MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving*

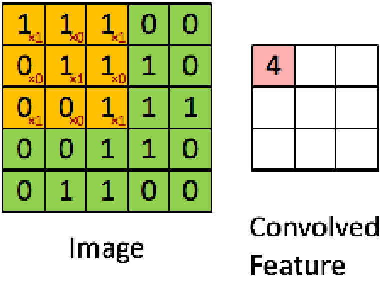
Marvin Teichmann, Michael Weber, Marius Zöllner, Roberto Cipolla and Raquel Urtasun

Presented by: Matt Angus

May 10, 2017

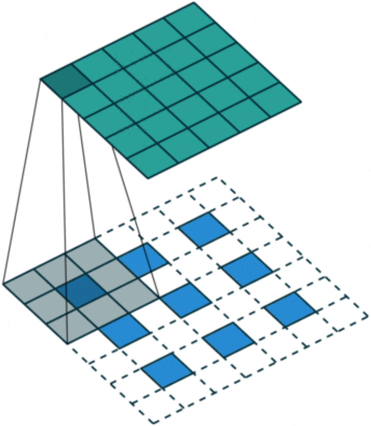
Overview

Convolution



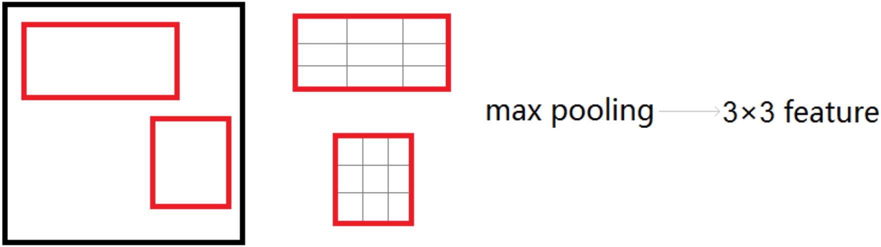
https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Transposed Convolution



https://github.com/vdumoulin/conv_ arithmetic

ROI-Pooling



http://www.bozhiyue.com/anroid/wenzhang/2016/0516/99944.html

New Layers

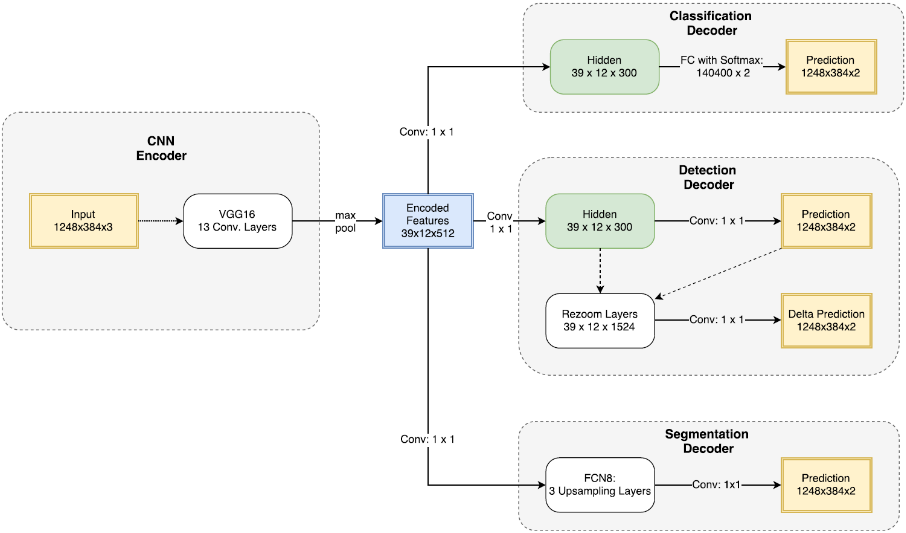
Rezoom

1. Apply ROI-Pooling to higher dimension features (e.g. 156×48)
2. Concatenate with extracted features (e.g. 39×12)
3. Apply 1×1 convolution

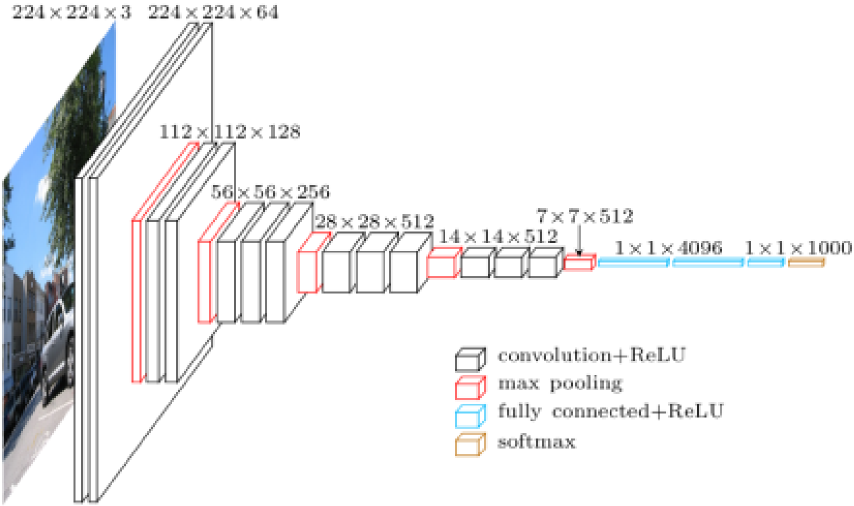
MultiNet Tasks

- Classification
  - Highway/Minor Road
- Detection
  - Bounding box
- Segmentation
  - Free space

Network Architecture

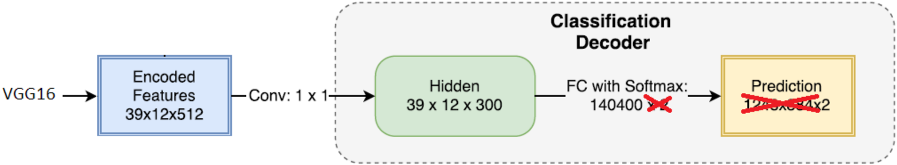


CNN Encoder (VGG16)

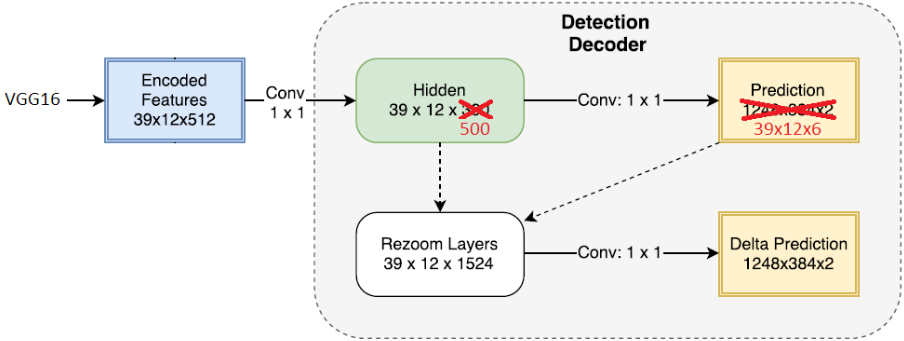


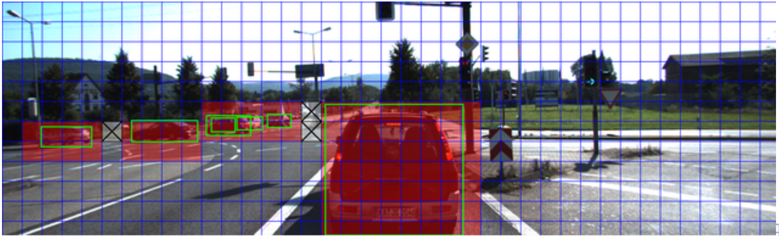
https://www.cs.toronto.edu/ frossard/post/vgg16/

Classification Decoder

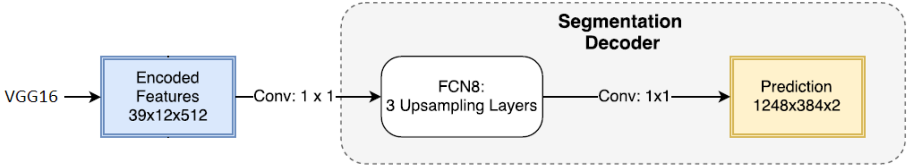


Detection Decoder





Segmentation Decoder



Loss Functions

p: prediction

q: ground truth

I: examples in minibatch

C: classes

δ: 1 if cell has positive confidence, 0 otherwise

Classification/Segmentation

$$\mathrm{loss}_{\mathrm{class}}(p, q) := -\frac{1}{|I|} \sum_{i \in I} \sum_{c \in C} q_i(c) \log(p_i(c))$$

Bounding Box

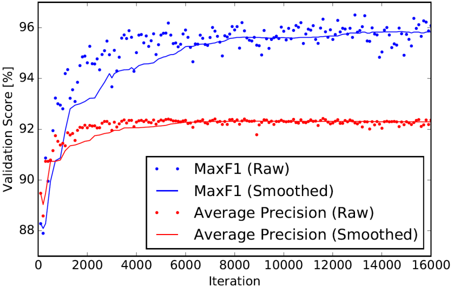
$$\mathrm{loss}_{\mathrm{box}}(p, q) :=$$
$$\frac{1}{|I|} \sum_{i \in I} \delta_{q_i} \cdot \left(|x_{p_i} - x_{q_i}| + |y_{p_i} - y_{q_i}| + |h_{p_i} - h_{q_i}| + |w_{p_i} - w_{q_i}|\right)$$

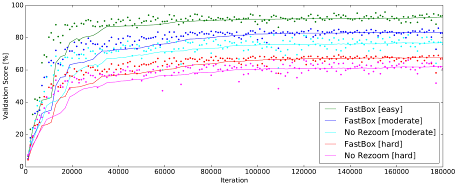
Training Strategy

- merge gradients computed by each loss function, with equal weight
- detection network requires more iterations
    - two updates with just detection, then one update with all
- 0.5 dropout probability on inner 1×1 conv.

Results

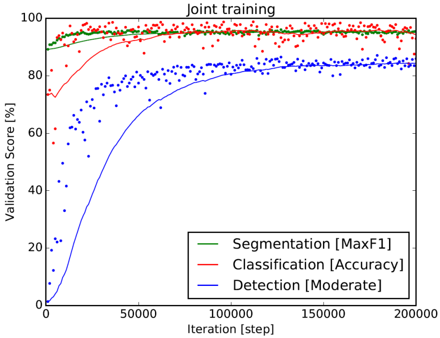
Validation Scores

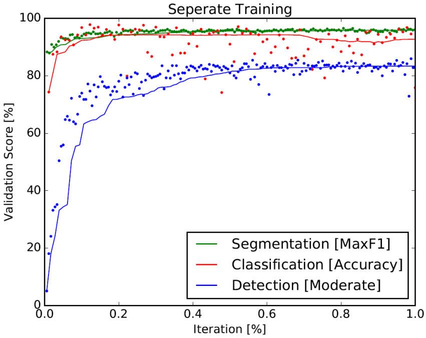




$$\text{MaxF1} = \max_{\tau}\left(2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}\right)$$

Joint Training





Task: Metric	seperate	2 losses	3 losses
Segmentation: MaxF1	**95.83 %**	94.98 %	95.13 %
Detection: Moderate	83.35 %	83.91 %	**84.39 %**
Classification: Accuracy	92.65 %	—	**94.38 %**

Results

Runtime

MultiNet	Segmentation	Detection	Classification
98.10 ms 10.2 Hz	94.6 ms 10.6 Hz	37.5 ms 27.7 Hz	35.94 ms 27.8 Hz

Limitations

- Intentionally overfit/Bad generalization
- Not much information gain from classification
- Only detects cars
- Not so good with dark shadows

Limitations

CamVid

Limitations

Cityscapes

Limitations

Shadows

Summary of Contributions

- Rezoom layer: increase performance, little cost
- Joint training: slight increase performance, slower convergence
- Runtime: 10 fps for all tasks

Questions?