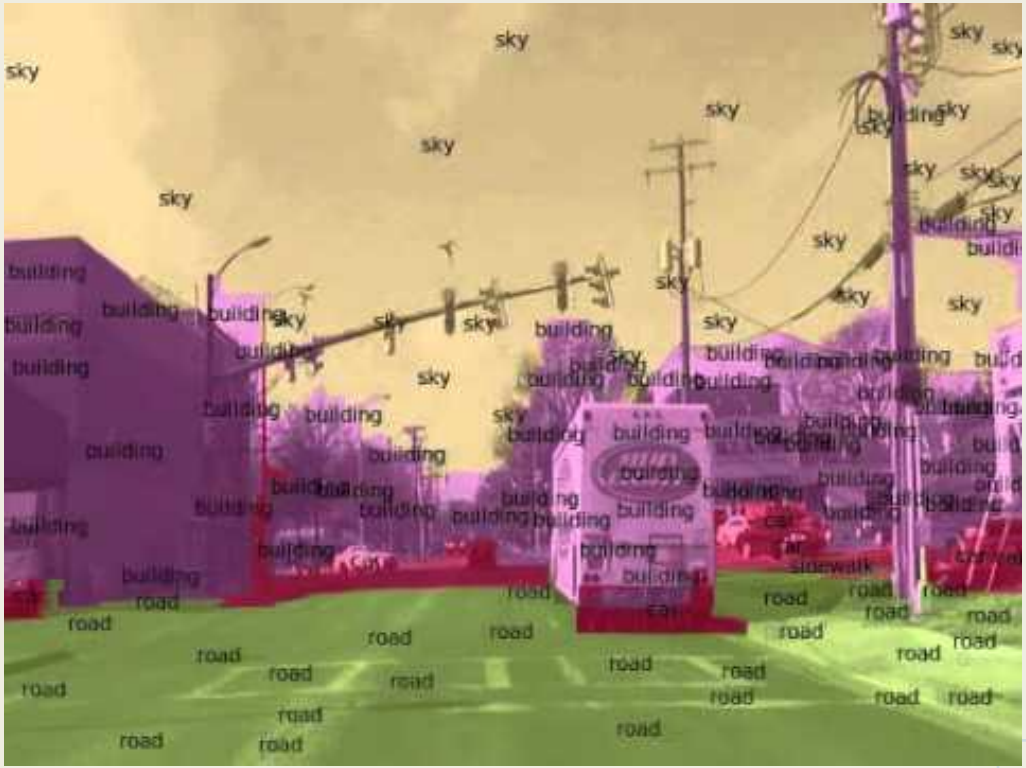
Object Detection Using Given Key Words

James Guevara and Ankit Gupta

Overview

1. Problem Statement
2. Conventional Approaches
3. Image Segmentation Using Deep Learning
4. Implementation Details
5. Multiscale Convolutional Neural Network
6. Results
7. Future Work

Real-Time Scene Parsing



Problem Statement

Original motivation:

- Necessary step for performing more complex tasks (i.e. robots grasping objects, etc.).
- Applications such as real-time scene parsing.

Given an image with a label for each pixel:

- Train a classifier to predict label for each pixel.
- User inputs a keyword and pixels with the corresponding label are highlighted.
- Using the model proposed by LeCun, *et al.*

Conventional Approaches

- Histogram-based methods
  - Histogram is computed from all the pixels in the image.
  - Peaks and valleys in the histogram are used to locate the clusters in the image.
  - Use color or intensity as measure.

Conventional Approaches

- Clustering methods
  - Use K-means algorithm to partition image into K clusters.
  - Assign each pixel in the image to the cluster that minimizes the distance between the pixel and cluster center.
  - Distance (either L2 or L1 norm) typically based on pixel color, intensity, texture, and location.

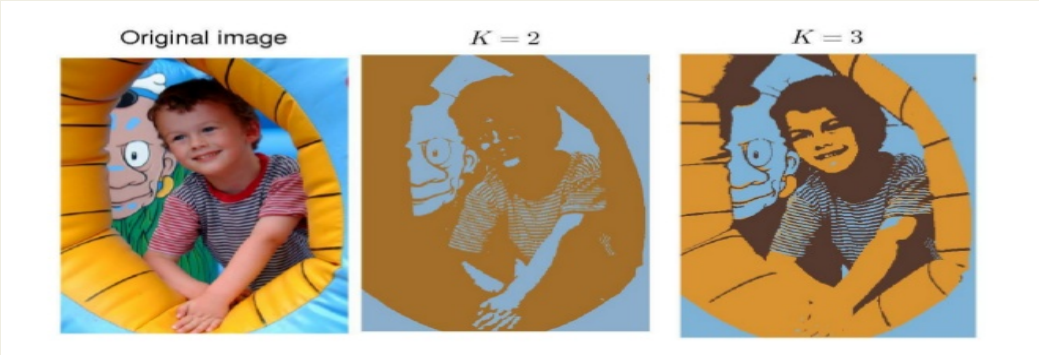
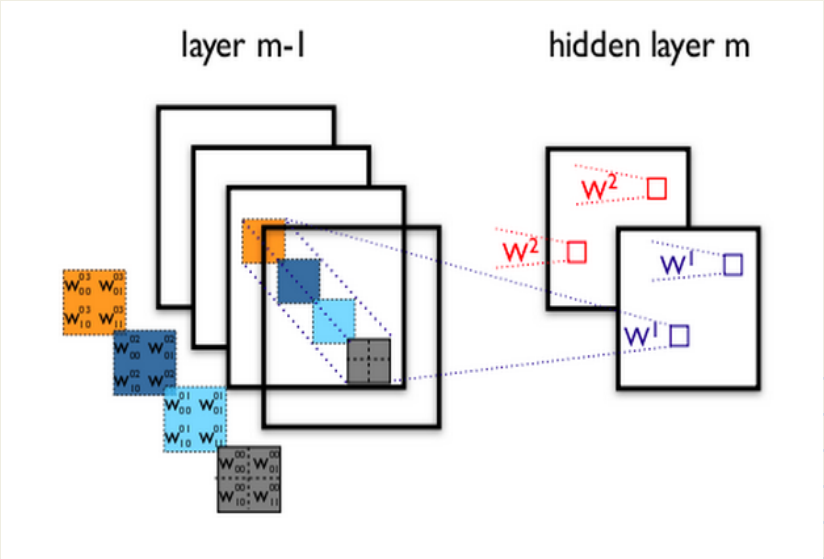


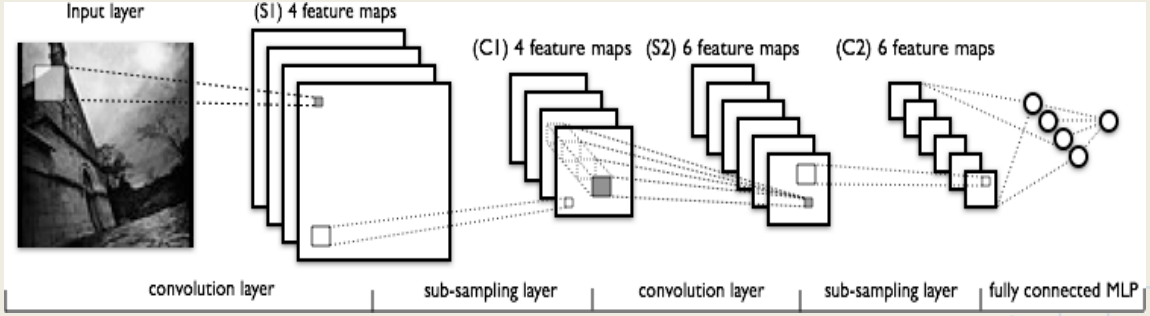
Image Segmentation Using Deep Learning

- Supervised learning model.
- Pixel label classification using CNNs.
- Feeding multiple scales of images into the CNN to obtain scale-invariant features.
- CNN filter layers perform feature extraction.
- Use logistic regression or MLP for classification based on accumulated features.

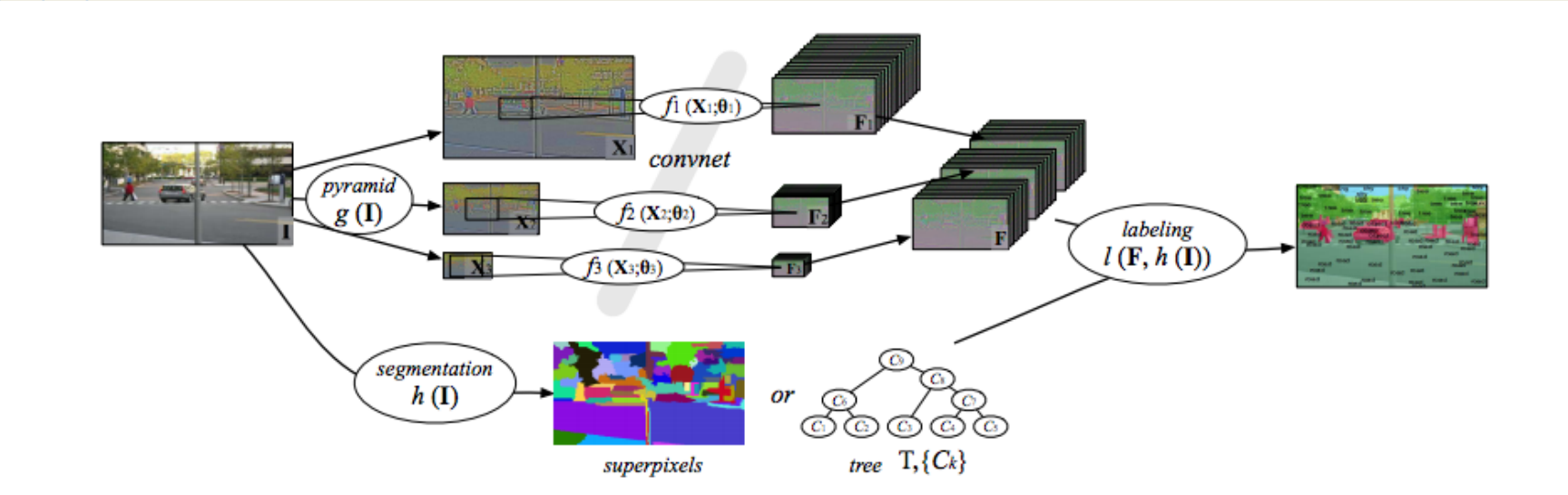


Implementation Details

- Preprocessing: Convert to YUV, zero-mean and unit variance.
- Sequence of CNN filters followed by max-pooling layers are used to compute feature maps for an image.
- Concatenate feature maps.
- Feed feature vector for each pixel into softmax function.
- 9 labels: sky, tree, road, grass, water, building, mountain, foreground objects, unknown.



Multiscale Convolutional Neural Network



- Construct a pyramid of input images by downsampling original image (achieve scale invariance).
- Training over half a million parameters on GPU.

Results

Models	**Validation Error**	**Test Error**
Single-scale model	30.785%	32.026%
Multi-scale model	28.561%	30.527%
LeCun *et al.* 2012 (single-scale)	-	34%
LeCun *et al.* 2012 (multi-scale)	-	22.2%

Results (Original Image)

Results (Single-Scale CNN)

Results (Original Labels)

Results (Original Image)

Results (Single-Scale CNN)

Results (Original Labels)

Results (Original Image)

Results (Single-Scale CNN)

Future Work

- More background datasets.
- Natural language processing.
- Speech recognition.
- Superpixels.
- Upsampling using interpolation.