Fine Grained Segmentation task for Fashion and Apparel

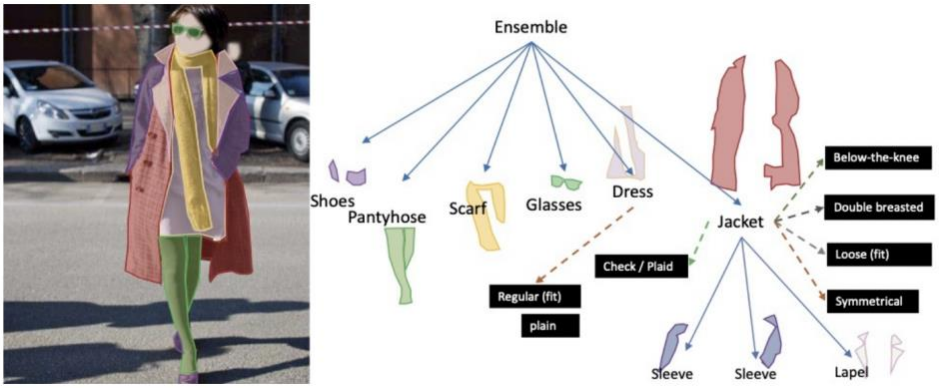
Shubham Mehta
Department of Statistics
University of Illinois, Urbana-Champaign
mehta45@illinois.edu

Anushree Vilas Pimpalkar
Department of Statistics
University of Illinois, Urbana-Champaign
avp4@illinois.edu

Umesh Karamchandani
Department of Statistics
University of Illinois, Urbana-Champaign
umeshk2@illinois.edu

Executive Summary

The goal of the project is to perform the segmentation of fashion apparel that will help with a key step towards automatic product detection to be able to accurately assign segmentations for a fashion image. The proposed task includes various semantic segmentation models, which can identify multiple fashion items in pre-defined categories (for example: top, scarf, dress, shoes). U-Net, SegNet and FCN Resnet50 models were implemented, each with the combination of losses and optimizers. The two losses that were used are sparse cross entropy (SCE) loss and focal loss; two optimizers that were used are SGD and Adam optimizer on 256x256 images. A comparison is made between these models to find the best performing model-loss-optimizer combination. Interestingly, the performance of FCN Resnet50 was the poorest among all with every loss-optimizer combination. SegNet model using SCE loss with SGD optimizer and U-Net model using focal loss with Adam optimizer turned out to be the best performing models for the given problem. Further, similar analysis was performed on 512x512 images for these two best performing models, however, no improvement in the accuracy or IoU was observed.



Abstract *— Fashion Industry is a huge industry globally, that finds its purpose in conception, production and promotion of style based on desire. With a plethora of options and fashion retailers, everyday competition, changing trends and consumer shopping preferences, it has become more difficult for the fashion industry professionals to keep a track of what preference do the consumer have for the fashion items to wear and how do they pair them with other items. There has been increasing attention in recent years for the visual analysis of clothing/fashion. Being able to recognize apparel products and associated attributes from pictures could enhance the shopping experience for consumers and increase work efficiency for fashion professionals. This project aims to perform the segmentation of fashion apparel that will help with a crucial step towards automatic product detection — to accurately assign segmentations for a fashion image, by implementing several Deep Learning Architectures using the iMaterialist dataset consisting of 45,000 images with 46 different clothing and apparel categories.*

Keywords—segmentation, UNet, SegNet, Resnet, IoU

I. INTRODUCTION

The fashion industry is globally valued at 3 trillion dollars, which is 2% of the world's Gross Domestic Product (GDP). It includes many different, smaller and more niche industries and is constantly evolving. With constant change in trends and innovation, the industry faces a lot of challenges to retain and acquire consumers/customers. With an increase in competition, constant consumer shopping preference shifts and quick fashion, it has become more difficult for the fashion industry professionals to keep a track of what preference do the consumer have for the fashion items to wear and how do they pair them with other items.

Apparel segmentation and recognition can significantly help the industry to face these challenges. With the methods of efficiently and inexpensively capturing the information of what type of clothing/fashion goods would people love to wear, and what other items would go together with any clothing, can help improve the business functionality and raise the efficiency of the industry.

In this project, various semantic segmentation models which can identify multiple fashion items in pre-defined categories (for example: top, scarf, dress, shoes) are presented. The report is categorized as follows: Initially, some related work on the topic and motivation are discussed. Then, data exploration, methodology/models, and details of the approach are described. The results from these models are illustrated and their performances are compared with respect to different analysis that were performed. The conclusion is provided in the last section of the report.

II. RELATED WORK

In 2014 a Simultaneous Detection and Segmentation (SDS) [1] task was introduced to detect all instances of an image category and mark the pixels belonging to each instance. They used convolutional neural networks to classify category-independent region proposals (R-CNN), and then introduced a new architecture for Simultaneous Detection and Segmentation. The work depended on a lot of regional proposals, and therefore required high computational resources.

In 2018, Learning semantic Segmentation with Diverse Supervision [2] method was introduced, which was a method to learn CNN-based semantic segmentation models from images with different types of annotations available for several computer vision tasks. Their method can be used with any CNN-based semantic segmentation networks. Their proposed methods used for the evaluation on the PASCAL VOC 2012 dataset and SIFT-flow benchmarks show improvement in the performance of the learned models using the diverse training data.

Context Aggregation Network [3], a dual branch convolutional neural network was proposed in 2021, which required significantly lower computational costs as compared to the state-of-the-art and maintained a good prediction accuracy. They designed a high-resolution branch for effective spatial detailing using the existing dual branch architectures for high-speed semantic segmentation, to seize long-range and local dependencies for accurate semantic segmentation. This was done with low computational overheads.

III. MOTIVATION

Apparel segmentation and recognition can significantly help the Fashion industry to face challenges like constant change in trends and innovation. With the methods of efficiently and inexpensively capturing the information of what

type of clothing/fashion goods would people love to wear, and what other items would go together with any clothing, can help improve the business functionality and raise the efficiency of the industry. The project aims to present various semantic segmentation models which can identify multiple fashion items in pre-defined categories.

IV. Details of the Approach

A. Method

1. **Data Acquisition:** The data was available from the Kaggle competition *'iMaterialist (Fashion) 2020 at FGVC7'*. This competition was sponsored by Google AI, CVDF, Fashionpedia and Hearst Magazine.
2. **Data Exploration and Pre-processing:** To understand the data better, images and their corresponding masks were explored. In addition to that, the class distribution, height, and width of the images were also evaluated. Some preprocessing steps like resizing the images to 256x256, 512x512; normalizing the pixel values; flipping the images and masks from left to right etc. were performed.
3. **Model Selection:** Appropriate models that were developed in the last 7-8 years and require moderate computational resources were selected. SegNet, U-Net and FCN Resnet50 were used for the semantic segmentation of the images.
4. **Model Training:** 12 models with 3 different architectures, 2 different loss function and 2 different optimizers were trained using 256 x 256 images. The top 2 models based on evaluation metrics on the test data were selected out of the 12 and these were trained again with 512x512 sized images with the hope of better efficiency.
5. **Discussion and Conclusions:** Analysis and discussion of results from different models were made.

B. Data Exploration

Fig. 1. Sample Images and True masks

The dataset used is iMaterialist (Fashion) 2020 which is publicly available on Kaggle website. The Data contains images of people wearing a variety of clothing types in a variety of poses and has 50000 clothing images (with both segmentation masks and fine-grained attributes) in daily life, celebrity events, and online shopping are labeled by both domain-experts and crowd workers for fine-grained segmentation. Figure 1 shows a sample of desired random outputs and Figure 2 illustrates the distribution of classes. It can be interpreted that there is a class imbalance where class neckline appears to be in 72.6% of the images and class leg warmer appears only 64 times in the data

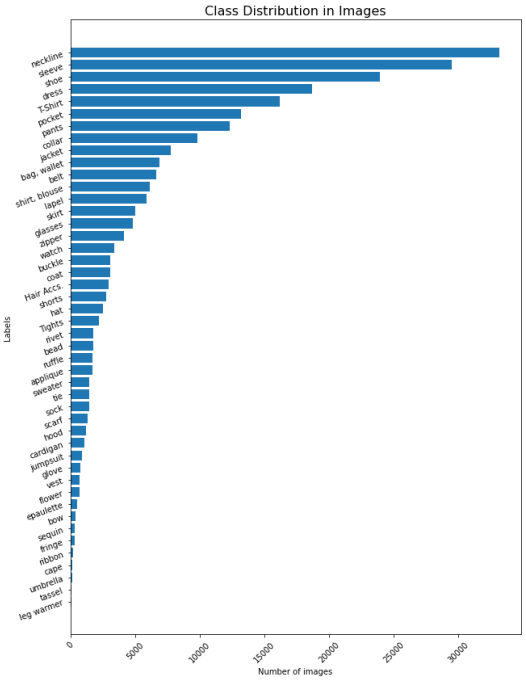


Fig. 2. Category Distribution

C. Models

1. **SegNet:** Segnet [4] uses fully convolution architecture for sematic pixel-wise classification. It has an encoder and coreesponding decoder network, followed by a pixel-wise classification layer. The decoder maps the low resolution feature maps to full input resolution feature maps for pixel wise classification. The decoder uses max unpooling to perform non-linear upsampling. This eliminates the need for learning to upsample.
2. **U-Net:** Unet [5] consists of a contracting path to capture the context and a symmetric expanding path which enables pixel-wise classification. The contracting path mimices the architecture of a convolutional network. It consists of the repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels gets doubled. Upsampling happens in the expanding path, where at everys step 2x2 up-convolution halves the numbe of feature channels.
3. **FCN Resnet50:** Fully Convolutional Networks (FCN) [6] are Convolutional Neural Networks (CNN) that take an input of arbitrary size and produce correspondingly-sized output. This is achieved by replacing the last few layers of ordinary CNN with fully convolutional layers to make an efficient end-to-end learning and inference. FCN ResNet50 uses the same architecture of ResNet50 with the difference that the final classifier layer is removed and all fully connected layers are replaced by convolutions.

D. Evaluation Metrics

1. **Accuracy:** In many classification tasks, accuracy is used as the default mtetric. However, since the data is class imbalanced, accuracy alone is not good enough. This is because a model which only predicts background can have a high accuracy, whereas the results would be completely wrong.
2. **IoU:** Intersection over Union measures the number of pixels that are common between the ground truth mask and predicted mak, divided by the total number of pixels present in both the masks. In this classification problem, binary IoU as used as anaother evaluation metric. Thus, there are only 2 classes, background, and non-background. The model will get a high IOU if it succeed in differentiating a category from the background. However, the model can still get a high IoU even if it predicts a wrong category.

Thus, both accuracy and IoU goes together. Only having high accuracy or IoU alone is not enough. To ensure that model has correctly differentiated a correct class from the background, both the metrics should be high.

E. Loss Function

1. **Sparse Cross Entropy Loss:** This is the most common type of loss for image segmentation tasks and computes the cross-entropy loss between the labels and predictions. As mentioned above since the above dataset is imbalanced, this presents a big problem, and especially in the segmentation tasks it is even harder to solve since the proportion of pixels belonging to the background is much higher than the proportion of pixels that belong to a non-background class.
2. **Focal Loss:** This loss presents a better approach to an unbalanced dataset like the above. It modifies the cross-entropy loss to reduce the impact of correct predictions and focus on incorrect examples. The hyperparameter gamma specifies how powerful this reduction will be. This loss is a robust loss function that aims to give more importance to classes that appear less and therefore makes a better choice of loss function for our problem. The focal loss is defined as:

$$FL(p_t) = -(1-p_t)^{gamma}\log(p_t)$$

F. *Optimizers*

1. Adam Optimizer: It is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. It computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients. [7]

2. SGD: This was the second kind of optimizer that has been tried in the project since as per the literature [8] SGD generalizes better than Adam in Deep Learning.

V. RESULTS

Different models with varied combinations of architecture, loss functions and optimizers were tested on 256 x 256 sized images, totaling 12 models as shown in the flow chart below.

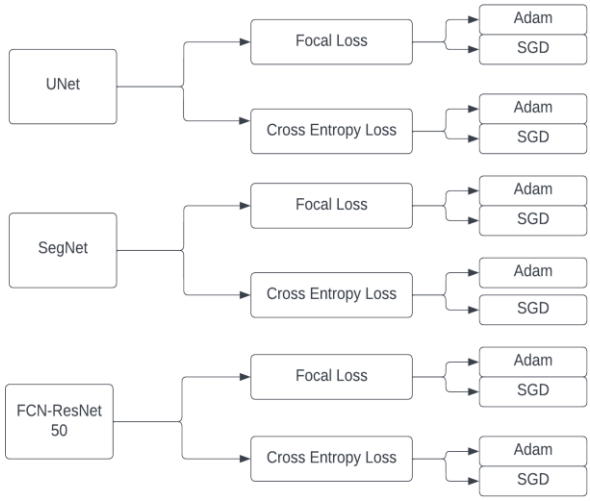


Fig. 3. 12 different architecture, loss and optimizer combinations

Given below are the results obtained by implementing Sparse Cross Entropy (SCE) loss and Focal loss with Adam optimizer and SGD optimizer over U-Net, SegNet and FCN Resnet50 models.

TABLE I. U-NET RESULTS (ACCURACY, IOU)

U-Net	Adam (Lr = 0.001)	SGD (Lr = 0.01)
SCE Loss	0.83 , 0.86	0.79 , 0.83
Focal Loss	**0.84 , 0.86**	0.78 , 0.81

TABLE II. SEGNET RESULTS (ACCURACY, IOU)

SegNet	Adam (Lr = 0.001)	SGD (Lr = 0.01)
SCE Loss	0.82, 0.85	**0.86, 0.88**
Focal Loss	0.80, 0.82	0.86, 0.87

TABLE III. FCN RESNET50 RESULTS (ACCURACY, IOU)

FCN Resnet50	Adam (Lr = 0.001)	SGD (Lr = 0.01)
SCE Loss	0.77, 0.67	0.78 , 0.82
Focal Loss	0.69,0.39	-

Given below are the predicted masks of the resulting model predictions. Going from left to right, they are demonstrated as follows: Original Image, True Masks, Predicted Masks: 1) Sparse Cross Entropy (SCE) loss with Adam optimizer 2) Sparse Cross Entropy (SCE) loss with Adam optimizer 3) Focal loss with Adam optimizer 4) Focal loss with SGD optimizer.

Fig. 4. Predicted masks for U-Net architecture

Fig. 5. Predicted masks for SegNet architecture

Fig. 6. Predicted masks for FCN Resnet50 architecture*

As demonstrated in Table 1, U-Net model with Adam optimizer and focal loss provided the best results across all U-Net models. Although there isn't a significant difference in the performance of the model with SCE loss, Figure 4 suggests better predictions with respect to the model with the Sparse Cross Entropy loss and Adam optimizer. Similarly, Table 2 demonstrates that SegNet model with SGD optimizer and SCE loss outperformed other SegNet models. This is further justified by the visualization of the predictions in Figure 5. Table 3 demonstrates the poor performance of FCN Resnet50 model for our dataset, which is also corroborated by the predicted masks in Figure 6.

Given below are the plots obtained through the implementation of above-mentioned model-loss-optimizer combinations for the validation set for U-Net and SegNet models owing to their better performance when compared to FCN Resnet50 model.

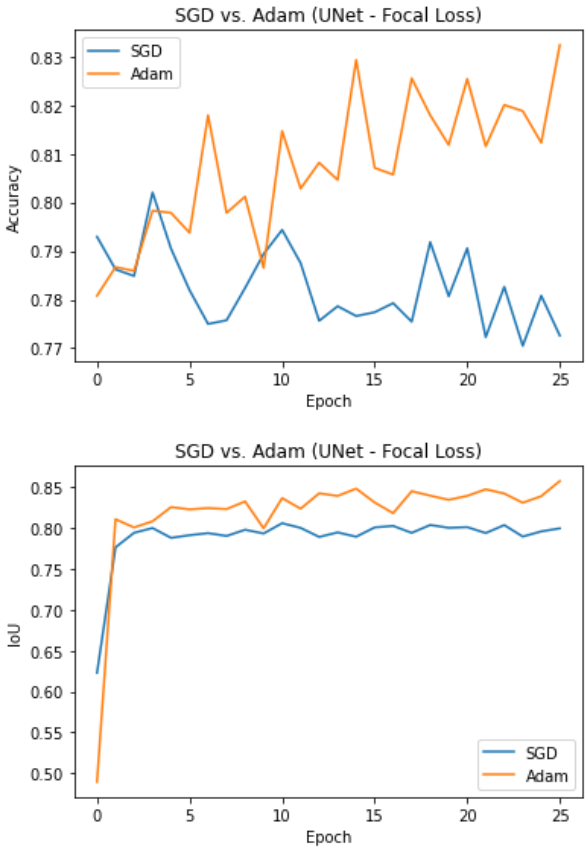


Fig. 7. U-Net: SGD vs Adam optimizer with Focal loss

Figure 7 represents the comparison of SGD and Adam optimizer in U-Net model with Focal loss. It is evident from the graphs that the Adam optimizer resulted in better accuracy and IoU score (although with ups and downs in accuracy throughout the epochs) for UNet, which can also be justified by the results mentioned in Table 1 and predicted masks in Figure 4.

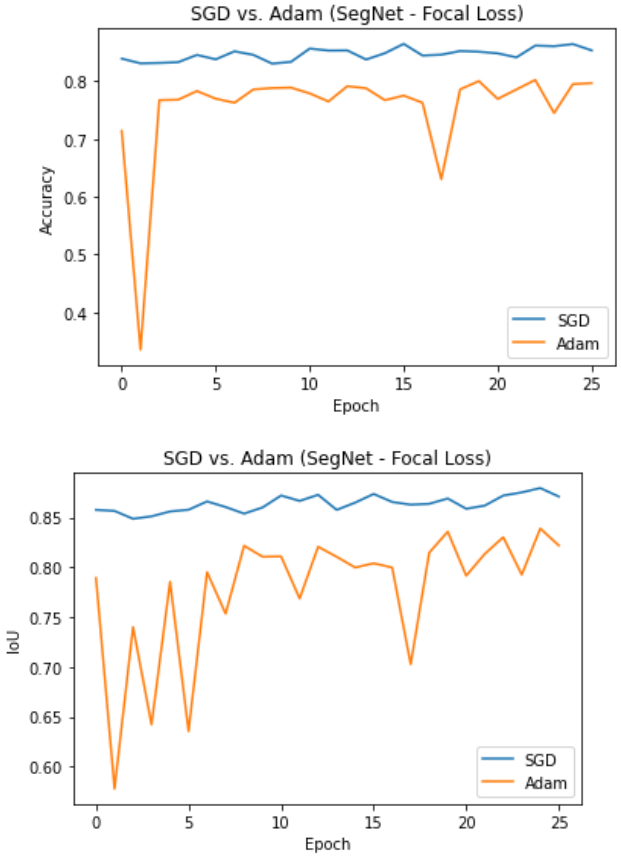
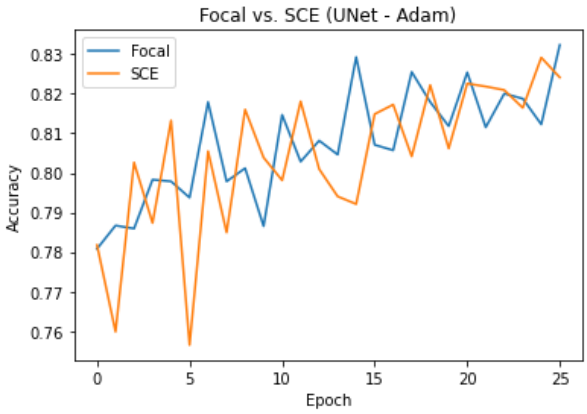


Fig. 8.

Fig. 9. SegNet: SGD vs Adam optimizer with Focal loss

Figure 8 represents the comparison of SGD and Adam optimizer in SegNet model with Focal loss. It is evident from the graphs that the SGD optimizer resulted in better accuracy and IoU score for SegNet, when compared to other combinations for this model. This can also be justified by the results mentioned in Table 2 and predicted masks shown in Figure 5.



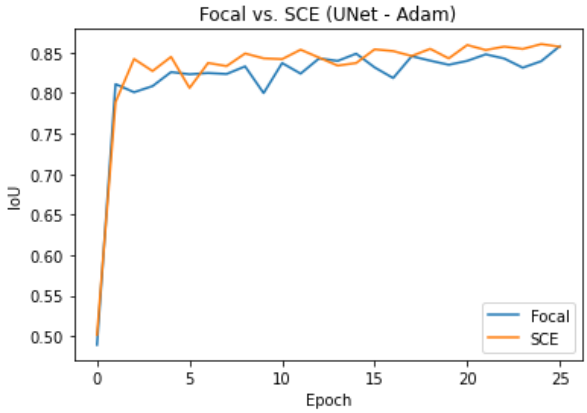
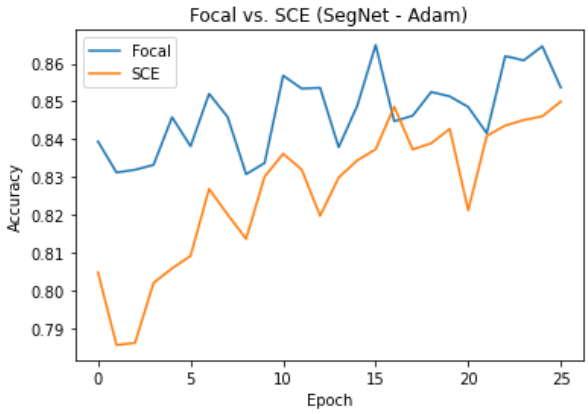


Fig. 10. U-Net: Focal loss vs SCE loss with Adam optimizer



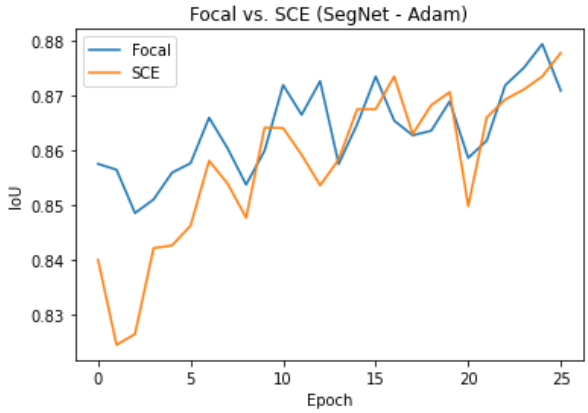


Fig. 11. SegNet: Focal loss vs SCE loss with Adam optimizer

Figure 9 represents the comparison of Focal loss and Sparse Cross Entropy loss implemented in U-Net model with Adam optimizer. It is evident from the graphs that both the losses did not perform in a significantly different way. This can also be justified by the results mentioned in Table 1 and visualization shown in Figure 4, where there is a slight difference in accuracy score and no difference in IoU score for these two settings.

Figure 10 represents the comparison of Focal loss and Sparse Cross Entropy loss implemented in SegNet model with Adam optimizer. It is evident from the graphs that both the losses show increasing accuracy and IoU score trend through the epochs, although Focal loss tends to perform slightly better than the SCE loss over the validation set. However, the performance of SCE loss is slightly better over Focal loss in the testing set as shown in Table 1.

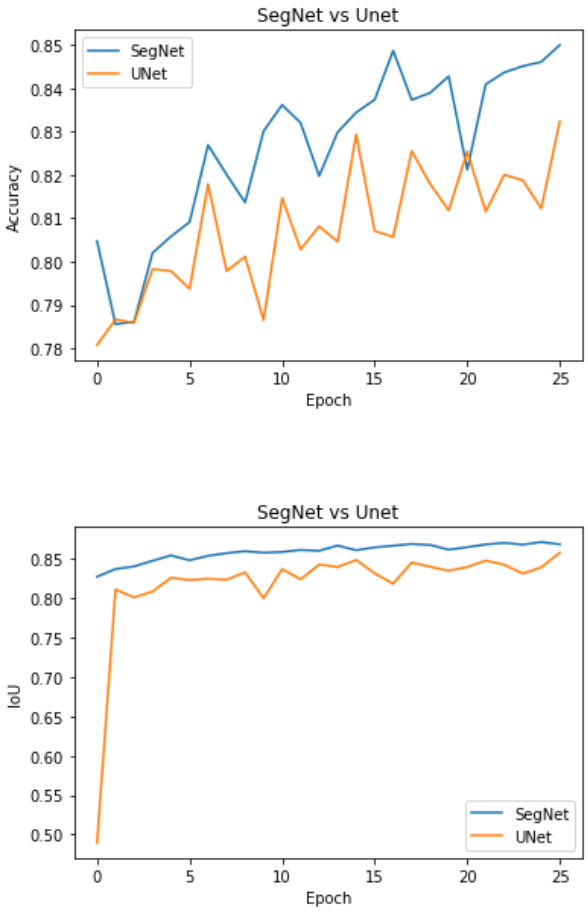


Fig. 12. SegNet (SCE - SGD) vs UNET (Focal – Adam)

Figure 11 represents the comparison of the two best model-loss-optimizer combination, which is SegNet model with Sparse Cross Entropy loss & SGD optimizer and the U-Net model with Focal loss & Adam optimizer. It is clear from the graphs that the SegNet model outperformed the U-Net model for the given combination over the testing dataset. Similar results for the testing set can be seen in Table 1 and Table 2.

The above analysis was performed on 256x256 images, which resulted in the best results for the SegNet-SCE-SGD and U-Net-Focal-Adam model combination. Similar analysis was performed on these two best model combinations over 512x512 images to compare the performances through both the settings. The table below shows the performance of these models on 512x512 images.

TABLE IV. SEGNET & U-NET ON 512X512 IMAGES

Model	Accuracy Score	IoU Score
SegNet-SCE-SGD	0.80	0.82
U-Net-Focal-Adam	0.79	0.81

Comparing Table 4 with Table 1 and Table 2, it is evident that the performance of both the models did not improve with 512x512 images. There is a significant decrease in both the accuracy and IoU score for these models when compared to those on 256x256 images. This could be because both 256x256 images and 512 x 512 images were run for similar number of epochs. Had a greater number of epochs were run for 512 x 512 images, results could have been better than 256 x 256 images; however, that could not be done because of high computational requirements.

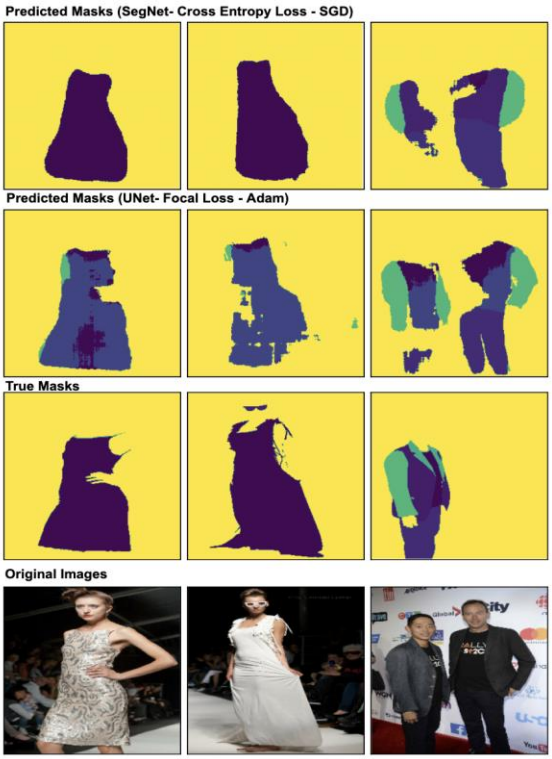


Fig. 13. Sample predicted masks

Figure 12 shows the final predicted masks using the top 2 architecure, loss and optimizer combinations. The predicted masks, although not state of the art, are good. More state-of-the-art architectures like Mask – RCNN could have given better segmentation results, however, it was not used because of the high computational requirements.

VI. Discussion and Conclusions

U-Net, SegNet and FCN Resnet50 models were implemented each with the combination of SCE loss and focal loss, along SGD and Adam optimizer on 256x256 images. When compared, the performance of FCN Resnet50 was the poorest among all with every loss-optimizer combination. SegNet model using SCE loss with SGD optimizer and U-Net model using focal loss with Adam optimizer turned out to be the best performing models for the given problem with the IoU score of 0.88 and 0.86, while the accuracy score obtained with these models were 0.86 and 0.84 respectively. Moreover, when similar analysis was performed on 512x512 images for these two best performing models, the accuracy score and IoU score obtained was (0.8, 0.82) and (0.79, 0.81) respectively. Therefore, suggesting no improvement in their performance.

As part of the future improvements, these models can be tuned to get the right set of hyperparameters. Moreover, a learning rate schedule could be incorporated to optimize the performance of the models. Different models like the state-of-the-art Mask RCNN and loss functions, other than the ones implemented here could be tried to further analyze the performance over the given dataset.

VII. Statement of Individual Contribution

Shubham worked on implementing U-Net model, updated the details of approach and some parts of result section in the report. Anushree implemented SegNet model and updated the cover page, abstract, introduction, related work, motivation, some part of results and conclusion section in the report. Umesh updated details of the approach, implemented FCN ResNet 50 model, and updated some parts of result in the report.

Shubham – 33.33%

Anushree – 33.33%

Umesh – 33.33%

VIII. Refrences

[1] B. Hariharan, P. Arbelez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. Lecture Notes in Computer Science, 2014.

[2] Linwei Ye, Zhi Liu, Yang Wang. Learning Semantic Segmentation with Diverse Supervision. Lecture Notes in Computer Science, 2018.

[3] Michael Ying Yang, Saumya Kumaar, Ye Lyu, Francesco Nex on Real-time Semantic Segmentation with Context Aggregation Network, vol. 178., ISPRS Journal of Photogrammetry and Remote Sensing (2021), pp.124–134.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, PAMI 2017

[5] O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, MICCAI 2015

[6] J. Long, E. Shelhamer, T. Darrell. Fully Convolutional Networks for Semantic Segmentation, CVPR 2015

[7] Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Hoi, Weinan E. Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning arXiv:2010.05627 [cs.LG]

[8] Diederik P. Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization arXiv:1412.6980 [cs.LG]

[9] https://github.com/aurora95/Keras-FCN/blob/master/models.py

https://coderoad.ru/