- Training a neural network by minimizing a surrogate loss that approximates the target evaluation metric, which may be non-differentiable.
- The surrogate is learned via a deep embedding where the Euclidean distance between the prediction and the ground truth corresponds to the value of the evaluation metric.
- Little attention has been paid to automate the process of designing the loss functions. ED which is used to compare pred with ground truth is non-differential. But the training losses used have only little correlation with eval metric.
- IoU loss was designed but it assumes BBs to be axis aligned. To address issues, surrogate loss was formed. Gradients of eval metric value with respect to the inputs are not required for learning the surrogate (embedding func). Model is tuned with LS.
- Task-specific hand crafted losses have been designed, but they don't generalize. Experiments have shown that learning the loss can outperform hand crafted loss.
- Architecture was manually designed. Surrogate was learned by 2 losses : one to match approx eval metric to actual metric (quality of approximation) ; other loss for stable gradient (quality of gradients) that worked for GAN.
- Source of training the surrogate was local+global. The surrogate parameters Φ are updated first while the model parameters Θ are fixed.
- Quality of approximation is measured by L1 distance which falls below 0.2 for LS-ED. Quality of gradients is judged by performance of $f_\Theta(x)$.
- Global approximation leads to a low quality of the approximation. This can be accounted to the domain gap between the data obtained from the random generator and the model.
- Local approximation leads to a higher quality of the approximation, but quality of gradients is low (surrogate over-fitting on samples obtained from the model and losing generalization capability). Global+local => High quality of Approx+Gradients.
- Scene Text Recognition (STR): Given an input image of a cropped word, the task of STR is to generate the transcription of the word. (a) transformation, (b) feature extraction, (c) sequence modelling, and (d) prediction.
- The transformation module attempts to rectify the curved or tilted text, making the task easier for the subsequent modules of the model. Ex : thin-plate spline (TPS).
- Feature extraction maps image to representation that focuses on the attributes relevant for character recognition, while the irrelevant features are suppressed. Vgg-16 / Resnet.
- Sequence modelling module captures the contextual information. The output character sequence is predicted from the identified features of the image.
- Softmax output of STR model is fed to LS. 1D convs are applied on z (|A|xL). Random generator samples word from corpus and applies random distortions. Domain gap arises as z is softmax output but z_r is one-hot.
- Scene Text Detection. Given a natural scene image, the objective is to obtain precise word-level rotated bounding boxes.It extends Faster-RCNN based object detector to incorporate rotations.It is trained on CE loss (classification) and smooth L1 (BB regress).
- A rotated bounding box is represented with six parameters, two for the coordinates of the centre of the box, two for the height and the width and two for cosine and sine of the rotation angle. The centre coordinates and the dimensions of the box are normalized with image dimensions to make the representation invariant to the image resolution.
- Random generator samples training label BB and distorts it.