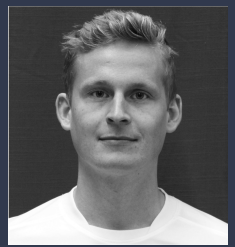
Comparing Objective Functions for Segmentation and Detection of Microaneurysms in Retinal Images

Medical Imaging with Deep Learning, Montréal, 6 - 9 July 2020

Jakob K. H. Andersen[1,3,*]
Jakob Grauslund [2,3]
Thiusius R. Savarimuthu [1]

[1] The Maersk Mc-Kinney Moeller Institute, University of Southern Denmark.
[2] Department of Ophthalmology, Odense University Hospital.
[3] Steno Diabetes Center Odense.

[*] jkha@mmmi.sdu.dk



SDU
UNIVERSITY OF SOUTHERN DENMARK

Region Syddanmark
Steno Diabetes Center Odense

OUH
Odense University Hospital

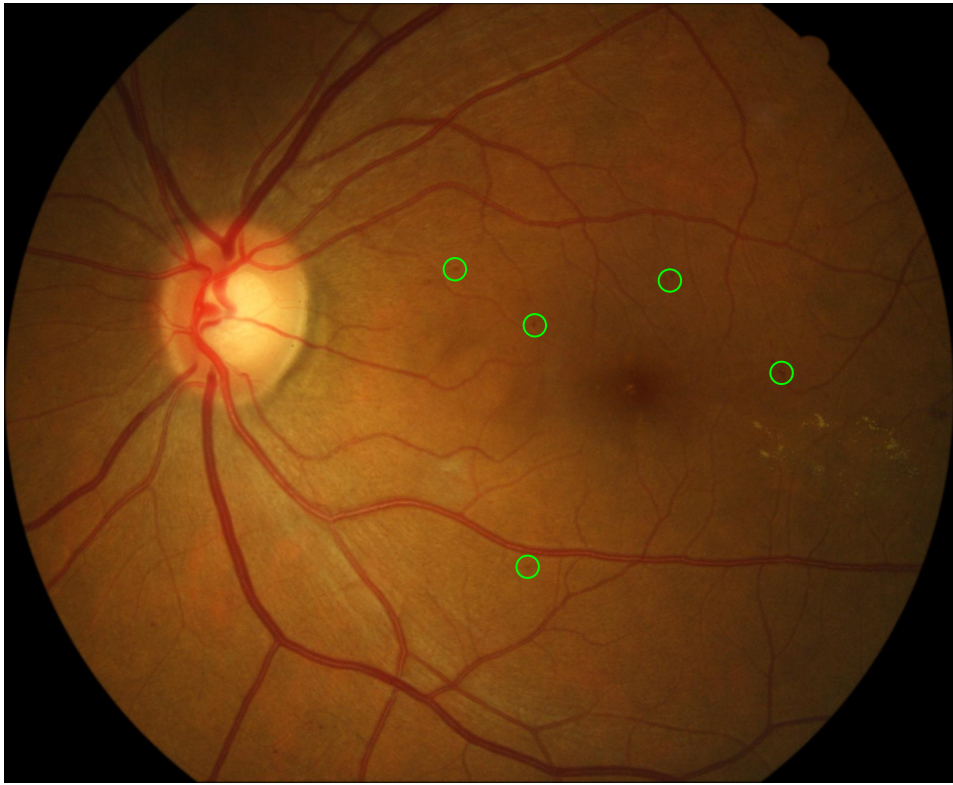
Introduction

SDU UNIVERSITY OF SOUTHERN DENMARK

Steno Diabetes Center Odense
Region Syddanmark

OUH Odense University Hospital

- Retinal microaneurysms (MAs) are the earliest sign of potentially sight threatening diabetic retinopathy (DR).
- MAs account for less than 0.5% of retinal images.
  - (hard to detect).
- MAs indicate the lowest level of DR severity (level 1)
  - ICDR scale 0 - 4 (Wilkinson et al., 2003).
- Deep neural networks have been successfully applied to binary classification of DR.
  - [0, 1] v. [2, 3, 4]
- Less success for full scale classification (Nielsen et al., 2019).
  - Possibly due to microscopic nature of MAs



Introduction

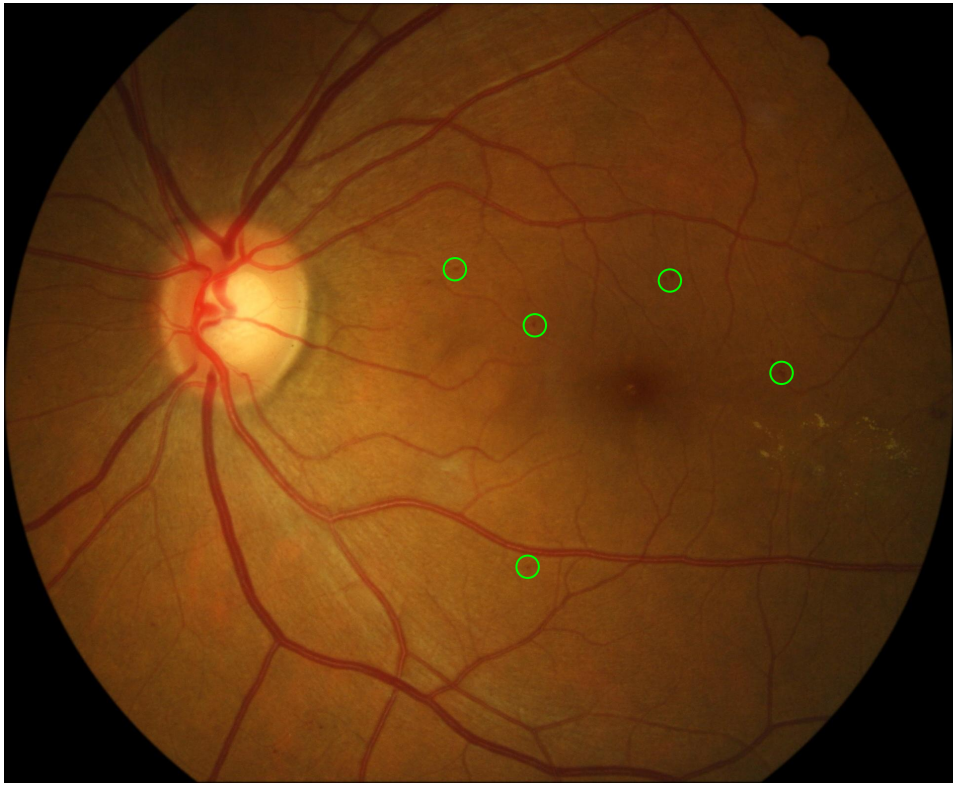
SDU UNIVERSITY OF SOUTHERN DENMARK

Steno Diabetes Center Odense
Region Syddanmark

OUH Odense University Hospital

- MA detection is labour intensive and resource demanding
  - Costly
- Automatic MA detection could lead to:
  - Decreased strain on medical professionals
  - Reduced healthcare expenditure
  - Automatic management of patient referral
    - Most diabetes patients elicit no signs of DR
  - Faster diagnosis
  - Fewer cases of DR related blindness
    - 90% can be eliminated by effective screening

<u>Improved full scale classification?</u>



Introduction

SDU UNIVERSITY OF SOUTHERN DENMARK

Steno Diabetes Center Odense — Region Syddanmark

OUH Odense University Hospital

- DNNs (e.g. U-nets) can learn to segment biomedical image features.
  - Learning network parameters can be affected by class imbalance

- Different loss functions have been proposed as alternatives to standard crossentropy loss (1).
  - Weighted crossentropy (2)
  - Dice loss (Sudre et al., 2017) (3)
  - Focal loss (Lin et al., 2017) (4)
  - Focal tversky loss (Abraham and Khan, 2018) (5)

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

$$CE(p_t, y) = -\alpha log(p_t) (2)$$

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

$$FL(p_t, y) = -\alpha (1 - p_t)^{\gamma} log(p_t) (4)$$

$$FTL_c(p_{i_c}, g_{i_c}) = \left(1 - \frac{\sum_{n=i}^{N} p_{i_c} g_{i_c}}{\sum_{n=i}^{N} p_{i_c} g_{i_c} + \alpha \sum_{n=i}^{N} p_{i_c} g_{i_c} + \beta \sum_{n=i}^{N} p_{i_c} g_{i_c}} \right)^{\frac{1}{\gamma}} (5)$$

Methods

SDU UNIVERSITY OF SOUTHERN DENMARK

Steno Diabetes Center Odense
Region Syddanmark

OUH Odense University Hospital

- Compare DNNs trained for segmentation of retinal MAs.
  - DNNs trained using different objectives to determine the best loss function for segmentation of MAs.
    - Residual U-nets for all experiments (Drozdzal et al., 2016, Zhang et al., 2018).
    - Trained using publicly available retinal images (E-ophtha: Decencire et al., 2013). (n=233)
  - Resulting network segmentation maps used for detection of individual MAs
    - E-ophtha (n=80)
  - As well as for image level detection.
    - Messidor (Decencire et al., 2013, Krause et al., 2017) (n=1287)

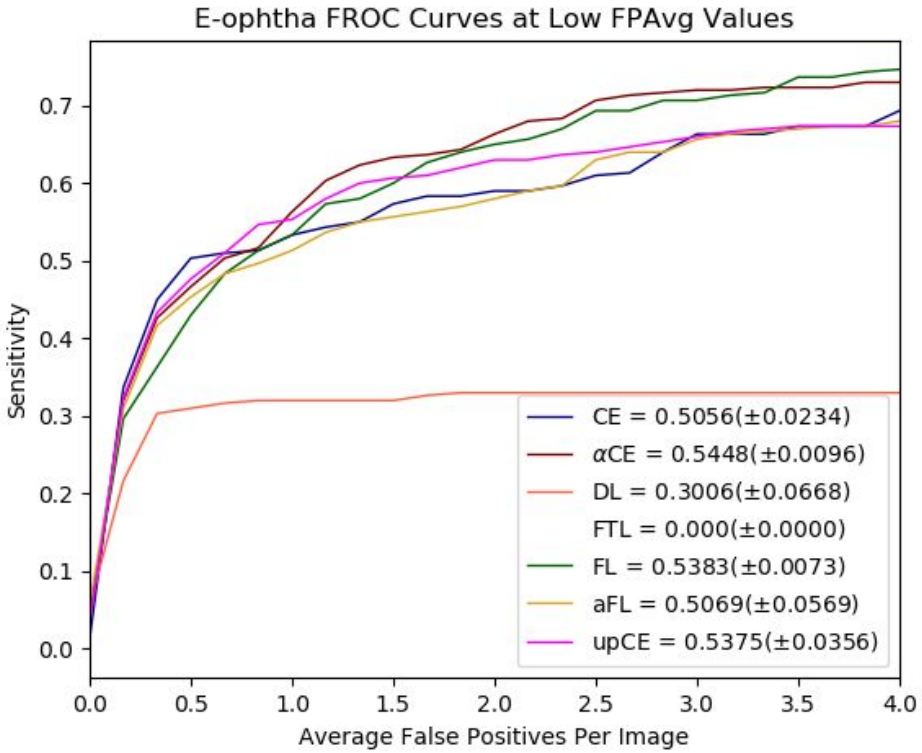
- Evaluation
  - MA detection
    - Free response ROC
      - Mean sensitivity at seven average false positive per image (FPAvg) thresholds of 0.125 , 0.25 , 0.5 , 1, 2, 4 and 8
      - Repeated measures ANOVA with Post-hoc Tukey test
  - MA segmentation
    - Average precision (AP)
  - Image level detection
    - Bootstrapped AUC (95% CI)
  - DR classification
    - Cochrans's Q and Post-hoc McNemar test

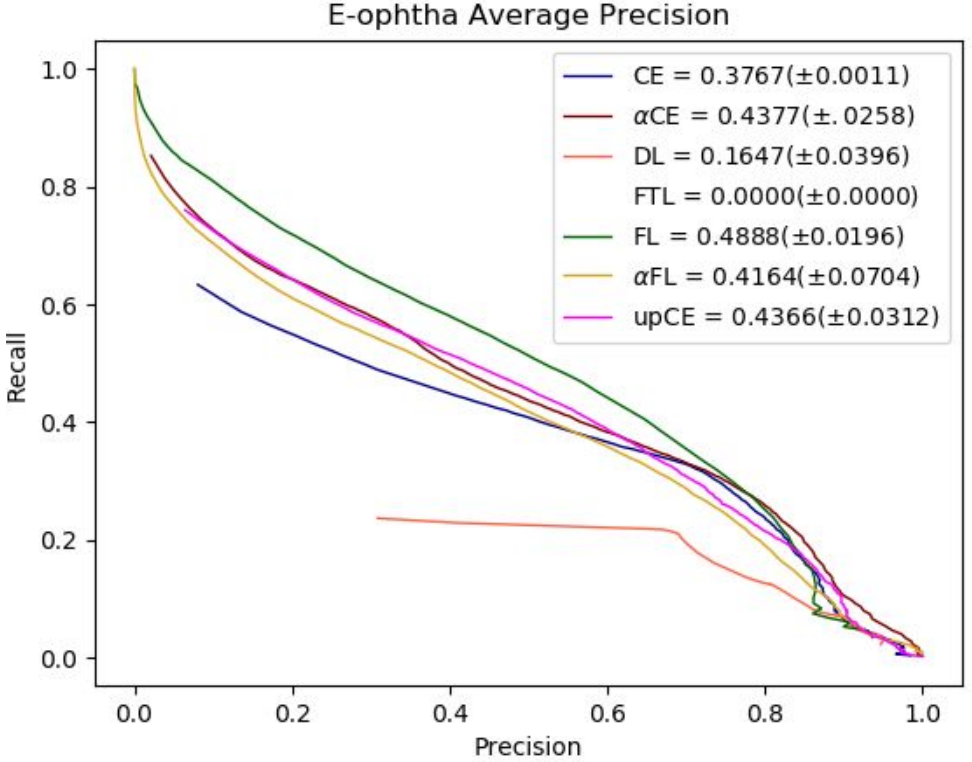
Results

SDU UNIVERSITY OF SOUTHERN DENMARK

Steno Diabetes Center Odense
Region Syddanmark

OUH Odense University Hospital





Results

Table 3: Average FROC, clinically relevant sensitivity scores and average pixel precision on E-ophtha MA test set.

Loss function (parameters)	FROC Score	1.08 FPavg Score	AP score
CE	0.5067(±0.0115)	0.5333(±0.0173)	0.3767(±0.0013)
αCE (α=0.9)	**0.5448(±0.0096)**	**0.5743(±0.0045)**	0.4377(±0.0258)
FL (γ=5)	0.5383(±0.0073)	0.5606(±0.0274)	**0.4888(±0.0196)**
αFL (γ=5, α=0.25)	0.5069(±0.0569)	0.5246(±0.0441)	0.4164(±0.0704)
DL	0.3006(±0.0668)	0.3219(±0.0782)	0.1647(±0.0396)
FTL (γ=1, α=0.7, β=0.3)	0.0000(±0.0000)	0.0000(±0.0000)	0.000(±0.0000)
CE + Upconvolution	0.5375(±0.0356)	0.5689(±0.0178)	0.4366(±0.0312)

Table 4: Bootstrapped AUC and ensemble prediction accuracy on E-ophtha and Messidor test sets.

Loss function	**E-ophtha AUC (95% CI)**	**E-ophtha Accuracy**	**Messidor AUC (95% CI)**	**Messidor Accuracy**
CE	0.978 (0.947 - 1.0)	0.8987	0.671 (0.670 - 0.731)	0.5501
αCE(α=0.9)	0.978 (0.948 - 0.997)	0.8734	0.715 (0.683 - 0.745)	**0.6161**
FL(γ=5)	**0.993 (0.980 - 1.0)**	0.8987	0.701 (0.675 - 0.731)	0.5990
αFL(γ=5, α=0.25)	0.984 (0.954 - 1.0)	**0.9493**	0.720 (0.691 - 0.753)	0.5648
DL	0.993(0.979 - 1.0)	0.9240	0.706(0.675 - 0.742)	0.5897
FTL(γ=1, α=0.7, β=0.3)	0.500 (0.500 - 0.500)	*0.6329*	0.500 (0.500 - 0.500)	*0.7902*
CE + Upconvolution	0.977(0.940 - 1.0)	0.9113	**0.730 (0.683 - 0.745)**	0.5951

Results

- FROC score of 0.5448 (±0.0096) using weighted CE.
  - Orlando et al., 2018:
    - 0.3683
  - Chudzik et al., 2018
    - 0.5620 (±0.2330) on 27 images
  - Savelli et al., 2020
    - 0.4795
- DL performs significantly worse ($p < 0.001$)
- Using the Focal Tversky objective we fail to detect any MAs
- AP of 0.4888 (±0.0196) using same objective

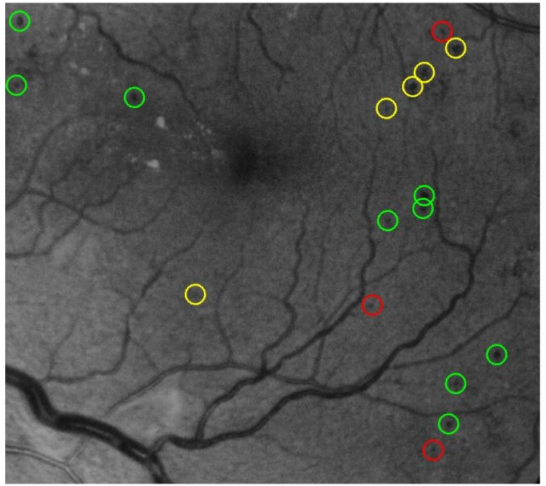
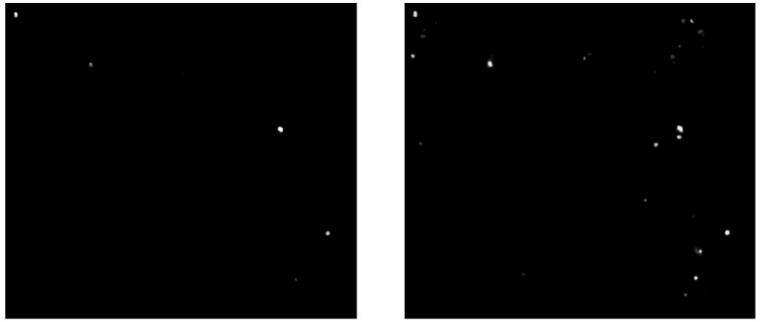


Figure 3: Preproccesed test image with indication of true positive (green), false positive (red) and false negative (yellow) predicted MAs at a clinical relevant 1.08 FPAvg threshold by network trained using weighted Crossentropy objective.

Results

- AUC of 0.993 (95% CI: 0.980 - 1.0) using the Focal loss on E-ophtha images.
- No significant difference for image level detection on the 80 image E-ophtha test set.
  - Excluding the Focal Tversky loss (AUC = 0.5)
- An AUC of 0.730 (95% CI: 0.683 - 0.745) ($p < 0.001$) using model with CE on Messidor images (adjucated ICDR scores)
- AUC of 0.9005 (0.882 - 0.918) (Original R0 vs. R1-R3)
  - Orlando et al., 2018:
    - 0.8932
- AUC 0.8932 (0.874 -0.912) (R0 & R1 v. R2 & R3)
  - Orlando et al., 2018:
    - 0.9374



(a) (b)

Figure 2: Example crops of output probability maps from the same image: (a) Dice loss and (b) weighted Crossentropy.

Conclusion

SDU UNIVERSITY OF SOUTHERN DENMARK

Steno Diabetes Center Odense
Region Syddanmark

OUH Odense University Hospital

- We detect MAs with high sensitivities at low FPAvg:
  - 0.5743 (±0.0054) at 1.08 FPAvg

- Losses based on the Crossentropy (weighted Crossentropy and Focal loss) perform at least as well or better than the DL and FTL despite these being designed to deal with unbalanced data.

  - Results suggest that it is important to benchmark new objectives against losses based on the Crossentropy as we achieve the best performance in all our test using these.

- MA detection can be used to detect low levels of DR
  - AUC of 0.730 (95% CI: 0.683 - 0.745)
    - ICDR level 0 v. level 1

  - AUC of 0.9005 (0.882 - 0.918)
    - R0 vs. R1-R3

  - AUC 0.8932 (0.874 -0.912)
    - R0 & R1 v. R2 & R3

SDU
UNIVERSITY OF SOUTHERN DENMARK

Region Syddanmark Steno Diabetes Center Odense

OUH
Odense
University Hospital

Thank you!