Sachit Gaudi
CSE 803 : Computer Vision
Semantic Segmentation

1 Deep Learning

This document is structured as follows: we will discuss different architectural choices, drawing inspiration from successful models like ResNet and VGGNet. Additionally, we will conduct an ablation study on these design options. In Section 2, we delve into the interpretability of the models.

The nature of the ablation study is to change one attribute and measure the impact. However we are not interested to study the impact of optimisation, So we keep the optimiser, scheduler and number of epochs constant throughout the experimentation.

2 Classification

ResNet has achieved success in enabling models to scale and handle increased depth through the utilization of skip connections for gradient propagation. On the other hand, VGGNet initially demonstrated success. In our approach, we incorporate concepts from both architectures. The primary distinction lies in how each architecture down scales to cover a broader receptive field, essential for aggregating information from all areas of image. While ResNet employs stride operators, VGG achieves this through Maxpool operations. For this specific task, we will discuss and evaluate the significance of these design choices in contributing to accuracy, We call the fusion of VGG and ResNet as VGG-ResNet.

Visualization plays a crucial role in discerning whether a model relies on spurious features or features integral to the object's form. Our aim is to gain insight into this visualization process by closely correlating certain digits with fashion classes. This approach allows us to discern whether the model is primarily focused on recognizing the digits themselves as opposed to the overall characteristics of the objects within the classes.

We can study various design choices like swapping the feature block with Transformer module.

We represent the model as $VGG - ResNet(a, b, c)$, where a denotes the block size within each downsize module. In this context, b signifies the number of channels in the initial block, the subsequent scaling by a factor of 2 for every subsequent block, following the ResNet paper's methodology. Given the constraints of a smaller image in later phases, more channels are needed in later blocks to effectively carry the same information forward. The first parameter (a) investigates how the model's performance varies with an increase in depth, while the second parameter (b) examines the impact of increasing width, controlled by the number of channels. The third parameter (c) specifies the downsizing mechanism for the image, achieved either through max-pooling or stride operations, both resulting in a halving of the image size.

Method	Train Accuracy	Test Accuracy
VGG-ResNet(2,32,maxpool)	99.04	92.41
VGG-ResNet(2,32,stride)	-0.28	-1.15
VGG-ResNet(2,32,stride) - skip	-0.31	0.42
VGG-ResNet(2,32,stride) - Batch Norm	-0.80	-0.41
VGG-ResNet(2,32,stride) - activation	-1.40	-1.79
VGG-ResNet(2,32,stride) with GelU	0.30	0.02
VGG-ResNet(1,32,stride)	-0.14	-1.54
VGG-ResNet(4,32,stride)	-0.05	-0.99
VVGG-ResNet(2,16,stride)	-0.19	-1.21
VGG-ResNet(2,64,stride)	+ 0.21	-0.77
VGG-ResNet(1,32,maxpool)	+0.32	-0.52
VGG-ResNet(4,32,maxpool)	-0.01	0.26
VVGG-ResNet(2,16,maxpool)	0.20	0.28
VGG-ResNet(2,64,maxpool)	-0.20	-0.16

Table 1: Ablation of various network choices.

It did not make any difference to change the parameter if the model is deep enough on the simple data the architecture choice will play a Little impact if the data is complex as in the Section 3 the architecture will play a huge difference.

To gain insights into what the model is learning, we will take the base model and freeze its features. Subsequently, we will train it on a surrogate task involving the prediction of class labels on data that combines Fashion-MNIST with MNIST, as illustrated in Figure 1. By examining the resulting attention maps, we aim to discern the model's focus.

Instead of relying on average pooling, which provides a general sense of where the image is located, we will employ a more detailed approach to understand the model's behavior. In this experiment, we will structure the data to establish a robust correlation between a specific class in MNIST (specifically, Digit 3) and a corresponding class in Fashion MNIST (for example, Pants). By doing so, we aim to unveil valuable insights into the interpretability and significance of the features learned by the model. This correlation-driven approach will enable us to examine how the model processes and relates features from these distinct datasets, shedding light on its capacity for meaningful feature extraction.

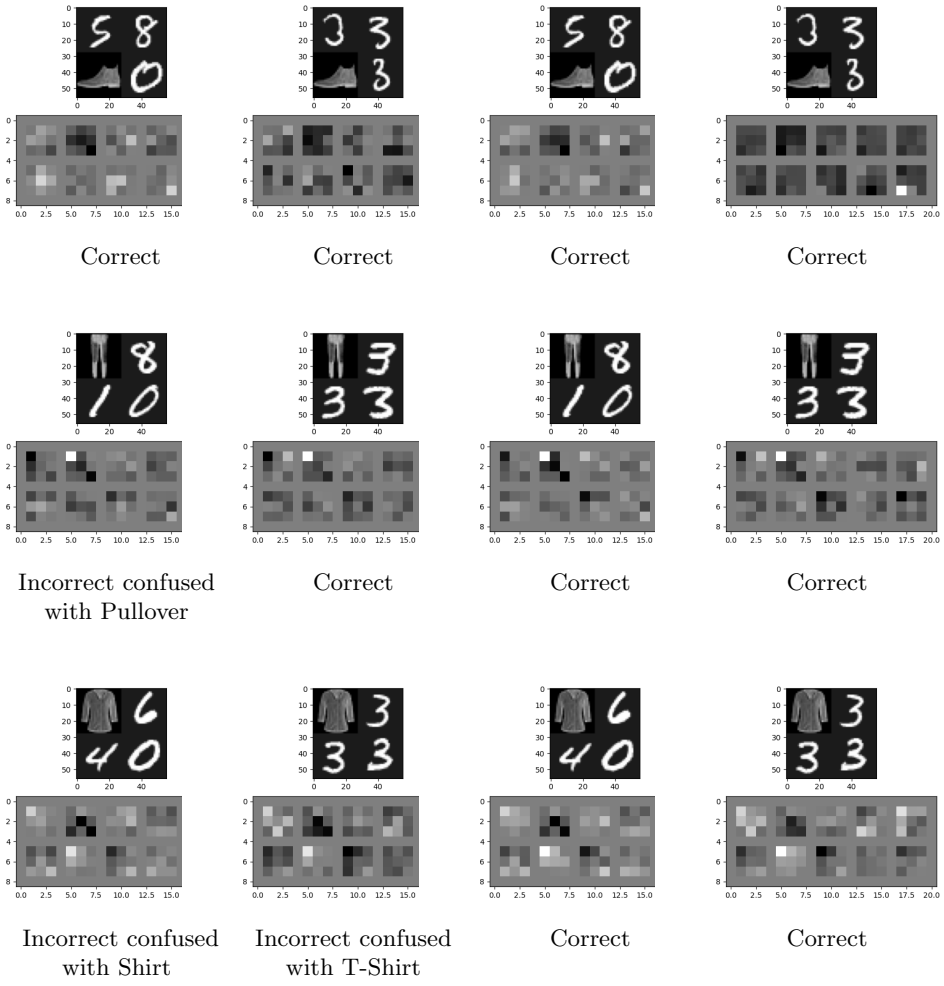


Figure 1: Attention Map: The first 2 columns are the attention maps of the model with finetuning of final layer but the last two columns indicate training the same architecture from scratch and the captions indicate whether the model predicted them correctly or not

The attention map provides a general indication of where the model focuses its attention to make decisions. Upon inspecting the images, we observe that due to the smaller size of the dataset, even with a high correlation with MNIST, the model tends to look in the corner where Fashion-MNIST is located. In the case of incorrect predictions, such as the confusion with a shirt, both the surrogate model and the base model make similar mistakes.

Specifically, for the incorrectly predicted case (row 3), both models exhibit low attention for the non-top row attention maps. However, the base model demonstrates high attention maps for class 0 (the first attention map) and class 6. This observation suggests that, in the misclassification scenario, certain features related to class 0 and class 6 might contribute to the confusion between the predicted class and the actual class, emphasizing the need for further analysis and potential model refinement.

The attention maps warrant careful scrutiny. Despite the attention map accurately capturing information in the (1,0) image, the classifier still faces confusion, misclassifying

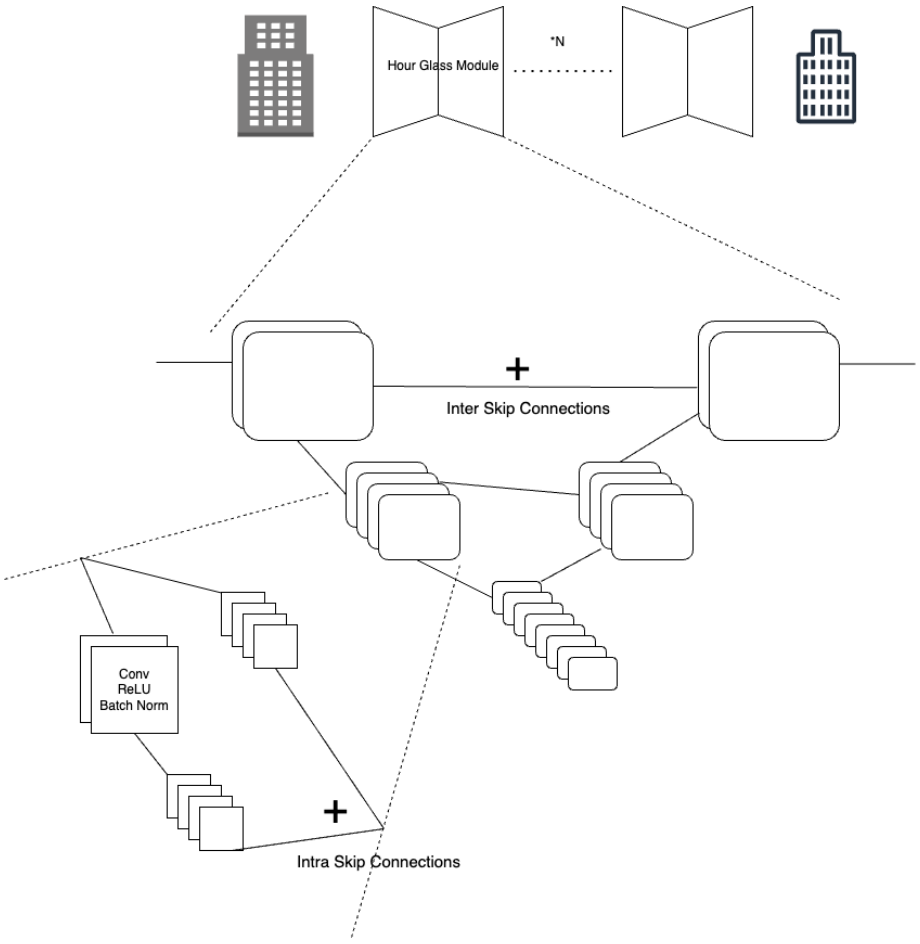


Figure 2: HourGlass(2,2) Architecture, where the first 2 indicates number of hourglass modules stacked together and next 2 indicates the number of hidden layers for the first Block in Hour glass, In the image number of curved rectangles.

it as a pullover. Although the attention maps offer valuable localization information, it's essential to consider that the model's decision-making involves complex interactions among various features.

By conducting two tests—freezing the base model and training from scratch—we can discern whether the base model is inherently struggling with the given samples. If both models exhibit difficulties in utilizing attention information, it may indicate that the sample is challenging or possesses correlations with background information that complicate the classification task. This dual-test approach helps provide a comprehensive understanding of the model's performance on specific instances and contributes to informed decision-making for model improvement.

3 Semantic Segmentation

The significant issue at hand is the imbalance in the data, particularly in the pixel space. This concern is addressed in the paper [2]. In addition to reporting the Mean Average

Precision (MAP), we introduce the Worst Group AP (WGAP) metric. Notably, when training the best model with just cross-entropy, WGAP shows a considerable decline. To mitigate this, we switch to using Weighted Cross-Entropy loss. Exploring various loss functions, including Dice loss, did not yield any improvements in MAP.

This shift in the loss function aims to better handle pixel-specific imbalances, as highlighted in the mentioned paper. By adopting Weighted Cross-Entropy, we strive to enhance the model's performance in scenarios where standard cross-entropy struggles due to data imbalances. The evaluation metrics, MAP and WGAP, provide a comprehensive understanding of the model's performance, especially in addressing the challenges posed by imbalanced data in the pixel space. We will have a +8% gain in the performance with the use of Weighted Cross-Entropy loss.

The model with the best configuration is listed below and the ablation student is performed on various parameters

The switch connection in the hour glass model plays a great deal in improving the MAP score(17%) , but however dude to very less skip connections in one block they don't play any major role, As there is a gap between train and test loss this necessitates the use of dropout of 0.2, as recommended which will bring the train and test models close to each other in accuracy which will increase the test accuracy by (5%)

To benchmark our model, we compare it with the Unet model. We achieve significant test gains. The use of concatenation instead of the sum of skip connections results in more parameters and potential overfitting, as evidenced by the drop in test accuracy compared to training accuracy.

The models are highly sensitive to BatchNorm because of the large weights with varying magnitudes. This leads to a shift in the distribution of neuron activations, and BatchNorm is crucial for correcting this to prevent issues such as exploding gradients.

Expanding the model width by increasing the number of channels by 1% resulted in better gains compared to making the model deeper by adding more Hourglasses, which yielded a decrease of 0.9%.

The standard model is defined as an Hourglass(2,32) with an initial layer of 32-dimensional hidden layers. It comprises 2 blocks of Hourglasses, and ReLU is employed as the activation function. Skip connections are incorporated within both the Upscale and Downscale blocks, as well as between the corresponding Upscale and Downscale blocks. It's noteworthy that the HourGlass(2*32) model, trained with Weighted Cross Entropy, is regarded as the standard model in this context.

Although there exist various hyperparameters to fine-tune, the comparison is conducted with standard semantic segmentation methods. Due to constraints in resources and time, the models are trained for 40 epochs using the ADAM optimizer, with a learning rate of 1e-3 and weight decay of 1e-4.

From performing various experiments on Table 2,

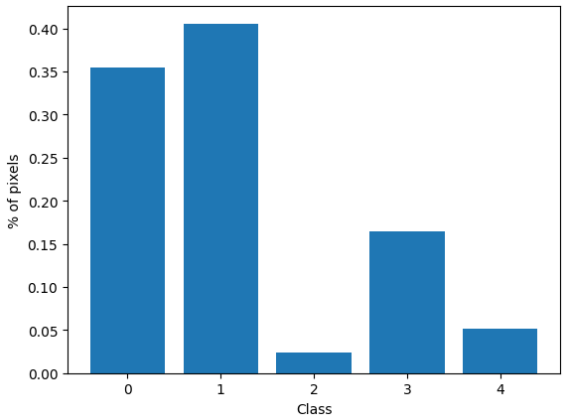
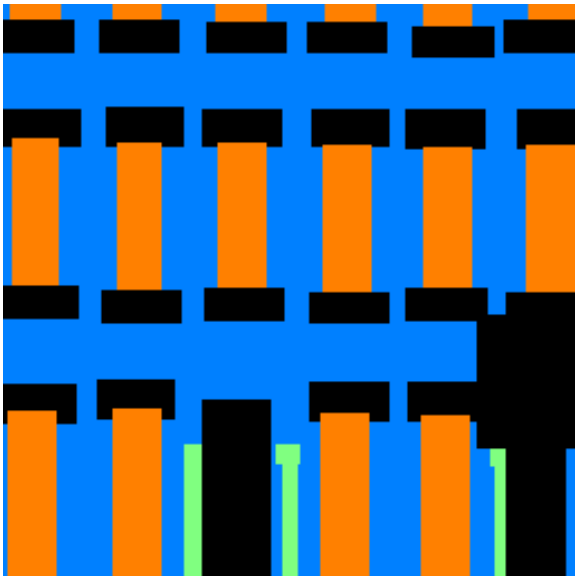


Figure 3: Class Imbalance

Method	Train MAP	Test MAP	WGAP
Standard Model	89.5	73.9	46.8
Unet (Benchmark) [1]	+2.08	-4.88	-5.37
Standard Model - intra skip	+1.11	+1.55	+0.63
Standard Model - inter skip	-12.3	-4.00	-17.00
Standard Model - Batch Norm	-9.2	-7.8	-16.50
Standard Model - activation	-50.97	-33.22	-37.5
Standard Model with Cross Entropy	+2.68	-0.61	-8.78
Standard Model with DICE Loss	-70.13	-54.59	-45.3
HourGlass(1*32)	+3.25	-0.40	-4.79
HourGlass(4*32)	- 2.80	-0.90	-6.8
HourGlass(2*16)	-3.27	-2.50	-5.80
HourGlass(2*64)	+4.30	+1.00	+1.00

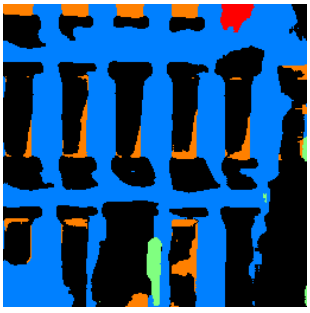
Table 2: Ablation of various network choices.

Example Unseen Image

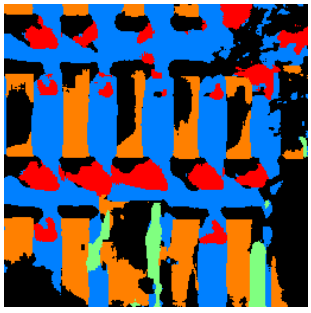


Ground Truth

Standard Model Prediction)



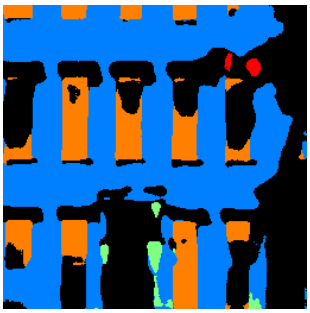
HourGlass(2*64)



Standard Model - Batch Norm

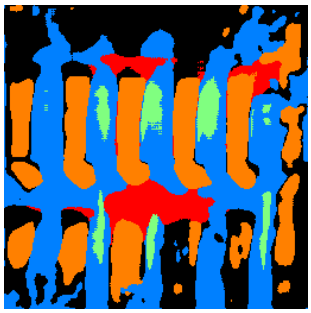
Standard Model with Cross Entropy

HourGlass(2*16)

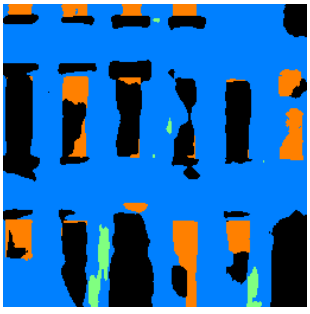


HourGlass(1*32)

Standard Model - Inter Skip

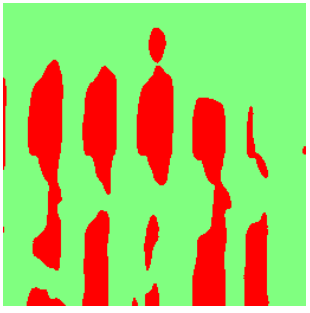


Standard Model - activation



Unet)

Standard Model - Intra Skip



Standard Model with DICE loss

Figure 4: Model Prediciton

4 Conclusion

For straightforward datasets like MNIST, model architecture doesn't significantly impact performance. Deeper and wider models work well, and the choice of non-linearity is inconsequential. Removing non-linearity doesn't affect performance due to the compensatory nature of the max-pooling step..

We can prove the theory that convolutions are translation invariant. If you move Fashion-MNIST image around the grid the performance of the frozen layer does not change, As there is a change in the size of image, one epoch finetuning of final layer is required.

Attention maps provide information on the localization of features within an image, aiding us in understanding the classes with which the model may be confused.

The spurious correlation did not impact the Fashion-MNIST classification, as Fashion-MNIST contains much richer edge information compared to MNIST.

Non-linearity is crucial in handling complex tasks, and architectural choices play a significant role. Batch normalization contributes to noteworthy performance boosts. The use of weighted variants for loss functions enhances worst group precision. Skip connections facilitate the smooth transfer of information from the encoder to the decoder. The inclusion of skip connections between encoder layers results in substantial performance gains. Notably, wider models demonstrate better improvement compared to deeper ones.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[2] Thanh-Dat Truong, Ngan Le, Bhiksha Raj, Jackson Cothren, and Khoa Luu. Fredom: Fairness domain adaptation approach to semantic scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19988–19997, 2023.