Semantic Segmentation

Compiled by

Dr. Venkat Ramana Peddigari

Samsung R&D Institute India, Bangalore

Dec 14, 2018

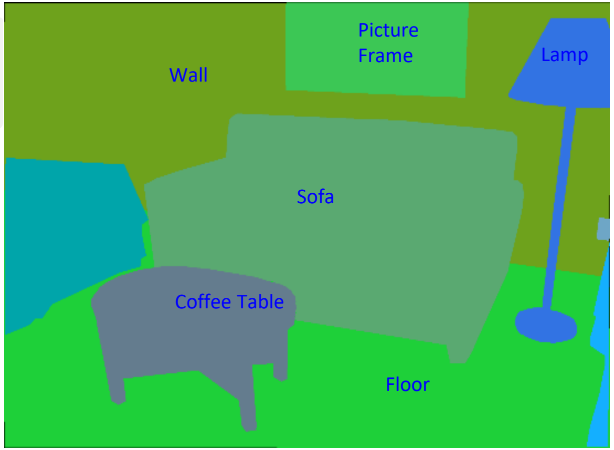
Contents

What ? - Definition

- Semantic Segmentation is a process of understanding an image at pixel level
  - Assigns a label or object class to each pixel in the image
  - Delineates the boundaries of each object class or label
  - Involves dense pixel-wise predictions unlike classification

Input Image

Object Classification



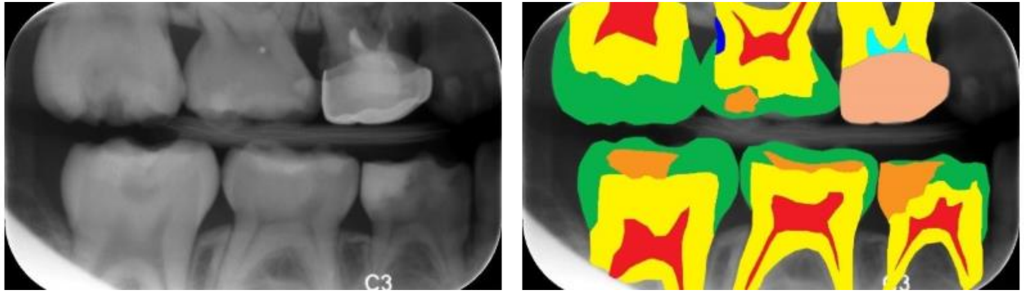
Semantic Segmentation

*Images Credit: ATL Team, Samsung Research Institute, Bangalore

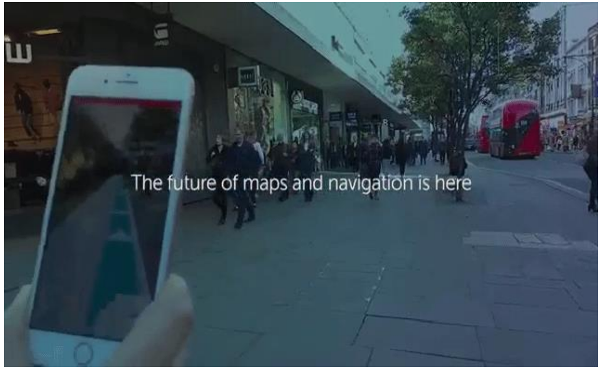
Where ? – Applications

- Semantic Segmentation is quite useful in various domains such as
  - Autonomous Driving
    - ✓ Delineates the exact boundaries of the road and curb
  - AR Navigation
    - ✓ Outlines the walking path in AR world
  - Medical Diagnostics
    - ✓ Automatic Detection of Dental Caries

Autonomous Driving



Medical Diagnostics

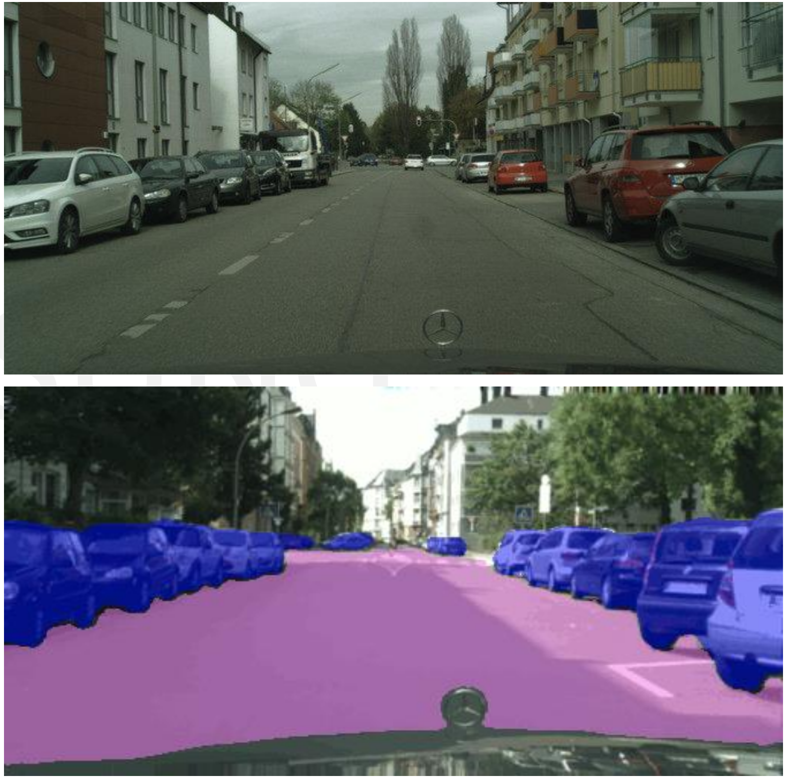


AR Navigation

*Source : https://www.youtube.com/watch?v=ATlcEDSPWXY, https://arxiv.org/abs/1701.08816, Blippar

How ? – Using Deep Learning

- Modeling semantic segmentation problem using deep learning broadly consists of following steps
  - Visual Representation
  - Naïve Architecture
  - Challenges
  - Available Datasets
    - ✓ PASCAL VOC 2012
    - ✓ COCO 2018
    - ✓ BDD100K
    - ✓ CamVid
    - ✓ Cityscapes
    - ✓ Mapillary Vistas
    - ✓ ApolloScape Scene Parsing

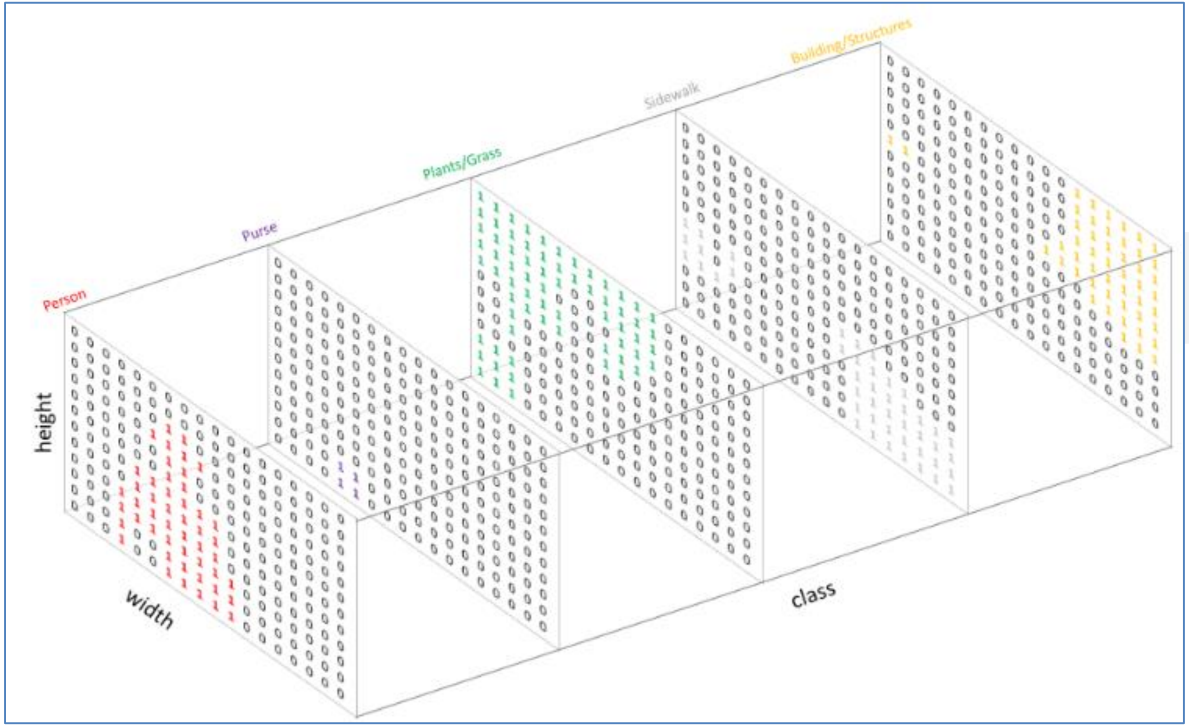


Sample Annotated Cityscapes Dataset

*Images Credit: CityScapes

Visual Representation

- Goal : Output a segmentation map where each pixel contains a class label
  - One-hot encoding for each possible class

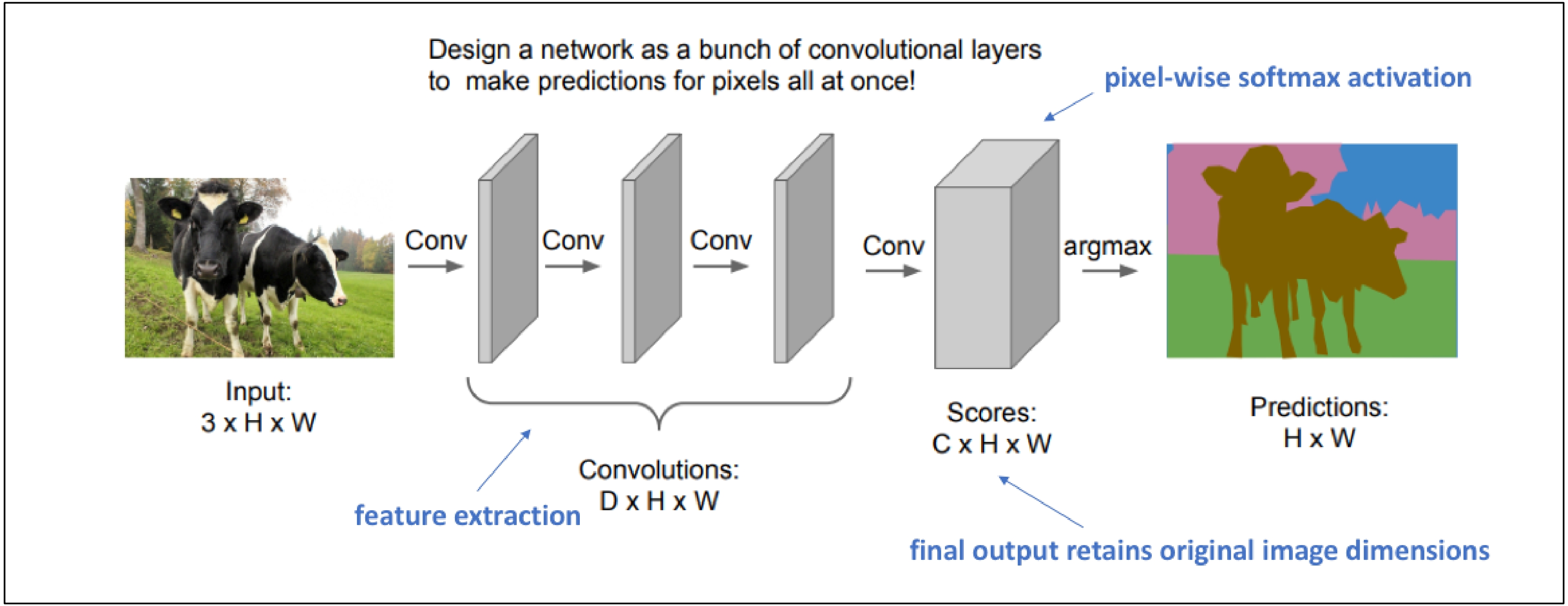


0: Background/Unknown
1: Person
2: Purse
3: Plants/Grass
4: Sidewalk
5: Building/Structures

Source: https://www.jeremyjordan.me/semantic-segmentation/#representing

Naive Architecture

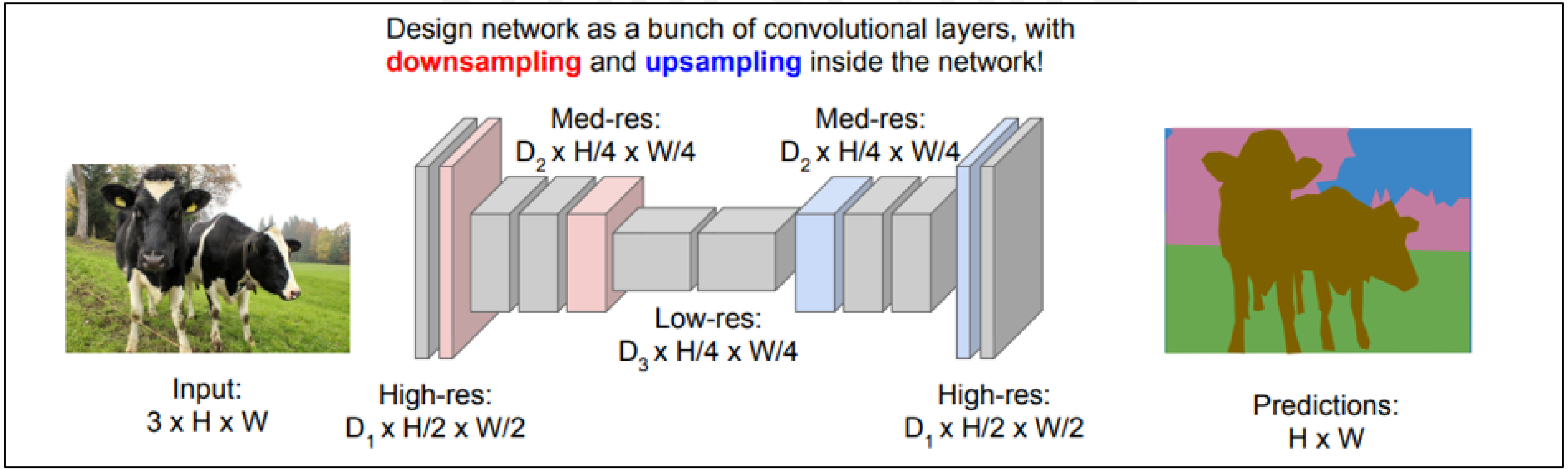
- A stack of convolutional layers with same padding to preserve dimension
- Learns a direct mapping from input to output pixel label through successive transformation of features



Source: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

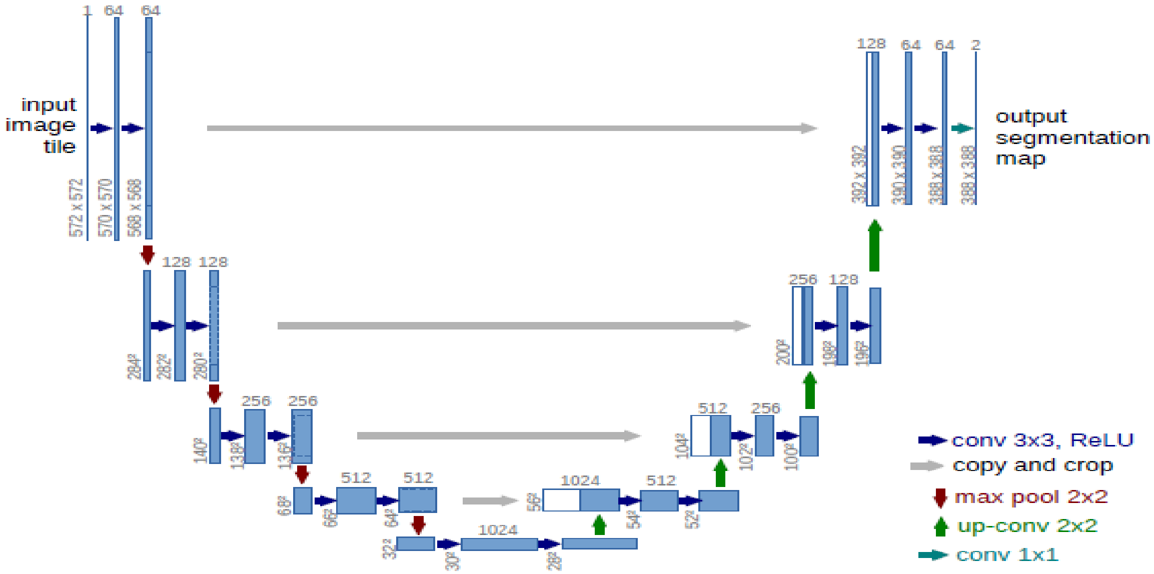
Challenges

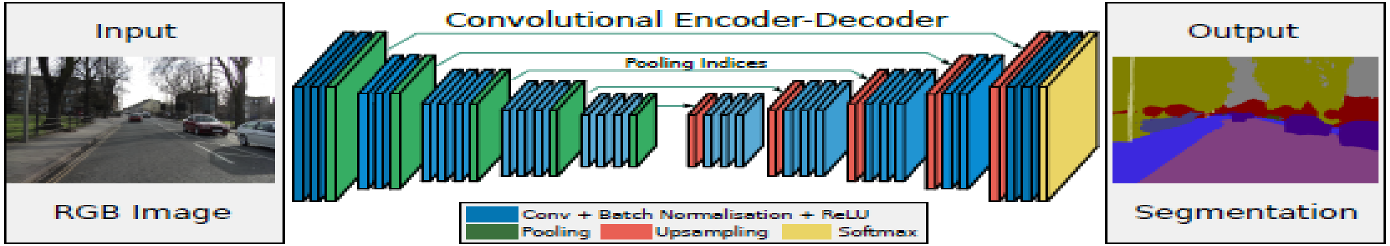
- Computationally very expensive to preserve image dimensions through entire network
- **Solution:** Encoder/Decoder Architecture
  - Low resolution feature mappings : Highly efficient to discriminate between classes
  - Downsample the spatial resolution of input i.e., **Pooling**
  - Upsample the feature representation to full resolution segmentation map i.e., **Unpooling**
  - Skip Connections between encoder and decoder layers



Source: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Encoder Decoder Architecture

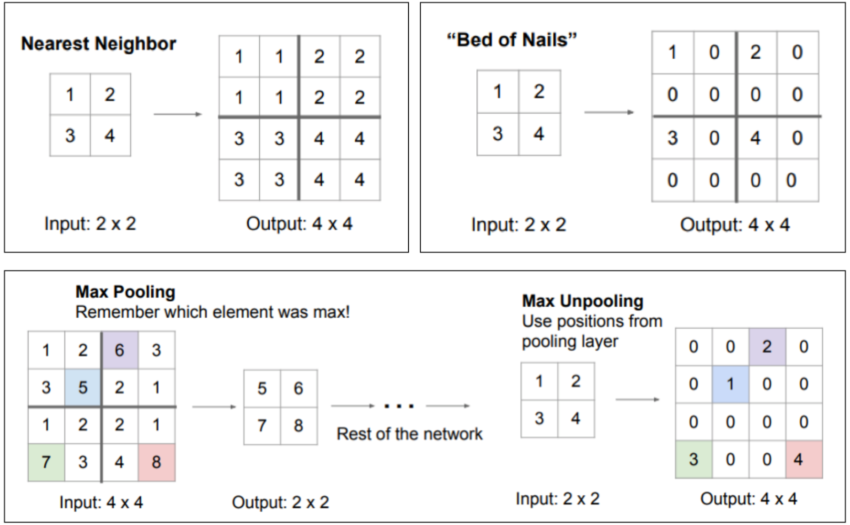


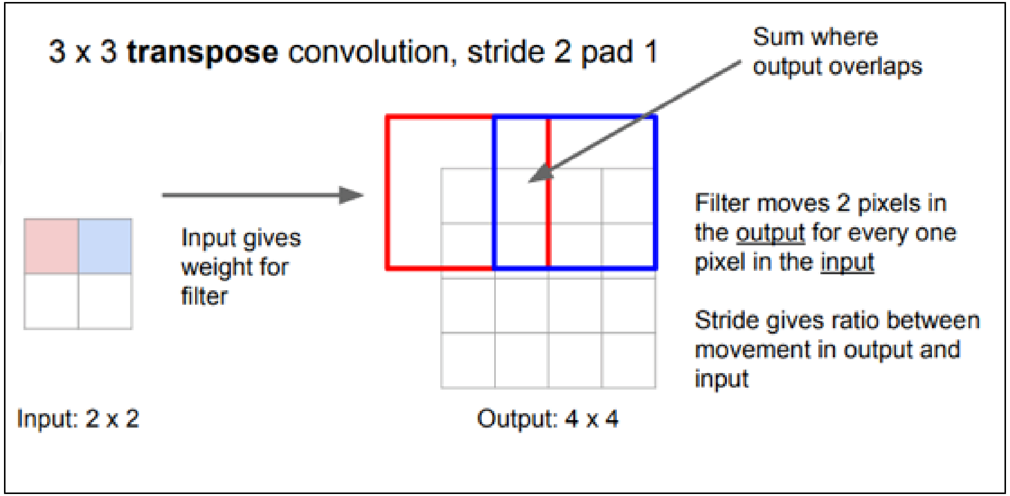


Source: Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

Methods of Upsampling - Unpooling

- Up-sample the resolution by distributing a single value into higher resolution
- Uses the indices from pooling layers

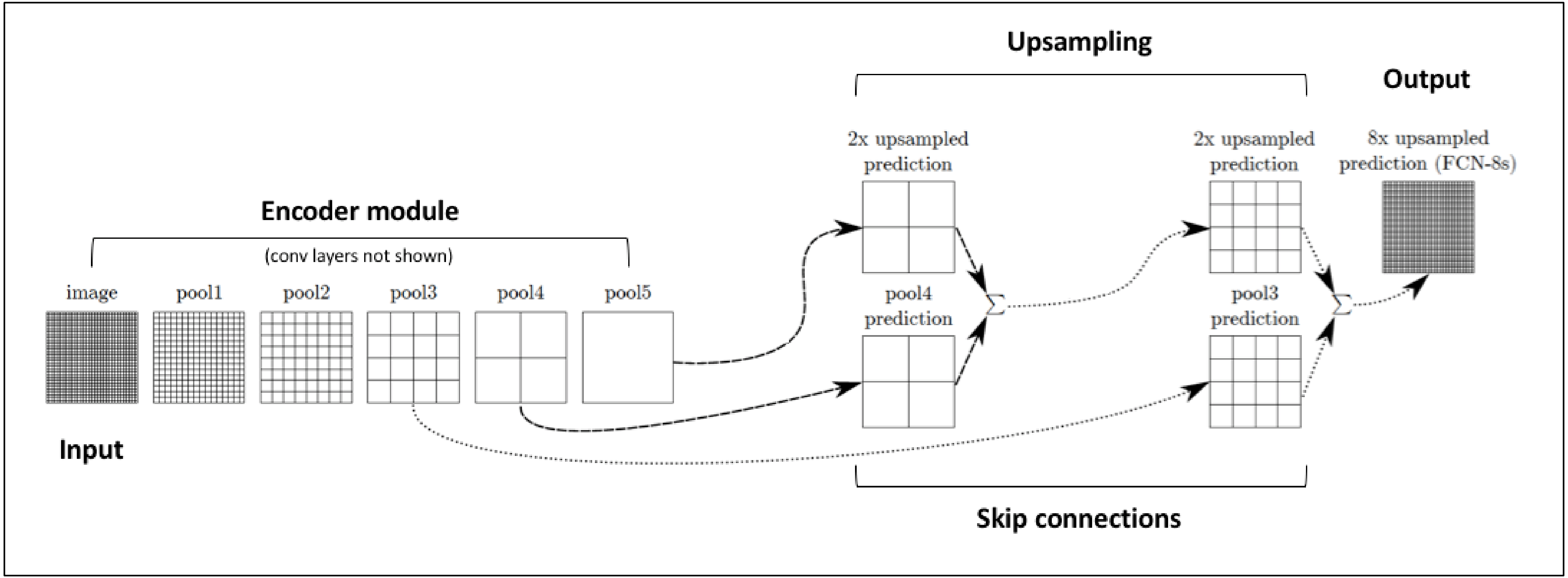




Source: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Adding Skip Connections

- Combines fine layers and coarse layers to ensure that the global structure is retained while making local predictions

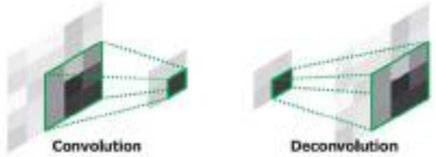


Source: Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Decoder Layers Visualization

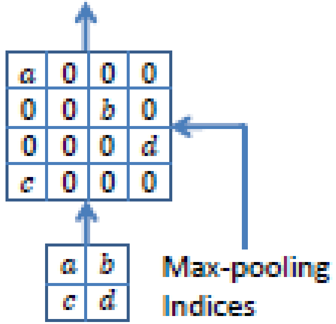
- Unpooling
  - Place activations to pooled location
  - Preserve structure of activations

- Deconvolution
  - Densify sparse activations
  - Bases to reconstruct shape

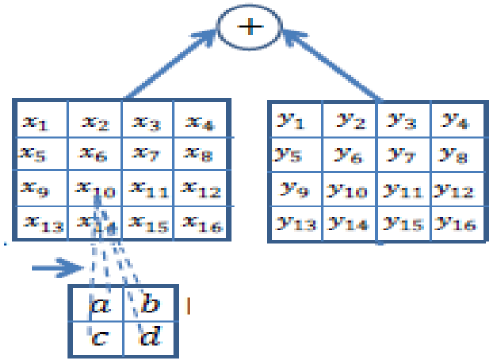


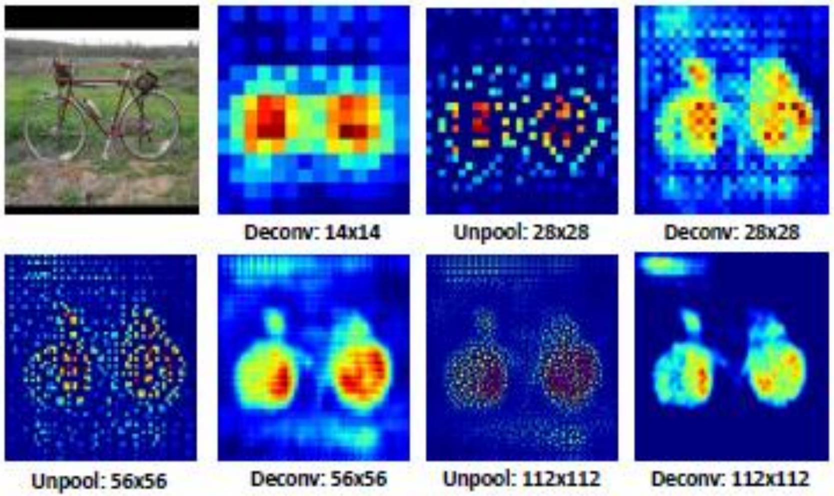
- ReLU
  - Same with convolution network

Convolution with trainable decoder filter



Deconvolution with Up-sampling

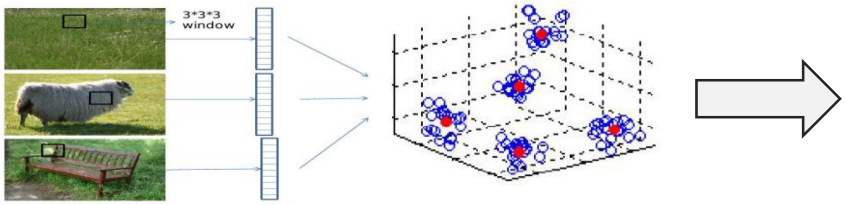




Source: http://cvlab.postech.ac.kr/~bhhan/class/cse703r_2016s/csed703r_lecture7.pdf

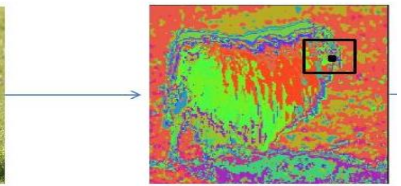
Different Deep Learning Approaches

- Texton Forest and Random forest based classifiers



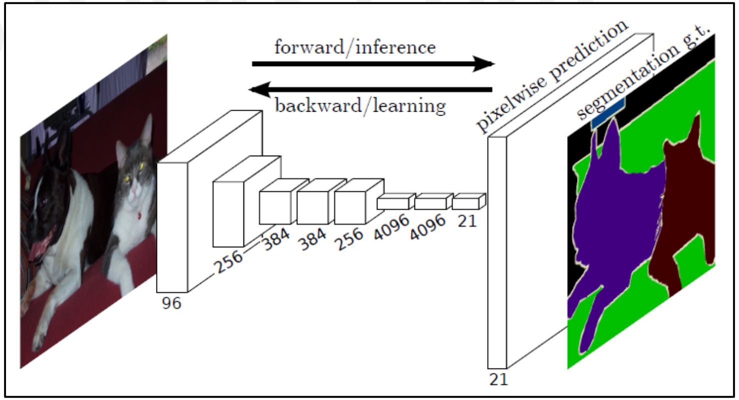
Training Phase

Test Image

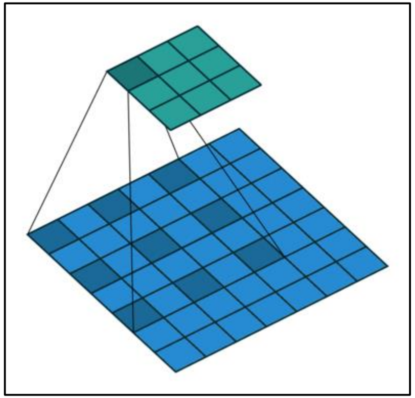


Texton map of the test image

- Patch based classification
- CNN based semantic segmentation
  - Encoder Decoder Architecture
- Available Network architectures
  - FCN
  - SEGNET
  - ENET
  - DeepLab v1 & v2



FCN Architecture

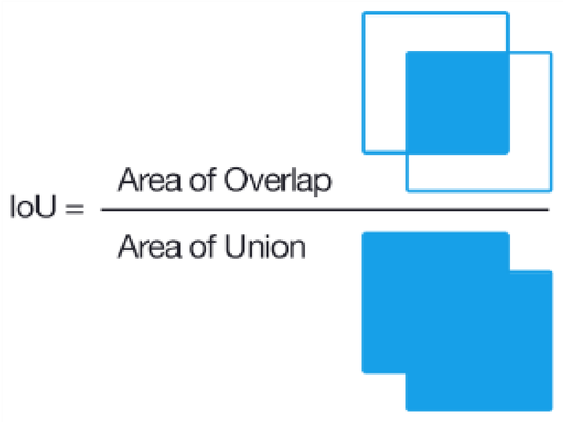


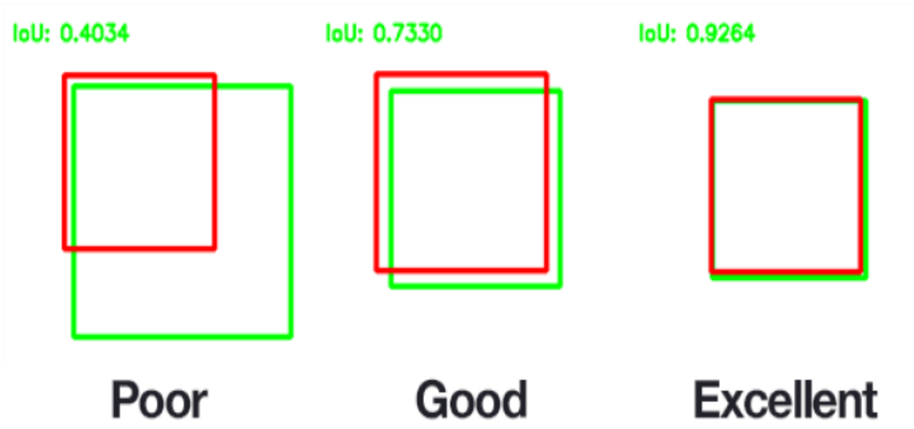
Atrous/Dilated Convolution

Source: Random Forest, FCN, Dilated Convolution

Quantitative Metrics

- Mean of Intersection over Union (mIoU) – Metric used for accuracy evaluation of methods





$$\mathrm{IoU}_i = \frac{\sum_n I_j}{\sum_n U_j}$$

$$\mathrm{mIoU} = \left(\frac{1}{N}\right)\left(\sum IoU_i\right)$$

where

IoU=Intersection over Union

mIoU =mean IoU

n =number of classes

N=number of images

I_j=Intersection of class j for an image

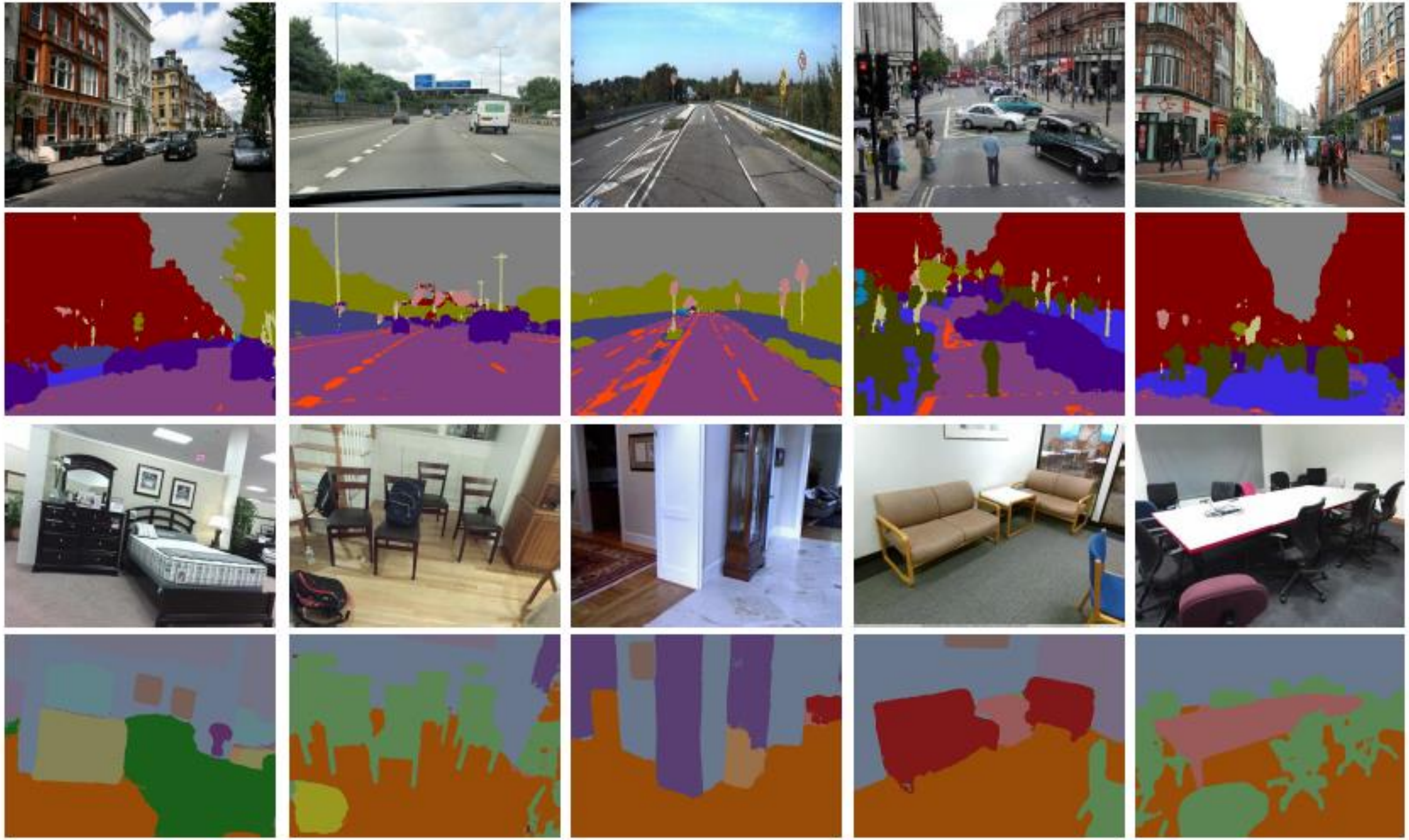
U_j=Union of class j for an image

Comparison Summary

Network architecture	Accuracy	Performance* (on PC)	Intended application
ICNET, Y2017	69.5	33 ms	Semantic Segmentation (High Resolution)
ENET, Y2016	58.3	13 ms	ADAS use case
PSP NET, Y2016	81.2	Very slow	ADAS use case
SEGNET, Y2016	57	60 ms	ADAS use case
UNET, Y2015	77.50		Medical use case
FCN, Y2014	70		Object segmentation

- *GPU with CUDA acceleration, is used for performance benchmarking
  - https://www.cityscapes-dataset.com/benchmarks/

Sample Outputs using SegNet



References

1. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
2. Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation." *Proceedings of the IEEE international conference on computer vision*. 2015.
3. Paszke, Adam, et al. "Enet: A deep neural network architecture for real-time semantic segmentation." *arXiv preprint arXiv:1606.02147* (2016).
4. Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *arXiv preprint arXiv:1511.00561* (2015).
5. Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2018): 834-848.
6. Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." *arXiv preprint arXiv:1511.07122* (2015).
7. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.
8. Zhao, Hengshuang, et al. "Icnet for real-time semantic segmentation on high-resolution images." *arXiv preprint arXiv:1704.08545* (2017).

Acknowledgements

- Extend sincere gratitude to following members for providing their valuable support and inputs
  - Dr. Lokesh Rayasandra Boregowda
  - Raj Narayana Gadde
  - Sujoy Saha
  - Suhas Shantaraja Palasamudram
  - Akankshya Kar
  - Vasanthakumar Rajagopalan
  - Sourav Kumar Bose
  - Anand Bhoraskar


THANK
YOU