SIS-CNN: Semantic Image Segmentation Using Convolutional Neural Networks

Muhammad Adeel Ahmed Tahir
School of Computer Science and Engineering
Xian Technological University
Xian, China
E-mail: adikhan0313@gmail.com

Zaryab Shaker
School of Computer Science and Engineering
Xian Technological University
Xian, China
E-mail: zaryabkhan0346@gmail.com

Feng Xiao
School of Computer Science and Engineering
Xian Technological University
Xian, China
E-mail: xffriends@163.com

***Abstract*—Semantic image segmentation is a vast area of interest for computer vision which has gained exceptional attention from the research community. It is the process of classifying each pixel in respective category. In this paper, we exploit the problem of scene understanding and perform the segmentation by combining different classification models as a feature encoder and segmentation models as a feature decoder. All of the experiments were performed on Camvid dataset. It covers a wide range of real-world applications such as autonomous driving, virtual/augmented reality, indoor navigation, etc.**

Keywords-Semantic Segmentation; Computer Vision; Scene Understanding; Classification Model; Segmentation Model

I. INTRODUCTION

Semantic segmentation [1] is the process of alloting class labels to each pixel in an image. Pixel-wise labels provides us better descriptions of images than bounding box labels. Concluding such labels is a much more challenging task because it involves extremely complex structured prediction problem. Semantic image segmentation [1] (pixel-level classification) is an immense area of interest for computer vision, machine learning [2], and deep learning [3] researchers with many challenges. It has a wide array of practical applications like remote sensing, autonomous driving, indoor navigation, video surveillance and virtual or augmented reality systems etc.

Nowadays Deep Learning techniques [4] provide state-of-the-art performance for image segmentation and classification as well as for detection tasks and captioning using Convolutional Neural Network models and have been mainly accelerating the recent breakthroughs in semantic segmentation using different combinations of CNN models such as VGGNet [5], AlexNet [6], and ResNet [7].

VGG[5] is an advanced object-recognition convolutional neural network model that supports up to 19 layers pre-trained on ImageNet [8] (achieves 92.7% accuracy) and performs efficiently on many datasets outside of ImageNet [8]. ResNet [7] is a deep neural network that has 150+ trainable layers. The modal achieves the highest accuracy in the 2015 ImageNet [8] dataset Challenge. U-Net [9] is a Convolutional Neural architecture designed to deal with biomedical images to solve the problem i-e what and where.

In this paper, we proposed a Segmentation Architecture by combining the two models i-e base model [5], [7] with our segmented model [9], [12] for segmentation. We use our base model as an object feature extractor and use the preceeding

segmentation model to segment the images based on extracted features. We use different models with the implementation of an encoder-decoder [14] having skip architecture [10] for segmenting the boundaries accurately.

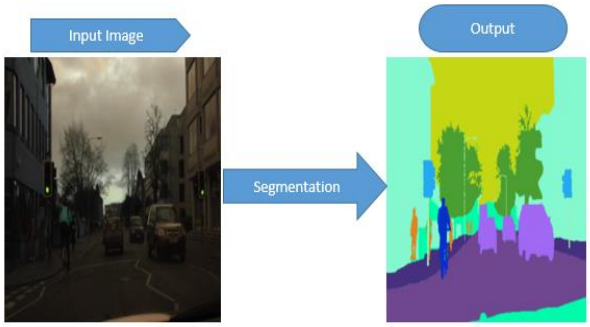


Figure 1. Semantic Image Segmentation

The second part of this paper involves a short survey for segmentation with CNN models. The third part describes the proposed methodology of our framework, the fourth part involves experiments results and graphs. Conclusion is in the fifth part and references are drawn in the last.

II. LITERATURE SURVEY

In recent research of computer vision and pattern recognition, CNN [11] capabilities are highlighted which solve challenging tasks like segmentation [13] and classification [23]. Recent progress in semantic segmentation are mainly enhanced by powerful DNN architectures [9], [12], following by the ideas of FCN's [13]. Different architectures have been developed in this context. Some of the deep learning-based works for semantic segmentation include Fully convolutional networks [13], Encoder-decoder based models [14], Multi-scale and pyramid network-based models [15], Dilated convolutional models [16], and DeepLab family [17], Recurrent neural network-based models [18], Attention-based models [19], etc. All of these approaches have in common that they generally rely on the powerful feature extraction provided by CNN's [5], [6], [7]. Following is a brief study of some of our concerned techniques.

In 2014, Long and Shelhamer et al. [13] presented the novel approach of FCNs for semantic segmentation. The approach represented the state-of-the-art in semantic segmentation and has since set the standard for future directions. FCNs [13] are trained end-to-end, provide a pixel-to-pixel prediction. They also use skip architectures [10] to combine semantic and appearance information. The authors have demonstrated 62.2% mean pixel (IU) on the PASCAL VOC 2011 dataset [24].

The work by Long and Shelhamer et al. [13] builds off of the concept of CNNs pioneered by Matan et al. [20], and the concept of jets pioneered by Koenderink and Van Doorn [21]. In 1991, Matan et al [20]. Used CNNs for recognizing an unconstrained handwritten multi-digit string. They presented a feed-forward network architecture. This is an addition to the work on recognizing isolated digits. In 1987, Koenderink and Van Doorn [21] used local jets to give rich representations of local geometry and semantics with filters on multiple scales. Since the work of Long and Shelhamer et al. [13], several other methods have been explored to improve the performance of semantic segmentation. [1]

In 2017, Chen and Papandreou [17] incorporated probabilistic graphical models in the form of fully Conditional Random Fields (CRF) to overcome poor localization. They proposed "DeepLab" system by applying the 'atrous convolution' with upsampled filters trained on image classification to the task of semantic segmentation for dense feature extraction and further extend it to atrous spatial pyramid pooling. They also combine ideas from DCNNs [22] and FCRFs [23] to produce semantically precise predictions and comprehensive segmentation maps. The proposed technique significantly advances the state-of-art in several challenging datasets, including PASCAL VOC 2012 [24] semantic image segmentation benchmark, PASCALContext [25], and Cityscapes [25] dataset.

Later, Zheng and Jayasumana [26] showed that unpacking dense CRFs into individual computations and joining them to the network yields further improvement. They combine the strengths of CNNs and CRFs [26] in a single deep network. Their formulation fully integrates CRF-based probabilistic graphical modeling with emerging deep learning techniques that are

capable of passing on error differentials from its outputs to inputs during back-propagation-based training of the deep network while learning CRF [26] parameters. The approach achieves a state-of-the-art on the popular Pascal VOC segmentation benchmark [24].

In 2015, Noh et al [27]. demonstrate a novel semantic segmentation algorithm by learning a deconvolution network that incorporates a learned deconvolution network for even better performance. Since coarse-to-fine structures of an object are reconstructed progressively through a sequence of deconvolution operations, it helps to generate dense and precise object segmentation masks. They further proposed an ensemble approach, which combines the outputs of the proposed algorithm and FCN-based [13] method, and achieved substantially better performance with the help of characteristics of both algorithms.

Losing the context information for images during segmentation was a problem until it was addressed by Yuantao Chen et.al [28] in the paper "improving semantic image segmentation based on feature fusion model". They proposed a feature fusion model with context features layer-by-layer. Firstly, an image pyramid is formed by pre proceesing the original images. Secondly, an image pyramid is inputted into the network structure by the initialization of feature fusion and expanding receptive fields using Atrous Convolutions. Finally, the score map of the feature fusion model had been calculated and sent to the conditional random field for further processing to optimize results. The approach on the PASCAL VOC 2012 and PASCAL Contex [25] t datasets had achieved better IU than the state-of-the-art works. The method has about 6.3% improved to the conventional methods.

III. PROPOSED METHODOLOGY

We started the problem by taking the camvid dataset [25]. Our segmentation task was carried out by a combining two different models. One is used as a base model and the other one is the segmentation model. Our base model is a feature extractor for a given image and pre-trained on the ImageNet [8] dataset. We fine-tuned our base model on our relative dataset and use it as an encoder [14] part for our segmentation task. We use the skip architecture [10] by taking the output from our concern layers.

The second one is our segmentation model which is used as the decoder [14] part for our architecture. It takes the output from certain layers in our base model through skip architecture [10] as its input. Then the segmentation model segments the image based on the features extracted by the base model.

A. Base Model

CNNs shows a state-of-art for image classification and recognition because of its high accuracy. The CNN follows a hierarchical model [29] that works on building a network, like a funnel, and finally gives out a fully-connected layer where all the neurons are connected to eachother and the output is processed. Our base model is used as a feature extractor having Input size of 224x224 pre-trained on ImageNet [8] with 1000 classes. We are taking classification features by removing fully connected nodes and fine-tune the model on specific layers. We transform the fully connected layers into convolution layers to produce a classification heatmap [13].

We get the image at the input layer and then initialize the weights to avoid layer activation outputs from exploding or vanishing during a feed-forward propagation [30]. After weight initialization, all of the weights 'w' multiplied by input 'x' are summed up and add a bias of 1 to allow units to learn an appropriate threshold. (1).

$$y = [x_1 w_1 + x_2 w_2 + \ldots\ldots + x_n w_n] + b$$

$$y = \sum x.w + \beta \qquad (1)$$

We add zero paddings and apply 3x3 kernels with a stride of 2 and apply max-pooling. A trick of 'shift and stitch'[13] in which the values are being max-pulled after doing the shifting and then we stitch the results into the original image. After that there implies a relu activation function 'R'(2).

$$R = \max(0,y)$$

$$\hat{y} = P(\psi) \quad (2)$$

Base model works as an encoder [14] for our architecture. Encoder-Decoder [14] module works as a backbone for semantic segmentation tasks. The encoder extracts features from the input image which is used to produce segmentation output. We get the abstract representations via downsampling. In downsampling, we decrease the number of pixels by getting only the pixels with features. This is done because we are facing memory limits on computer and to reduce processing time. The result of using a pooling layer and creating downsampled or pooled feature maps is a summarized version of the features detected in the input.

B. Skip Connections improve Segmentation Details

We use skip connections [10] in our architecture, skip some layers in the neural network and feed the output of one layer as the input to the other layers instead of just passing to one next layer. By using a skip connection [10], we provide an alternate path for the gradient (with backpropagation). It makes it easier to estimate good weight values for the architecture to obtain better generalization performance. After cascading a set of CNN weights 'w', biase 'b', and non-linear layers to the input 'x', we extract image features 'x_f' from each 'n' layer is defined by

$$x_{fn} = \widehat{y_n} \quad (3)$$

all outputs 'x_{fn}' are passed to the segmentation architecture through Skip Connections[10].

C. Segmentation Model

Image segmentation with CNNs, involves feeding segments of an image as input to the segmentation model [9], [12], which labels the pixels. Our segmentation module consists of different CONV layers that receive the input from different levels of the base model [5], [7]. It involves upsampling of the images via deconvolution (also known as a transposed layer).

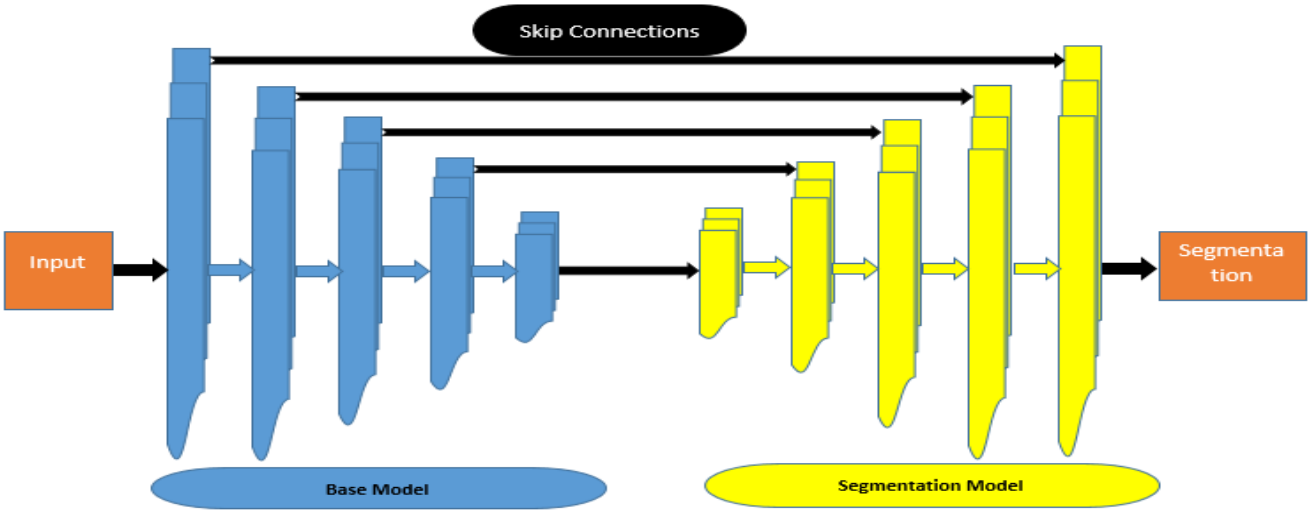


Figure 2. An encoder- decoder based CNN architecture with different combinations of different models for semantic segmentation.

Let 's_n' be the decovolution layer in our segmentation model receiving the inputs from base model. Deconvolution [27] layers need to be stacked very deeply which increases computations and memory allocation. So we use 1x1 conv [31] where the stride is 1 without bias. It gives us faster computation with less information loss by reducing the dimensions of the previous layer and also adds more non-linearity to enhance the potential representation of the network. Input samples 'x_{fn}' are average pooled and passed through the 1x1 conv layer [31]. Applying batch normalization, we regularized our model to avoid the need of dropout. It also reduce the training

epochs and get higher accuracy. This is done before utilization of Relu activation function. So,

$$s_n = \chi o \nu \varpi(x_{fn}) \quad (4)$$

After that there is a concatenation layer 'C' which concatenates all the inputs receives from previous model in a linear form as well as from skip connections [10] which are then concatenated and pass through 1x1 conv layer with batch normalization and relu function 'R'.

$$X = \chi o \nu \varpi \left(\sum s_n\right) \quad (5)$$

It passes the results to output layer 'Z' having a softmax activation function 'SOF'.

$$Z = \Sigma O \Phi(X) \quad (6)$$

Where 'Z' is the segmented image.

IV. EXPERIMENTS AND RESULTS

1) Dataset: We are using the Camvid [25] dataset consists of 701 original images of 360x480p. The images are divided into 3 sets *having* 367 training images, 233 test images and 101 validate image. We make annotations for each image in the original dataset. After that, data augmentation is performed for training set. The images are flipped vertically and horizantly, make 2 more images for each image. So the total number of training images is 1,101 with RGB colors.

2) Models: Pre-trained classification and segmentation models are fine tuned and combined which works as an encoder and decoder part for our architecture. By using transfer learning, we adopt VGG[5] and Resnet[7] with pre-trained weights on ImageNet as our base (encoder) module whereas U-net[9] and PSP-net[12] as our segmentation (decoder) module. We are training our model based on the combination of VGG_U-net, VGG_PSP-net, ResNet_U-net, ResNet_ PSP-net.

3) Training Setup: We are doing traing against loss and accuracy. Our loss is difference between predicted and actual value defined by 'L'

$$\Lambda = (\hat{y} - \psi)$$

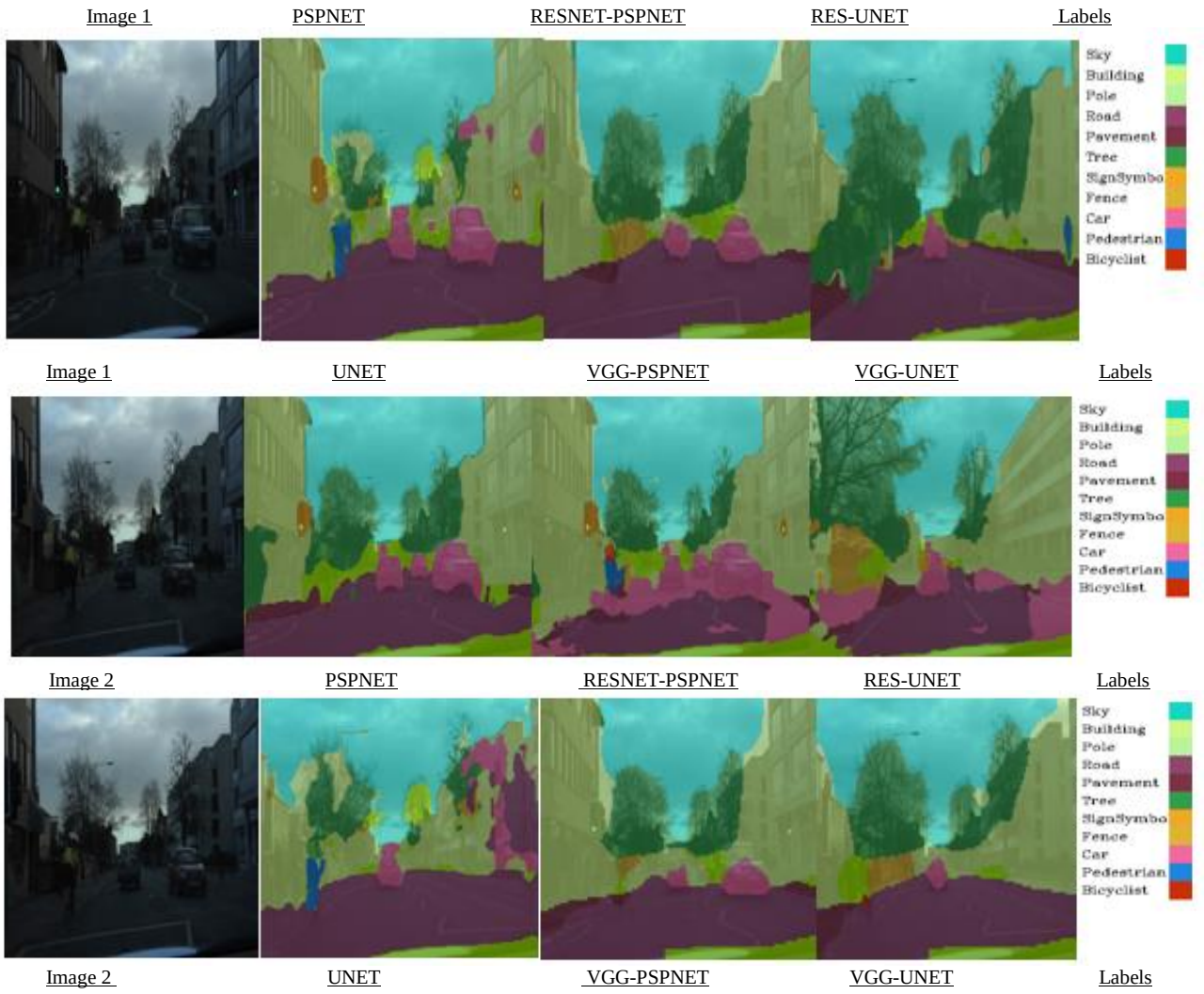
Weights are updated according to the following relation for backpropogation to minimize the loss

$$w_n = w_o - n \frac{\delta L}{\delta w_o}$$

Where 'n' is learning rate, n=$2e^{-4}$, 'w_n' is the new weight and 'w_o' is the old weight. Batch normalization is used to avoid the need of dropout and serves us to regularize the model. We use adam optimizer to minimize the loss value. We do training for only 5 epochs due to limited resouces with 512 steps on each epoch. After every epoch the model will save its learned weights and it will also validate itself by using the given validation set. It takes around 13 seconds per step (6904s per epoch). During training the model, after each epoch the model will evaluate the performance based on the validation set. To check the overall performance of the model we use the test set with newly images for model prediction.

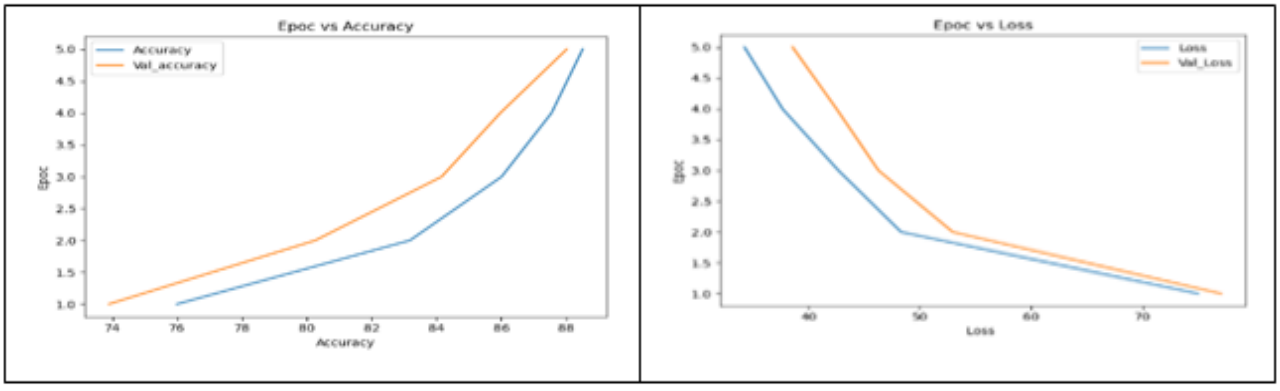
A. RESULTS

Results shown below describes about the model performance. Due to less resources these results were carried out on simple laptop(core i7, 8gb ram).Results are satisfactory as the model was trained only for 5 epochs.

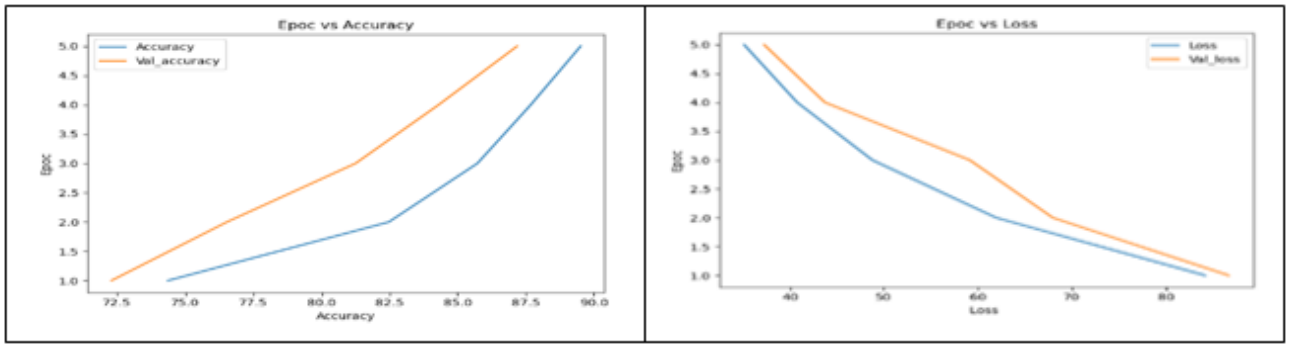


B. *Graphs and Comparisons*

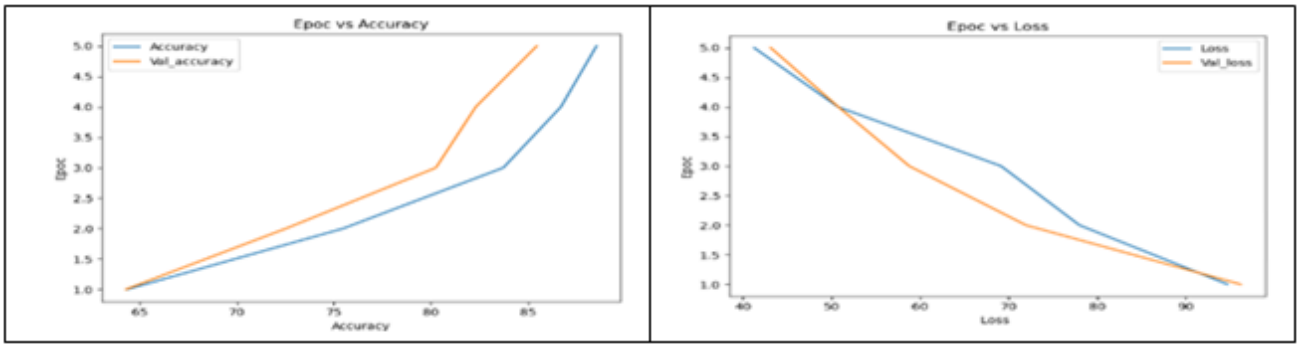
The Loss and Accuracy graphs for each model against each epoch



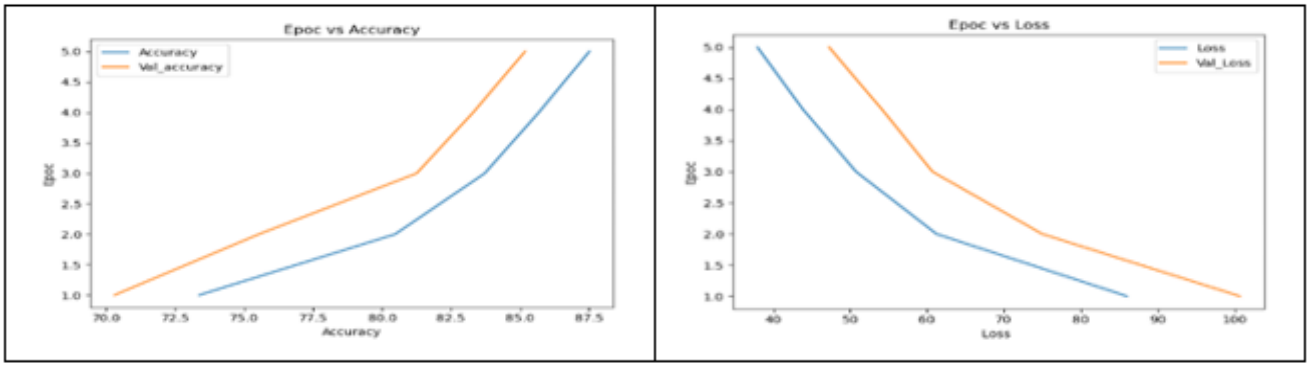
PSPNET



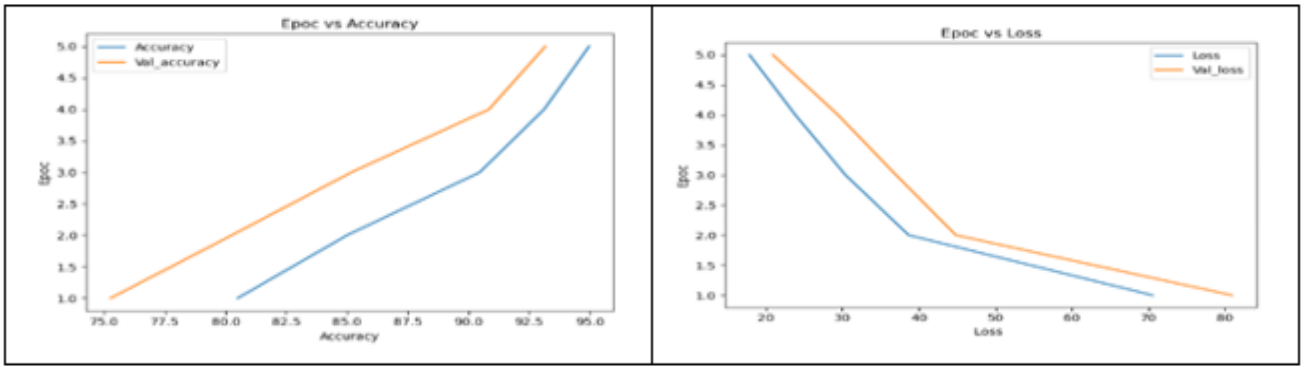
RESNET_UNET



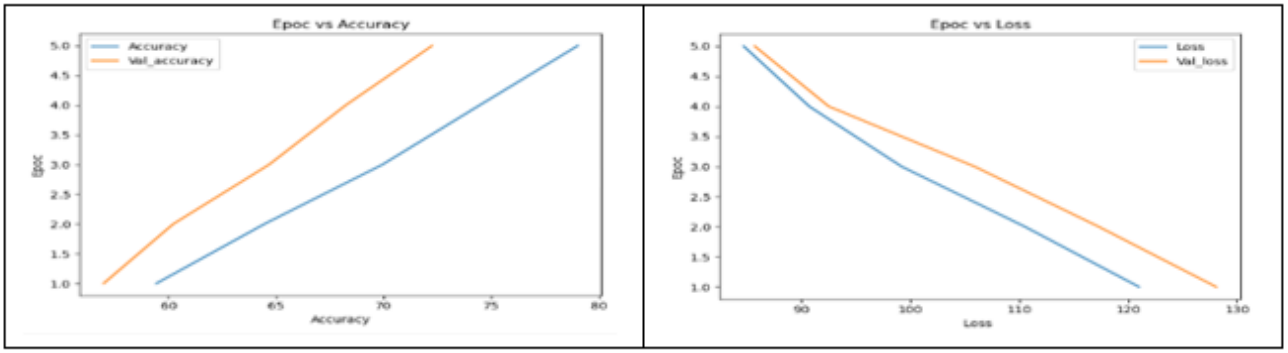
VGG_PSPNET



UNET



RESNET_PSPNET



VGG_UNET

Results and graphs show that when Resnet is combined with pspnet, it gives better results for segmentation in our problem.

V. CONCLUSION

In this work, we have demonstrated the concept of combining different pre-trained classification models and segmentations models for the semantic image segmentation. We developed an end-to-end trainable model that achieved good performance and results on camvid dataset as compared to the level of resources available. We trained our model for only 5 epochs due to limited number of resources. Moreover, if this model is trained on large dataset with a large number of epochs, more accurate and precise results will be achieved that can be used in many real-world applications like autonomous driving.

REFERENCES

[1] P.Wang, P. Chen, Y. Yuan, D. Liu, "Understanding convolution for semantic segmentation," IEEE, 2018.
[2] MI. Jordan, TM. Mitchell," Machine learning: Trends, perspectives, and prospects," Science, 2015.
[3] Y. LeCun, Y Bengio, G.Hinton, "Deep learning," nature, 2015.
[4] A Garcia-Garcia, S Orts-Escolano, S Oprea, "A review on deep. Learning techniques applied to. Semantic segmentation," TPAMI, 2017.
[5] Karen. Simonyan, Andrew. Zisserman, "Very deep convolutional networks for large -scale image recognition," Department of Engineering Science," University of Oxford, 2015.
[6] A Krizhevsky, I Sutskever, GE. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in.neural information, 2012.
[7] Kaiming He, Xiangyu Zhang, S Ren, J Sun, "Deep. Residual. Learning for Image Recognition," Proceedings of the IEEE, 2016.
[8] J. Deng, W. Dong, R. Socher, LJ Li, K Li," Imagenet: A large-scale hierarchical image database," IEEE conference, 2009.
[9] Ronneberger, P Fischer, T Brox, "U-net: Convolutional networks for biomedical image segmentation," International Conference on Medical, 2015.
[10] D. Wu, Y. Wang, ST. Xia, J. Bailey, X. Ma, "Skip connections matter: On the transferability of adversarial examples generated with resnets," ICLR conference paper, 2020.
[11] R. Yamashita, M. Nishio, RKG. Do, K. Togashi, "Convolutional neural networks: an overview and application in radiology," Insights into imaging, 2018.
[12] H. Zhao, J. Shi, X. Qi, X. Wang, "Pyramid scene parsing network," Proceedings of the IEEE, 2017.
[13] J. Long, E. Shelhamer, T. Darrell, "Fully convolutional networks for .semantic segmentation," Proceedings of the IEEE, 2015.
[14] LC. Chen, Y. Zhu, G. Papandreou, "Encoder-decoder with.atrous separable convolution for semantic.image segmentation," Proceedings of the IEEE, 2018.
[15] J. Kang, S. Kim, KM. Lee, "Multi - modal/multi-scale convolutional neural network based in-loop filter design for next generation video codec" IEEE (ICIP), 2017.
[16] G. Lin, Q. Wu, L. Qiu, X. Huang, "Image super-resolution using a dilated convolutional neural network," Neurocomputing, 2018.
[17] LC. Chen, G. Papandreou, I. Kokkinos, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," zzZZ IEEE, 2017.
[18] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, "Recurrent neural network based language model," International Speech Communication Association, 2010.
[19] P. Badjatiya, LJ. Kurisinkel, M. Gupta, "Attention-based neural text segmentation,"ECIR, 2018.
[20] O. Matan, C. J. Burges, Y. LeCun, and J. S. Denker, "Multidigit recognition using a space displacement neural network," NIPS, 1992.

[21] JJ. Koenderink, AJ. Van. Doorn, "Representation of local geometry in the visual system," Biological Cybernetics, 1987.

[22] G. Papandreou, LC. Chen, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," Proceedings of the IEEE, 2015.

[23] B. Zhang, C. Wang, Y. Shen, Y. Liu, "Fully connected conditional random fields for high-resolution remote sensing land use/land cover classification with convolutional neural networks," Remote Sensing, 2018.

[24] M. Everingham, L. Van. Gool, CKI. Williams, "The pascal visual object classes (voc) challenge," International Journal of Computer Vision, 2009.

[25] M. Cordts, M. Omran, S. Ramos, "The cityscapes dataset," Future of Datasets in Vision, 2015.

[26] S. Zheng, S. Jayasumana, "Conditional random fields as recurrent neural networks," Proceedings of the IEEE, 2015.

[27] H. Noh, S. Hong, B. Han, "Learning deconvolution network for semantic segmentation," Proceedings of the IEEE, .2015.

[28] Y. Chen, J. Tao, L. Liu, J. Xiong, R. Xia, J. Xie, "Research of improving semantic image segmentation based on a feature fusion model," Journal of Ambient Intelligence and Humanized Computing, 2020.

[29] D-Roy, P-Panda, K-Roy, "Tree-CNN: a hierarchical deep convolutional neural network for incremental learning," Neural Networks, 2020.

[30] PM-Shakeel, S-Baskar, R-Sampath, "Echocardiography image segmentation using feed forward artificial neural network (FFANN) with fuzzy multi-scale edge detection (FMED)," International Journal of Signal and Imaging Systems Engineering," 2019.

[31] DP. Kingma, P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," NIPS 2018.