Comparing Objective Functions for Segmentation and Detection of Tiny Lesions in Retinal Images

Editors: Under Review for MIDL 2020

Abstract

Retinal microaneurysms (MAs) are the earliest signs of diabetic retinopathy (DR) which is the leading cause of blindness in the western world. MAs independently predict the risk of sight threatening DR and early detection is important to identify patients at risk. Detection and segmentation of retinal MAs present a particular challenging problem due to a large class imbalance with MA pixels accounting for less than 0.5% of the retinal image. Extreme foreground-background class imbalance can adversely affect the learning process in DNNs by introducing a bias towards the most well represented class. Recently, a number of objective functions have been proposed as alternatives to the standard Crossentropy loss in efforts to overcome this problem. In this work we investigate the influence of the network objective during optimization by comparing Residual U-nets trained for segmentation of MAs in retinal images using seven different objective functions; weighted and unweighted Crossentropy loss, Dice loss, weighted and unweighted Focal loss, Focal Dice loss and Focal Tversky loss. Three networks with different seeds are trained for each objective function using optimized hyper-parameter settings on a dataset of 382 images with pixel level annotations for MAs. The instance level MA detection performance is evaluated as the average free response receiver operator characteristic (FROC) score calculated as the mean sensitivity at seven average false positives (FPAvg) per image thresholds on 80 test images. The image level MA detection performance is evaluated as the average AUC on the same images as well as a separate test set of 1200 images. Segmentation performance is evaluated as the average pixel precision (AP). The unweighted Crossentropy loss and Focal loss outperforms all other losses for instance level detection achieving FROC scores of 0.5067(±0.0115) and 0.5062(±0.0045). The Focal loss has the highest pixel precision with an AP of 0.4254(±0.0096). For image level detection both objective functions in their unweighted form perform significantly better compared to using all other objectives. AUCs of 0.9450(±0.0080) and 0.8351(±0.0039) on the two test are achieved using the unweighted Crossentropy loss, while AUCs for the unweighted Focal loss was 0.9375(±0.0074) and 0.8253(±0.0042) respectively.

Conclusion: Despite the promise of using training objectives designed to deal with unbalanced data, the standard Crossentropy loss perform at least as well or better than all other objective functions in our experiments for lesion level and image level detection for small retinal MAs. While a number of newer objective functions have been introduced and shown to improve performance for unbalanced datasets compared to the Dice loss in recent years, our results suggest that it is important to also benchmark new losses against the Crossentropy or Focal loss function, as we achieve the best performance in all our test using these objectives.

Keywords: Semantic Segmentation, Detection, Diabetic Retinopathy, Diabetes, Retinal Imaging.
1. Introduction

Segmentation of image features provides a good basis for disease detection in medical images. Doctors make diagnostic decisions based on the presence of diseased tissue or other anatomical changes, so it makes sense that algorithms for automatic diagnosis should do the same. Image classification with deep neural networks (DNN) suffer from a lack of interpretability which can be problematic in the context of medical image analysis and computer aided diagnosis. Hence, using semantic information for disease detection is a logical approach. DNNs work well for biomedical image segmentation, with an example being the U-net architecture (Ronneberger et al., 2015) and the several variations thereof which have since been proposed.

Diabetic retinopathy (DR) is the most common micro-vascular complication of diabetes which is most common metabolic disease in world. Microaneurysms (MAs) are small changes occurring in the retinal vascularity and are the earliest sign of DR. MAs appear as small red dots in the retinal tissue and represent only a very small portion of the retina, making them very difficult to detect. As MAs independently predict the risk of sight threatening DR (Pappuru et al., 2019), early detection is important to identify patients at risk. Presence of MAs is also used for planning screening intervals for patients (Grauslund et al., 2018). Automated retinal image analysis have been an area of research for many years (Nørgaard and Grauslund, 2018) but has recently gained increased interest due to the improved performance of DNNs. Many automated retinal image analysis methods using DNNs perform binary classification, e.g. non-referable versus referable DR as opposed to more fine scaled classification with multiple levels of DR disease severity. In the binary classification setting, many DNNs have shown results comparable to human experts (Nielsen et al., 2019), but results also demonstrate that these methods have yet to achieve the desired diagnostic accuracy for full scale DR grading. This suggests that DNNs are able to detect the macroscopic abnormalities indicative of more severe levels of DR, but fail to recognise more subtle, microscopic lesions such as MAs which would enable DNNs to better distinguish between different levels of DR disease severity. In patient screening contexts the presence of MAs corresponds to the lowest level DR (Wilkinson et al., 2003) and MA detection can thus be used to predict this level of disease. The ability to detect low levels of DR is important because it enables the algorithm to recognize signs of the disease before it has severe effects on patient health. In the case of DR it can also be used for managing the referral of patients in screening settings and decrease the workload on clinicians as patients with low levels of DR might need to be screened more often compared to those showing no signs. Semantic information and visual interpretability is important for the adoption of computer assisted diagnostic methods. When using DNNs for medical image analysis interepretability is often a challenge due to the black box which is said to encapsulate DNNs(Adadi and Berrada, 2018). A simple approach to solving this problem is to use semantic image information as part of the diagnostic pipeline. In this work we demonstrate this approach by training DNNs for segmentation of retinal MAs and subsequently using the segmentation output for classification of DR based on the presence or absence of MAs in an image. This approach of using the output probability maps for classification resembles classification by simple logistic regression. The advantages of this approach is the semantic information yielded by
the segmentation DNNs and the ability to adjust the model’s sensitivity. Another appealing quality of this method is that it requires significantly fewer images compared to classification DNNs which generally requires thousands of images in order to converge. Although, as data annotation for segmentation is significantly more labour intensive, the advantage is perhaps not as pronounced as one could have liked.

2. Methods

We compare DNNs trained for segmentation of retinal MAs. The resulting network segmentation maps are used for detection of individual MAs as well as for image level detection. In order to optimize segmentation performance of the MAs, we train several DNNs using different objective functions to determine the best objective for segmentation of these small lesions. As skip connections have been shown to improve performance of DNNs for biomedical image segmentation (Drozdzal et al., 2016), we utilize a Residual U-net (Zhang et al., 2018) for all our experiments, which is a U-net based architecture with addition of short residual skip connections in the encoder and decoder. The models were trained using publicly available retinal images from the E-ophtha database with pixel level annotations for retinal MAs (Decenci`ere et al., 2013). The dataset consists of 382 images of which 149 contains one or more MAs. The remaining 233 images contain no MAs. Images were randomly split into training, validation and test sets of 252, 50 and 80 images respectively. The images in the validation set were used to tune hyperparameters of the network and objective functions. After training, individual MA detection as well as image level detection of MAs and pixel segmentation performance was evaluated on the 80 images in the test set. Additionally, image level MA detection was performed on another test set of 1200 images from the Messidor database (Decenci`ere et al., 2014). Evaluation of individual MA detection was performed using the free response receiver operator characteristic (FROC) score which is calculated as the mean sensitivity at seven average false positive per image (FPAvg) thresholds of $\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4$ and $8$. MAs were counted as true positives if a single pixel of predicted MAs overlapped with MAs in the ground truth segmentation map. Segmentation performance was evaluated as the average pixel precision (AP). Image level detection was defined as correctly detecting any MAs in the images of the test sets at each threshold of a network’s probability map. If a MA was detected in an image and the ground truth image contained MAs the detection was counted as a true positive and false positive otherwise. The Messidor images are divided into four different levels; 0, 1, 2 and 3 with 546, 153, 247 and 254 images respectively. Each levels corresponds to an increasing severity of DR. In order to evaluate the networks abilities to detect mild forms of DR we also performed test for detection of only level 1, which is defined as images containing between 1 and 5 MAs.

Prior to training, all images were pre-processed using a common pre-processing scheme by performing contrast limited histogram equalization of the color images and extracting the green channel and finally cropping the image borders around the retina. The training set was artificially increased through data augmentation by sampling 128 by 128 pixel crops from the images using a sliding window approach. If crops contained any MAs pixels the
original crops were included in the training set along with five augmented crops. The augmentations consisted of gamma adjustment, flipping and flipping plus warping. If the crop did not contain any MAs, the crop was either kept as is or underwent one of the five augmentations. This yielded a total of 30574 crops for training. The models were implemented and trained using the Keras deep learning framework \citep{Chollet2015} with tensorflow backend on different Nvidia GPUs (GeForce GTX 1080, GeForce RTX 2080, TITAN X Pascal), each training one model at a time. Seven different objective functions were used in our experiments; The Crossentropiy loss (CE), Focal loss (FL) \citep{Lin2017}, Dice loss (DL) \citep{Sudre2017} and Focal Tversky loss (FTL) \citep{Abraham2018} as well as $\alpha$ weighted forms of CE ($\alpha$CE) and FL ($\alpha$FL) and the unbalanced form of the FTL.

The standard objective for training segmentation DNNs is the average pixelwise CE (Equation (1)), where $p_t$ is the probability assigned to each pixel belonging to the correct class $y$.

$$CE(p_t, y) = -\alpha \log(p_t)$$ (1)

The FL (Equation (2)) extends on the standard CE by addition of a focusing parameter $\gamma$. The idea is to differentiate between easy and hard examples and focus learning on examples with low probabilities for $y$. A weighting parameter $\alpha$ can be added to both the CE and FL formulations.

$$FL(p_t, y) = -\alpha (1 - p_t)^\gamma \log(p_t)$$ (2)

The DL can be formulated as in Equation (3) where $p_{ic}$ is the $i$th pixel probability and $g_{ic}$ is the corresponding ground truth pixel for class $c$. The objective is to maximize the overlap of the probability maps and ground truth segmentation maps. In practice, a small value $\epsilon$ is added to the numerator and denominator in order to avoid division by zero in case of empty segmentation maps.

$$DL(p_{ic}, g_{ic}) = 1 - 2 \times \frac{\sum_{n=1}^{N} p_{ic} g_{ic}}{\sum_{n=1}^{N} p_{ic} + \sum_{n=1}^{N} g_{ic}}$$ (3)

The FTL Equation (4) is based on the Tversky index which is a generalization of the Dice loss but with the add-on of weighting parameters that allows for balancing false positive and false negative examples. $p_{ic}$ is the probability that pixel $i$ is of the MA class $c$ and $p_i \bar{c}$ is the probability that pixel $i$ is of the non-MA class, $\bar{c}$ and the same is true for the ground truth $g$.

$$FTL_c(p_{ic}, g_{ic}) = \left(1 - \frac{\sum_{n=1}^{N} p_{ic} g_{ic}}{\sum_{n=1}^{N} p_{ic} g_{ic} + \alpha \sum_{n=1}^{N} p_{ic} g_{ic} + \beta \sum_{n=1}^{N} p_{ic} g_{ic}}\right)^{\frac{1}{\gamma}}$$ (4)

We trained models both with and without the class balancing parameters in the loss functions except in the case of the FTL where $\alpha$ and $\beta$ where set to 0.5. In the case of the FTL, setting the $\alpha$ and $\beta$ parameters to 0.5 turns the Tversky index into the Dice coefficient, meaning the loss function essentially becomes the Focal Dice loss (FDL). For each objective function we performed hyper-parameter search for initial learning rate setting and dropout probability. Afterwards, we trained three networks per setting for each of the four objective
Retinal Lesion Segmentation and Detection

Figure 1: (a) Mean FROC curves at low FPAvg and FROC scores (±sd) and (b) precision-recall curves and average precision scores for each objective function on E-ophtha dataset functions using three different seeds for parameter initialization using a batch size of 24. The model parameters from the epoch with the lowest loss on the validation set was used for evaluation on the test set. Models where trained with stochastic gradient descent using the Adam optimization algorithm with default parameters. We adopted the recently proposed Lookahead (Zhang et al., 2019) in the optimization step as early experimentation indicated that it helped with model convergence, although we did not perform rigorous experiments. Lookahead was implemented with its standard parameters using the python package keras-lookahead (Loo). We experimented with learning rates of 1e^{-3}, 1e^{-4} and 1e^{-5} and with using no dropout as well as dropout of 0.2, 0.5 and 0.7. For the Focal loss we experimented with γ of 1.5, 2, 3 and 4. For the FDL and FTL we experimented with using the value suggested in (Abraham and Khan, 2018) of $\frac{4}{3}$ as well as 2 and 4. The weighting parameter α was set to 0.9 for the MA class and 1 - α for the background class in the αCE and αFL. The α parameter was set to 0.7 and β to 0.3 in the FTL. We used seeds of 1, 2 and 3 respectively for initialization of model parameters except for an instance of the FDL objective where a seed of 5 was used as using a seed of 2 failed to converge on the training set and a seed of 1 failed to converge when using the FTL objective. For our final experiments we used a learning rate of 1e^{-4} and dropout of 0.5 except when using FTL and FDL objectives where a dropout of 0.2 was used.

3. Results

In this section we summarize the results of the best performing models for each of the objective functions. We show the results of individual MA detection and pixel level segmentation performance on the E-ophtha MA test set for the best parameter setting averaged across the three runs (Figure 1). In Table 1 we show the mean AUC for image level MA detection of the best performing individual MA detection models for each objective function on the two test sets. The results in Figure 1 and Table 1 show that the standard CE objective function has the highest average performance for lesion level and image level detection and
that the addition of the focusing parameter $\gamma$ has a negligible effect for the unweighted form of this objective. Changing the value of $\gamma$ also has minimal effect on model performance for the FL objective as the mean FROC score for each of the four values of $\gamma$ were 0.5000($\pm0.0095$), 0.4939($\pm0.0172$), 0.4944($\pm0.0106$) and 0.5062($\pm0.0045$) respectively. The DL and FTL based loss functions have significantly lower FROC scores compared to the CE and FL objectives and the sensitivities are lower at all FPAvg thresholds as well. Weighting the CE also has negative effect on the performance. For mean AUC on the two test sets there is a smaller but not insignificant difference between the two unweighted CE and FL objectives and the $\alpha$CE and $\alpha$FL, DL, FDL and FTL. The asterisk in Table 1 indicate that a fourth seed with a value of 5 was used as the model initialized with a seed of 2 failed to converge on the training set and produced all zero probability maps for the model trained with FDL and the same happened using a seed of 1 for the model trained using the FTL. The mean AUC on the Messidor data is lower than on the the E-ophtha data for all objective functions. The difference between the CE and FL is also negligible on the Messidor dataset, and both of them perform significantly better than the DL, FDL and FTL objectives. Table 1 also show that results of both FTL and FDL objectives have a higher variance compared to the other objectives for mean AUC on the E-ophtha dataset. For detecting images of level 1 in the Messidor dataset defined as retinas with between 1 and 5 MAs the model trained using the CE achieves identical mean AUC of 0.8351($\pm0.0039$). In Figure 1 we see that the addition of the focusing parameter $\gamma$ in the FL does improve segmentation performance, achieving and AP of 0.4254($\pm0.0096$).

Table 1: Average AUC on E-ophtha and Messidor test sets

<table>
<thead>
<tr>
<th>Loss function</th>
<th>E-ophtha AUC($\pm$sd)</th>
<th>Messidor AUC($\pm$sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>0.9450($\pm0.0080$)</td>
<td>0.8351($\pm0.0039$)</td>
</tr>
<tr>
<td>$\alpha$CE($\alpha=0.9$)</td>
<td>0.8988($\pm0.0133$)</td>
<td>0.7665($\pm0.0111$)</td>
</tr>
<tr>
<td>FL($\gamma=4$)</td>
<td>0.9375($\pm0.0074$)</td>
<td>0.8253($\pm0.0042$)</td>
</tr>
<tr>
<td>$\alpha$FL($\gamma=4$, $\alpha=0.9$)</td>
<td>0.9154($\pm0.0088$)</td>
<td>0.8099($\pm0.0224$)</td>
</tr>
<tr>
<td>DL</td>
<td>0.9162($\pm0.0016$)</td>
<td>0.7596($\pm0.0056$)</td>
</tr>
<tr>
<td>FDL($\gamma=\frac{4}{3}$)*</td>
<td>0.9097($\pm0.0248$)</td>
<td>0.7265($\pm0.0100$)</td>
</tr>
<tr>
<td>FTL($\gamma=\frac{4}{3}$, $\alpha=0.7$, $\beta=0.3$)*</td>
<td>0.9018($\pm0.0320$)</td>
<td>0.7299($\pm0.0095$)</td>
</tr>
</tbody>
</table>

4. Discussion

In this work we have demonstrated the use of DNNs for segmentation and detection of MAs in retinal images. We trained Residual U-nets to optimize seven different loss functions in order to determine which objective was best suited to deal with the large class imbalance between MA and background pixels. The CE loss which is the standard objective and is used in training of most DNNs perform better on average than the other objectives; $\alpha$CE, $\alpha$FL, DL, FDL and FTL for instance level and image level detection of MAs. These alternative loss functions have all been proposed as means to deal with large class imbalance in detection and segmentation datasets. Nevertheless, none of them are able to improve the performance of our network compared to the standard objective except for pixel segmenta-
tion where the FL achieves better results. Even so, this does not lead to better detection performance. The unweighted CE loss and FL outperforms all other losses for instance level detection achieving FROC scores of 0.5067(±0.0115) and 0.5062(±0.0045. For image level detection, both objective functions in their unweighted form perform significantly better compared to using all other objectives. AUCs of 0.9450(±0.0080) and 0.8351(±0.0039) on the two test are achieved using the unweighted CE loss, while AUCs for the unweighted FL was 0.9375(±0.0074) and 0.8253(±0.0042) respectively.

In the FL objective, the focusing parameter $\gamma$ does not directly account for class imbalance. Rather its purpose it to focus the loss on examples with low probabilities and less emphasis on examples with high probabilities for the correct class. In theory this works on imbalanced data-sets, as the class with fewer examples is expected to be the harder examples which will have lower probabilities. The same mechanism is applied in the Focal Tversky loss, where the focus is based on the Tversky index of the misclassified examples. As shown in Figure 1 and Table 1, introducing a large weighting parameter $\alpha$ in the CE based loss functions also does not have a positive effect on our results. These results are somewhat surprising as we expected improvements over the standard CE loss when using objectives specifically developed to deal with highly imbalanced datasets. In (Sudre et al., 2017), they introduce the generalized Dice loss and compare the performance of this objective to a sensitivity and specificity loss and the weighted CE loss on a dataset with similar imbalance to ours. They conclude that the difference between using the different losses is minimal on datasets with moderate imbalance, but that the generalized Dice loss is more stable when the imbalance increases and they argue that the choice of loss function is crucial when training DNNs. Generally, our results indicate that the choice of objective function is important, but contrary to their results, it seems that the standard CE is better suited for our task. In (Abraham and Khan, 2018) they compare their Focal Tversky Loss to the DL and the unbalanced Tversky loss using different network architectures on two datasets. On average, they improve results by 0.062 points for Dice coefficient and 0.0255 and 0.0625 for pixel level precision and recall respectively by using the Focal Tversky objective, although in one case precision is better when using the Dice loss. The Dice coefficient is an overlap measure and it is shown that the FTL can improve results for this particular metric but we did not observe similar improvements for the AUC and only minor difference in FROC scores when using the FTL compared to the DL, while performance was significantly lower compared to the CE loss. As they do not compare their method to the standard CE or FL it is unclear how using these objectives would affect their results.

DNNs can struggle with generalization, meaning that performance is better on data from the same distribution compared to data from a different distribution, i.e. different population demographics, equipment setups, operators etc. in the case of medical imaging. This is also clear from the results of Table 1 comparing the mean AUC on the E-ophtha and Messidor datasets. In our work we can really only use our models for detection of level 1 DR in the Messidor dataset, as the definition for higher levels also contain other imaging bio-markers such as hemorrhages and neovascularization. Similar to our work, Orlando et al. experimented with an ensemble based approach for red lesion detection in retinal fundus images (Orlando et al., 2018) by combining DNNs with hand engineered features and
random forest. For lesion level detection on the E-ophtha dataset this method achieves a FROC score of 0.3683 when combining all three. Their DNN alone achieves a FROC score of 0.3057 which is lower than our best performing network using CE loss with a score of 0.5067 (± 0.0115). Their algorithm is evaluated on the full E-ophtha dataset and trained on images from two other databases. They also report the sensitivity at a FPAvg value of 1 which they describe as being clinically significant. Here, their method achieves a sensitivity of 0.3680 using the combined features and the DNN alone has a sensitivity of 0.2894. In comparison, our model has a sensitivity of 0.5402 (± 0.0036) at this threshold. As in our work, they apply the algorithm to image-level detection of MAs and DR. For detection of MAs in Messidor images (n=1200) they achieved an AUC of 0.8932 for the combined approach and the DNN alone achieved AUC of 0.7912. As such, their combined approach yields a significantly higher AUC on the Messidor data compared to our best model with an AUC of 0.8038. On the E-ophtha dataset an AUC of 0.9031 is reported for the combined approach and AUC of 0.8374 for the DNN alone. Seeing as there is a difference in data used for training, comparing our method to the ensemble by Orlando et al. can be problematic. They do not train on E-ophtha data and achieve lower scores in all comparable tests on this dataset compared to our approach. Conversely, they achieve better results on the Messidor data. This underlines the problem of generalization when using DNNs for medical image analysis. The data used for training in their work is perhaps more similar to the Messidor data than to the E-ophtha data used in our experiments. A more recent work by Chudzik et al. (Chudzik et al., 2018) train a DNN for detection of MAs. Using an iterative freezing approach they fine-tune a DNN using the Dice loss objective function. They report a FROC score of 0.562 (± 0.233) on 27 images from E-ophtha dataset but the variance of the results and the number of images used for testing renders the results unfit for comparison. In (Orlando et al., 2018) significant improvements are achieved by combining hand crafted features with DNN features. They employ a simple DNN architecture, and in comparison to our DNN it achieves significantly lower scores on both the E-ophtha and messidor data, these results along with those by (Abraham and Khan, 2018) where a deeper architecture lead to an increase in both precision and Dice coefficient indicate that more complex architectures and feature engineering can lead to improved results.

5. Conclusion

Despite the promise of using training objectives designed to deal with unbalanced data, the standard Crossentropy loss perform at least as well or better than all other objective functions in our experiments for lesion level and image level detection for small retinal MAs. While a number of newer objective functions have been introduced and shown to improve performance for unbalanced datasets compared to the Dice loss in recent years, our results suggest that it is important to also benchmark new losses against the Crossentropy or Focal loss function, as we achieve the best performance in all our test using these objectives.

Acknowledgments

References


François Chollet et al. Keras. [https://keras.io](https://keras.io), 2015.


