DMSM)
   3. use MELM + GRNN
3. An entities recognition model that identifies celebrities and landmarks
4. A classifier for estimating the confidence score for each output caption

# Evaluation Matrix

Metrics such as BLUE, METEOR and CIDEr are handy for fast development and tuning

Reference:

<https://ai.google.com/research/ConceptualCaptions>