SIIM 2017 Scientific Session
Analytics & Deep Learning Part 2
Friday, June 2 | 8:00 am – 9:30 am

Towards a Better Method of Gradient Descent Based on the Continuous Kappa Loss Function: A Diabetic Retinopathy Example

Njenga P. Wakahiu, MSc, Behold.ai; Jing Chao Lin; Jeet Samarth Raut

Hypothesis

For Computer Aided Detection (CADe) algorithms employing deep learning in medical images, the degree of severity of an abnormality is important factor when performing inference during classifications. Weighted Kappa coefficient is a metric which quantifies inter-rater concurrence (agreement or conversely disagreement) between two raters or judges who classify N items into K mutually exclusive classes. The kappa coefficient statistic is a more robust measure than aggregate percent difference since it takes into account the possibility of one of the rates guessing at random. We posit that a loss continuous kappa loss function based on quadratic weighted kappa metric (QWK) is more appropriate for supervised learning in medical images since they it takes into account how much an expert (for example a trained ophthalmologist, radiologist or a consortium of them) is in agreement with a classification algorithm. Intuitively and by way of example, when grading the severity of disease such as diabetic retinopathy (DR), the order of the labels matters since mild DR is closer to no diabetic retinopathy than proliferative DR.

Introduction

Diabetes is the most common cause of blindness for people aged 20-70 in the world.[1] With developing countries such as India, which had an estimated 50.8 million cases in 2010, and an expected 87 million cases in 2030, it is a major cause for concern.[2] Recent advancements in the field of deep learning have led to research highlighting the ability to automate the detection of diabetic retinopathy. The results had both high sensitivity, and specificity.[3] Object recognition/computer vision tasks have improved greatly with the use of Convolutional Neural Networks for image feature extraction.[4]

CNNs & CNN Layers

Traditional Neural Networks are composed of neurons arranged in a series of layers. They receive as input a vector which is transformed through a series of hidden layers. The components of the input vector are fed into every neuron. The neurons in the hidden layers are fully connected to all the neurons in the previous layer and function independently of those in the same layer. The last (output) layer functions is a class score function.[5] ConvNets take advantage of the fact the inputs are images and sensibly constrains the architecture. Specifically, the layers of a ConvNet are arranged in a 3D manner i.e. H*W*D. Each Neuron is connected to a small volume of neurons in the previous layer instead of being fully connected. Input layer: Receives as input the raw pixel values of the image including (usually) the 3 color channels.

Convolution layer: The CONV layer's parameters consists of a set of learnable filters. Each filter is small (spatially). Each filter is connected to local regions of input but extends over the full depth of the input volume. During the forward pass, we convolve (slide) the filter across the entire image producing a 2D activation map of that filter. We compute the dot product between the filter weights and the input. Intuitively, the network will learn filters that activate when they see a feature at that spatial position. Stacking these activation maps along the depth of the input volume gives full output volume. Every

entry in the output map can be interpreted as an output neuron that only looks at a small region in the input map and shares parameters with regions in the same activation map. Each neuron is connected to only a small local region of the input volume in the spatial extent and fully connected along the depth axis. The spatial extent is a hyperparameter called the receptive field of the neuron. The extent of this connectivity along the depth axis is always constant and equal to the depth.
Rectified Linear Units (RELU) layer: Applies an element wise activation function such as max(0,x) i.e. a thresholding at 0. This leaves the size of the input volume unchanged.

Pool layer: Performs downsampling along the spatial (w*h) dimensions. It is commonly inserted between CONV-layers to reduce the size of the representation and hence the number of parameters and computation and to control overfitting. A commonly used pooling layer is the MAX-pooling with filters of 2*2 and with stride 2 downsizing the input volume by factor of 4.

Loss function:
Supervised learning involves a two component system.
A score function that maps the raw data to scores

A loss function that quantifies the agreement between predicted scores and ground truth labels. Training is an optimization problem in which the loss functions is iteratively optimized with respect to the parameters of the loss function. Typically, the softmax function (or normalized exponential function) is used as the last layer of neural networks. Such networks are usually trained under a log loss (cross-entropy) regime, resulting in a nonlinear variant of multinomial logistic regression. The output of the softmax function is used to represent a categorical distribution i.e a probability distribution over K possible outcomes. It is the gradient log normalizer of the categorical probability.

However, the softmax does not have a way of capturing the agreement between two judges. For medical image classification gradient descent based on the continuous Kappa loss function, provides a better statistic of the disparity between two judges. It is based on the quadratic weighted kappa. The QWK measures the difference between two ratings. It typically ranges between 0 - for completely random agreement between raters and 1 - complete agreement between raters. The metric may be negative if there is less agreement between raters than chance would allow. In the case of DR, it is calculated between scores generated by the machine learning classification method in question and human expert raters.

Methods

A Large Set of high-resolution retina images taken under a variety of imaging conditions was obtained from the Kaggle Diabetic Retinopathy competition. The ground truth is obtained from clinicians who rated the presence of diabetic retinopathy in each image on a scale of 0 to 4. Each image is characterized by a tuple as show.

0 - No DR
1 - Mild
2 - Moderate
3 - Severe
4 - Proliferative DR

The images are obtained from different models and types of cameras resulting in different visual appearances between the right and left eye. Some of the retina images appear as they would anatomically (For the fight eye: macula on the left and optic nerve on the right eye). Others appear as they would through a through a microscope condensing lens.

Two classifiers we trained; one based on the cross entropy loss and another based on the continuous weighted kappa. The classifiers we first trained using the cross entropy loss in the first 10 epochs since the continuous kappa can be unstable for the large loss values.

Figure 1

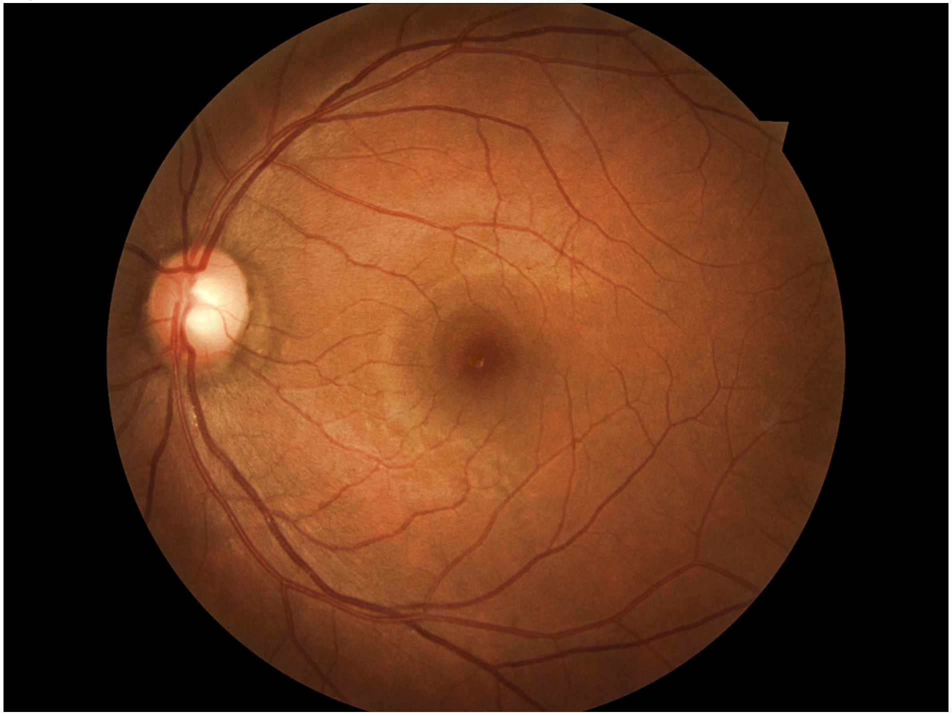

Figure 1. Left

Figure 2

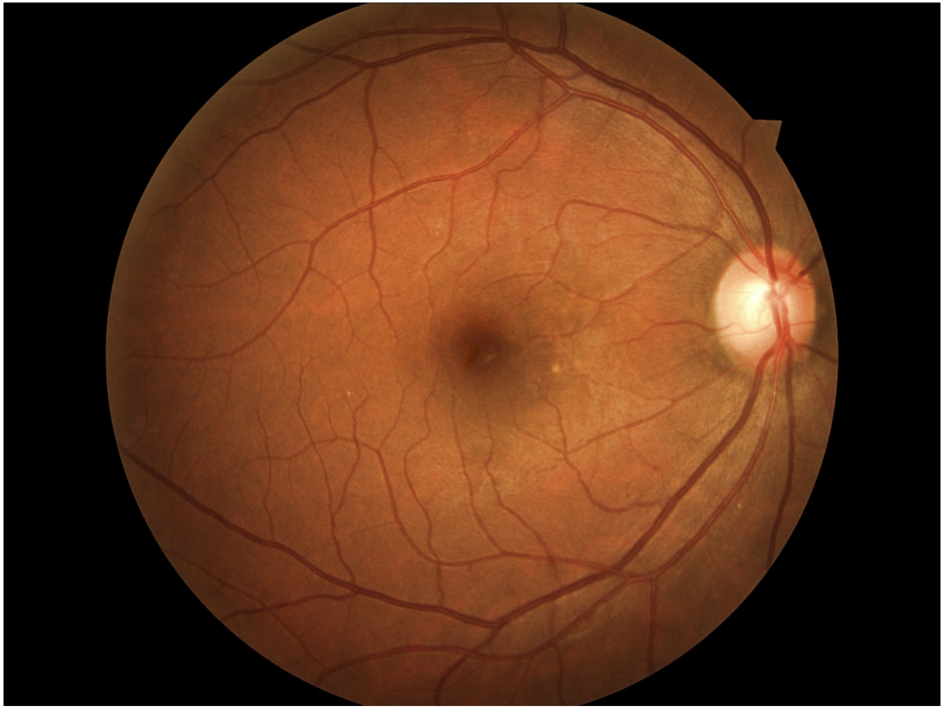

Figure 2. Right

Figure 3

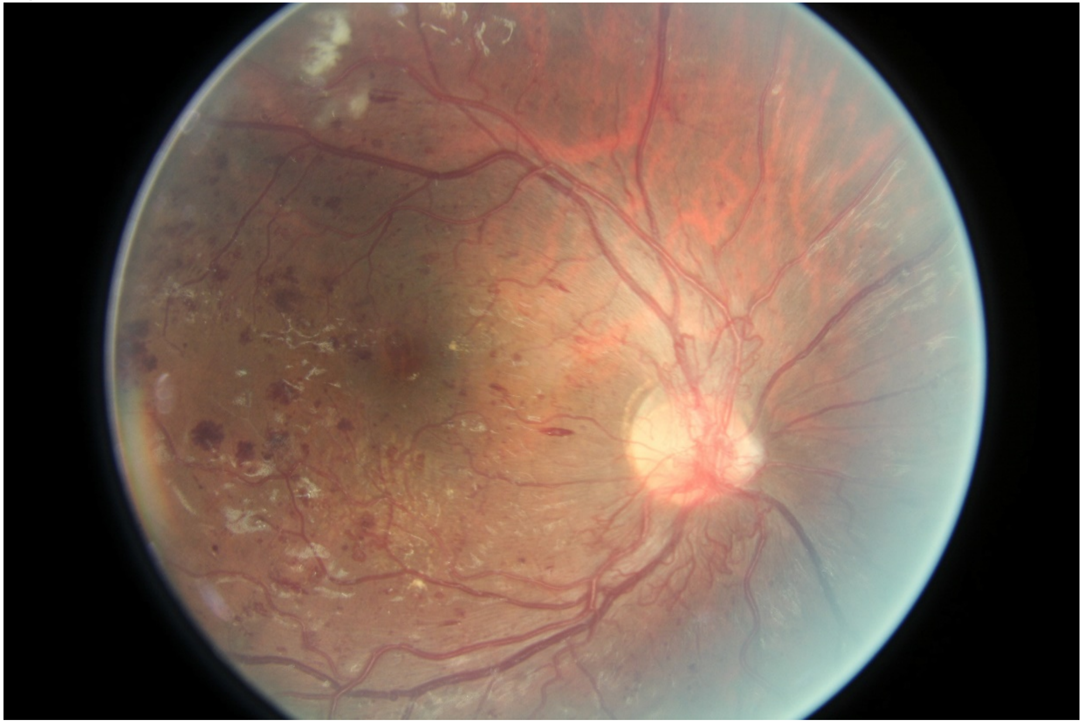

Figure 3. Left

Figure 4

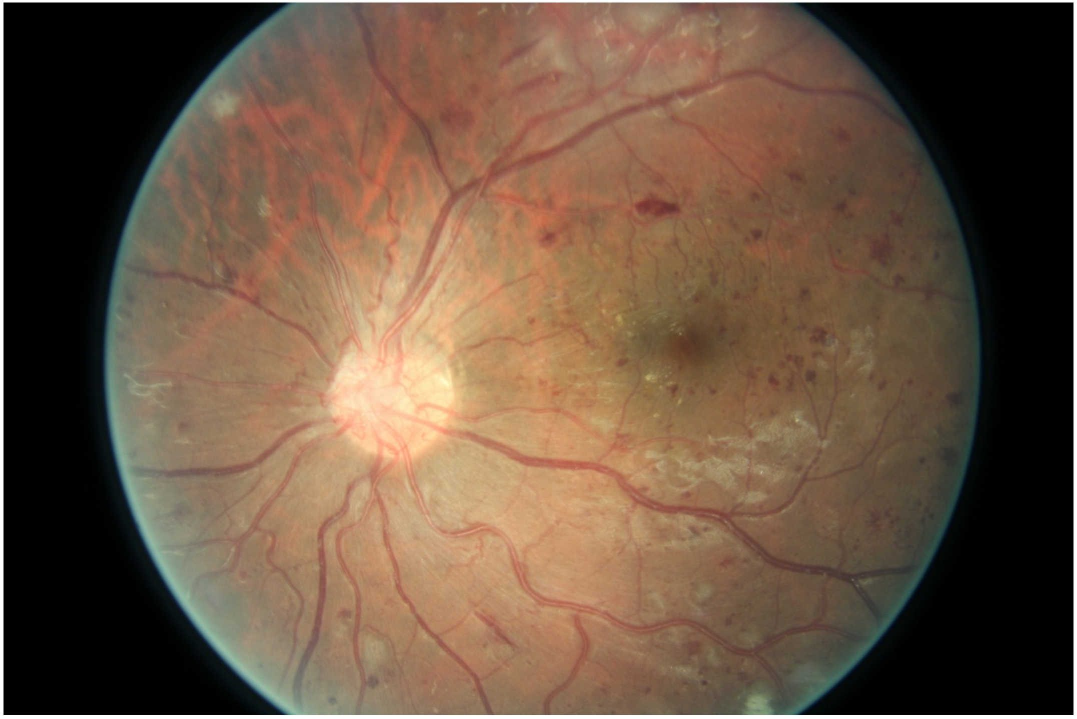

Figure 4. Right

Results

The classifier training based on the continuous weighted kappa resulted in an area under the curve (AUC) of 0.76 (mean 6 standard deviation). The classifier based on the cross entropy loss resulted in AUC of 0.73 on the Kaggle Diabetic Retinopathy detection challenge dataset.

Discussion

For images to be classified based on a scheme in which there is a measure of distance, gradient descent based on the continuous Kappa loss function, provides a better statistic of the disparity between two judges and therefore a better method for gradient descent. The continuous kappa loss function is set up so that the classifier “wants” to perform an inference that is as close an expert (usually a human specialist) as possible. This is because for medical images, the degree of severity of an abnormality is important when performing detection. Notice that it’s sometimes helpful to anthropomorphize the loss functions as we did above to provide a more intuitive understanding for the reason the continuous kappa loss function is apt for classification of medical images.

Conclusion

Employing a continuous weighted kappa loss function provides a higher discriminating power than training under a log loss (or cross-entropy) regime alone.

References

1. Lee, Ryan, Tien Y. Wong, and Charumathi Sabanayagam. "Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss." Eye and Vision 2.1 (2015): 1.
2. Shaw, Jonathan E., Richard A. Sicree, and Paul Z. Zimmet. "Global estimates of the prevalence of diabetes for 2010 and 2030." Diabetes research and clinical practice 87.1 (2010): 4-14.
3. Gulshan, Varun, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." JAMA (2016).
4. LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.
5. Haykin, Simon, and Neural Network. "A comprehensive foundation." Neural Networks 2.2004 (2004).

Keywords

deep learning, computer-aided detection, loss function, continuous kappa loss function, quadratic weighted kappa