HourGlass vs U-Net: A Comparative Study in Image Segmentation

Ahmad El Acham and El Hassan Hajbi

December 1, 2023

1 Introduction

In the field of computer vision, image segmentation plays a crucial role in various applications such as medical imaging, autonomous vehicles, and object recognition. In this report, we aim to explore and compare two popular convolutional neural network architectures for image segmentation: HourGlass and U-Net. The primary goal is to implement both architectures, analyze their convergence behavior, and evaluate their performance on the tictoc dance sequences dataset.

2 Experiment Setting

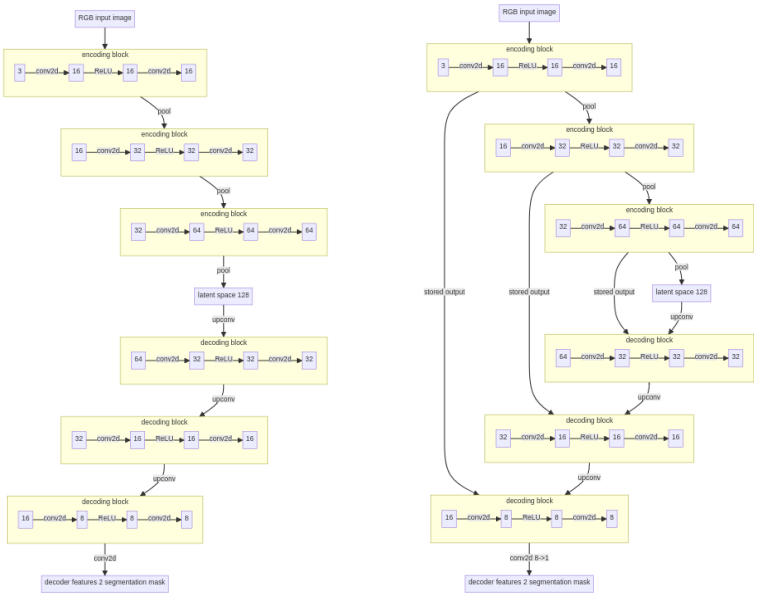
2.1 Dataset

The dataset used for this experiment is derived from the tictoc dance sequences. It includes images and corresponding masks stored on Ensimag machines in the `/matieres/5MMVORF/04-dataset` folder. The training code is provided in the `train.py` file.

2.2 Architectures

1. **HourGlass Network**: The HourGlass architecture is characterized by its unique design, creating a bottleneck in the network to reduce the number of features for encoding the abstract space of people segmentations. The encoder consists of three blocks with a channel configuration of (3, 16, 32, 64), utilizing pooling layers to progressively reduce the feature map size. The decoder loop employs `ConvTranspose2D` layers to upscale features, following the reverse channel configuration of (64, 32, 16, 8). This design allows the HourGlass Network to capture hierarchical features and spatial information effectively.

2. **U-Net**: U-Net distinguishes itself by its use of skip connections, enabling the direct concatenation of feature maps from the encoder to the corresponding layers in the decoder. The encoder follows a channel configuration of (3, 16, 32, 64), similar to the HourGlass architecture. However, the decoder adopts a reverse channel configuration of (64, 32, 16, 8). The unique aspect of U-Net lies in its ability to preserve fine-grained details by incorporating information from earlier encoding stages directly into the decoding process. This mechanism enhances the segmentation performance, especially in tasks requiring precise localization.



(a) Hourglass architecture (b) U-net architecture

Figure 1: Architectures of Hourglass and U-net

3 Part 1: HourGlass Convolutional Neural Network

3.1 Experiment and Results

1. **Convergence Observation**: The HourGlass network exhibits a reasonable convergence behavior. Training it multiple times may reveal variations in convergence speed, which will be documented.

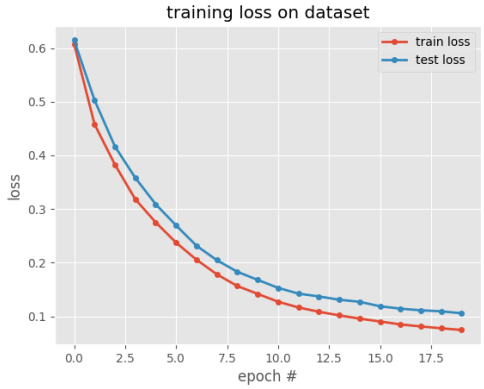


Figure 2: Convergence of Hourglass with 3 blocks

2. **Impact of Varying the Depth:** This study investigates the effects of varying the depth in an hourglass architecture on convergence and performance. Three depth configurations were explored: Configuration with 1 block : $3 \rightarrow 16 \rightarrow latentspace(128) \rightarrow 16 \rightarrow 8 \rightarrow 1$ representing one block of encoding and decoding, $3 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow latentspace(128) \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow 1$ representing four block of encoding and decoding, and$3 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow latentspace(128) \rightarrow 256128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow 1$ representing five block of encoding and decoding. Trained over 20 epochs with consistent hyperparameters. The results reveal that:
   (a) Configuration with 1 block : $3 \rightarrow 16 \rightarrow latentspace(128) \rightarrow 16 \rightarrow 8 \rightarrow 1$: exhibited faster convergence and lower loss than our initial hourglass.
   (b) Configuration with 4 block : $3 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow latentspace(128) \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow 1$:
      i. Converged at 20 epochs but with a reduced rate compared to the shallower architecture.
      ii. Resulted in lower performance.
   (c) Configuration with 5 block : $3 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow latentspace(128) \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow 1$:
      i. Failed to converge in 20 epochs.
      ii. Resulted in significantly lower performance.
   (d) These findings suggest an optimal depth for the hourglass architecture, with $3 \rightarrow 16 \rightarrow 16 \rightarrow 8$ striking a balance between depth and convergence speed for superior performance.
   (e) Deeper architectures did not consistently improve outcomes and, in some cases, can hinder convergence. which might be resumed as an over-encoding of the images.

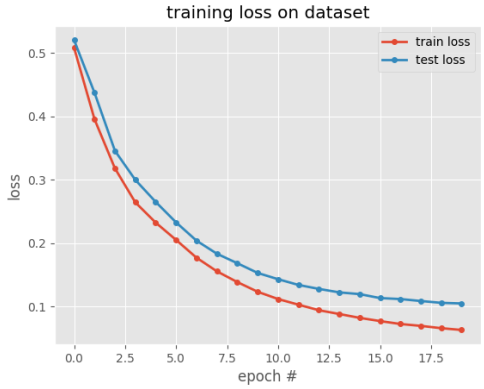


Figure 3: training and test loss of Hourglass with 1 blocks

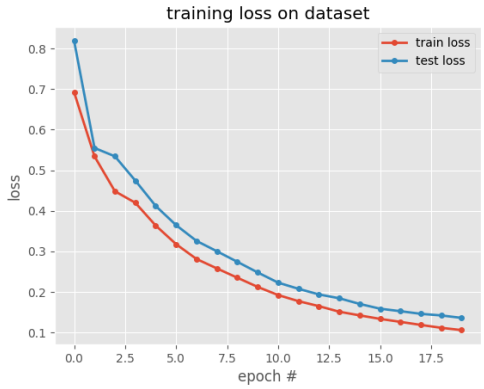


Figure 4: training and test loss of Hourglass with 4 blocks

Figure 5: training and test loss of Hourglass with 5 blocks

4 Part 2: U-Net Convolutional Neural Network

4.1 Experiment and Results

1. **Convergence Observation**: Similar to the HourGlass network, the U-Net architecture will be trained to observe its convergence behavior.

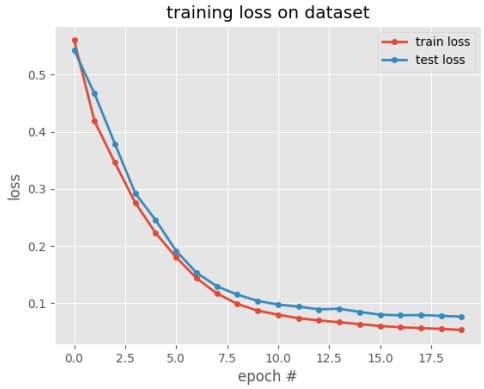


Figure 6: Convergence of U-Net

2. **Impact of Varying the Depth:** for Unet architecture, we followed the same study and investigated the same configurations, to provide insights about the effect of varying number of encoding blocks on our segmentation task of this specific dataset

   (a) Configuration with 1 block : $3 \rightarrow 16 \rightarrow latentspace(128) \rightarrow 16 \rightarrow 8 \rightarrow 1$. The model starts (first epochs) with a little loss and converges

easily.

(b) Configuration with 3 block : $3 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow latentspace(128) \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow 1$: The model still converges easily, but we start noticing that the loss is a little bigger than for the 1 block scenario and it converges slower.

(c) Configuration with 4 block : $3 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow latentspace(128) \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow 1$: The Unet model still converges but the loss is increasing and speed of convergence is decreasing.

(d) Configuration with 5 block : $3 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow latentspace(128) \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow 1$: Note that the idea behind this experiment is to remark the difference between Unet and Hourglass and remember that the only difference in architecture is provided by Unet's skip connection (contenation of some encoding features for decoding). Skip connections in Unet are employed to address the challenge of information loss during down-sampling and up-sampling in image segmentation tasks, and may also help the model avoid vanishing or exploding gradient, thus continue to learn the weights and converge at some point. In fact, for this scenario, hourglass couldn't converge, while Unet did at a point, even if the loss (mostly for first epochs) is becoming bigger.

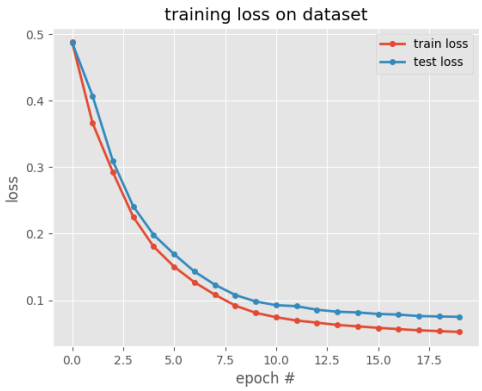


Figure 7: training and test loss of Unet with 1 blocks

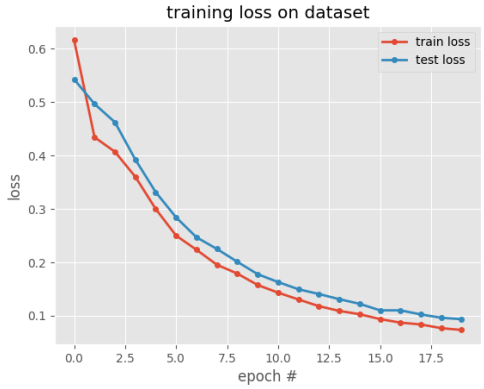


Figure 8: training and test loss of Unet with 4 blocks

Figure 9: training and test loss of Unet with 5 blocks

3. **Comparison with HourGlass**: We have conducted a comprehensive comparative analysis between the HourGlass and U-Net architectures, focusing on their convergence speed and segmentation performance with the tictoc dance sequences dataset. Upon reviewing the results, a discernible distinction emerged. While both architectures exhibit strong performance, we notice a small but noticeable difference. Specifically, the U-Net architecture demonstrates slightly faster convergence compared to the HourGlass architecture. Additionally,with the use of 'BCEWithLogitsLoss' indicates consistently lower losses for both training and test datasets within the U-Net model. This subtle yet significant difference underscores the nuanced advantages of the U-Net model in terms of both training efficiency and generalization on the specified dataset.

5 Conclusion

In conclusion, the comparative study between HourGlass and U-Net for image segmentation on the tictoc dance sequences dataset revealed nuanced differences. The HourGlass architecture demonstrated reasonable convergence behavior, with shallower configurations outperforming deeper ones. Conversely, U-Net, distinguished by its skip connections, exhibited good convergence and showcased the remarkable ability to handle information loss during down-sampling and up-sampling. This unique feature contributed to U-Net's effectiveness, allowing it to converge effectively even with increased depth, and it demonstrated slightly faster convergence compared to the HourGlass architecture. While skip connections are a distinguishing factor in U-Net, their impact on faster convergence is likely influenced by the synergy between architecture and skip connections, not solely their presence. These connections play a pivotal role in preserving fine-grained details and addressing information loss, making U-Net well-suited for precision-focused segmentation tasks. The choice between architectures should prioritize this advantage, especially in scenarios where handling information loss is crucial. Recommendations for selecting the appropriate architecture should consider the specific demands of the segmentation task at hand, acknowledging the efficiency and advantages offered by U-Net's skip connections.