Evaluating Semantic Scene Segmentation Models

A Comparative Analysis of GAN, FCNN, and Watershed Approaches on ADE20K Dataset

Chenlin Li (z5527155)
Yuji Mao (z5522942)
Haitao Wu (z5516805)
Junwei Zhang (z5448642)
Zhongtao Du (z5504303)

Introduction

Semantic understanding of visual scenes is one of the biggest enigmas in computer vision. Scene parsing, which is the process of recognizing and distinguishing objects and other features in an image, plays a significant role in scene comprehension. For this study, our objective is to evaluate the GAN model as the best option for semantic segmentation. To ensure the generalizability and robustness of our model, we used ADE20K, a large-scale image dataset that allows us to evaluate and validate our results, to compare the outcomes of non-deep learning techniques, simple deep convolutional neural network (ConvNet) approaches, and GAN.

Literature Review

Why ADE20K

ADE20K is a dataset with densely annotated images, which means every pixel has a semantic label and the labels are diverse and unrestricted [1]. The images in this dataset are meticulously segmented manually, encompassing a wide range of object components, scenes, and categories. Compared to other datasets, ADE20K not only has a large set of classes with condensed and consistent annotations labeled in diverse scenes but also contains independently segmented object instances with associated object parts [1]. According to these outstanding features of ADE20K dataset, it facilitates advancement in highly detailed and contextually accurate scene parsing when dealing with image content removal and scene synthesis.

Other datasets have been addressed because of their limited number of objects in sparely annotated images, and some objects can not be commonly encountered in real-world situations. Moreover, others have a narrow range of object classes and predefined vocabulary, with some of them having noisy labels at object labels, which increases the difficulty of training [1]. As a solution of these challenges noted, ADE20K includes high-resolution imagery with comprehensive annotations and a wide range of scenes. Therefore, scene parsing and instance segmentation as two significant benchmarks is established with ADE20K on pixel-wise scene interpretation, which requires models to have strong capability for better performance in various scenes [2].

Method

To understand the reason for GAN being the best approach for semantic segmentation, we conducted a comparison of the outcomes of non-deep learning techniques, simple deep convolutional neural network (ConvNet) approaches, and GAN on ADE20K. We chose mean Intersection over Union (mIoU) as the metric of comparison, which is calculated by the overlapping region between predicted and ground-trued segmentation divided by the area of union between the two [3]. Given the extensive scale of the ADE20K dataset with over 25,000 annotated images, we opted to train our model using the 150 most frequently annotated classes, ensuring efficient training while maintaining a representative sample of common objects and scene elements.

Our initial approach leverages the watershed transform, a straightforward and mathematically intuitive morphological method for scene segmentation. While watershed segmentation effectively distinguishes overlapping or adjacent objects by treating the image as a topographic surface, it has significant limitations: it tends to produce over-segmented regions, is highly sensitive to noise, struggles to detect low-contrast boundaries, and has difficulty identifying thin structures. Although methods such as marker imposition, prior information integration, and geodesic reconstruction have been explored to mitigate these issues, watershed segmentation remains challenged by noise sensitivity (i.e. reflection of water) and the inability to reliably capture subtle boundaries and details [3].

After the watershed segmentation sheds light on separate touching or overlapping objects that treat the image as a topographic surface, we are now shifting the focus to simple deep-learning techniques to improve the performance. We attempted to use Convolutional Networks (CNN) to make up for the defects of watershed transform. Our fully convolutional model's ability to predict dense outputs from arbitrary-sized inputs, makes it seem to be an ideal model for semantic segmentation especially for this dataset [4]. With an image of input size h * w, we transformed it into a three-dimensional array, multiplying d as the color channel dimension. Replacing fully connected layers with convolutional layers and training them end-to-end allows the network to process images of a wide range of sizes directly from pixels to segmentation outputs using back-propagation [2] With this expectation, we tried to combine layers of the feature ordering and modifying spatial prediction of the output by reducing pooling layer strides and fine-tuning all layers in the whole net for optimization. Although our CNN model eliminates the drawbacks of the watershed model as its over-segmentation and poor noise resilience, it lacks the ability to capture complex distributions and retain details in output quality.

In order to further improve the model's accuracy, we shifted the focus on Generative Adversarial Network model (abbreviated as GAN) after exploring the traditional thresholding and machining learning. This deep learning model consists of two neural networks, the generator and the discriminator, which are trained in a process of competition. In our model, we utilized a standard 5-layer CNN as the generator and introduced a 5-layer CNN as the discriminator, along with a refined training loop. The generator's objective is to create realistic data samples based on random noise, while the discriminator's job is to distinguish between real data samples from the training set and fabricated data samples generated from the generator. This setup forms an adversarial game where the generator tries to "fool" the discriminator, and the discriminator tries to correctly classify real from generated samples. In the training loop, we calculated the adversarial loss in GAN using binary cross-entropy and the generator's pixel loss with cross-entropy loss, while evaluating the final image generation quality using mIoU. Under the constant confrontation of generator and discriminator, the model's performance exceeds that of a single CNN.

Result

Watershed

In our initial approach, the model was highly sensitive to noise and had limited resilience to low-contrast edges. To improve our preset thresholding model's performance, we attempted to apply Gaussian blur to the current image (kernel size is set to 5X5) to reduce noise interference. To further extract the boundary features, a gradient calculation of the image was carried out and we further converted the gradient map to the grayscale map along with a binary operation of threshold 65 to generate the initial binary map of the advancing region [3]. We have tuned the parameter and threshold of the foreground region after the distance exchange (taking 0.6 times the maximum value of the distance transform) to determine the true foreground. Similarly, an expansion operation is iterated 3 times to determine the background. With these modified steps, we have stabled mean IoU between 0.09 and 0.15, based on the image chosen for testing.

CNN

Our FCN model performs slightly better than watershed with a mean IoU about 0.25. From the image generated, we observed that the predicted mask appears to capture the general structure of the input while it fails to capture details accurately. For instance, the windows and the bed have blurry or non-distinct boundaries, indicating that the model may not be handling edge precision well. These limitations may lead to a lack of sharpness in the segmentation, possible causes could be insufficient training data, incompetent model architecture complexity or the deficit emphasis on specific class features of loss function used for training.

GAN

After fine-tuning the parameters and adjusting the model structure, the mean IoU of GAN yielded around 0.33. Initially, our model had lower mean IoU compared to watershed and CNN, which was only around 0.1. We tackled the issue by adding a 0.3 dropout rate in the discriminator. This approach mitigates overfitting by incorporating dropout, along with adjustments to the generator's learning rate to prevent the discriminator from becoming overly specific and dominant in evaluating the generator's outputs.

Discussion

Reflecting on our research, we aimed to understand why the GAN model outperformed the FCNN model for semantic segmentation. The primary strength of GANs lies in their adversarial training mechanism. Unlike FCNNs, the generator in a GAN is trained to produce outputs that are indistinguishable from real data, driven by adversarial loss. This process encourages the generator to produce more realistic and consistent segmentation boundaries, thereby improving overall quality. GANs also excel at refining boundary details, as the discriminator enforces high fidelity around object edges and complex structures, resulting in sharper and more accurate segmentation.

Our experiments showed that the GAN model achieved a higher mean Intersection over Union (mIoU) compared to both watershed and FCNN methods, demonstrating the improved ability of GANs to handle complex segmentation tasks. By integrating dropout layers and adjusting the learning rates, we managed to overcome overfitting and balance the adversarial dynamics, which led to enhanced model performance.

Despite these successes, there were still challenges to overcome. The GAN occasionally struggled with distinguishing similar object boundaries, leading to visual inaccuracies. This may have been influenced by the disparity in learning capability between the generator and discriminator, with the discriminator initially dominating, making it hard for the generator to converge effectively. In addition, our computational limitations forced us to use reduced dataset sizes, which may have restricted our model's ability to fully optimize. Nonetheless, the GAN approach significantly improved segmentation quality, confirming its effectiveness over traditional CNN methods for complex scene parsing.

Conclusion

In this study, we investigated various approaches to scene parsing and semantic segmentation using the ADE20K dataset. Our goal was to compare non-deep learning techniques, traditional convolutional neural networks, and Generative Adversarial Networks (GANs) to determine the most effective method for semantic segmentation.

Our initial traditional approach of watershed transformation demonstrated significant limitations regarding poor handling of low-contrast edges and low noise sensitivity. To overcome these limitations and produce more consistent segmentation results, we trained the convolutional network (FCNN) to capture the general structure of the image with clear high-contrast boundaries. While it has improved results upon the watershed approach, FCNN struggles with boundary precision and complex feature extraction, which leads to further exploration on the deep learning model GAN.

Our findings suggested that GAN delivers better outcomes regarding semantic segmentation, especially for images and objects with sharp boundaries. Nonetheless, issues like efficient training and computational constraints need to be resolved. Further research could be conducted on optimizing model architecture with a balanced training dynamic. Explorations on more advanced segmentation with more extensive dataset like attention mechanisms or transformer-based models may also lead to enhanced results for challenging scene parsing tasks.

References:

[1] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene Parsing through ADE20K Dataset,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 633-641.

[2] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic Understanding of Scenes through the ADE20K Dataset," *arXiv preprint* arXiv:1608.05442, 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1608.05442

[3] J.-B. Kim and H.-J. Kim, "Multiresolution-based watersheds for efficient image segmentation," *Pattern Recognition Letters*, vol. 24, no. 1-3, pp. 473-488, 2003. [Online]. Available: https://doi.org/10.1016/S0167-8655(02)00270-2

[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 (CVPR)*, Boston, MA, USA, 2015, pp. 3431-3440. [Online]. Available: https://arxiv.org/abs/1411.4038