Introduction to Semantic Segmentation

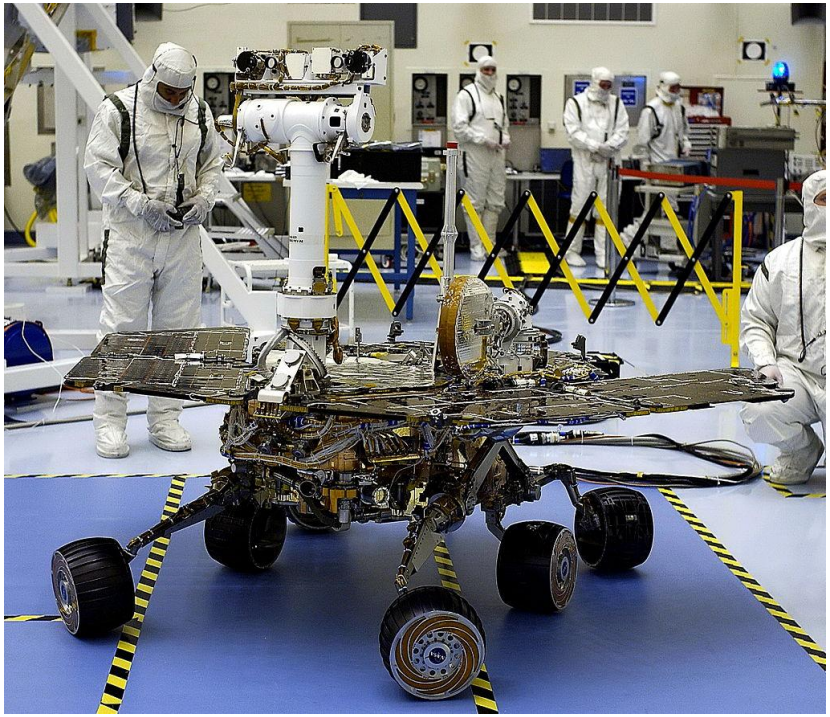
Sergei Belousov
Machine learning R&D Engineer

Internet of Things Group

Agenda

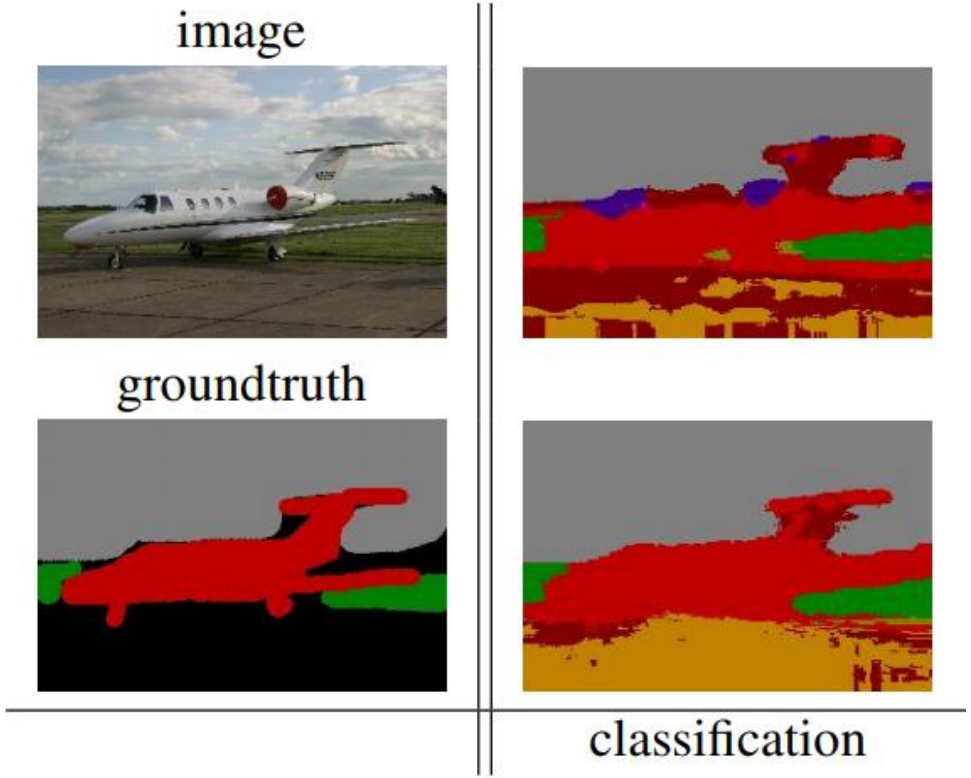
- Problem formulation
- Evaluation Metrics
- Datasets
- CNN
- Loss functions

In ancient time



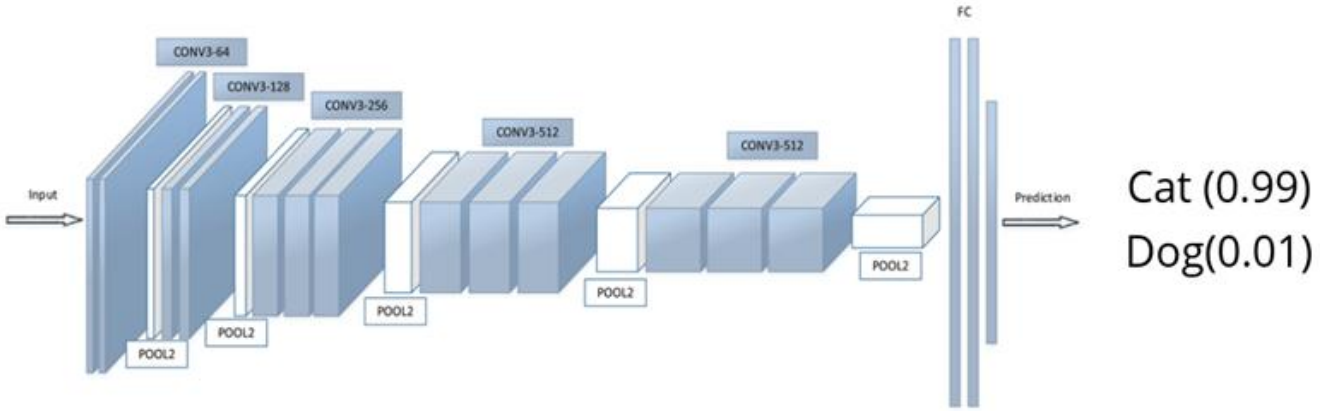
In ancient time

In ancient time

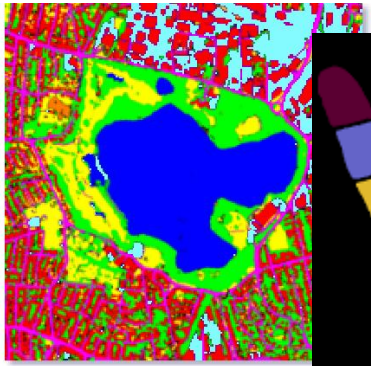


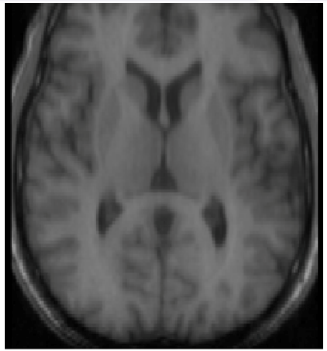
In ancient time

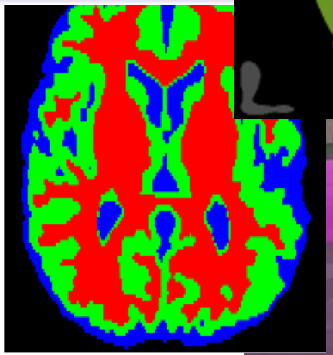
In recent time

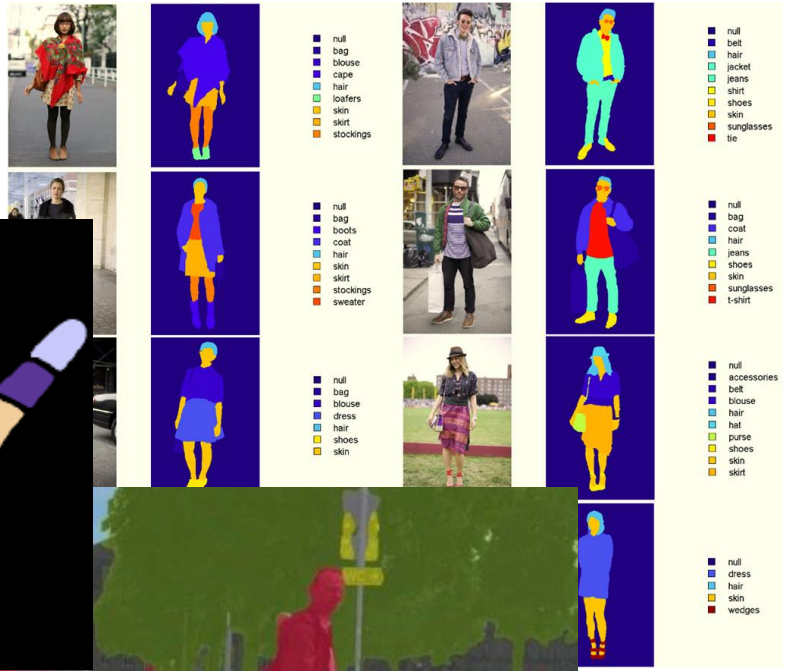


In recent time









Problem formulation

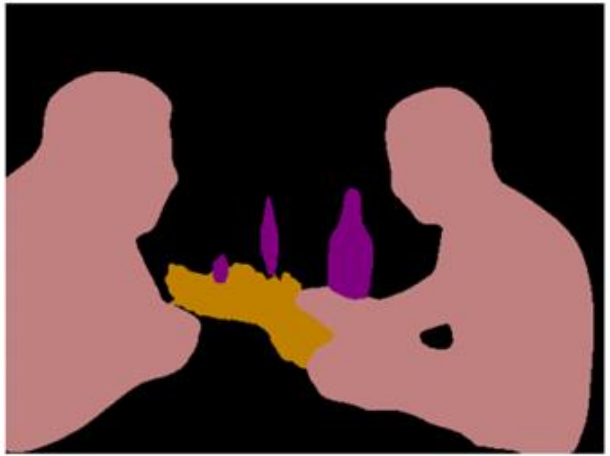
Input:

$$I \in R^{C*H*W} - input\ image$$
$$L \in [l_0, \dots l_n] - set\ of\ valid\ labels$$

Output:

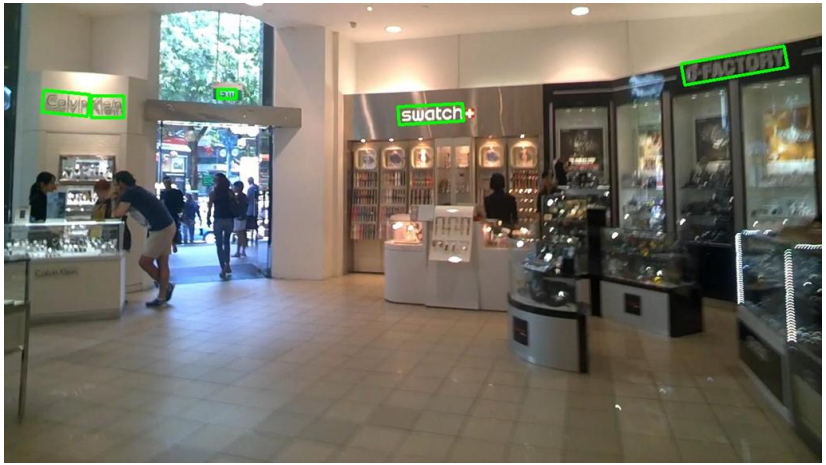
$$M \in L^{H*W} - labels\ mask$$

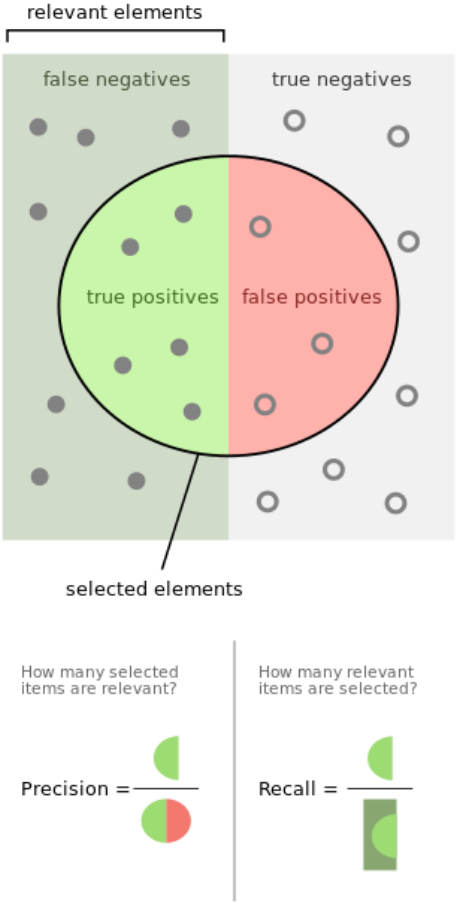
Evaluation metrics



Evaluation metrics

$$accuracy = \frac{TP+TN}{TP+TN+Fp+FN}$$





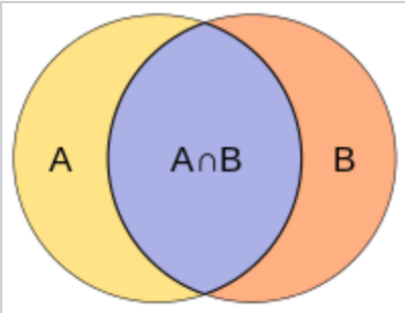
Evaluation metrics

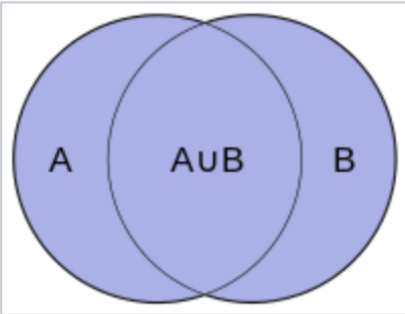
$$Dice(A,B) = 2\frac{|A \cap B|}{|A| + |B|} = \frac{2TP}{2TP + FN + FP}$$

$$IOU(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FN + FP}$$

$$IOU = \frac{Dice}{2 - Dice}$$

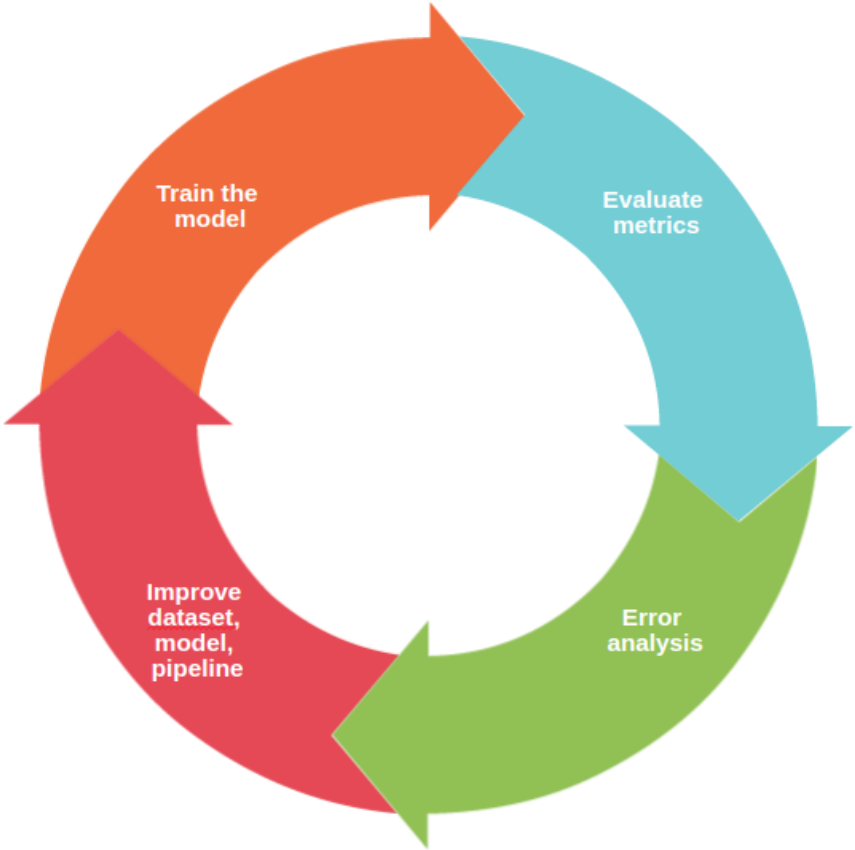
$$Error_{total} = c_0 FP + c_1 FN$$





Intersection and union of two sets A and B

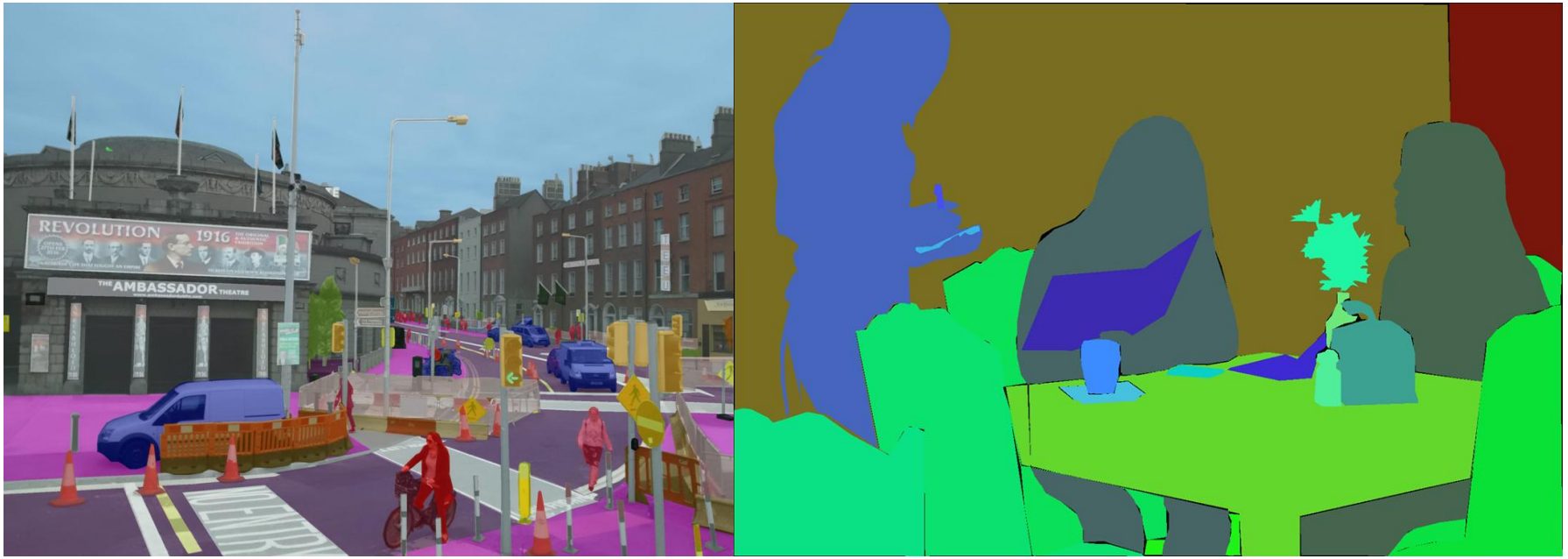
Lifecycle


Train the model
Evaluate metrics
Error analysis
Improve dataset, model, pipeline

Data is the oil of the 21st century

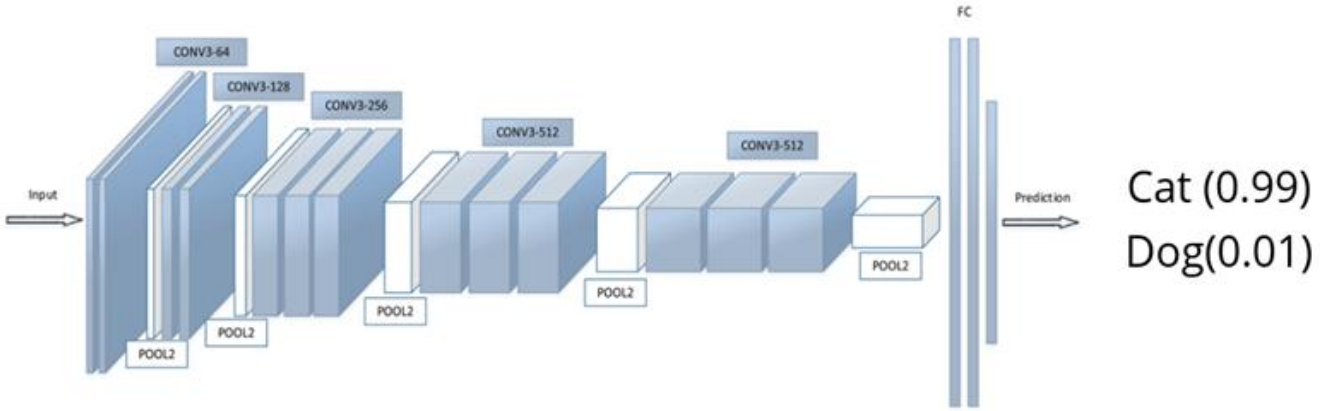
Dataset	Labeled Images for Training	Classes
KITTI	200	34
VOC PASCAL 2012	2913	21
Cityscapes	3478	34
BDD100K	8000	19
ADE20K	20210	3169
Mapillary Vistas	20000	66
ApolloScape	147000	36
WAYMO	600000	?

Data is the oil of the 21st century

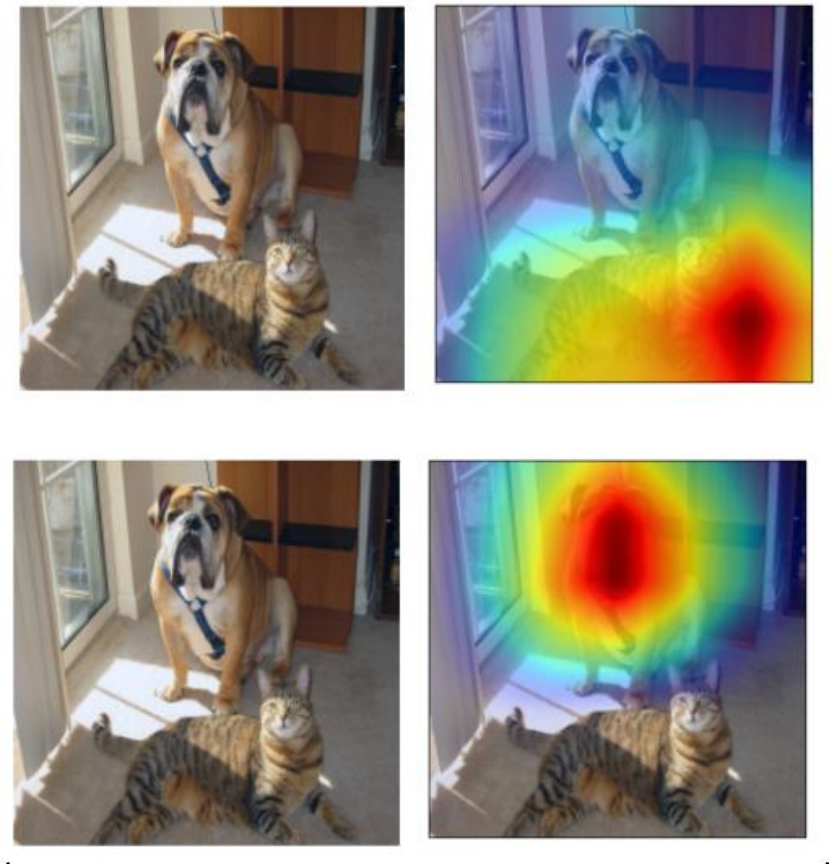


Data is the oil of the 21st century

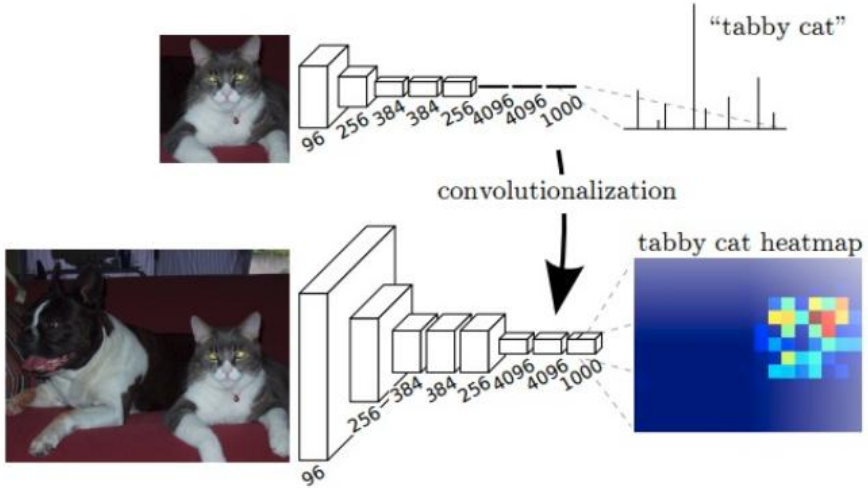
CNN

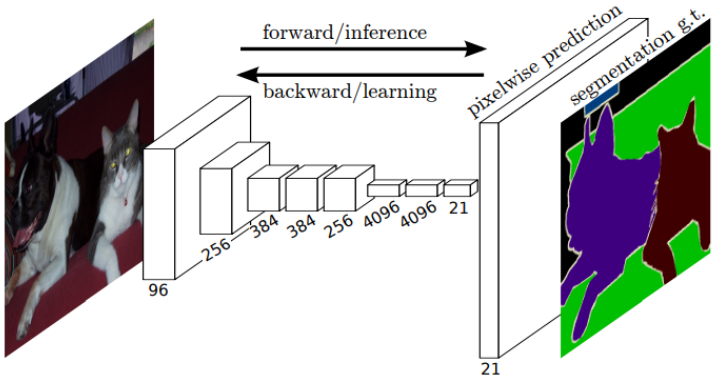

Input
CONV3-64
CONV3-128
CONV3-256
CONV3-512
CONV3-512
POOL2
POOL2
POOL2
POOL2
POOL2
FC
Prediction
Cat (0.99)
Dog(0.01)

CNN

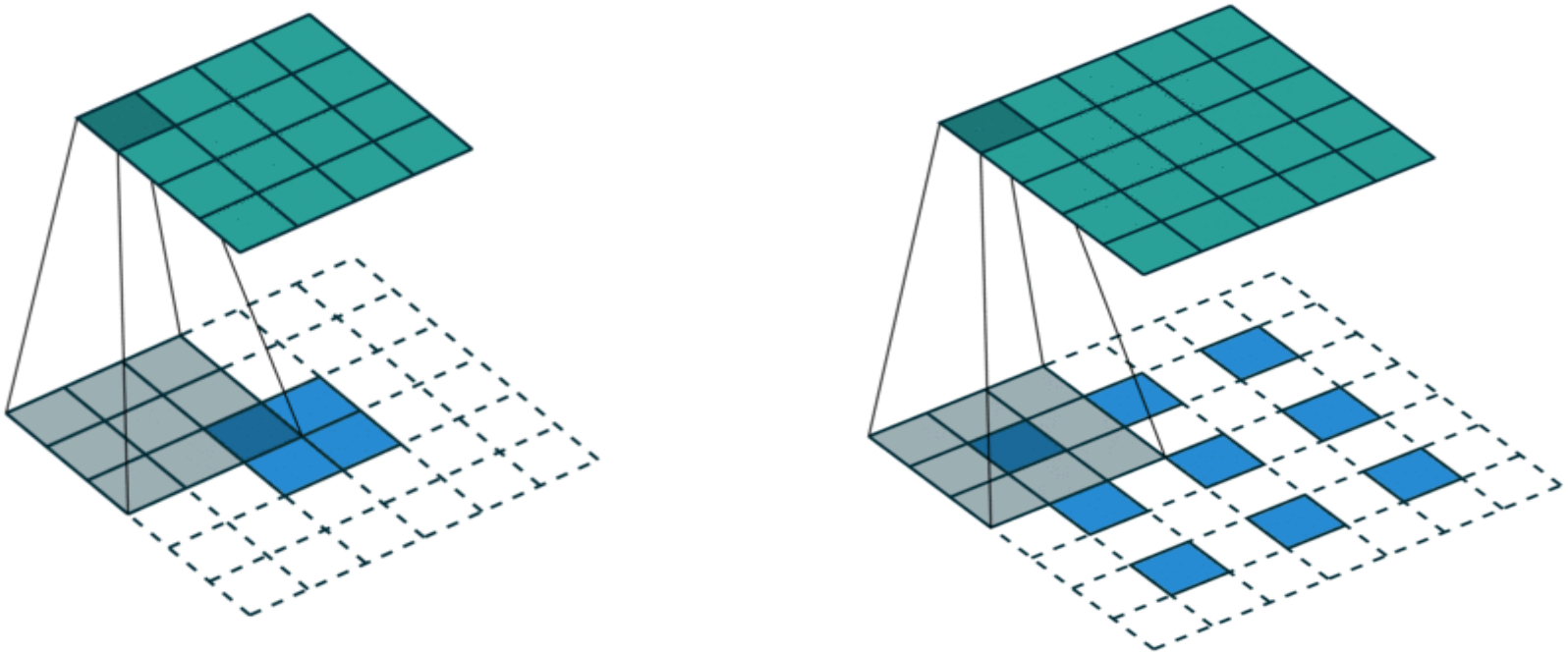


CNN: FCN

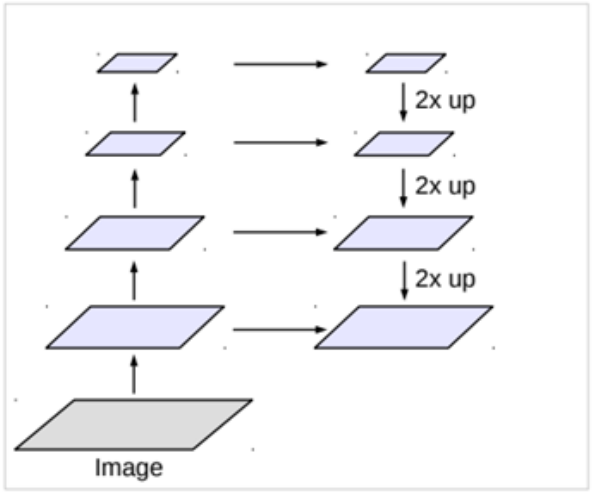

"tabby cat"
96
256
384
384
256
4096
4096
1000
convolutionalization
tabby cat heatmap
96
256
384
384
256
4096
4096
1000


forward/inference
backward/learning
pixelwise prediction
segmentation g.t.
96
256
384
384
256
4096
4096
21
21

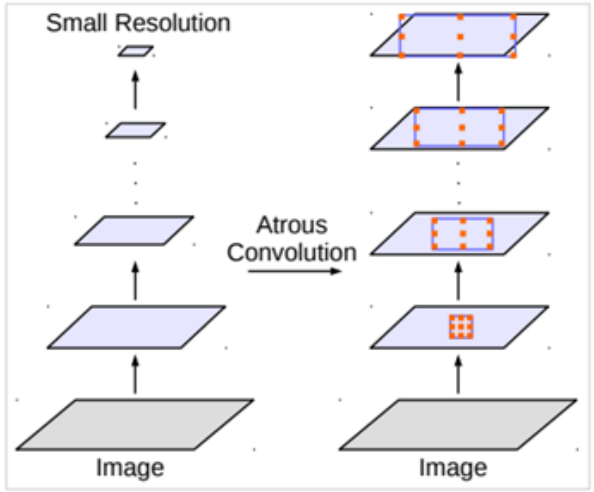
CNN: Deconvolution



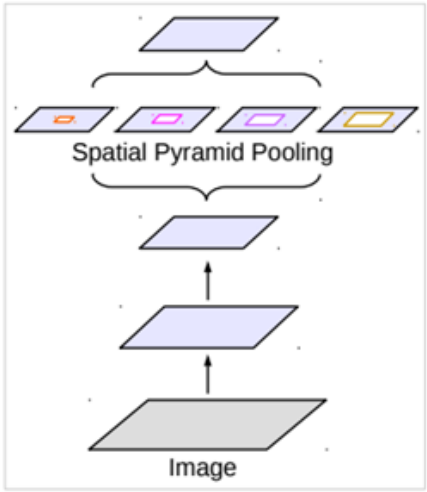
CNN: Architectures to capture multi-scale context



(b) Encoder-Decoder

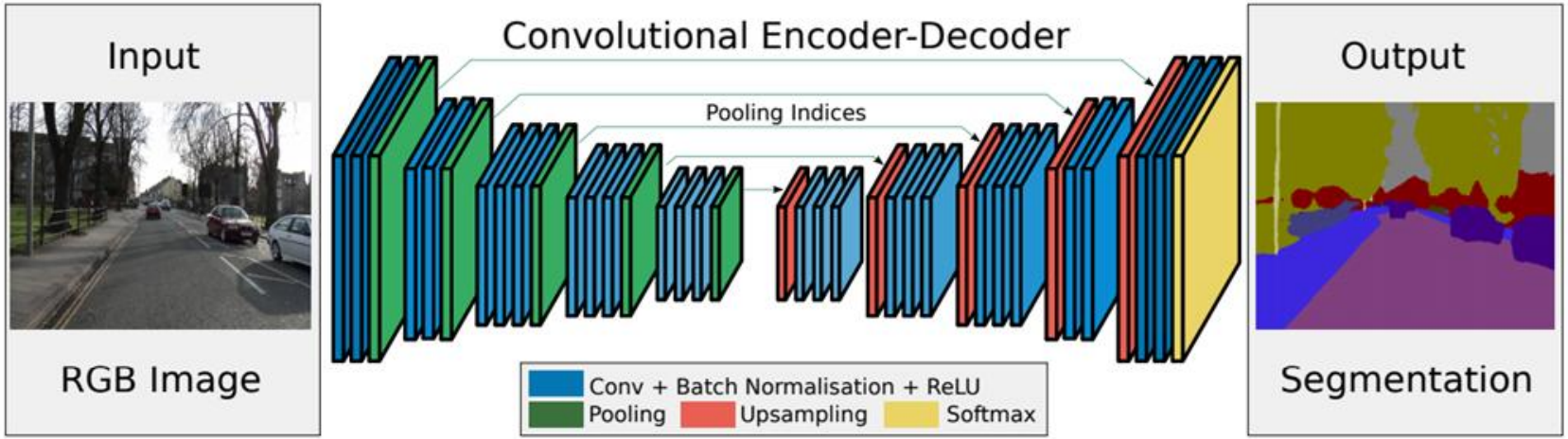


(c) Deeper w. Atrous Convolution

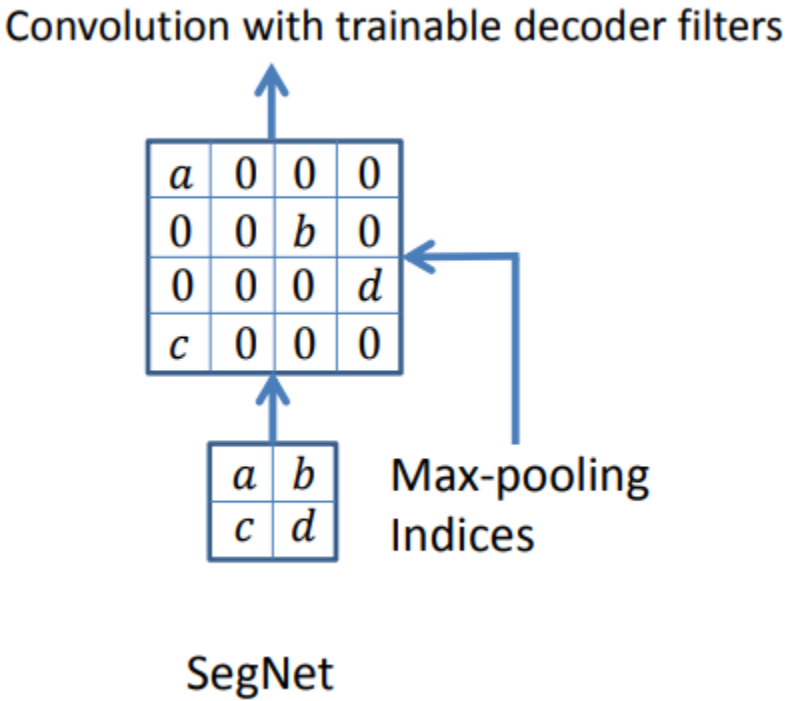


(d) Spatial Pyramid Pooling

CNN: Encoder-Decoder

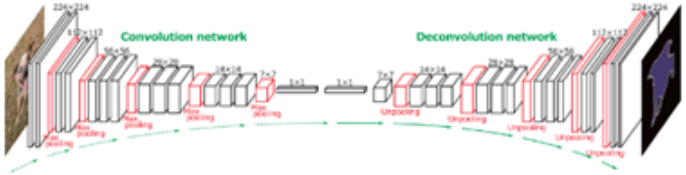


CNN: Encoder-Decoder

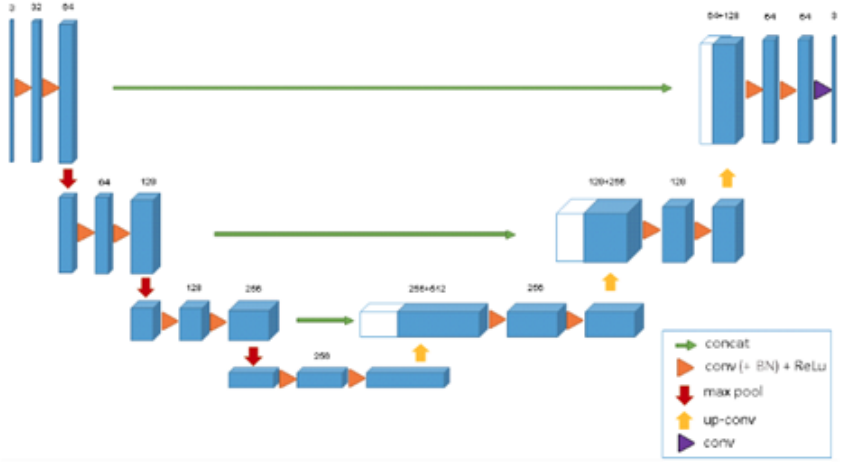


CNN: Encoder-Decoder

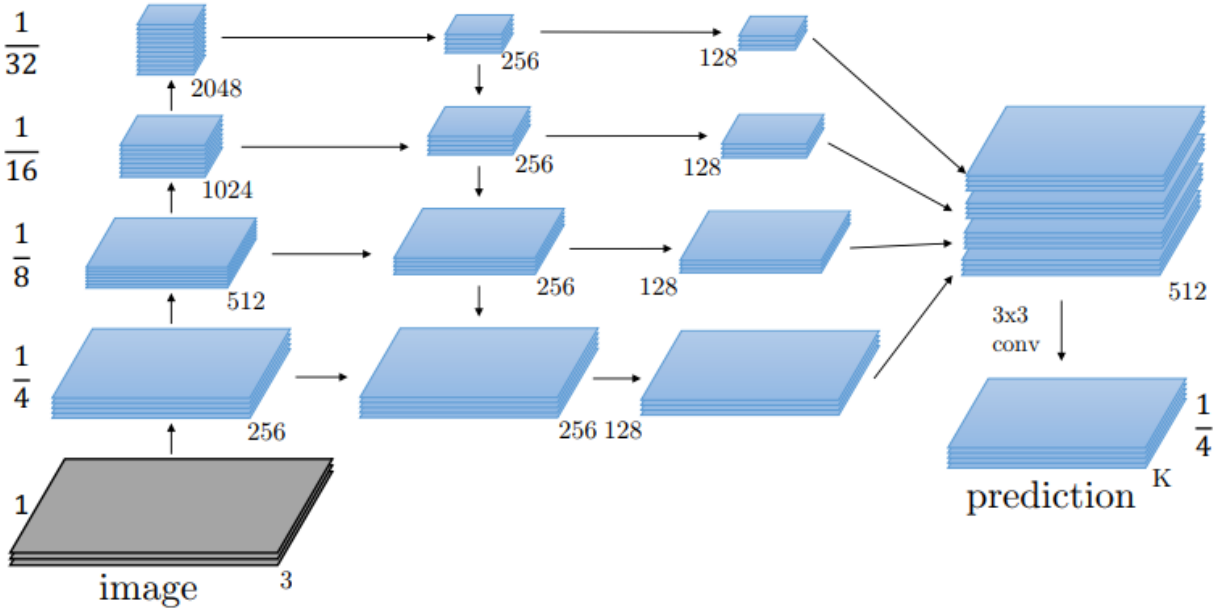
SegNet



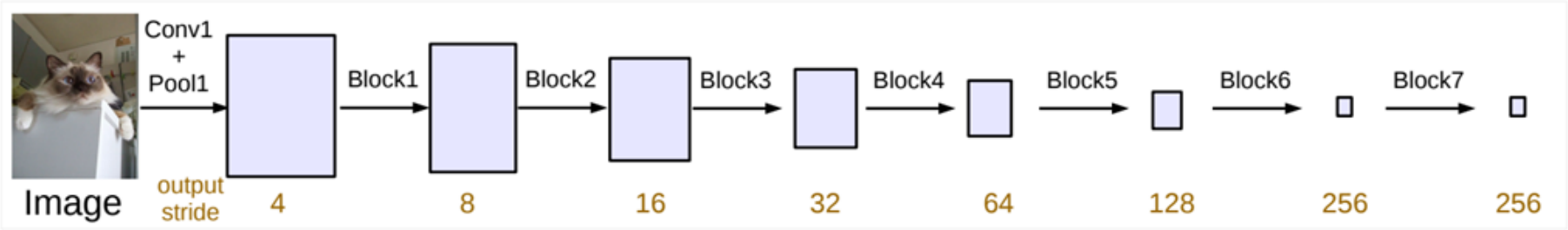
UNet



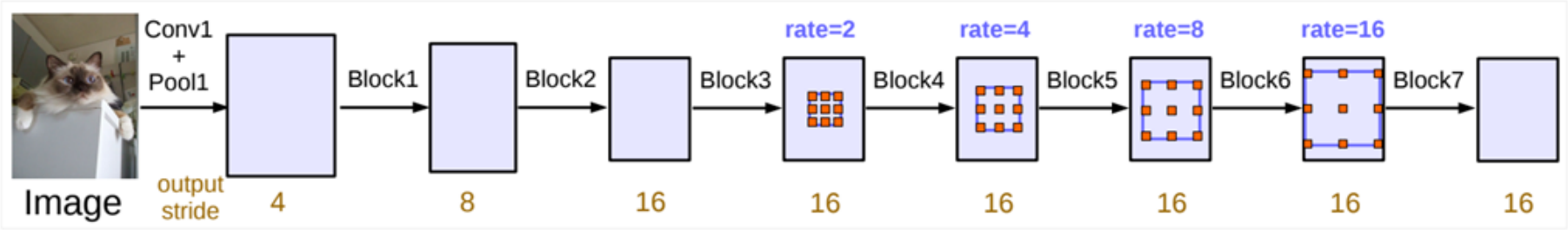
CNN: Encoder-Decoder



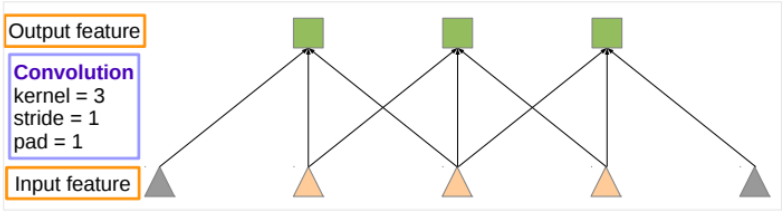
CNN: w. Atrous Convolutions



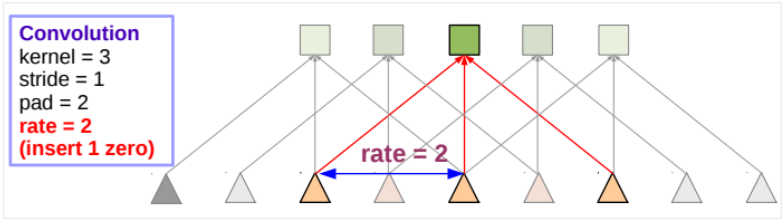
(a) Going deeper without atrous convolution.

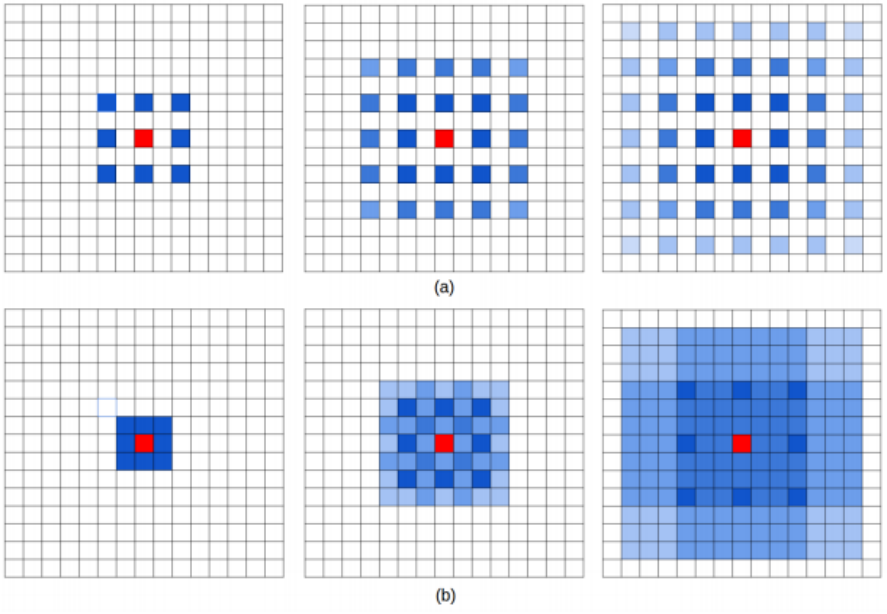


CNN: w. Atrous Convolutions

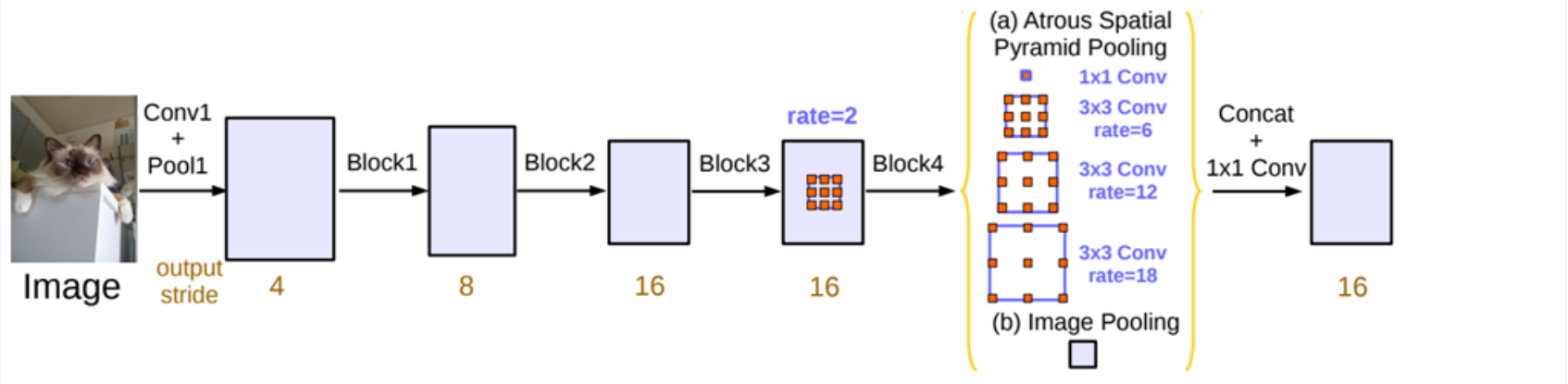


(a) Sparse feature extraction

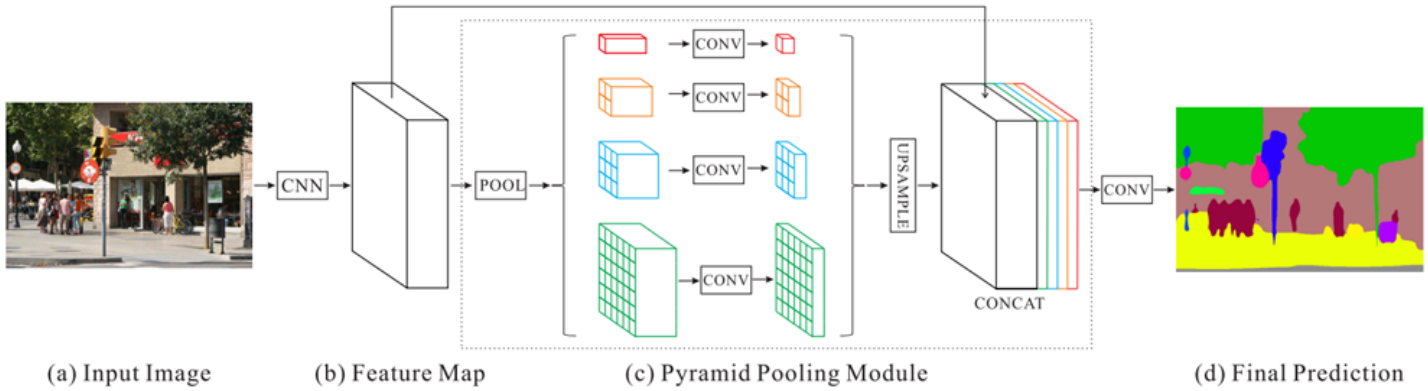




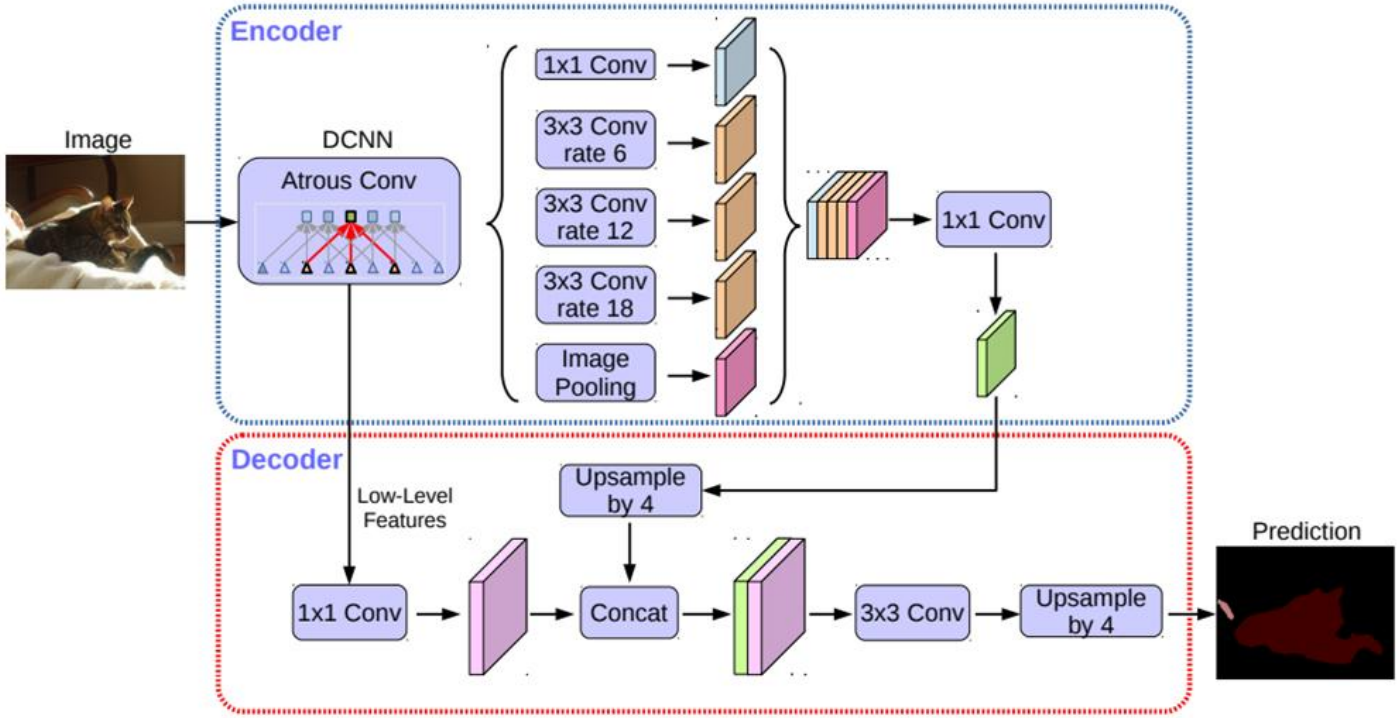
CNN: Spatial pyramid pooling



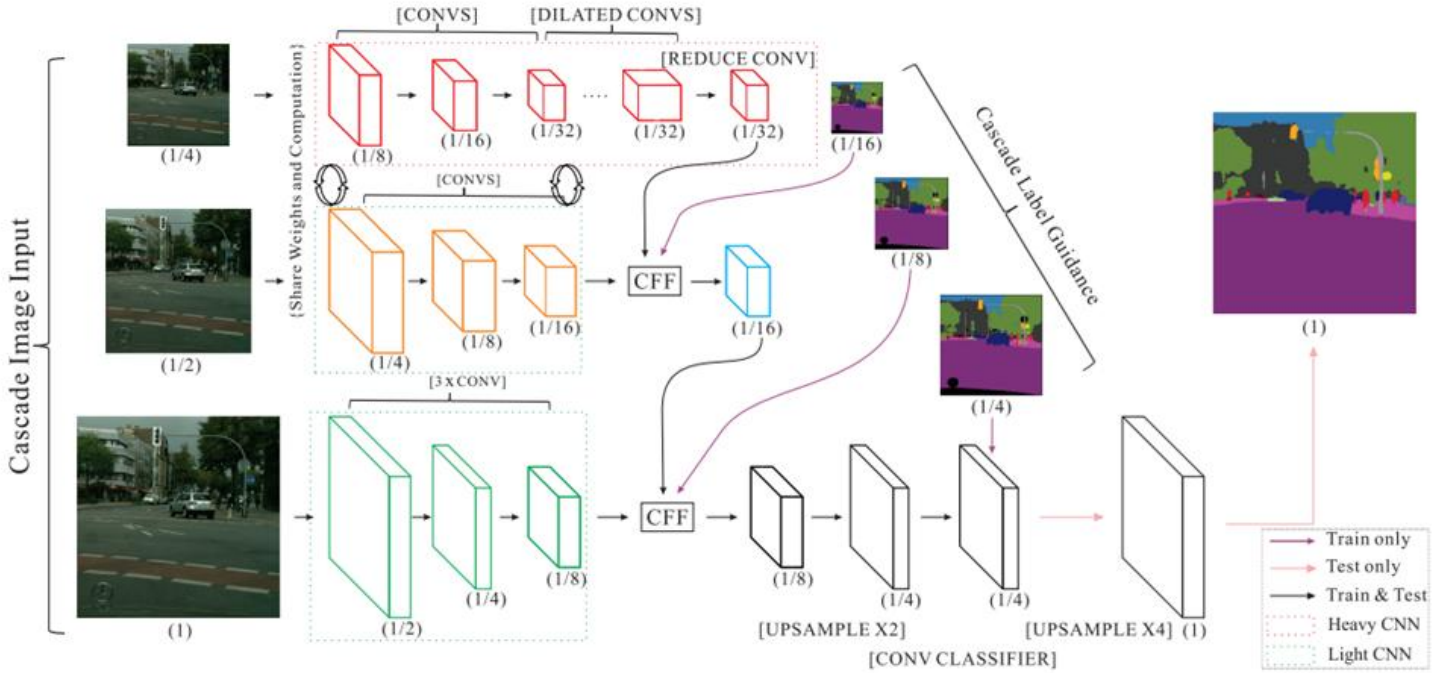
CNN: Spatial pyramid pooling



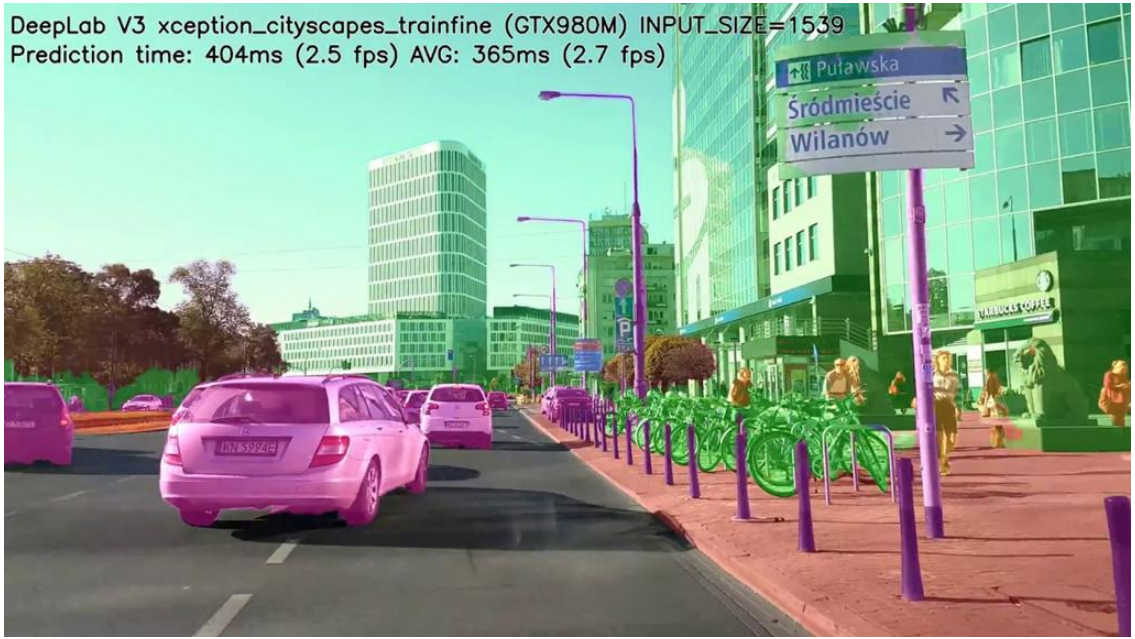
CNN: All included



CNN: ICNet

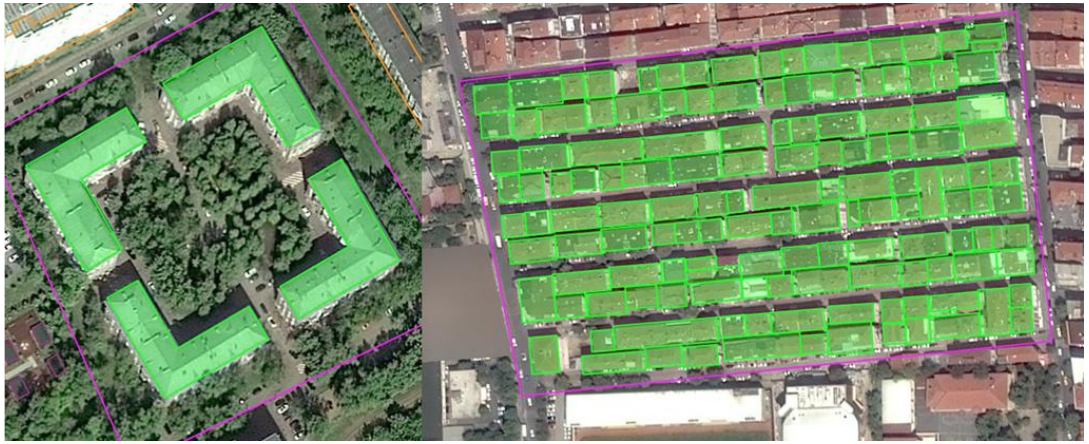

Cascade Image Input
(1/4)
(1/2)
(1)
{Share Weights and Computation}
[CONVS]
[DILATED CONVS]
[REDUCE CONV]
(1/8)
(1/16)
(1/32)
(1/32)
(1/32)
[CONVS]
(1/4)
(1/8)
(1/16)
[3 x CONV]
(1/2)
(1/4)
(1/8)
CFF
CFF
(1/16)
(1/16)
(1/8)
(1/4)
Cascade Label Guidance
(1)
(1/8)
(1/4)
(1/4)
[UPSAMPLE X2]
[UPSAMPLE X4]
(1)
[CONV CLASSIFIER]
Train only
Test only
Train & Test
Heavy CNN
Light CNN

Results

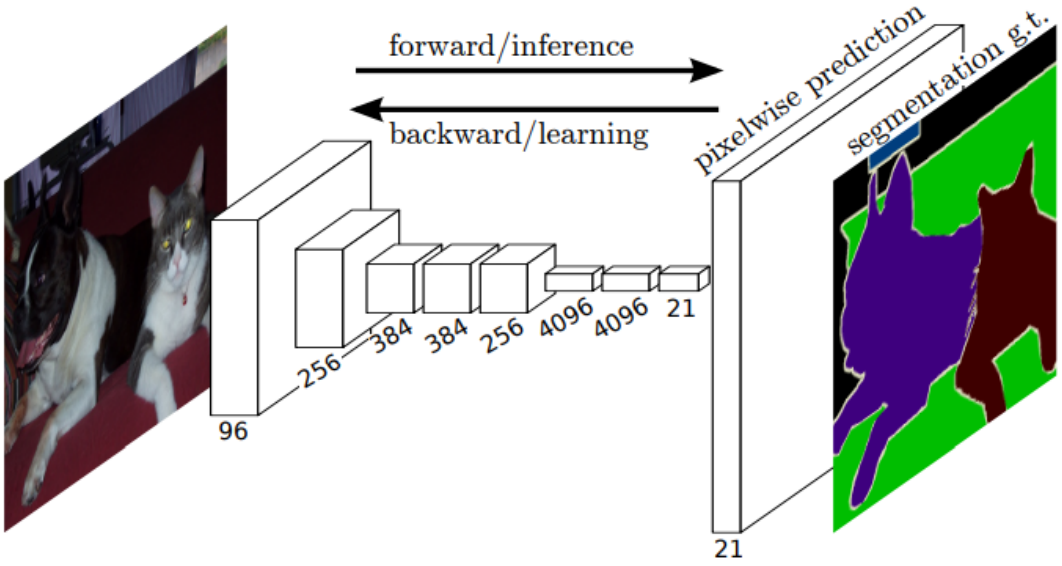

DeepLab V3 xception_cityscapes_trainfine (GTX980M) INPUT_SIZE=1539
Prediction time: 404ms (2.5 fps) AVG: 365ms (2.7 fps)
Puławska
Śródmieście
Wilanów

Results

Results



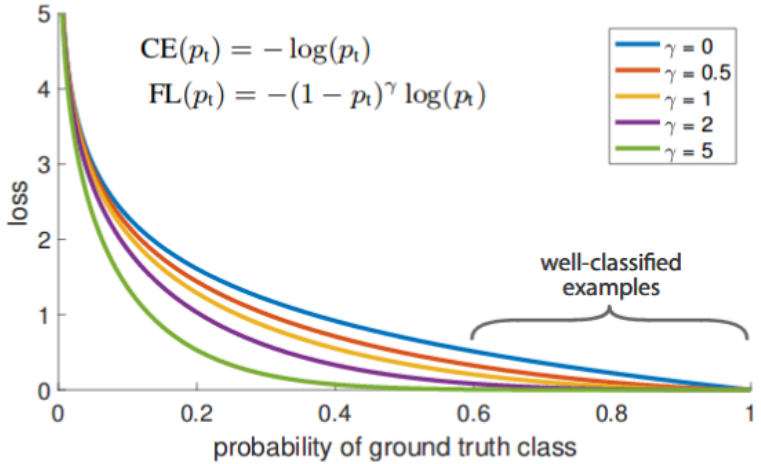
Loss functions



Loss functions

$$L_{CE}(p, y) = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

$$L_{Focal}(p, y) = -\sum_{c=1}^{M} y_{o,c} * (1 - p_{o,c})^{\gamma} * \log(p_{o,c})$$

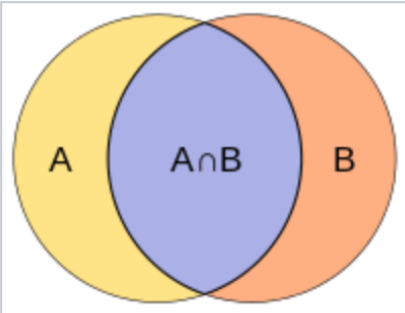


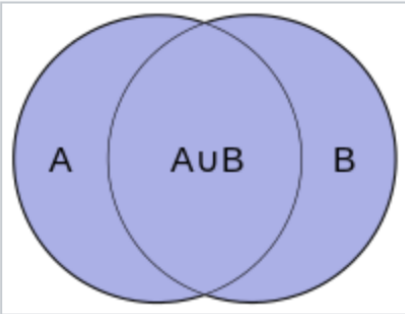
Loss functions

$$Dice = \frac{2TP}{2TP+FN+FP} = \frac{2|A \cap B|}{|A|+|B|}$$

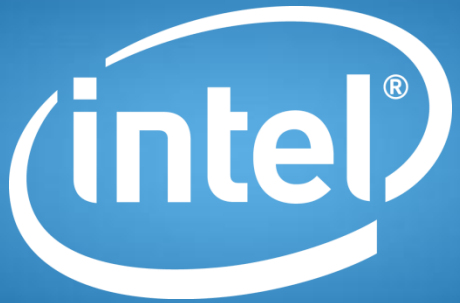
$$Dice(p, y) = \frac{2 * \sum_i^N p_i y_i}{\sum_i^N p_i + \sum_i^N y_i}$$

$$L_{Dice}(p, y) = 1 - \frac{2 * \sum_i^N p_i y_i}{\sum_i^N p_i + \sum_i^N y_i}$$





Intersection and union of two sets A and B


intel®

- UNet: https://arxiv.org/abs/1505.04597
- DeepLab: https://arxiv.org/abs/1606.00915
- DeepLabV3: https://arxiv.org/abs/1706.05587
- DeepLabV3+: https://arxiv.org/abs/1802.02611
- SegNet: https://arxiv.org/abs/1511.00561
- FCN: https://arxiv.org/abs/1411.4038
- Grad-CAM: https://arxiv.org/abs/1610.02391

- https://github.com/mrgloom/awesome-semantic-segmentation
- Kaggle: https://www.kaggle.com/
- ODS (@bes): https://ods.ai/ https://opendatascience.slack.com
- Deep Learning Book: https://www.deeplearningbook.org/