Question:Explain the difference between the task of classification and segmentation, explain why there might be conflicts between the two tasks.

Answer:

- The difference between classification and segmentation is: segmentation is dense prediction and classification is not dense prediction. Segmentation needs to label every pixel in an image and classification just needs to identify the overall category of the whole image.
- There might be conflicts because classification focuses on global information and segmentation relies on local information. Classification is coarse inference and segmentation is detailed pixel-level inference. Semantic segmentation faces an inherent tension between semantics and location: global information resolves what while local information resolves where.

Question: Introduce how FCN addresses the conflicts. Then introduce different versions of FCN, and explain how they balance the trade-off.

Answer:

- FCN re-architect and fine-tuned classification nets to direct, dense prediction of semantic segmentation. It converts fully connected layers to convolutional layers and allows the networks to handle inputs of any size so that they can produce spatially detailed outputs for segmentation. Besides, FCNs incorporate in-network upsampling and skip connections to refine the spatial precision of the output.
- FCN-32s: It predictes semantic labels at a stride of 32 pixels. Its output is relatively coarse compared to other FCN models.
- FCN-16s: It has the skip connections that incorporate features from an intermediate layer and it predicts semantic labels at a stride of 16 pixels.
- FCN-8s: It includes earlier features in the network and achieves segmentation at an 8-pixel stride.
- FCN-32s actually favors computational efficiency by working with coarser outputs. That makes it faster and less consumption but it would lose detail. FCN-16s try to contain details as well as maintain a reasonable computational load. It has improvement on spatial accuracy and detail of the segmentation. FCN-8s have the highest spatial accuracy and have an increased computational complexity.

Question: Compare the evaluation metrics of pixel accuracy and IU introduced in the paper. Also compare mean IU and frequency-weighted IU.

Answer:

- Pixel accuracy is the ratio of correctly predicted pixels to the total number of pixels in the image and IU is the ratio of the intersection and the union of the predicted and ground truth pixels. Pixel accuracy focuses on the overall accuracy across the entire image and it doesn't consider the interaction between different classes. It makes pixel accuracy highly sensitive to label imbalance. If there is a class dominated in the image, the pixel accuracy may be high just by predicting the dominated class. It makes pixel accuracy not that reliable when the label is imbalanced. At the same time, intersection over union is calculated by each class individually so it wouldn't be affected that much as the pixel accuracy.
- The mean IU is the average IU calculated as the ratio of the intersection and the union of the predicted and ground truth pixels, averaged over all classes. Frequency-weighted IU is

weighted by the frequency of each class and calculated as the sum of the IU for each class multiplied by the proportion of pixels of that class in the dataset, normalized by the total number of pixels. The main difference is that the mean IU is an unweighted average and treats each class equally and frequency-weighted IU adjusts the importance of the class according to its presence in the dataset.

Question: Comment on the limitations of FCN and potential rough directions for further improvements.

Answer:

- The limitation of FCN still can be the limited usage of the contextual information. It uses skip connection to try to combine different levels' features but this structure may still can't do very well in capturing and leveraging long-range dependencies and complex contextual information. Besides, there is still loss of resolution. FCNs using upsampling layers to recover resolution lost during the pooling process. However, it might not fully recover the fine details and the spatial information lost during the convolution and pooling may be challenging to restore sometime.
- I think the rough directions for this is trying to enhance contextual understanding and improve the upsampling techniques. Besides, it's still a good direction to try to balance spatial accuracy and the consumption.