Weakly Supervised Learning for Road Scene Understanding

Donglu Wang

Supervisor: Dr Jose M Alvarez

Outline

- Introduction
- Methods
  - Convolutional Neural Networks
  - Convolutional Auto-Encoders
  - Principal Component Analysis
  - Image Segmentation
  - Markov Random Fields
- Results
- Discussion
- Future Work

Introduction

- Road detection (monocular videos or images)
- Key to autonomous driving systems
- A challenging computer vision problem (varying weather and illumination conditions, cars and pedestrians)

Introduction

- Existing road detection methods highly involve handcrafted features
- Examples: edges, intersections, SIFT, etc.
- Not general, not easy to compute

- Goal:
  - Learn adequate features automatically
  - Tractable running time, generalizability

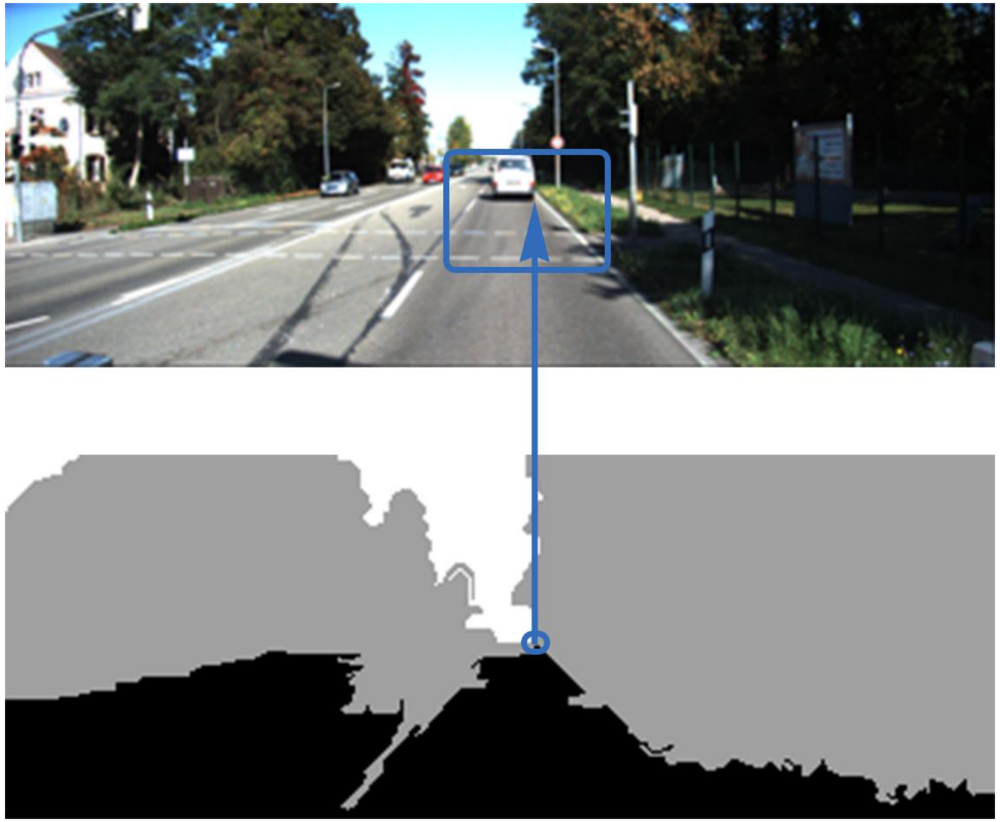
Methods

- Classification, three categories: road, sky, vertical

- One label is required for each pixel

- For each pixel, use its surrounding area as input (e.g., a 32*32 patch)

Methods



Methods

- Based on Convolutional Neural Networks
- Unsupervised feature learning
  - Principal Component Analysis, Convolutional Auto-encoders
- Unsupervised learning methods to improve performance
  - Image segmentation, Markov Random Fields

Data

- KITTI “Road” category
  http://www.cvlibs.net/datasets/kitti/raw_data.php?type=road

- Generate (noisy) labels by using 3D reconstruction tools
- Label each part of image into one of three categories: ground (road), vertical, sky
- Automatic Photo Pop-up (Hoiem et al.)
  http://web.engr.illinois.edu/~dhoiem/projects/popup/

- Manually labelled 60 images
- Label noise ratio: 0.1274

Data

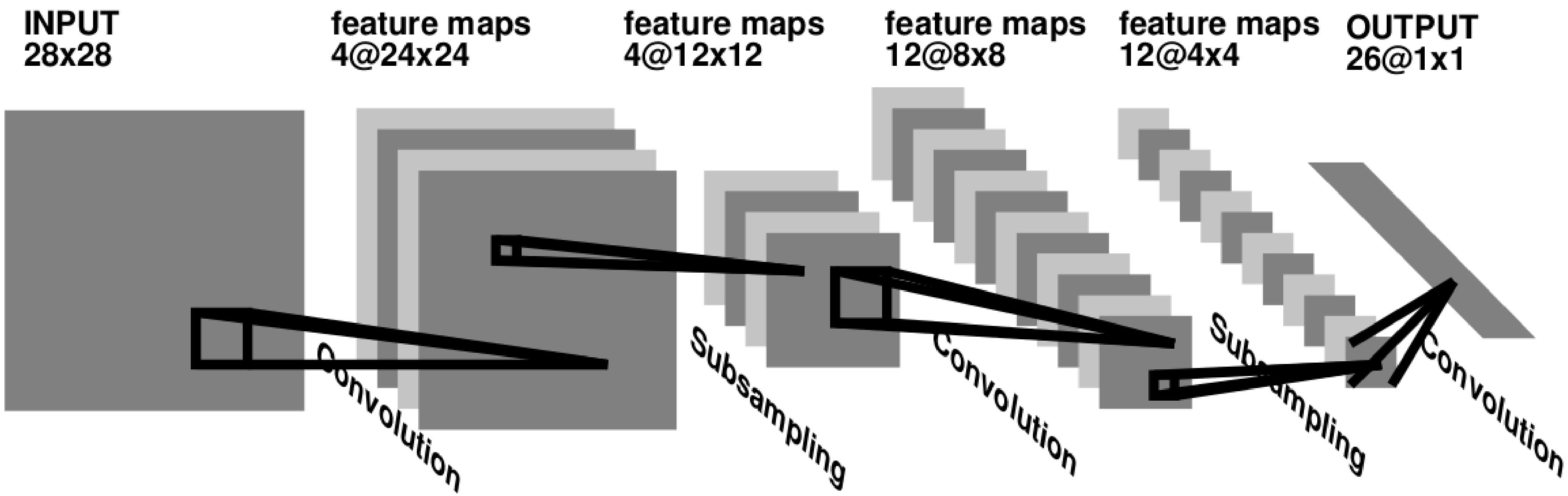
Original Image

Generated Labels (noisy)

Manual Labels

Convolutional Neural Networks

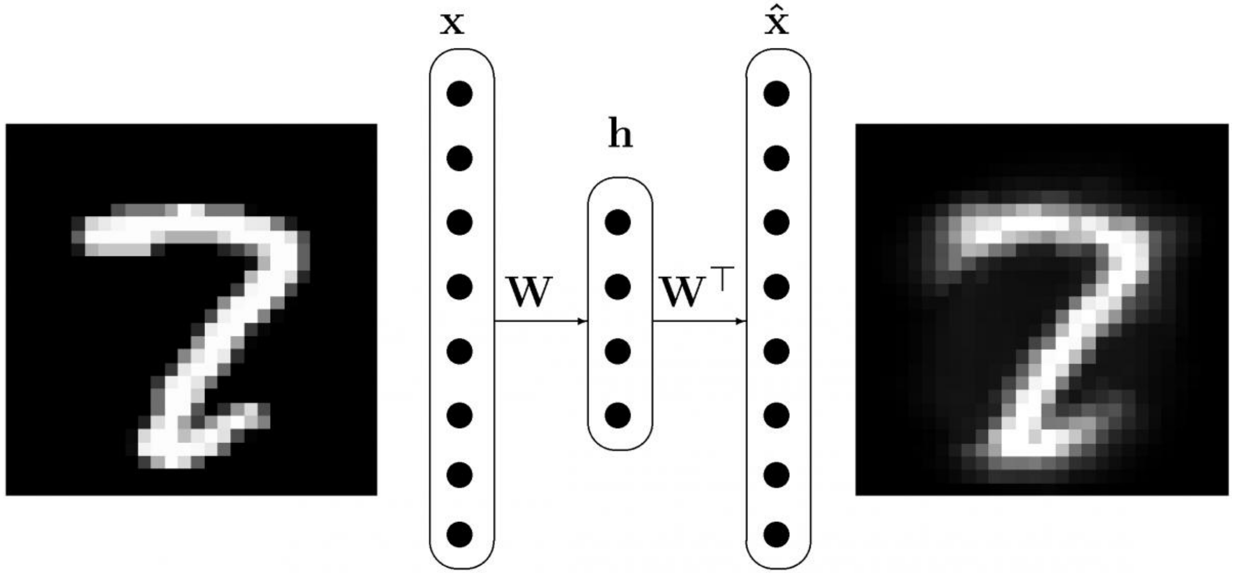
- Discrete convolution of receptive fields and kernels
- Learn local, translation invariant features
- Deep architecture, higher-order features



(Lecun et al. 1995)

Convolutional Auto-Encoders

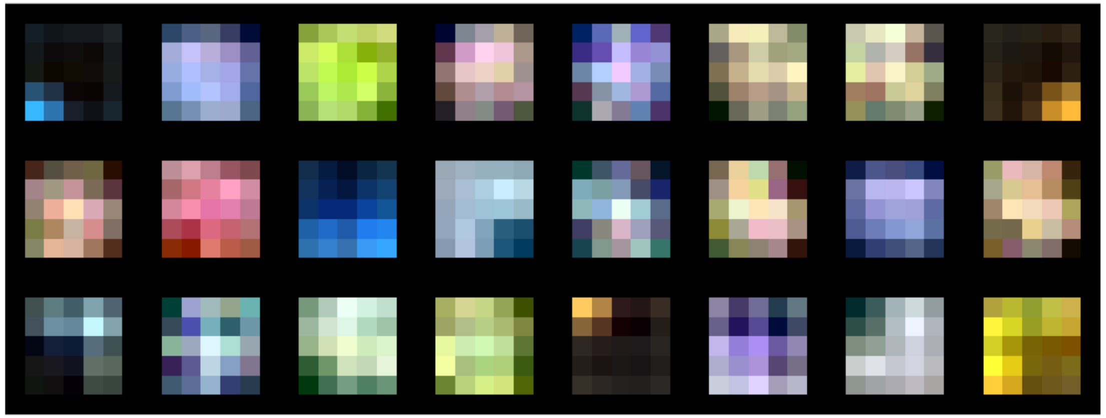
- Limited amount of labels, unsupervised feature learning
- Similar architecture as convolutional neural networks
- Use input data as target output
- Reconstruct input data from hidden representations



(Lemme et al. 2010)

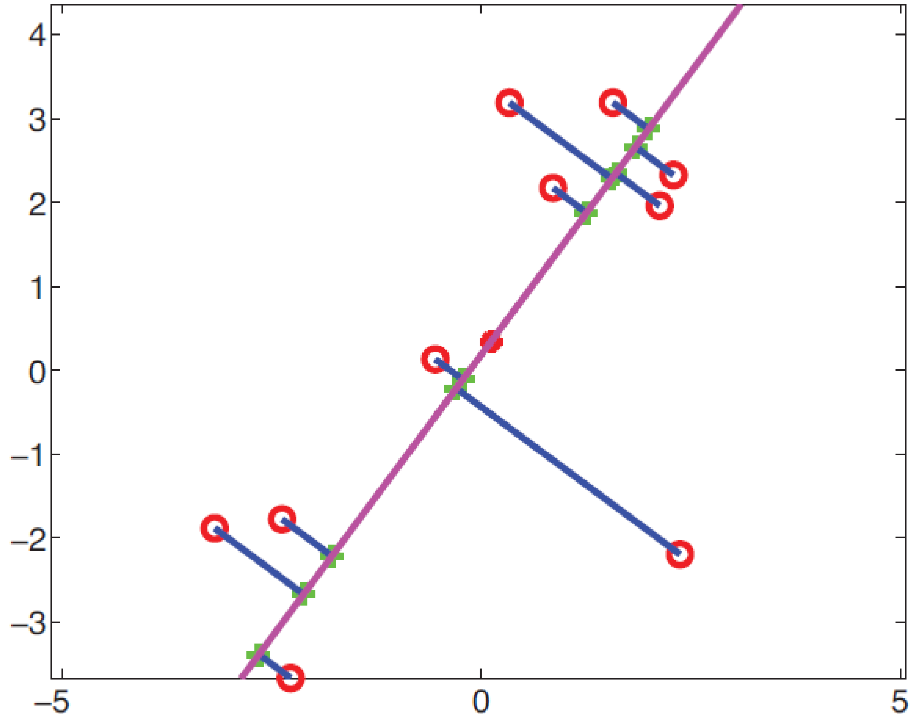
Convolutional Auto-Encoders

- By visualizing kernels and comparing reconstruction errors, help to adjust the architecture of convolutional neural networks



Principal Component Analysis

- Orthogonal projection of the data onto a subspace, such that the variance of the projected data is maximized



(KP Murphy 2012)

Principal Component Analysis

- Bases of subspace are features
- Less correlated with each other

- ZCA Whitening:
  - Features have the same variance
  - No dimensionality reduction
  - Rotate the data to be as close as possible to the original input data

Image Segmentation

- So far, pixels are predicted independently
- Pixel by pixel, high computational cost
- Segment images into super-pixels, use the centroid of each super-pixel to predict

Image Segmentation

- Benefits:
  - Speed up prediction process substantially
  - Avoid boundary points
  - Neighbouring pixels share the same label

Markov Random Fields

- Image denoising on super-pixel level
- Assume correlation between neighbouring super-pixels

Markov Random Fields

Markov Random Fields

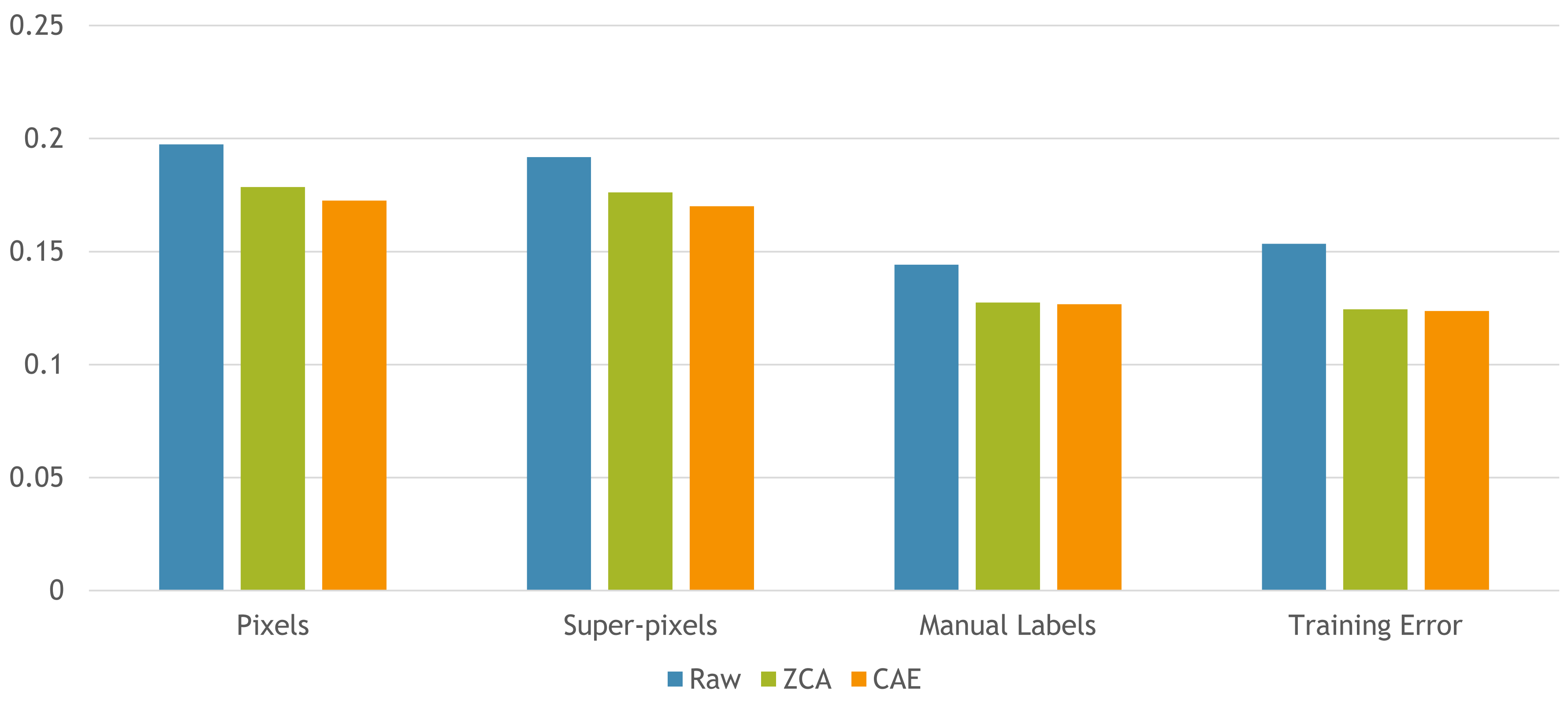
- Drawbacks:
  - Assumption does not hold
  - E.g. pedestrians on the road

Results

	Pixels	Super-pixels	Manual Labels	Training Error
Raw	0.1974	0.1918	0.1442	0.1535
ZCA	0.1786	0.1762	0.1274	0.1245
CAE	0.1726	0.1701	0.1267	0.1237

- Training data: 264 images and generated labels, 5 random patches from each class per image, in total 264*5*3 = 3,960 patches
- Test data one (pixels): 205 images and generated labels, 100 random patches per image, in total 205*100 = 20,500 patches
- Test data two (super-pixels): 205 images and generated labels
- Test data three (super-pixels on manual labels): 20 manually labelled images from visually different scenes

Results



- Best practice: CAE(pre-train) + CNN(train) + Super-pixels(predict)

Results

	Manual Labels
Raw	0.1173
ZCA	0.1320
CAE	0.1049

Discussion

- Acceptable results with a small amount of training data

- Effective approach
- Resistant to noise
  - 0.1274 training noise vs. 0.1267 test error on manual labels
- tractable running time, generalize well

- Limitations:
  - Patch size, lose information of the whole image
  - Correlation between consecutive images is not considered

Future Work

- Increase the quality and quantity of training data
- Capture the correlation between consecutive images in videos
- Apply on other problems

Thank You