SCENE SEGMENTATION USING UNSUPERVISED LEARNING

Rohan Shukla, Saumya Hetalbhai Mehta, Meghana Boinpally

ABSTRACT

The project aims to investigate the use of Unsupervised learning techniques for the problem of image segmentation. We investigate the use of both K means and convolutional neural networks, and both as an ensemble. We aim to improve the existing K means techniques and apply a CNN-based approach for improving feature extraction and clustering functions so that pixels with comparable features, spatially continuous pixels be allocated to the same label, and the number of unique clusters can be increase.

Index Terms— Unsupervised Learning, KMeans, CNN, OpenCV

1. INTRODUCTION

Image processing is a critical part in many recent machine learning technologies. Segmentation is the process of extracting or identification of distinguishable regions in an image and partitioning it. This is performed based on the properties of pixels such a color, proximity, intensity and so on. Segmentation using supervised learning works on a dataset of trained features here. Any fault in segmentation will led to inaccurate extraction of features which results in wrong prediction of the decision support systems. Hence, our project will focus on an investigation of various latest unsupervised image segmentation techniques performed in the field of deep learning.

1.1. Background

For decades, image segmentation has been a focus of computer vision research. Object detection, text recognition, image compression and more recently image captioning are all applications of image segmentation. Image Segmentation has been performed in many different fields and ecosystems. Medicine and Biology are one of the most crucial as well as diverse field, where image segmentation is used to determine pathological lung segmentation, detecting brain tumors and other important problems.

1.2. Motivation

Our project aims to answer the following research questions:

- Can unsupervised convolutional neural networks learn enough structure from data to generate good quality segments?
- Is spatial continuity important to generate good quality clusters?
- Can we improve results from CNN and GMMs using K-means?

The rest of this paper is organized as follows. We first review related works in Section 2. The architecture of our network is described in Section 3, description of our data in Section 4, and experimental methods are demonstrated in Section 5. Section 6 discusses the evaluation metrics, and the experimental results are demonstrated in Section 7. The conclusions are drawn in Section 8 and the future scope is presented in section 9.

2. RELATED WORKS

Kanezaki et al. [1] studied the use of convolutional neural networks (CNNs) for unsupervised image segmentation. Similar to supervised image segmentation, the proposed CNN architecture assigns labels to pixels that identify the cluster to which the pixel belongs. Once a target image is loaded, the pixel labels and feature representations are optimized concurrently, and their parameters are updated using gradient descent. To identify a reasonable label assignment solution. the suggested approach minimizes the combination of similarity loss and spatial continuity loss.

Unsupervised segmentation is performed by Xia et al.[5] by estimating segmentation from an input image and then recovering the input image using the estimated segmentation. As a result, identical pixels are assigned to the same label, even when the boundary of each segment is not estimated.

Croitoru et al. [6] proposed an unsupervised segmentation in the context of segmenting the main foreground objects in single images. Their method uses deep neural network techniques to conduct binary foreground/background segmentation and is based on deep neural network techniques. The suggested unsupervised learning system contains two paths, one for the teacher and one for the learner. The system is built to learn from multiple generations of teachers and students. In every generation, the teacher does unsupervised object discovery and then transfers the objects to the student pathway for training. Multiple students are trained using different deep network architectures at each generation to ensure greater diversity. At one iteration, students assist in the training of a better selection module, resulting in a more powerful teacher pathway at the following iteration.

3. TECHNICAL DESCRIPTION

We aim to develop a end-to-end network of unsupervised image segmentation for differentiable clustering. A CNN-based algorithm to optimize feature extraction functions and clustering functions so that pixels of similar features can be assigned to same label, spatially continuous pixels be assigned to same label and making the number of unique clusters large**.** Develop a spatial discontinuity loss function that mitigates the limitations of fixed segment boundaries. The we aim to test the effectiveness of the proposed approach on several benchmark datasets of image segmentation

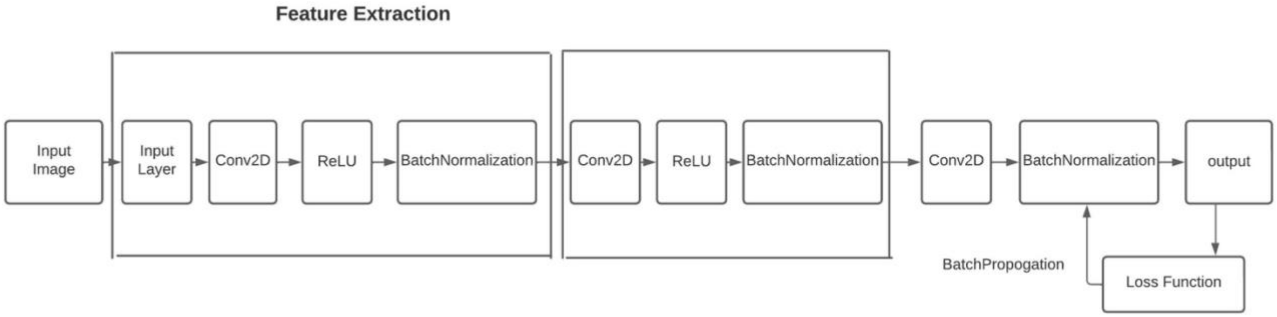


Figure 1: Model Architecture

4. DESCRIPTION OF DATA

We identified the following dataset for our problem statement

BSD500: The Berkeley Segmentation Dataset has about 12,000 hand-labeled segmentations of 1,000 dataset images. The dataset consists of all the grayscale and color segmentations for the images. The images are divided into a training set images, and a testing set images. They also have a human generated figure-ground labeling for a subset of these images

.

5. METHODS

5.1 K-Means Clustering

We implemented the K-means clustering algorithm for unsupervised image classification. Since we are dealing with colored images that have 3 dimensions based on RGB, we don’t treat them the same way as conventional data points. We convert this into 2-dimensional data. Then we convert it into float values as k-means algorithm only takes that as an input in OpenCV. We convert the image into data frame before passing it to the model.

We observed with an increase in the value of K, the image captures more minute details because the K-means algorithm can classify more classes or clusters of colors.

We use Expectation-Maximization algorithm to generate soft clusters for image segmentation and stop when the difference between previous and current likelihood is below a threshold. The results from GMM are comparable to the results from CNN when the prior distribution was set correctly. We used K-Means algorithm to generate a prior distribution of means. The optimial value of K is decided using elbow method. This prior distribution is then used to calculate the initial covariance matrix and priors. We then calculate the hard beliefs from soft clusters by using Maximum APosteriori (MAP) inference. The segmented image is generated by using these hard beliefs.

5.2 Convolutional Neural Networks (CNN)

Model architecture follows the method proposed by Kanezaki et al. Convolutional filters for feature extraction and differentiable processes for feature clustering are used in the proposed CNN architecture, which allows for end-to-end network training. Using backpropagation of the suggested loss to the normalized responses of convolutional layers, the proposed CNN provided cluster labels to image pixels and modified the convolutional filters to achieve better cluster separation. The segmented image is generated by using Maximum APosteriori (MAP) inference on the ouput of final convolutional layer.

5.2.1. Network Architecture

The model consists of convolutional layer followed by a chain of convolutional layers followed by a convolutional layer. Figure 1 shows the model architecture. The output of each convolutional layer is batch normalized. The kernel size of the first and hidden convolutional layers is 3x3 , with a stride 1, padding 1 and the last convolutional layer has a 1x1 kernel. The learning rate is initially set to 0.05 and then decayed using a learning rate scheduler.

5.2.2 Loss function

We use a weighted average of discontinuity loss(spatial continuity) and cross entropy loss(feature similarity) asfollows:

$$L = \underbrace{L_{\mathrm{sim}}(\{\boldsymbol{r}'_n, c_n\})}_{\text{feature similarity}} + \mu \underbrace{L_{\mathrm{con}}(\{\boldsymbol{r}'_n\})}_{\text{spatial continuity}},$$

Discontinuity Loss is the loss measure of spatial continuity of pixels within a cluster. This loss helps the CNN model ensure clusters are more spatially continuous.

8

Feature similarity Loss ensures that the pixels with similar feature vectors are grouped in the same cluster.

5.2.3. Experimental setup:

- Used weighted average for cross entropy loss and discontinuity loss, with 0.5 as μ.
- Used Xavier initialization improve the initialization of neural network weighted inputs.

- Used Batch Normalization to normalize outputs.
- Implemented a learning rate scheduler with reduction in learning rate by a factor of 10 every 50 epochs.
- Used Adam optimizer with momentum of 0.9.
- Used L1 loss for spatial discontinuity along y and z direction.

Cluster Number	0	1	2	3	4	5	6	7	8
Best Segment Match	0	0	0	0	0	0	0	1	1
IOU score	0	0.36	0.26	0.003	0.35	0.003	0.003	0.01	0.9

Table 7.1 Cluster – Segment match with highest IOU scores for CNN

	GMM	CNN	KMEANS	CNN+KMEANS	GMM+KMEANS
25098	0.63	0.74	0.81	0.47	0.60
3096	0.78	0.81	0.61	0.37	0.53
35058	0.36	0.5	0.23	0.37	0.45
22090	0.72	0.74	0.33	0.63	0.46
23025	0.69	0.81	0.37	0.64	0.63

Table 7.3 Cluster – Comparative results of five images from BSD300

5.3 Gaussian Mixture Models

Our second proposed method was Gaussian Mixture models. We use Expectation-Maximization algorithm to generate soft clusters for image segmentation and stop when the difference between previous and current likelihood is below a threshold. The results from GMM are comparable to the results from CNN when the prior distribution was set correctly. We used K-Means algorithm to generate a prior distribution of means. The optimal value of K is decided using elbow method. This prior distribution is then used to calculate the initial covariance matrix and priors. We then calculate the hard beliefs from soft clusters by using Maximum APosteriori (MAP) inference. The segmented image is generated by using these hard beliefs.

6. EVALUATION METRICS

We report cluster wise Intersection Over Union (IOU) scores from the image with segments from the ground truth instead of IOU score of the image with the ground truth. We then use the maximum IOU score per cluster which would correspond to the best ground truth match for a given cluster. We also report a mean IOU score over the final maximum IOU scores to get an IOU score for segmented image and the ground truth.

An IoU score is also known as Jaccard index, which was used as the evaluation metric in our segmentation model. The formula for jaccard index was:

$$\mathrm{J(A,B)} = \frac{|A \cap B|}{|A \cup B|}$$

Here,

A = Cluster of the segmented image
B = Cluster of Ground Truth Image

7. RESULTS

In our proposed methodology, a comparative study was performed between the algorithms. The results of the Intersection over Union (IoU) scores on the images of the BSD300 dataset was evaluated.

On the image 3096 of BSD300 dataset, for an in-depth analysis, we perform a cluster wise analysis of the IoU scores, with the ground truth clusters and find the top clusters which contribute for the IoU scores.

In the table 7.1, we can see that the best segment match for the cluster 8 is 0.9, which indicates, that the object segmented is 90% like the actual hand-segmented object in the ground truth.

Similarly, in the table 7.2, we observe through the results that the best segment match between the ground truth and the segmented clusters is 0.36, which is very inferior compared to the results from CNN.

Cluster Number	0	1	2
Best Segment	1	0	0
IOU score	0.59	0.11	0.88

Table 7.2 Cluster – Segment match with highest IOU scores for GMM

Table 7.3 shows us the comparative study for the 5 images selected. Table 7.4 shows us the average values across the images in the BSD300 dataset.

In the table 7.4, the BSD300 dataset IoU scores are seen. It resembles the subset of 5 images whose scores are reported in the table 7.3. We observe that CNN is consistent and superior to other algorithms, followed by GMM. Gaussian

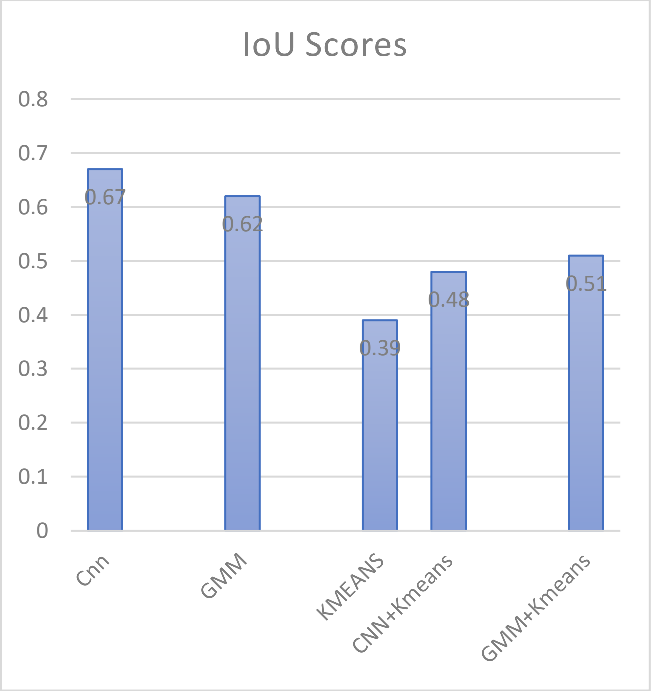
CNN	0.67
GMM	0.62
KMEANS	0.39
CNN+Kmeans	0.48
GMM+Kmeans	0.51

Table 7.4 Cluster – Segment match with highest IOU scores the BSD300 data

Mixture Models are still better, since an IoU score between 0.5 and 0.7 is considered good, but we observe that the traditional K Means fails to give better accuracies.

Here, we also notice that in the ensemble approach, where we first got the segmented image from the CNN, and then used that image into a less intensive K Means approach, we could see that the score of IoU decreased, rather than increasing. The same happened in the Gaussian Mixture Models, where again the IoU scores dropped below the already segmented scores.

The graph below an overall summary of the above table.



In figure 1, we show the results of segmentation of the image, when the loss function of Cross Entropy Loss was used. Here, the clusters are segmented, but not efficiently. We notice that there are clusters within clusters, which should not be the case.

Figure 2 shows the ideal ground truth image for the same image, which was segmented manually by the dataset creators.

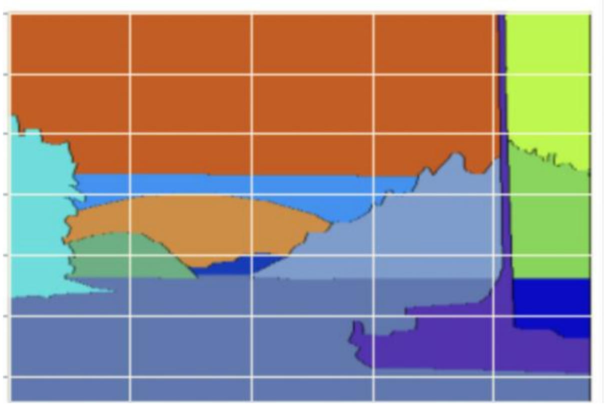


Figure 1: Cross Entropy Loss (22090-BSD300)

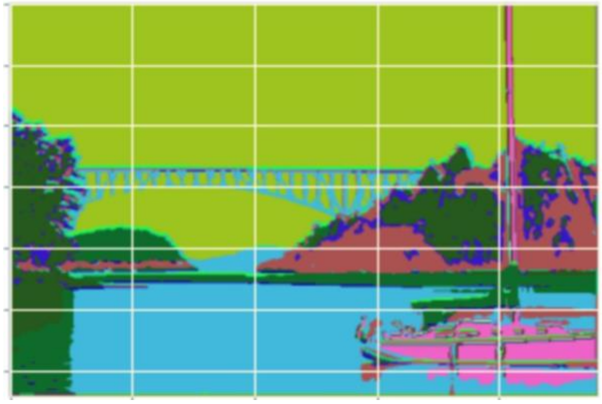


Figure 2: Ground Truth (22090-BSD300)

Thus, to overcome this non-efficient segmentation, we introduce Spatial Discontinuity loss function, whose results are seen in Figure 3. Here, we notice that the segmentation is much better. Although it is not perfect, objects are more distinctly identified. We can see that the boat, mountain and the trees are all identified separately and, rather than identifying more smaller segments within each object, as it was seen with the Cross Entropy Loss function.

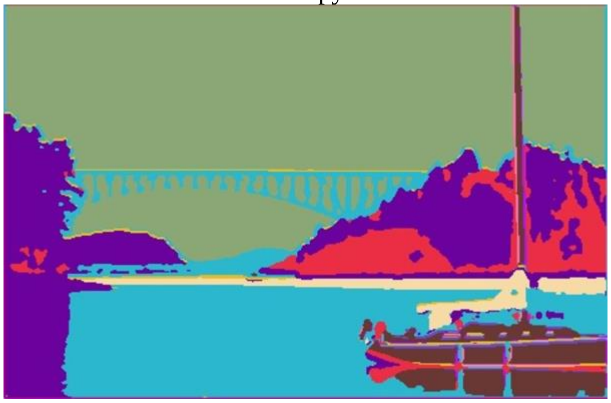
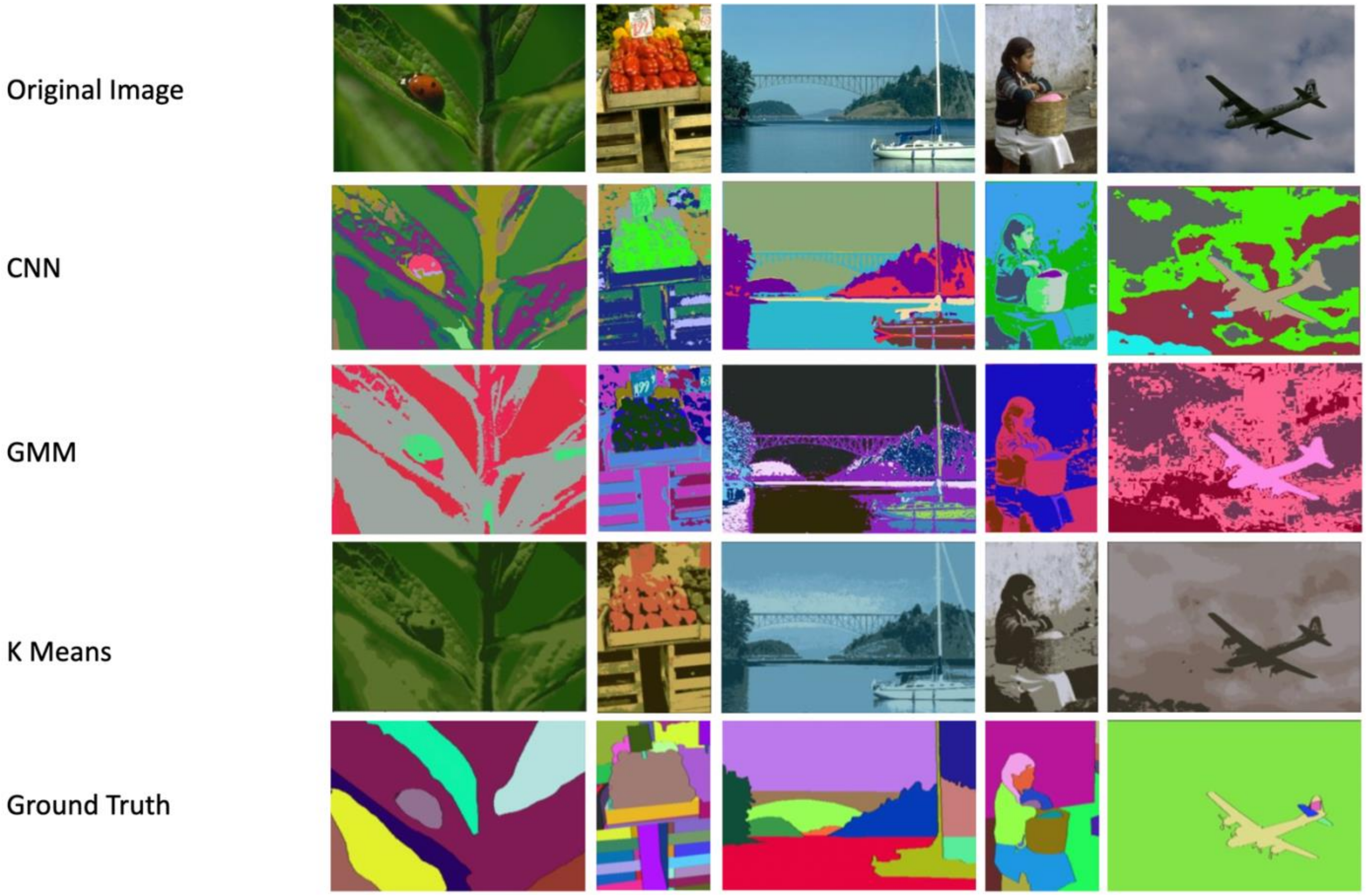


Figure 3: Spatial Discontinuity Loss Function (22090-BSD300)



In Figure 4, the comparative study for the segments is shown, where the original image, the segmented image and the ground truth is shown. We observe that the segmentation result aligns on the naked eye align with the IoU scores, shown in Table 7.3.

8. CONCLUSION

We presented a comparison of CNN architecture and its self-training and Gaussian Mixture models, as well as both as an ensemble process that enables unsupervised image segmentation. Investigated the importance of using Spatial Discontinuity Loss function over Cross Entropy Loss function to achieve better separation of clusters. We also presented a new metric for evaluating the Jaccard index (IoU) based on masked layers of pictures from various segmented clusters. The proposed method's effectiveness was demonstrated by experimental results on the BSDS500 benchmark dataset.

9. FUTURE DIRECTION

We anticipate that our method will be valuable in situations where labeled pixelwise supervision is difficult to achieve, where new data sets may necessitate extensive re-labeling for semantic segmentation algorithms to perform properly. Furthermore, our method could be improved in the future by incorporating various loss functions at different stages in the CNN. Using image enhancement techniques to improve the quality of the image. We also propose methods such as, pixel refining, employing edge density algorithms for segmentation priors for further research**.**

10. REFERENCES

[1] Kim, W., Kanezaki, A., & Tanaka, M. (2020). Unsupervised learning of image segmentation based on differentiable feature clustering. IEEE Transactions on Image Processing, 29, 8055-8068.

[2] Kanezaki, A. (2018, April). Unsupervised image segmentation by backpropagation. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1543-1547). IEEE.

[3] Ouali, Y., Hudelot, C., & Tami, M. (2020, August). Autoregressive unsupervised image segmentation. In European Conference on Computer Vision (pp. 142-158). Springer, Cham.

[4] Liu, Z., Xiang, B., Song, Y., Lu, H., & Liu, Q. (2019). An improved unsupervised image segmentation method based on multi-objective particle swarm optimization clustering algorithm. Comput. Mater. Continua, 58(2), 451-461.

[5] Xia, X., & Kulis, B. (2017). W-net: A deep model for fully unsupervised image segmentation. arXiv preprint arXiv:1711.08506.

[6] Croitoru, I., Bogolin, S. V., & Leordeanu, M. (2019). Unsupervised learning of foreground object segmentation. *International Journal of Computer Vision*, *127*(9), 1279-1302.

[7] Sun, Z. H., Zhou, W. H., & Zhang, W. (2010, August). Vehicle detecting in traffic scenes with introduction of subtractive clustering algorithm. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery* (Vol. 2, pp. 616-619). IEEE.