Improving OCT B-Scan classification feature attribution maps with adversarial training

Nathalia Spier, MSc[1]; Ghazal Ghazaei, PhD[2]; Alexander Urich, PhD[2]; Eva Hoeck, MSc[2]; Gary C. Lee, PhD[3]; Niranchana Manivannan, PhD[3]

[1]Carl Zeiss AG, Oberkochen, Germany; [2] Carl Zeiss AG, Munich, Germany; [3]Carl Zeiss Meditec Inc, Dublin, CA, USA.

Poster # 3528259

PURPOSE

Feature attribution methods provide insight into the decision-making process of convolutional neural networks by highlighting pixels that strongly influence the classification decision. Training a network using adversarial examples causes it to emphasize the most relevant image features, resulting in more focused feature maps in comparison to conventional training.

In this study, we investigated if a neural network trained with adversarial attacks produces more robust feature maps for an OCT B-scan classifier.

METHODS

- We trained an Xception network both conventionally and with adversarial attacks for a **binary classification task** (i.e. B-scan is either normal or abnormal). A B-scan is considered "abnormal" if was graded to contain at least one retina pathology as presented by Yu et al. IOVS 2020; 61(9).
- Five different feature attribution methods – **Grad-CAM, SmoothGrad2, VarGrad2, Integrated Gradients, and Vanilla Gradients** – were used to generate feature maps indicating how much each feature in the model contributed to the predictions. The **maps were qualitatively compared** for the two types of training regimes.
- The dataset consisted of **61,058 B-scans** from 478 independent eyes for training and **15,338 B-scans** from 120 eyes for testing.
- The percentage of samples where pathology is present/absent for training and test set was, respectively, 41%/59% and 38%/62%.
- All B-scans were acquired with CIRRUS™ HD-OCT 4000 or CIRRUS™ HD-OCT 5000 devices (ZEISS, Dublin, CA).

CONCLUSIONS

The results consistently show that **models trained with adversarial attacks yield better feature attribution maps**. There is a known trade-off between adversarial robustness and accuracy (as observed in this study), however, certain applications may benefit from incurring a slight drop in accuracy in order to obtain improved feature attribution maps.

Email: nathalia.spier@zeiss.com

Disclosures: NS(E): Carl Zeiss AG, Oberkochen, Germany; GG(E), AU(E), EH(E): Carl Zeiss AG, Munich, Germany; GL (E), NM(E) : Carl Zeiss Meditec Inc, Dublin, CA, USA

RESULTS

- Model trained with **adversarial examples displays clearer** and more focused feature maps, as shown in Figure 1.
- The conventionally trained model obtained an accuracy of 96.14% and AUC of 0.992, while the adversarial model obtained an accuracy of 94.74% and AUC of 0.989.

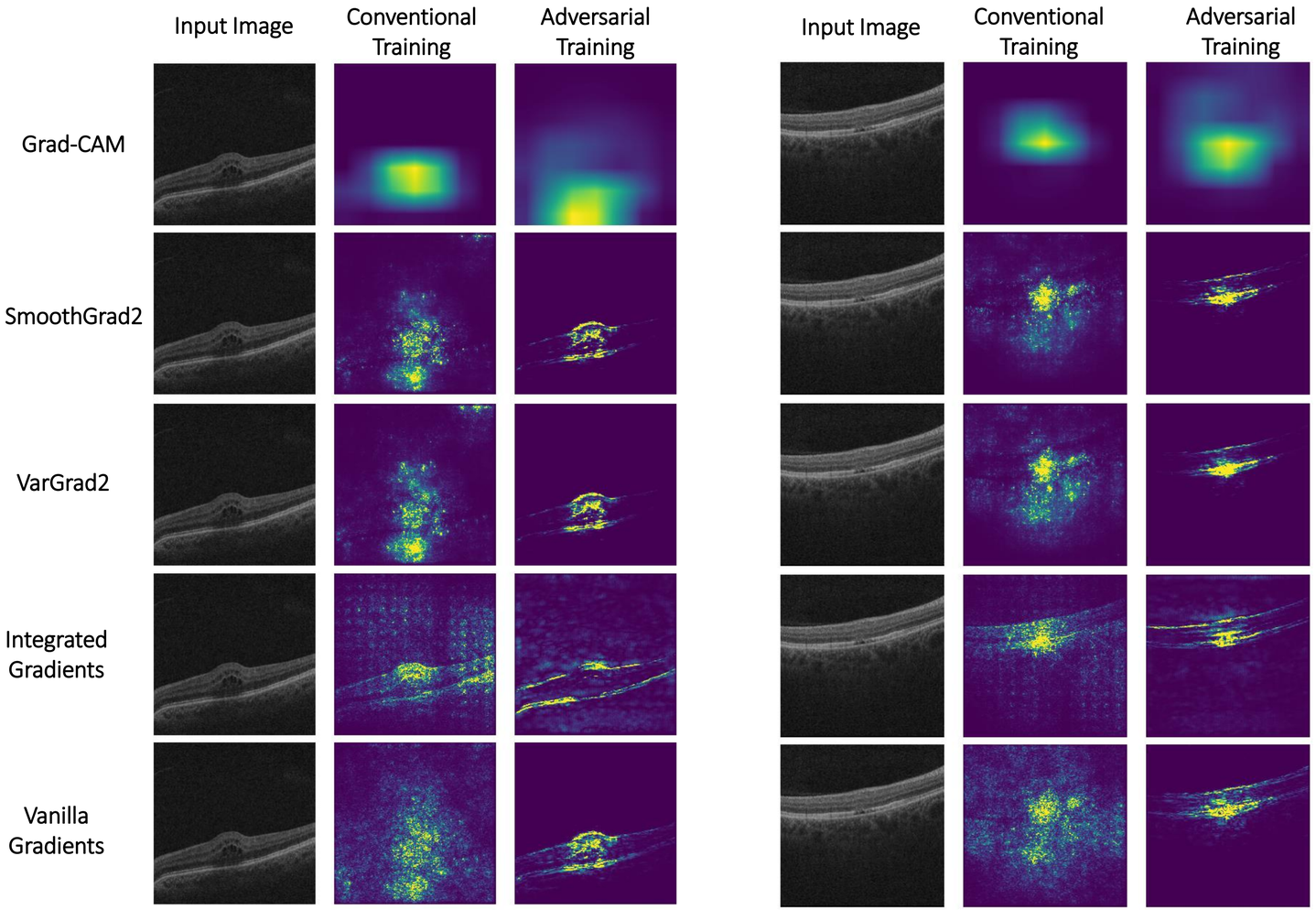

Figure 1: Impact of adversarial training on two different OCT B-scans. Left column: B-Scan image. Middle column: Feature map with conventionally trained model. Right column: Feature map with model trained with adversarial attacks.