Background replacement using UNet architecture

Iordachescu Anca-Mihaela and Patilea Catalina Camelia

Supervisor: Emanuela Haller

University of Bucharest - Faculty of Mathematics and Computer Science

Introduction

Replacing the background is a challenging task in image editing and a well known computer vision task, task that is divided into two parts: background removal which involves training a human soft-segmentation network and adding a new background using alpha blending.

Dataset

The dataset we used in order to solve the task is called 'Matting Human Datasets', developed by AISegment.com, and is currently the largest portrait matting dataset, containing more than 34,000 images and corresponding matting results [1]. It comes in two parts: 600x800-pixel half-length portrait images in the format jpg and the respective RGB masks, used to extract the alpha map, in the format png.

Due to the fact that this dataset is too large, we selected a small amount(about 2700) of photos for our model. The new dataset was preprocessed by adjusting the image size to 320x240 (scaled in both dimensions at percentage equal to 60) and by rescaling using MinMax scaler. The data was augmented by flipping horizontally and small-angle rotating.

Model

The model we used is U-Net, a convolutional neural network used especially for image segmentation. The first time U-Net was used was in 2015, applied to process biomedical images [2], being able to localize and distinguish objects, by classifying pixels.

The architecture is made of two symmetric parts:

- **the encoder** - the contracting part that consists of repeated application of down convolutions and each of them being followed by a rectified linear unit(RELU) and max pooling operation.
- **the decoder** - the expansive part that consists of repeated up-convolutions applied to the layers and concatenations with features.

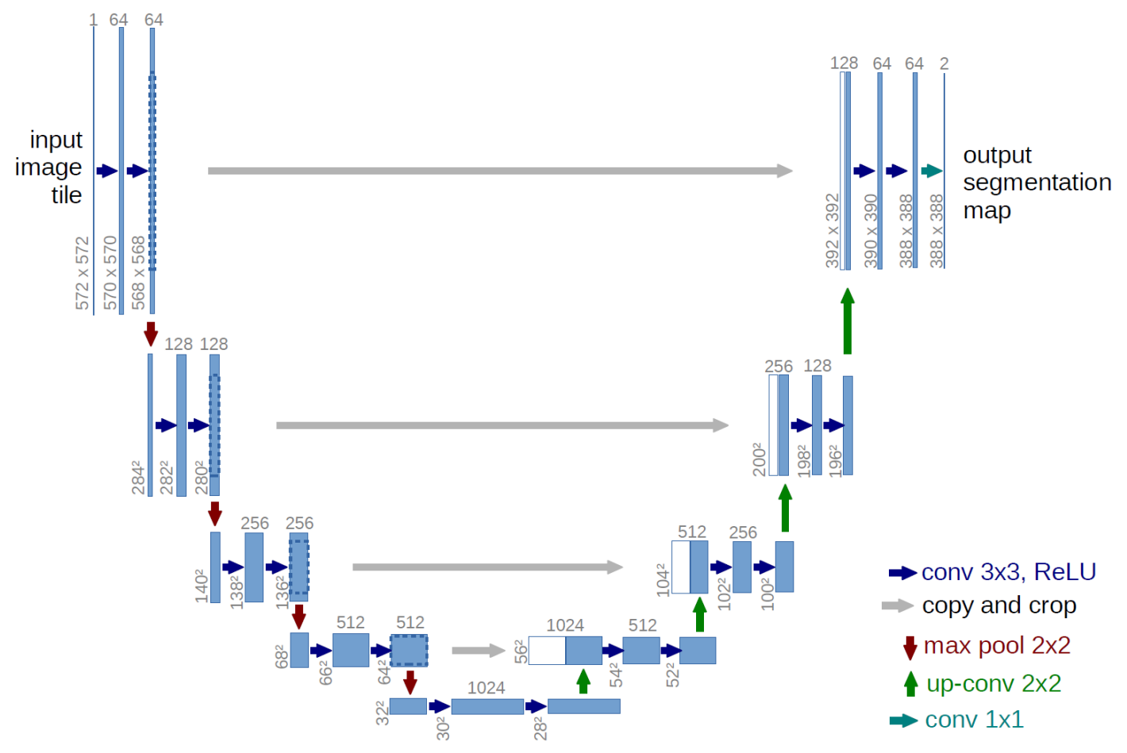

Fig. 1: UNet architecture

The encoder of the U-Net architecture is ResNet50, a variant of ResNet which is composed of 48 Convolution layers, one MaxPool layer and one Average Pool layer.

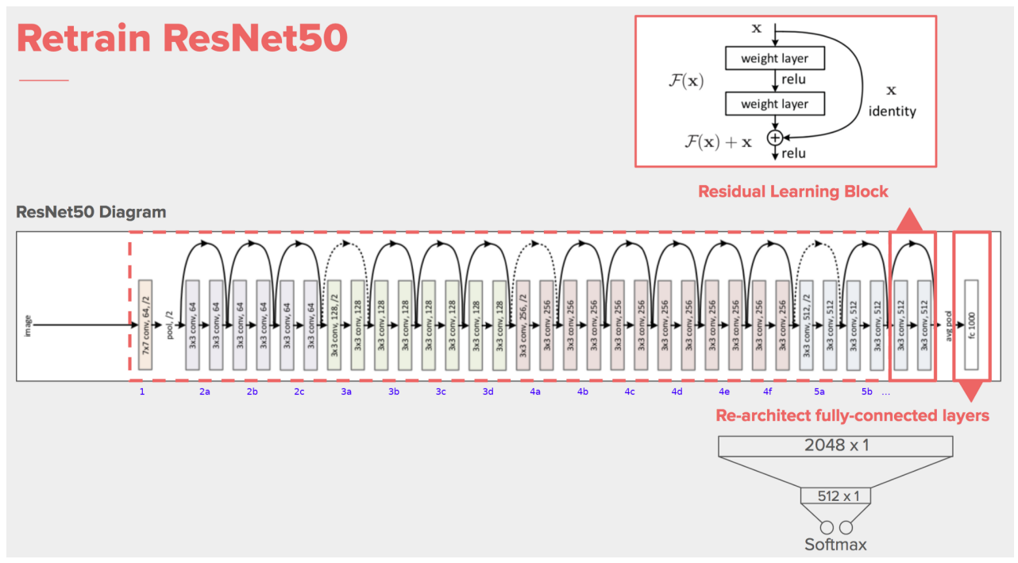

Fig. 2: ResNet50 Architecture

Loss function and Metrics

The loss we use to evaluate our model is L1 Loss Function. This loss is used to minimize the error which is the sum of the all the absolute differences between the true value and the predicted value.

Formula for the L1 loss, where yt and yp are the true and predicted labels:

$$\sum_{n=1} |y_t - y_p|$$

Another metric we used in order to evaluate how well our algorithm models the dataset is dice coefficient.

The dice coefficient is a statistic developed in 1940's to measure the overlap between two samples. This measure is between 0 and 1, where 1 means that the overlap is perfect and complete.

In order to formulate a loss function which can be minimized, we use 1-Dice known as the soft Dice loss because we directly use the predicted probabilities instead of thresholding and converting them into a binary mask [3].

Formula for the soft dice coefficient, where yt and yp are the true and predicted labels:

$$1 - \frac{2 * \sum_{n=1} y_t * y_p}{\sum_{n=1} y_t^2 + \sum_{n=1} y_p^2}$$

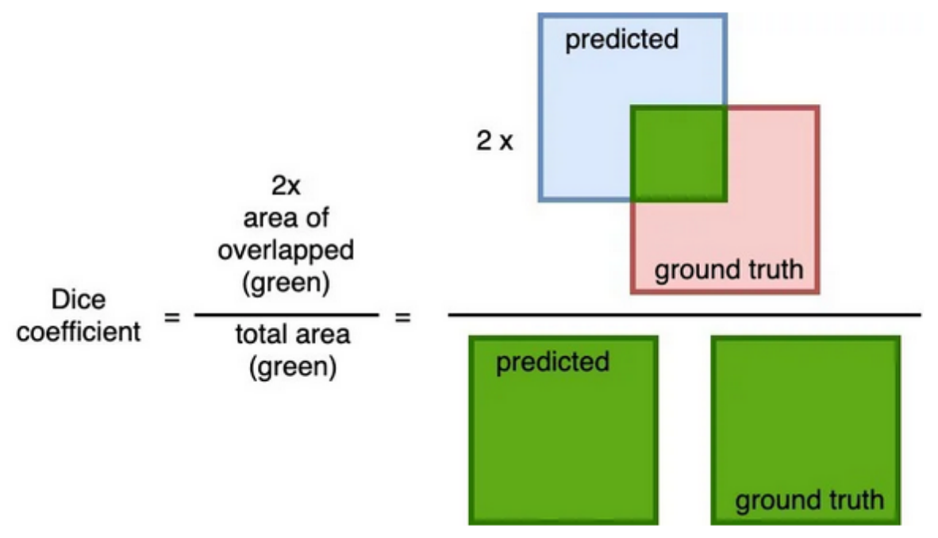

Fig. 3: Dice Loss function

Experiments

Our dataset was split in this manner: 80% train - 20% validation.

The model was trained using the Adam optimizer with a learning rate of 0.005, batch size of 128 and number of epoches of 40.

We use two metrics to evaluate our model: L1 Loss and Soft Dice Coefficient.

Based on the two plots showed on Fig.4 our model seems to perform well on our task.

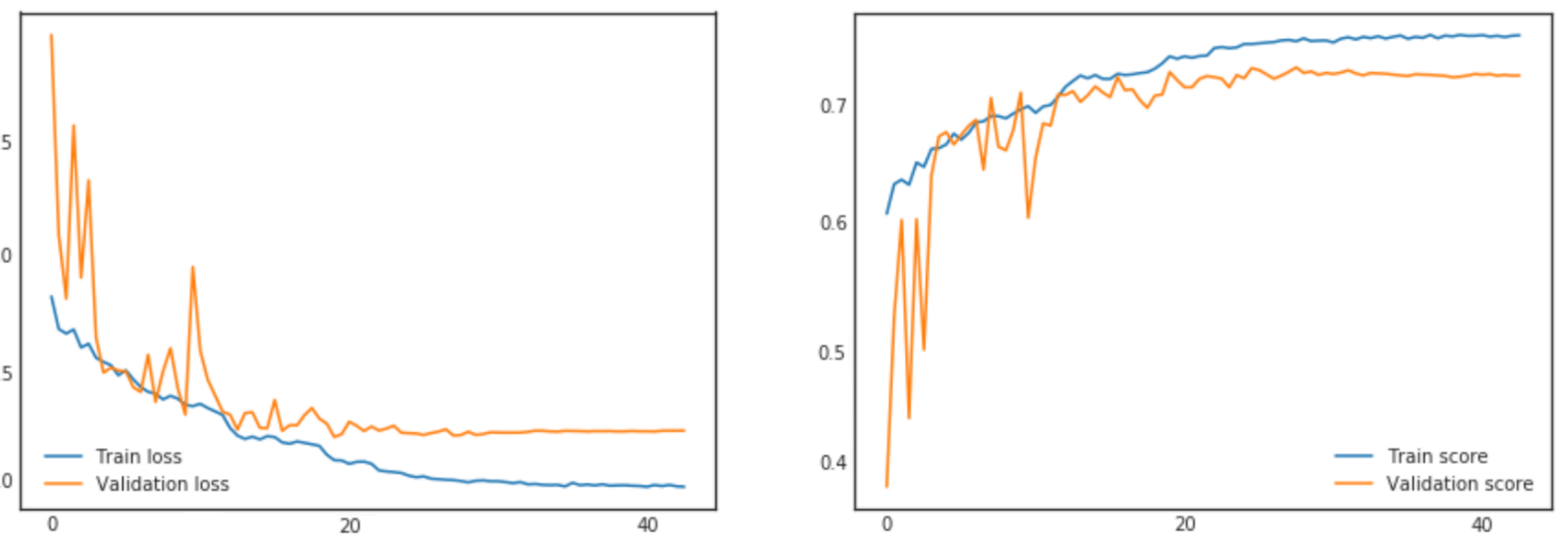

Fig. 4: L1 Loss and Dice coefficient

Evaluate model performance:

- Training loss - 0.174
- Validation loss - 0.421
- Training dice coefficient - 0.751
- Validation dice coefficient - 0.687

Background replacement

Once we have subtracted the background from the original image, we have to replace it with a new background.

In order to perform this task we use alpha-blending, the process that composes two source images into one output based on alpha channel. The alpha coefficient, which is the pixel's matting estimation, is considered to be in the interval [0,1][4].

This process needs three arguments: the original photo on which we apply background replacement, the segmentation map obtained using our U-Net model known as alpha coefficient and the new background.

The formula for alpha-blending:

$$finalImage = originalImage * \alpha + newBackground * (1 - \alpha)$$

Note that a value of 0 in the alpha matrix means that the pixel is fully transparent and does not provide any coverage information while a value of 1 means that the pixel is fully opaque.

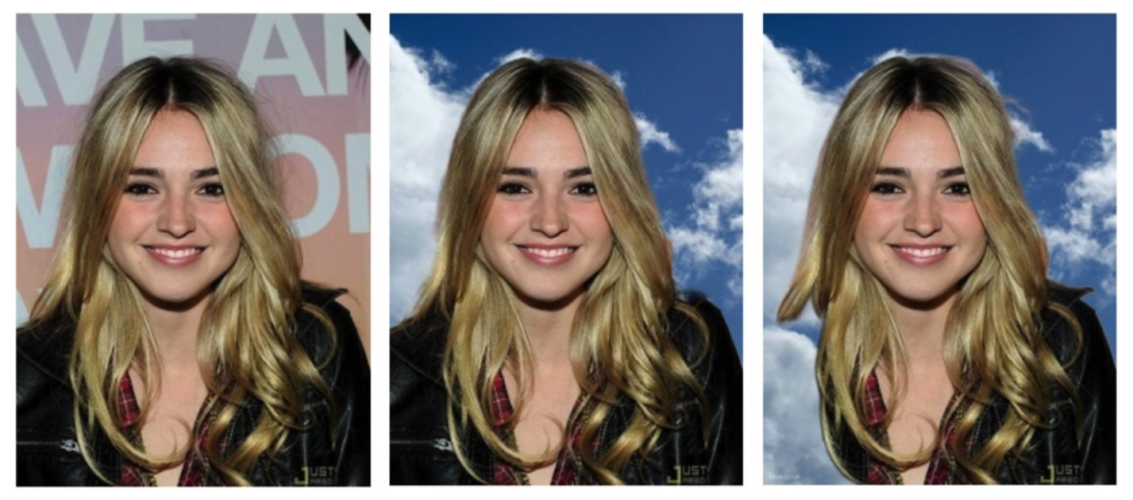

Fig. 5: (a) - Original photo. (b) - Ground truth. (c) - Prediction.

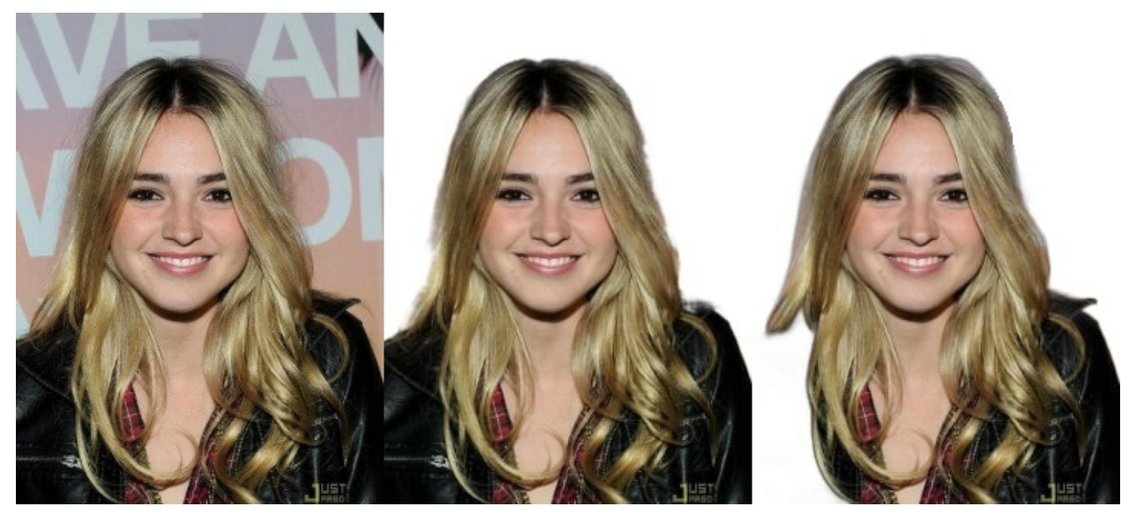

Fig. 6: (a) - Original photo. (b) - Ground truth - matting. (c) - Prediction - matting.

Conclusion

Due to ResNet50 used as the encoder, the model showed some promising results initially but was not able to improve much further with more training epochs because of the fact that out dataset is small compared to the original dataset.

We can improve our model by increasing the size of our dataset and by data augmentation(adding images from the dataset with new background).

References

[1] Laurent H. *AISegment.com - Matting Human Datasets*. URL: https://www.kaggle.com/laurentmih/aisegmentcom-matting-human-datasets/.

[2] T. Brox O. Ronneberger P. Fischer. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: (2015), p. 8.

[3] E. Tiu. *Metrics to Evaluate your Semantic Segmentation Model*. URL: https://towardsdatascience.com/metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2.

[4] *Transparency Handout*. URL: http://web.cse.ohio-state.edu/~parent.1/classes/581/Lectures/13.TransparencyHandout.pdf..