Background-related Visual Question Answering

Team member: Zihang Jiang, Runyu Wang, Jiachen Xu
Instructor: Prof. Osama Alshaykh

Introduction

Visual Question Answering is a research area about building a computer system to answer questions presented in an image and a natural language. For any VQA problem, it has two main focus: image and question.

Our product is to answer the **background** related question about image through training our own model and dataset. It can recognize the specific background object. Not only just figure out what type of background it is. For example, it can recognize the background is the Eiffel Tower instead of just recognized it is a building or tower.

Q: Where is the image taken?
Colosseum

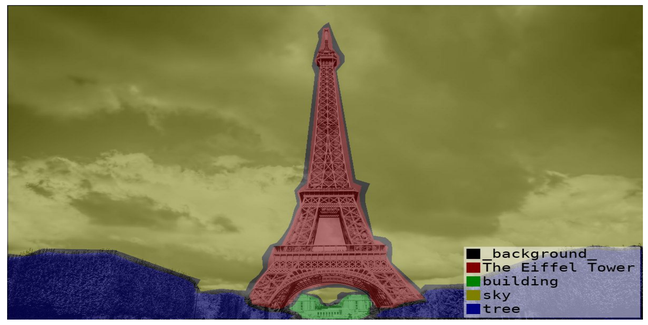
Q: What's in the background?
The Great Wall, Vegetation, Sky

Q: Which place is the picture in?
The Eiffel Tower

Q: Where is the camel?
Pyramid

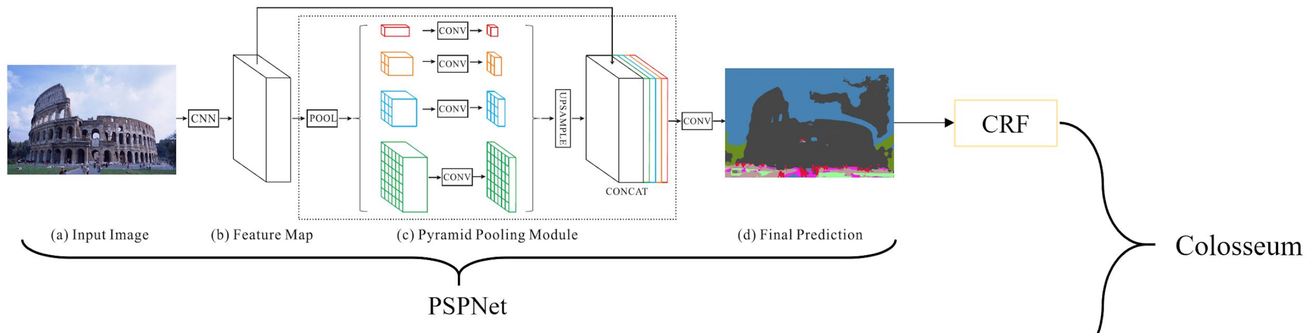
Dataset

We collected over one thousand images from internet about famous and landmark buildings then used **Labelme** to label all of them.
Combined our dataset with Pascal VOC finally including **32** classes.



Model Architecture

PSPNet: Pyramid Scene Parsing Network, using ResNet-50 as backbone
CRF: Conditional Random Field, validating the edges of objects and making the areas smoothly
LSTM: Long Short-Term Memory, extracting question information



Experiment

Trained our model with 160 epoches and three V100 GPUs on SCC in 7 hours.

The the accuracy of our model reaches amazingly **92%**. We don't need such a high mIoU to precisely describe the shape of object. Our aim is to know the background type, which is perfectly accurate.

Backbone	Val mIoU	Accuracy
PSPNet + CRF	53%	92%