

ELEG 5491: Introduction to Deep Learning

Semantic Segmentation in Computer Vision

Prof. LI Hongsheng

Office: SHB 428

e-mail: hsli@ee.cuhk.edu.hk

web: <https://dl.ee.cuhk.edu.hk>

Department of Electronic Engineering
The Chinese University of Hong Kong

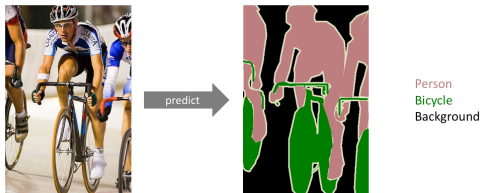
March 2023

Outline

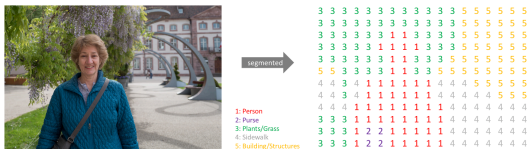
- 1 Semantic Segmentation Overview
- 2 Segmentation Datasets and Evaluation Metrics
- 3 Semantic Segmentation Networks

The Problem of Semantic Segmentation

- Given an input image, semantic segmentation aims at assigning each pixel a pre-defined class labels



- Ground-truth representation: each pixel stores a ground-truth class label

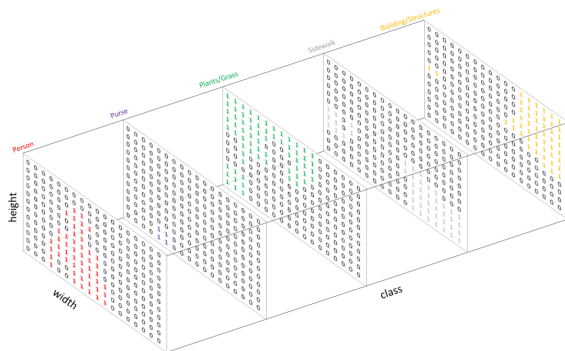


Input

Semantic Labels

Interpretation of 2D Class Maps and Pixel-wise Cross-entropy Loss

- The target class maps should be of size $H \times W \times N$, where H, W are height and width of the input image, and N is the number of pre-defined classes



- For each pixel, it requires to conduct a multi-class classification
- The multi-class cross entropy loss is therefore used for each pixel. In other words, the number of samples in a mini-batch would be (number of pixels \times number of images)

Interpretation of 1×1 Convolution

- The 2D feature maps of size $H \times W \times C$ can be viewed as a series of C -dimensional feature vector at each spatial location (x, y) to describe the image contents centered at (x, y)
- 1×1 convolutions can be considered as a fully-connected layer applied to each of the C -dimensional feature vector for local feature transformation

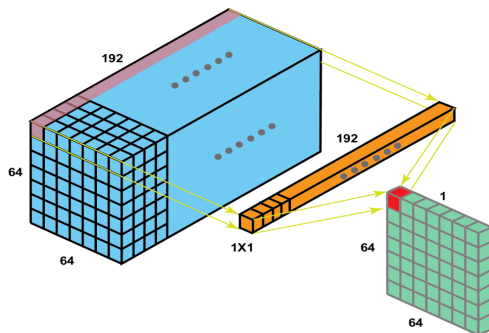
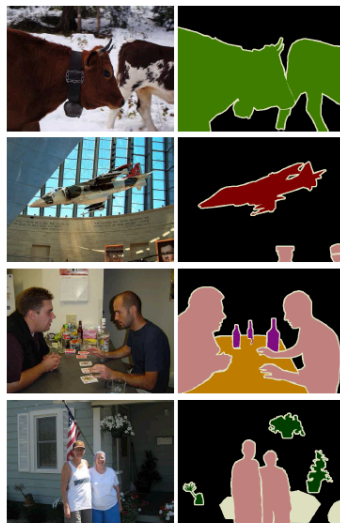


Figure: $H = 64, W = 64$ and $C = 192$. 1×1 convolution can be considered as applying a fully-connected layer to each 192-dimensional feature vector for feature transformation.

PASCAL VOC 2012 Dataset

- 1,464 images for training, 1,449 for validation, and 1,456 for testing
- 20 foreground objects classed and one background class



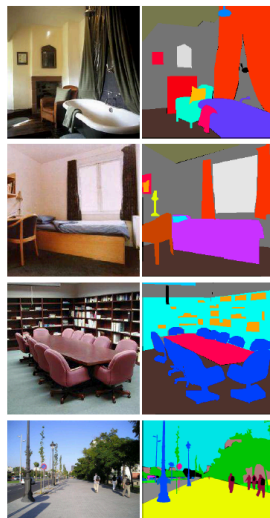
(a) Image

(b) Ground Truth



ADE20k Dataset

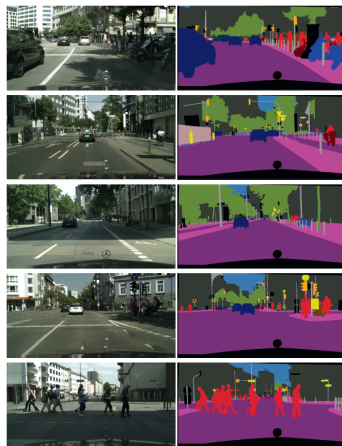
- 20K images for training, 2K images for validation, and 3K images for testing
- Used for ImageNet Scene Parsing Challenge 2016
- This dataset is more complex and challenging with 150 labeled classes and more diverse scenes



(a) Image (b) Ground Truth

Cityscapes Dataset

- Images of driving scenes from 50 cities
- 5,000 high quality pixel-level finely annotated scene images, which is divided into 2,975/500/1,525 images for training, validation and testing
- 30 classes, and 19 classes among them are used for evaluation.



(a) Image

(b) Ground Truth

Evaluation metrics

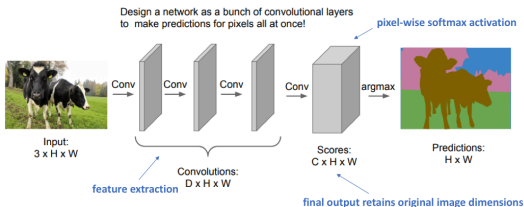
- Naively, one can use pixel accuracy as the evaluation metric. However, as there generally exist a large number of background pixels, accuracy is rarely used nowadays
- Currently, the most popular metric is mean of class-wise intersection over union (mIoU). The following equation is used for each class separately and their IoUs are averaged to obtain mIoU

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

- There are also other metrics, such as F1 score (mainly for binary segmentation), iIoU (used for Cityscapes), etc.

Constructing Semantic Segmentation Networks

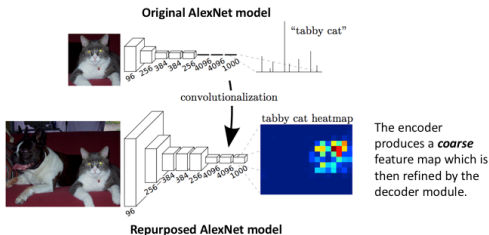
- A naive approach for constructing a network for segmentation would be to simply stacking a number of convolutional layer without any downsampling operations



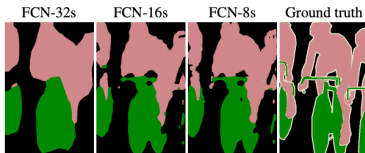
- On the one hand, such a design makes each pixel in the final feature map have very small receptive field, which cannot result in satisfactory performance
- On the other hand, general CNN for whole-image classification gradually decreases the spatial size but increases the feature channels to use longer feature vectors for encoding image contents
- However, such a mechanism cannot be achieved in the above design because if the spatial size is maintained throughout the whole network, increasing the feature channels would occupy the limited GPU memory

Fully Convolutional Network (FCN)

- Fully Convolutional Network (FCN) converts the image classification network to tackle the task of image segmentation. However, the last 2D feature maps has $1/32$ spatial size as the original input image

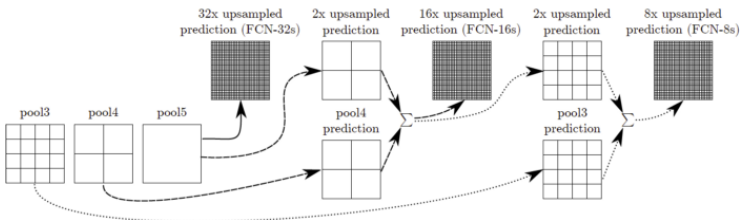


- Direct upsampling the $1/32$ segmentation results show over-smoothed segmentation boundaries

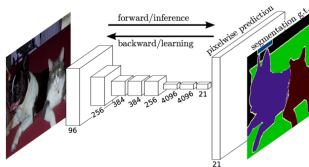


Fully Convolutional Network (FCN)

- The last pool5 (1/32) predictions are upsampled to 1/16 and summed with pool4 (1/16) predictions. The summed 1/16 predictions are upsampled to 1/8 and summed with pool3 (1/8). The network outputs 1/8 segmentation maps and upsample the results to the original resolution

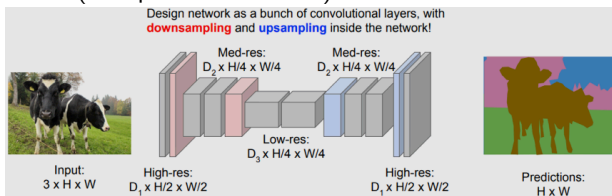


- The full network is trained with a pixel-wise cross entropy loss



Encoder-decoder Architecture

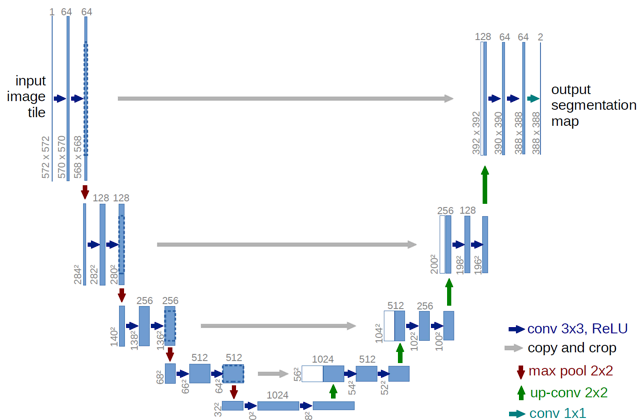
- However, the FCN obtains the segmentation result by upsampling feature maps of each size once, and the summed feature maps only go through a single layer of linear classifier (fully-connected layer)
- FCN therefore can only achieve mediocre performance
- One popular category of approaches is the encoder-decoder architecture, whose encoder gradually downsamples the spatial size but increases the feature channel number, and the decoder gradually upsamples the spatial size but decrease the feature channel number
- Two choices for upsampling the feature maps into higher spatial resolution: bilinear interpolation (more frequently adopted nowadays) and deconvolution (transposed convolution)



- However, as the encoded features inevitably lose some spatial resolution and the final results still might have over-smoothed boundaries

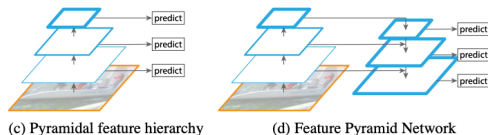
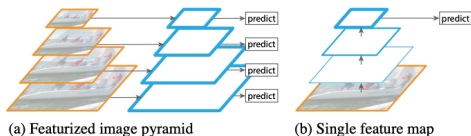
U-Net

- U-Net establishes shortcuts between feature maps of the same size of encoder and decoder
- In this way, the high-resolution information in the feature maps of the encoder would be fused with those of the decoder via channel-wise concatenation



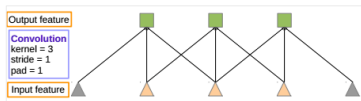
Feature Pyramid Network

- The Feature Pyramid Network (FPN) was originally proposed for object detection, but it can also create multi-scale feature maps as the U-Net does. The only difference is that FPN use addition as shortcut connections while U-Net uses concatenation as shortcut connections
- Unlike object detection, we only use the topmost feature maps, which is followed by an MLP to generate the resulting label map



The DeepLab Series: DeepLabv1

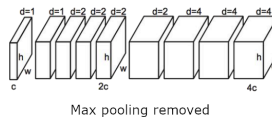
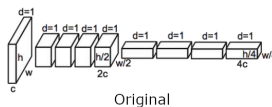
- A VGG-16 backbone pre-trained on ImageNet with fully connected layers being adopted
- Original VGG network downsamples the topmost feature maps to 1/32
- **Plausible Solution:** Remove the last two 2x max-pooling layers so that the topmost feature maps is 1/8 of the input size
- **Problem:** The un-maxpooled feature maps would mis-align with the pre-trained kernels of ImageNet
- **Solution:** Astrous convolution (dilation greater than 1) is introduced. Removing the first 2x max pooling makes the following convolution layers have dilation 2. Removing the second 2x max pooling makes the following convolution layer have dilation 4



(a) Sparse feature extraction

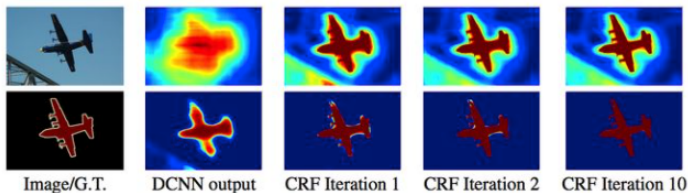


(b) Dense feature extraction



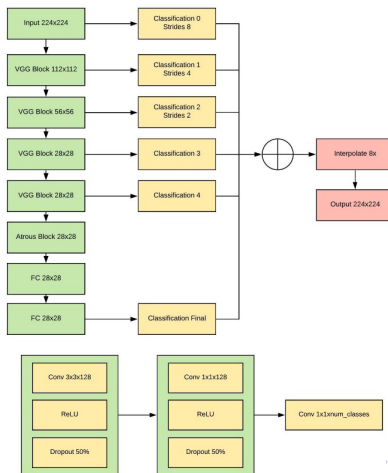
DeepLab v1: bilinear interpolation

- Use bilinear interpolation to upsample the $1/8$ feature maps into the original resolution
- Use continuous Conditional Random Field model to encourage the label maps follow the RGB pixel values: the more similar and closer a pair of pixels are, the more likely their labels are



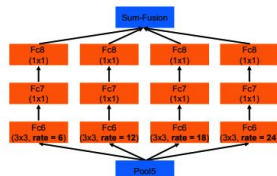
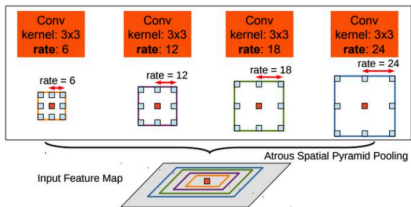
DeepLab v1: Multi-scale Prediction

- Similar to FCN, the multiple classification heads predict class label maps from feature maps of multiple depths and scales
- The multi-scale predictions are summed to generate the final label map



DeepLabv2

- DeepLabv2 introduces Atrous Spatial Pyramid Pooling (ASPP) before the classification head to include convolutions of different dilations to integrate context of different scales
- Reason: it is discovered as the sampling rate becomes larger, the number of valid filter weights become smaller

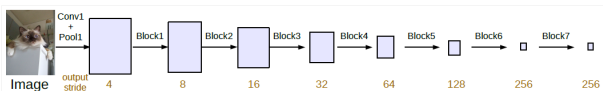


(a) DeepLab-LargeFOV

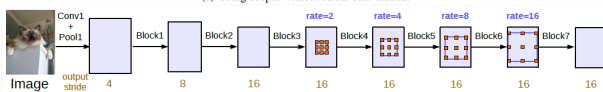
(b) DeepLab-ASPP

DeepLabv3

- **Multi-grid strategy:** Going deeper with atrous convolution after block3. Keep the stride constant but with large receptive field

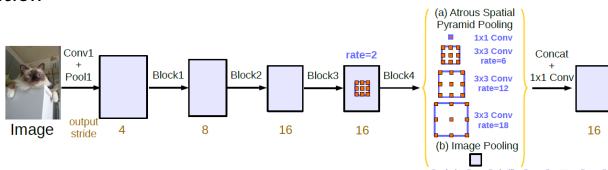


(a) Going deeper without atrous convolution.



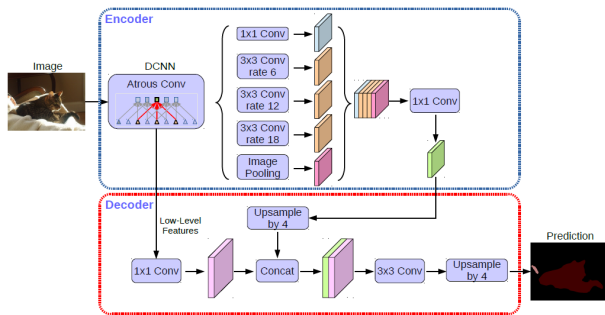
(b) Going deeper with atrous convolution. Atrous convolution with $rate > 1$ is applied after block3 when $output_stride = 16$.

- Use Batch Normalization into ASPP module
- Introduce a global average pooling to integrate the global contextual information



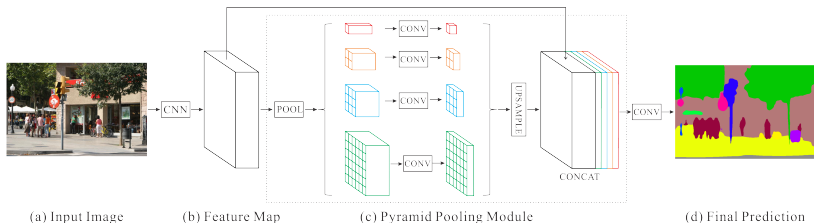
DeepLabv3+: Encoder-decoder Structure

- **Encoder.** Use DeepLabv3 as the encoder
- **Decoder.** The encoder features are first bilinearly upsampled by a factor of 4 and then concatenated with the corresponding low-level features
- There is 1×1 convolution on the low-level features before concatenation to reduce the number of channels
- After the concatenation, applying a few 3×3 convolutions to refine the features followed by another simple bilinear upsampling by a factor of 4
- Much better comparing bilinearly upsampling 16x directly



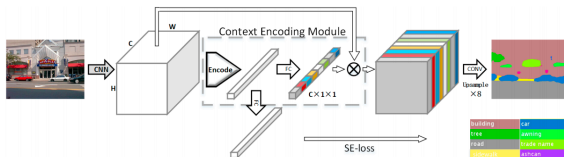
PSPNet: Pyramid Scene Parsing Network

- First use CNN to get the feature map of the last convolutional layer
- A pyramid parsing module is applied to harvest different sub-region representations, followed by upsampling and concatenation layers to form the final feature representation
- Our pyramid pooling module is a four-level one with kernel sizes and stride of 1×1 , 2×2 , 3×3 and 6×6 average pooling
- The final representation carries both local and global context information
- The representation is fed into a convolution layer to get the final per-pixel prediction

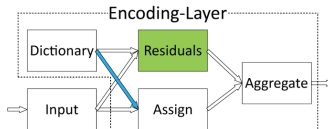


EncNet: Context Encoding for Semantic Segmentation

- Segmentation is always about capture contextual information for predicting pixel-wise labels
- EncNet proposed to encode global context for improving the classification performance

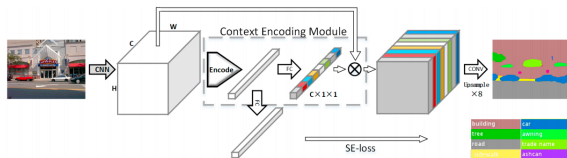


- **Encoder layer** aims at capturing the global context. It learns an inherent dictionary. Residuals are pairwise differences between visual features of the input and the dictionary codewords. Weights are assigned based on pairwise distance between descriptors and codewords. Finally, the residual vectors are aggregated with the assigned weights

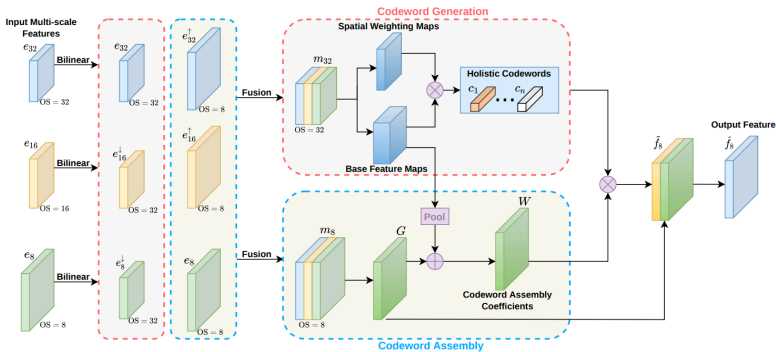


EncNet: Context Encoding for Semantic Segmentation

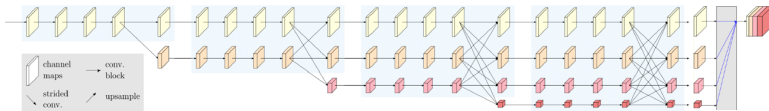
- Channel attention (re-weighting).** The encoded global context features is processed by a fully connected layer and a softmax function along the channel dimension to obtain a C -dimensional weighting vector, which sums up to 1
- It is element-wisely and spatially multiplied with each spatial pixel of the feature maps
- Semantic Encoding Loss.** The encoded semantics is processed by a fully connected layer and a sigmoid function to make individual predictions for the presences of object categories in the scene. It is trained with binary cross entropy loss



More advanced networks: EfficientFCN and HRNet



Efficient FCN (Liu et al. ECCV'20)



HRNet (Sun et al. CVPR'19)

References

- <https://www.jeremyjordan.me/semantic-segmentation/>
- Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." CVPR 2015.
- Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional networks for biomedical image segmentation." MICCAI 2015.
- Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature Pyramid Networks for Object Detection." CVPR 2017
- Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440. 2015.
- Zhao, Hengshuang, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. "Pyramid scene parsing network." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881-2890. 2017.
- Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).
- Zhang, Hang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Amrbrish Tyagi, and Amit Agrawal. "Context encoding for semantic segmentation." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 7151-7160. 2018.

References

- Liu, Jianbo, Junjun He, Jiawei Zhang, Jimmy S. Ren, and Hongsheng Li. "EfficientFCN: Holistically-guided Decoding for Semantic Segmentation." In European Conference on Computer Vision, pp. 1-17.
- Wang, Jingdong, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu et al. "Deep high-resolution representation learning for visual recognition." IEEE transactions on pattern analysis and machine intelligence (2020).
- Chen, Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." IEEE transactions on pattern analysis and machine intelligence 40, no. 4 (2017): 834-848