Internship Technical Report - Semantic Segmentation

Jessica Klebe
j.s.klebe@gmx.de

May 11, 2021

1 Introduction

This report summarizes the work for task 2. The aim was to build and train a net such that it predicts the class or background for each pixel of the entered image. Therefore, important aspects that the net needs to cover are on one hand the correct classification and on the other hand the correct detection of the area of the lesion. First the method, second the experiments and finally the results are presented in this report.

2 Method

In the following the method of the net's evaluation is described.

2.1 Scores

Different scores have been implemented in order to evaluate the performance of the network.

- **Pixel Similarity between Prediction and Mask [PixSim]**: Determines how many pixels of the output and the mask are same and returns the average value of the set.
- **Intersection over Union [IoU]**: Returns intersection over union of the set. The IoU is computed of the concatination of all predictions in order to avoid division by zero.
- **Jaccard Score [Jaccard]**: Returns the Jaccard Score of the set. The Jaccard score is determined based on the IoU values.
- **Number of predictions that predicted at least 90% of the image correctly [N_90]**: Relative number of predictions of the set where at least 90% of pixels have been predicted correctly.
- **Number of predictions where lesion is predicted correctly for at least 90% of pixels [N_lesion_90]**: Relative number of predictions of the set where at least 90% of the lesion (without background) is predicted correctly.
- **N_lesion_50**: Relative number of predictions of the set where at least 50% of the lesion (without background) is predicted correctly.

The early stopping method has been implemented once on the IoU and once on the Jaccard score.

2.2 Auxiliary values

In order to understand how the prediction of background might affect the scores, it might be interesting to consider additional values which will be referred to as auxiliary values.

- **PixSim_black**: Average number of correctly identified pixels if the prediction was purely background.
- **N_90_black**: Number of images where the lesion consists of less than 10% of the pixels of the image.

2.3 Progressive Growing

In another approach progressive growing has been implemented.

3 Experiments

In this section the model, its implementation and the methods of data augmentation are described.

3.1 Model and Implementation Details

The model has a Unet architecture with ResNet34 as the encoder. The cross entropy loss function and AdamW as an optimizer with a learning rate of 0.00003 have been used. Again, the learning rate has been adjusted using a cosine annealing scheduler where the number of update steps is $T_{max} = 1037$ and corresponds to one epoch.

The dataset is split into a training, validation and test set - the same sets as for the first task are used for the semantic segmantation network. Early stopping is implemented; if a certain parameter (IoU or Jaccard; cf. Section 2.1 for information on the parameters) does not improve over 10 epochs, the training period is finished. After a total number of 100 epochs the training is interrupted even if the early stopping condition has not been met.

3.2 Data augmentation

3.2.1 Training set

The images are transformed into tensors; subsequently the color values of the images are transformed using `ColorJitter(0.4,0.4,0.4)` and normalized. Then, the masks are transformed into tensors. Finally, images and masks are simulaneously transformed: they are resized to images of size $256px \times 256px$, then they are randomly cropped to size $224px \times 224px$, they are randomly horizontally and/or vertically flipped and random affine transformations and distortions are applied.

3.2.2 Validation and Test sets

Certrain data augmentation procedures - like flipping or transformations - are only necessary to improve training of the network. Hence, transformations applied on data of the validation and test sets differ from the ones applied on the training set: First, images are transformed to tensors and normalized. Subsequently, masks as transformed to tensors. Finally, both images and masks are resized to $256px \times 256px$ and center cropped to a size of $224px \times 224px$.

4 Results and Conclusion

4.1 Early stopping based on Jaccard

First, the net has been trained using the early stopping method in dependence on the Jaccard Score (cf. Section2.1 for description of the score), i.e. if the Jaccard score has not improved for 10 epochs, the training has been terminated. The results of the score on the validation and test sets and the auxiliary values are illustrated in Tab. 1.

4.1.1 Validation set

Table 1: Scores and auxiliary values for the validation and test sets where the network has been trained with an early stopping condition based on the Jaccard store.

Score or Auxiliary	Validation	Test
PixSim	0.8842	0.8788
PixSim_black	0.6469	0.6339
Jaccard	0.3056	0.1046
IoU	0.5875	0.5195
N_lesion_90	0.6750	0.6683
N_lesion_50	0.8195	0.8135
N_90	0.7836	0.7506
N_90_black	0.1176	0.1187

The PixSim value states that on average 88.42% of the pixels of the validation set were predicted correctly. However, since from this value one cannot deduce how much the background accounts to the final

result, the value alone does not offer sufficient information in order to assess the quality of the net adequately. Instead, it should be considered in combination with PixSim_black = 0.6469 which means that 64.69% of the pixels are predicted correctly if the prediction consisted solely of background. The difference between 88.42% and 64.69% is notable, nevertheless, these values suggest further investigation about the net's quality should be conducted.

The next score considered is the Jaccard score with Jaccard = 0.3056. Since the Jaccard score for one class is 0 if for the respective class IoU < 0.65 and since there are 7 classes, a Jaccard score of 0.3056 implies that 3 classes reached an IoU value above 0.65. The IoU value for each class is depicted in Tab. 2. It illustrates that for the classes akiec, df and nv the IoU scores are above 0.65. The remaining classes have an IoU value between 0.4 and 0.6. The classes bcc and mel contain malicious lesions which generally have to be removed via surgery. A low IoU value for these classes is a particularly insufficient performance. It would be crucial that the net at least identifies a malicious lesion as a malicious one such that the removal of the lesions can be initiated. Hence, if the net would confuse the classes bcc and mel with each other it would be a more desirable scenario than having the net confusing a malicious lesion with a benign one. Therefore, it would be interesting to investigate whether the IoU score improves if bcc and mel are merged into one class - a malicious lesions group. Unfortunately, the corresponding IoU value reaches only a value of 0.4315. This value suggests that the net also has problems classifying malicious lesions if they are grouped into one group. This is an undesirable result for clinic applications. The overall IoU score is 0.5875.

Since the main goal is to predict the lesions correctly, it would be interesting to know how many instances have predicted at least 90% of the lesions without background correctly (=No_lesion_90) and how many instances have at least predicted half of the lesion correctly (=No_lesion_50). 67.50% of the instances have predicted 90% of the lesion correctly. This value is improvable. 81.95% have predicted at least 50% of the lesions correctly. This shows that the majority of lesions are detected at least partially as the correct class. Since the difference between No_lesion_90 and No_lesion_50 is rather large one cannot draw the assumption that once the class has not been predicted correctly, a large amount of the lesion is detected.

Table 2: IoU values for the predictions on the validation and test sets. Early stopping is based on the Jaccard score.

Class	Validation	Test
akiec	0.7539	0.4789
bcc	0.4185	0.5038
bkl	0.5550	0.4574
df	0.6696	0.4991
mel	0.4274	0.4574
nv	0.7155	0.4916
vasc	0.5727	0.7325

In order to establish a connection to the first task where just a ResNet34 net for classification has been implemented, the values No_lesion_90 and No_lesion_50 are compared with the accuracy of the ResNet34 net. The accuracy on the validation set for the ResNet34 was 0.8774 (mean recall = 0.8077, balanced accuracy =0.8342, for the confusion matrices please refer to the first report). While No_lesion_50 almost reaches the accuracy of the ResNet34 classification net, No_lesion_50 is substantially smaller.

What cannot be deduced from the score No_lesion_90 is how the net is prone to errors regarding false prediction of areas that are background. Consequently, the score No_90 is introduced. It says the net has predicted 78.36% of all instances for at least 90% of the image correctly. This value is a bit more than 10%-points higher than No_lesion_90, suggesting that some errors also occur in the background, however, the majority of errors seems to happen in the detection of the actual lesion. Again, No_90 needs also to be compared to the cases where the prediction as solely background in order to understand how much the background contributes to this value. No_90_black has a value 0 0.1176, hence, 11.76% of values are predicted for at least 90% correctly if the prediction was only background. This implies that the prediction is sufficiently far

away from simply predicting background.

4.1.2 Test set

The same analysis of the scores can be repeated on the test set. The PixSim and PixSim_black scores are similar to the ones of the validation set, hence, the conclusions drawn from above apply here as well. Since the values on the validation set and test set are comparable, it can be deduced that a good generalization has been achieved. However, the performance scores of the ResNet34 net (task 1, simple classification) are decreased (accuracy = 0.8489, mean recall = 0.7638, balanced accuracy = 0.7330) such that the performance difference between the semantic segmentation and ResNet34 nets is also decreased. Neverthelesse, the ResNet34 net still performs better than the Unet such that the conclusion that further improvement on the Unet needs to be done is still drawn.

The Jaccard score has only a value of Jaccard = 0.1046 and, hence, one class can have yielded a value above 0.65. It is expected that the value is lower on the test set since the early stopping method optimized the Jaccard score based on the validation set. The average IoU score is 0.5195. The IoU values for each class are depicted in Tab. 2. It is interesting to observe though that this time none of the classes (akiec, df, nv) that have previously yielded a value above 0.65 reaches the Jaccard condition. Instead, only the class vasc meets the Jaccard condition and scores above 0.65. While akiec, a class that is potentially not as dangerous as bcc or mel but still not a benign lesion, has met the Jaccard condition on the validation set, one cannot rely on its prediction on the test set. Again, if the two most malicious classes (bcc and mel), who both entail removal, were considered as one class, an IoU score of only 0.4491 is reached, suggesting the very malicious classes are not confused with each other but other errors happen to occur.

For the scores No_lesion_90, No_lesion_50 and No_90 the values are similar to the validation set. Again this suggests a good generalization.

Another conclusions can be drawn when comparing how the Jaccard and IoU values have changed from the validation to test set while the No_lesion_90 and No_lesion_50 scores have kept similar values: The success rate of the detection of the lesion itself appears to stays the same, nevertheless the Jaccard and IoU scores decrease in value. Consequently, not the intersection becomes smaller but the union becomes larger. This indicates that the net mistakenly predicts at least parts of the background as lesion. Fig. 1 depicts two examples where the net mistakes the background with the lesion and the lesion with the background. On average the background takes the majority of the pixels, hence, these kind of mistakes lead to a substantial increase in the union.

In conclusion, while most scores suggest that the net needs improvement, it is at least capable of detecting parts of the lesion and classifying at least parts of it correctly. The Jaccard score and IoU values of each class suggest that particularly the malicious classes that are important to detect in order to initiate a removal of the lesions are not sufficiently predicted by the net. Consequently, further improvement of the net is required.

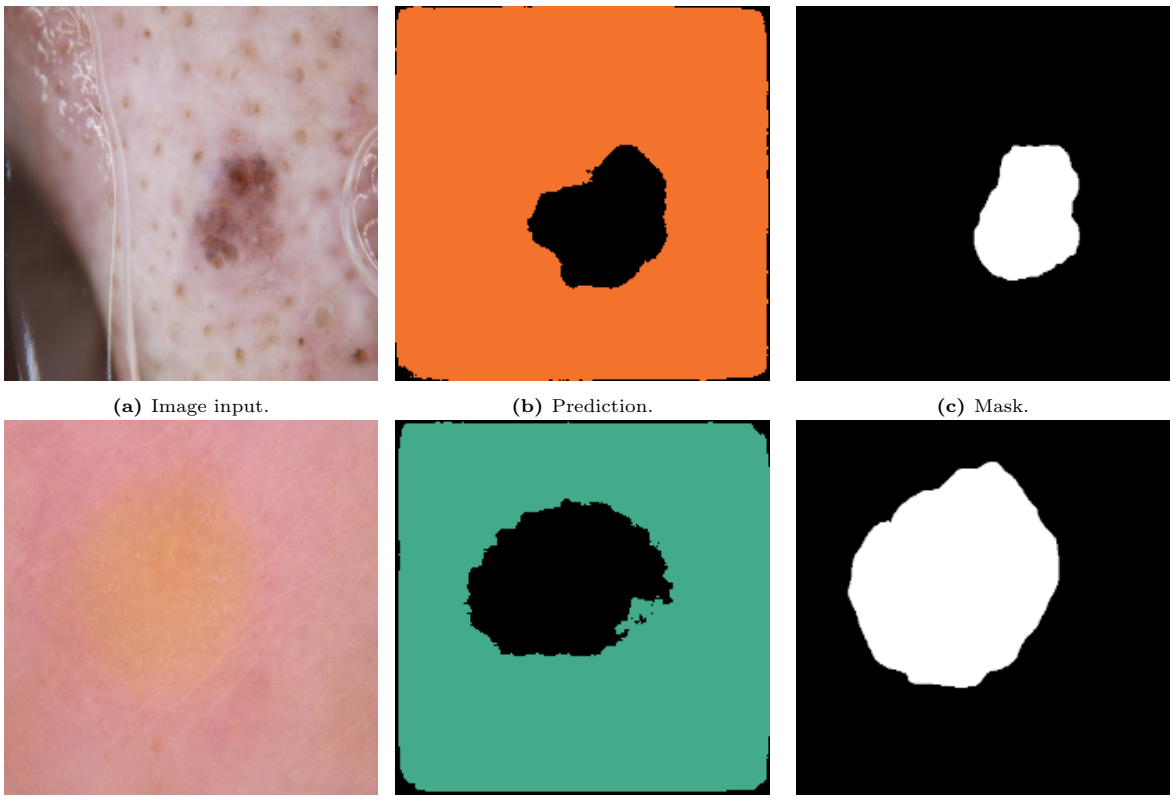
4.2 Progressive Growing

Progressive Growing has not improved the results. Hence, a detailed analysis is waived in this report.

4.3 Early stopping based on IoU

Second, the net has been trained using the early stopping method in dependence on the IoU score (cf. Section 2.1 for description of the score). If the IoU score has not improved for 10 epochs, the training has been terminated. Since the scores are similar to the first attempt, the results of the scores on the validation and test sets are illustrated in the appendix in Tab. 3.

An interesting result drawn from this experiment would be however, that on the validation set the classes akiec and nv reached IoU values above 0.65 while on the test set only the class nv met this condition. All other values are again between 0.4 and 0.6. In combination with the first attempt, this suggests that there is some kind of variation which classes score high on the Jaccard score. Therefore, if it will be possible to train several nets that are good in



(a) Image input. **(b)** Prediction. **(c)** Mask.

(d) Image input. **(e)** Prediction. **(f)** Mask.

Figure 1: Comparision of the image, prediction and mask for two examples where the net confuses background and lesion. The green color indicates that the net determined these pixels to be benign keratosislike lesions and the orange color indicates that the net determined these pixels to be a vascular lesion.

predicting different classes, the average of the models might help improve the overall correct prediction rate.

5 Further ideas

While for the first task (classification with a ResNet34) balanced success rate scores have been considered, balancgin has not been implemented yet for the task on the semantic segmentation net. Hence, the implementation could be a next step.

Furthermore, as shown above in Fig. 1 in some instances the net seems to flip background and lesion. It would be interesting to determine under which condition this happens. If the error could be erased, it might improve the net's success rate significantly. It appears that the net makes this mistake particularly when the background does not consists entirely of smooth skin but of a distortion of background. However, this assumptions that was made solely taking a look at the images and corresponding predictions by hand needs further investigation.

In addition, one could evaluate whether under the condition that a lesion has been detected correctly the class has been determined correctly. This might give some further insight into the net's susceptibility to errors.

Finally, if it is possible to train several nets that each predict different classes correctly, it might improve the net's performance to take the average of these nets. Consequently, as a next step one could also try to train various nets that perform on certain classes well.

Table 3: Scores and auxiliary values for the validation and test sets where the network has been trained with an early stopping condition based on IoU.

Score or Auxiliary	Validation	Test
PixSim	0.8887	0.8772
PixSim_black	0.6469	0.6339
Jaccard	0.2141	0.1032
IoU	0.5961	0.5103
N_lesion_90	0.6580	0.6479
N_lesion_50	0.8285	0.8015
N_90	0.7896	0.7461
N_90_black	0.1176	0.1187

6 Appendix