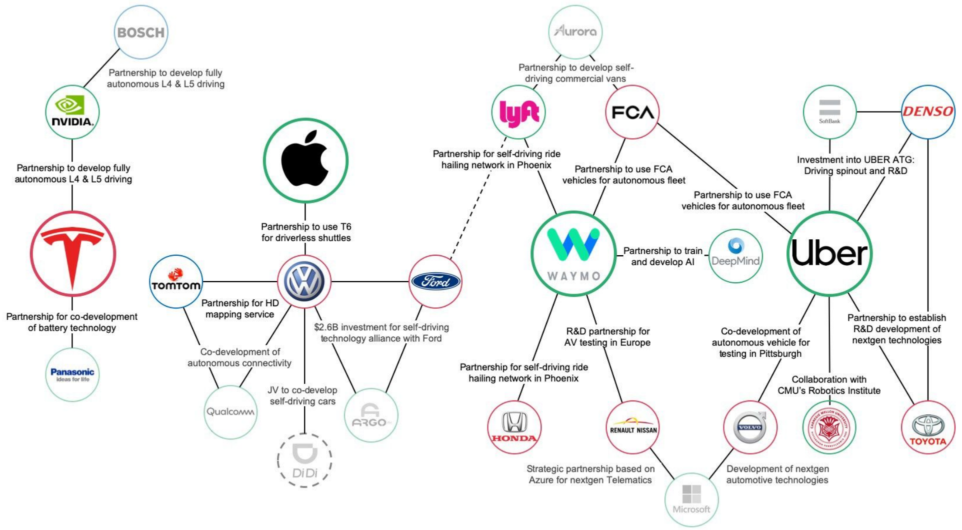
Semantic Segmentation for Autonomous Vehicles

Background

Autonomous vehicles are a massive industry (~$20 billion)

Scene understanding is necessary for safe and consistent driving



https://medium.com/@firstmilevc/avlandscape-8a21491f1f54

Scene Understanding

Semantic segmentation - per pixel image/video understanding

Helps to identify what is in the road and surroundings to be used in optimal control

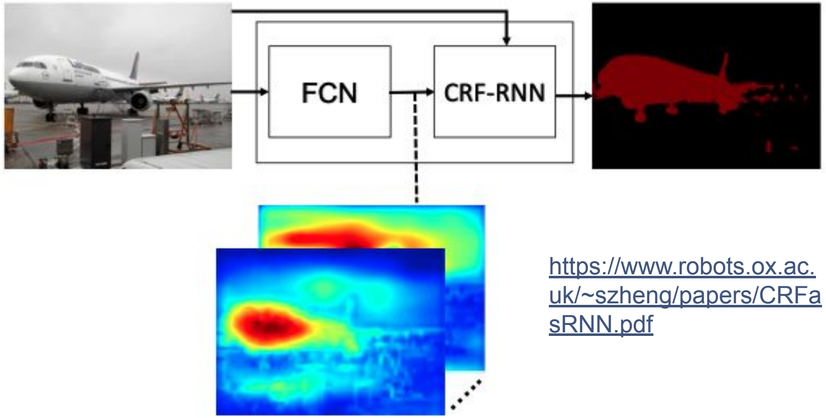
https://www.tesla.com/autopilotAI

Why ML/DL can help

No prescribed method in determining a class for given pixels in an image

Success with Deep Learning - data not always abundant

Conventional methods were poor performers (now used as refinement layers)

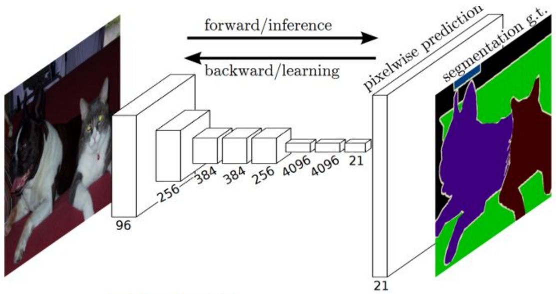


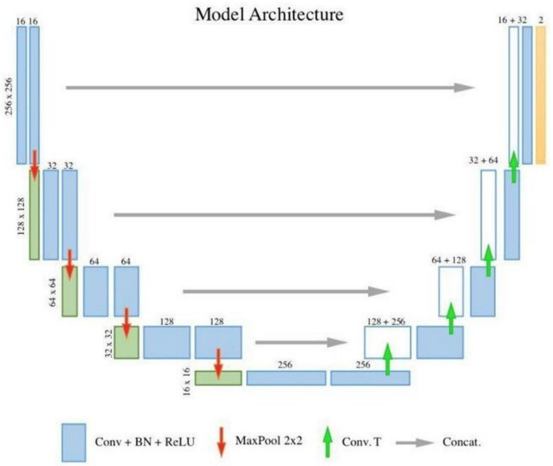
https://www.robots.ox.ac.uk/~szheng/papers/CRFasRNN.pdf

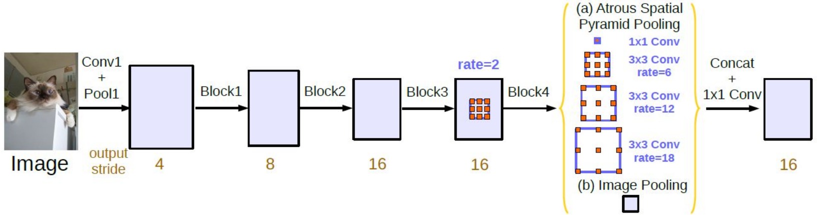
Literature Review

Semantic Segmentation: FCN, U-Net, DeepLab, GCN … plus many others

➔ CNNs! (mainly)



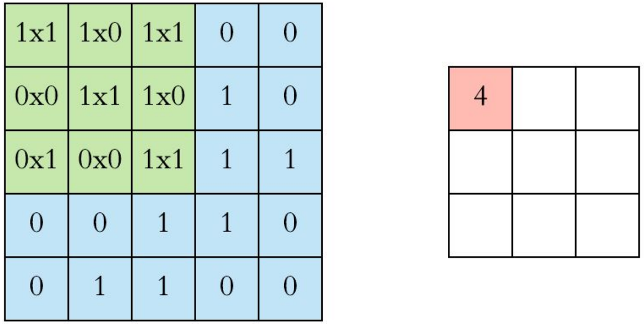


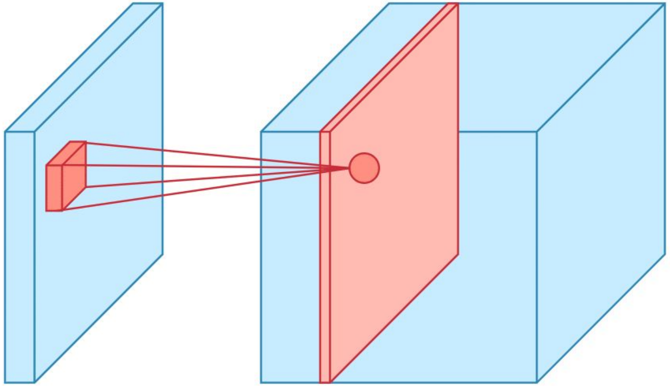


https://towardsdatascience.com/semantic-segmentation-popular-architectures-dff0a75f39d0
https://towardsdatascience.com/review-deeplabv3-atrous-convolution-semantic-segmentation-6d818bfd1d74

Feature Extraction

Convolutional layers extract features automatically from the data by what causes large output values - *filters* out poor indicators





https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Dataset details

CityScapes data subset - tackling small data problem

- ~2600 images for training
- ~300 images for validation
- ~500 for test

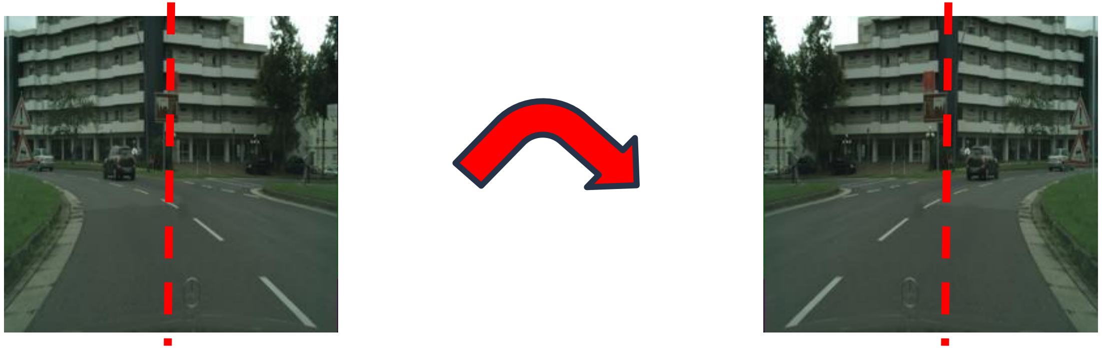
Labels were attached to data (right) and not categorical

https://www.kaggle.com/dansbecker/cityscapes-image-pairs

Our approach

Methods for dealing with small data - fewer parameter networks, skip connections, transfer learning, data augmentation, etc.

Evaluation: Intersection Over Union, **Pixel Accuracy, Manhattan Score**

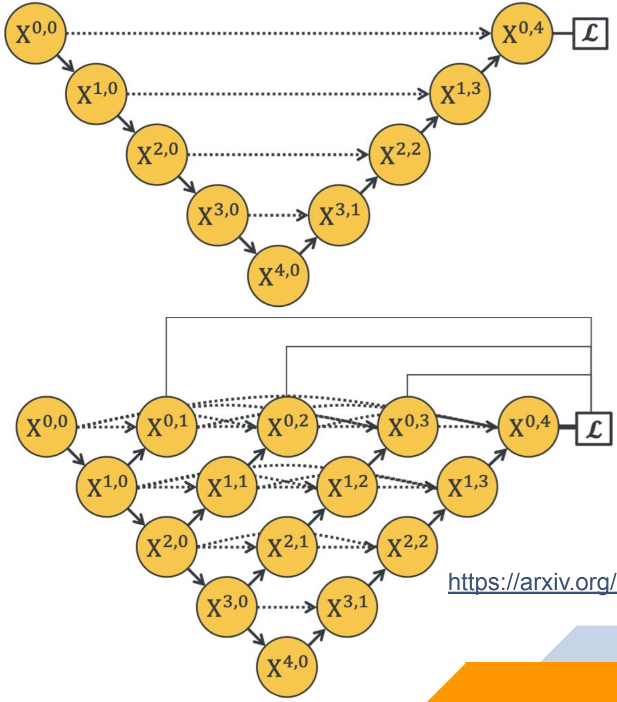


Models - U-net/U-net++

Encoder and Decoder structure

- Convolutional layers
- Pooling
- Upsampling

Skip connections to counter vanishing gradient/poor learning

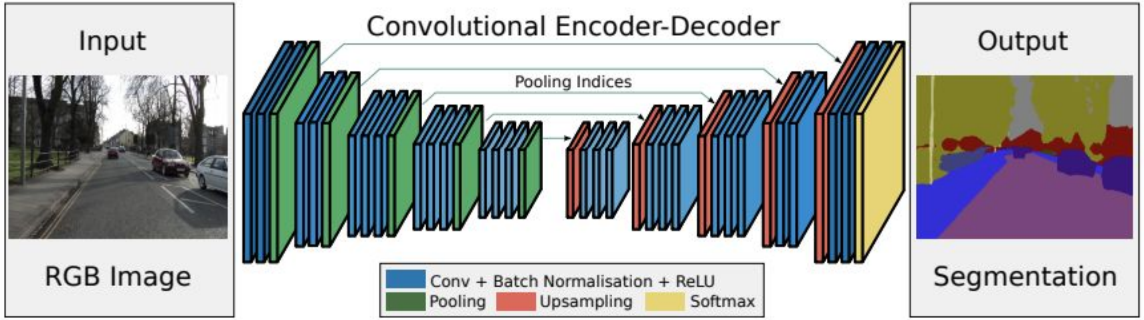


https://arxiv.org/pdf/1912.05074.pdf

Models - SegNet

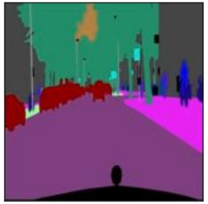
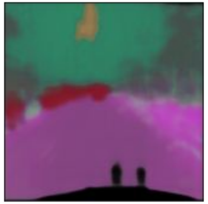
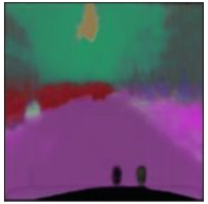
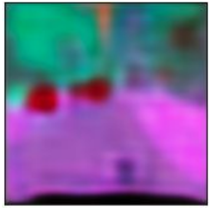
Store max-pooling indices: location of the maximum feature value of each pooling window

Replace U-net skip connections with passing these indices



https://arxiv.org/pdf/1511.00561.pdf

Results - same filters/kernels

Metric	Label	U-net	U-net++	SegNet
Image				
Training Method	-	From scratch	From scratch	From scratch
Parameters	-	~543k	~641k	~278k
Pixel Accuracy	-	45%	62%	22%
Manhattan Score per pixel	-	0.21	0.17	0.52

Metric	SOTA
mIoU	86%
Parameters	~6M

Observations

Not ideal to use MSE for segmentation

Models trained from scratch did OK

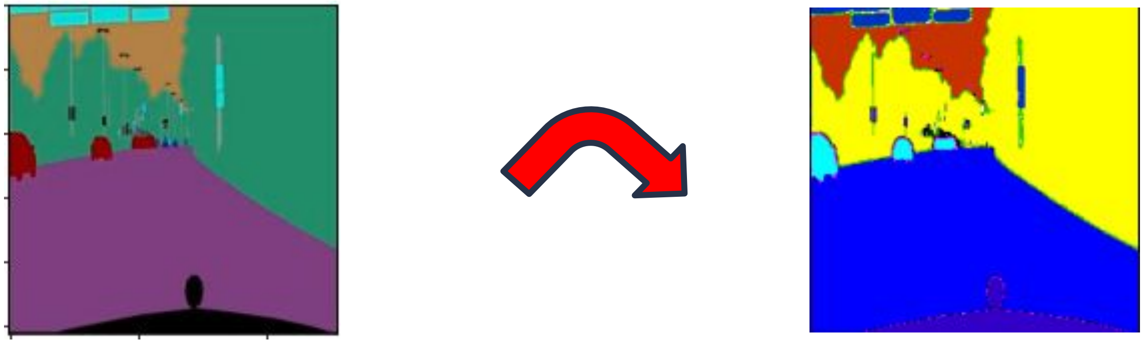
Computational power/availability was limited for extending and perfecting results

Further improvements ideas

Transfer learning with pre-trained FCN

Labeling for different loss functions/metrics

Further data augmentation



References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[3] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, Springer, 2018, pp. 3–11.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[5] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.