Byron Dowling

Cover Letter for Discussion Leader

CSE-60876 (Research Methods)

**Paper**: CYBORG: Blending Human Saliency into the loss improves Deep Learning

**General Topic**: Incorporating Human Saliency into deep learning to increase generalization.

**Specific Behavior or Activity Studied**: The authors describe a method for improving deep learning by collecting annotations from human subjects and blending them into the loss function. This method is then deployed to the task of synthetic face detection.

**Specific Research Questions**:

- “Can deep learning models achieve greater generalization if their training is guided by reference to human perceptual abilities?”
- How can we implement human saliency into the loss function in a practical manner?

**Challenges**:

- Generative Adversarial Networks such as the StyleGAN family have made the task of detecting synthetically generated faces a difficult one.
- If human accuracy is only at random chance level, will this input prove to be useful?

**Paradigm**: The authors take a pragmatic approach in their methodology by replicating an experiment where human subjects are asked to judge pairs of face images and select which image from the pair, they believe is real or synthetic. However, the authors require that participants annotate regions to support their decision. The correctness of the results depends not only on whether the annotator’s decision was correct, but also that the overall percentage of correct selections is better than random chance.

**Problem**: The current state of training deep learning models is more passive than active. Collecting large, annotated data sets and pushing them through deeply connected neural networks to see what comes out the other side is largely the common practice. However, many of these models fail to generalize to unseen data that they are attempting to detect, despite large training sets.

**Importance**: There are many domains where obtaining large quality data sets is simply not possible. It is crucial to develop strategies to increase generalization of model performance from limited data sets. Additionally, involving humans in the learning process can potentially lead to an increase in trust in the correctness of model decisions.

**Claims**: The authors create a loss function called CYBORG loss that compares the model saliency and the saliency of the human annotation and penalizes large differences. The claim is that this loss function should help “coach” the model on the parts of the image that are salient to the problem. Additionally, the authors claim that this loss function is agnostic to model architecture and will increase generalization on unseen data.

**State of Knowledge**: The authors draw inspiration from previous works on the issues of synthetic image detection and using human perception in biometrics. Specifically, the authors cite that while spatial saliency and human saliency have been explored, they have not been directly compared and blended into an overall loss function. Additionally, many attempts at synthetic face detection models are tuned to specific specialized attention modules, however the authors strive to show that CYBORG loss is agnostic to model architecture.

**Evidence**: The authors first compare the accuracy of participants with no annotation requirements vs the results of participants when asked to make annotations. The results showed that requiring annotations increases classification accuracy. Once established, CYBORG Loss was tested against traditional classification loss, models trained with face segmentation masks, and a model with a seven times larger data set across multiple DNN Architectures. The authors then calculate the AUC, TPR, and FPR over the validation set which consists of images from multiple GANs not seen in training.

**Story Structure**: The authors discuss how the current state of deep learning does not always provide increased generalization on unseen data, even with the use of large data sets. Prior work is discussed showing the potential benefits of combining human saliency and model saliency. The authors demonstrate that the requirement of annotations in a classification task is shown to perform better than random chance. Furthermore, this human saliency is crafted into a loss function that is added to multiple DNN architectures and tested against various strategies. Finally, the authors compare the AUC, TPR, and FPR statistics of CYBORG loss against these others strategies showing that CYBORG loss increases generalization in validation testing.