Abstract

This paper proposes a technique for training a neural net- work by minimizing a surrogate loss that approximates the target evalu- ation metric, which may be non-differentiable.
We learn the surrogate by deep embedding where euclidean distance between the prediction and ground truth corresponds to the evaluation metric. Post - Tuning setup.

Introduction

Edit Distance - defined by counting unit operations (addition, deletion, and substitution) necessary to transform one text string into another.
ED and IOU are non - differentiable,
The metric is approximated via a deep embedding, where the Euclidean distance between the prediction and the ground truth corresponds to the value of the metric.
The mapping to the embedding space is realized by a neural network, which is learned using only the value of the metric.
In this paper, the focus on a post-tuning setup, where a model that has con-
verged on a proxy loss is tuned with LS.
LS - ED, LS - IOU

Learning Surrogates via Deep Embedding

The tech-nique proposed in this paper addresses the cases when metric e(z, y) is non-differentiable by learning a differentiable surrogate loss denoted as $\hat{e}\Phi(z, y)$. The learned surrogate is realized by a neural network, which is differentiable and is used to optimize the model.

$\hat{e}\Phi(z, y) = ||h\Phi(z) - h\Phi(y)||2$

Learning the surrogate

It is a supervised task which requires 3 major components.

1. Model Architecture
2. Training Loss - there are 2 objectives i) learned surrogate corresponds to the evaluation metric -$\hat{e}\Phi(z, y) \approx e(z, y)$ ii) first order derivative of learned surrogate with respect to prediction z shld be close to 1.
   Parameters of h(phi) are learned by minimizing the loss(the linear combination of the above 2).

3. Sources of Training Data: Global, local, local - global approximation

Algorithm

Aspects for evaluating the surrogates are - quality of approximation of metric, and quality of gradients.

Quality of Approximation - we see it by L1 distance

Quality of Gradients - We rely on the improvement or the decline in the performance of the model $f\Theta(x)$ to judge the quality of the gradients.

Scene Text Recognition

The state-of-the-art architectures for scene text recognition can be factorized into four modules (in this order): (a) transformation, (b) feature extraction, (c) sequence modelling, and (d) prediction.

Post-tuning with LS-ED is investigated for two different configurations of STR models

TPS-ResNet-BiLSTM- Attn. - SOTA