Note: Reducing Information Bottleneck for Weakly Supervised Semantic Segmentation

sgc

May 29, 2022

1 Introduction

This paper mainly proposed 2 ideas. The first method is designed to reduce the information bottleneck in the final layer of the DNN by retraining the DNN without the last activation function. Additionally, we introduce a new pooling method that allows more information embedded in non-discriminative features. The main contributions of this work are:
Authors figure out that the bottleneck occurs mostly in the final layer of the DNN. And authors proposed a method to reduce the bottleneck by simply modifying the existing scheme.

2 Method

2.1 Motivation

Previous works show that although the final feature map of the classifier contains rich information on the target object, the final classification layer filters out most if it.
And they made a toy experiment to analyze the bottleneck problem.

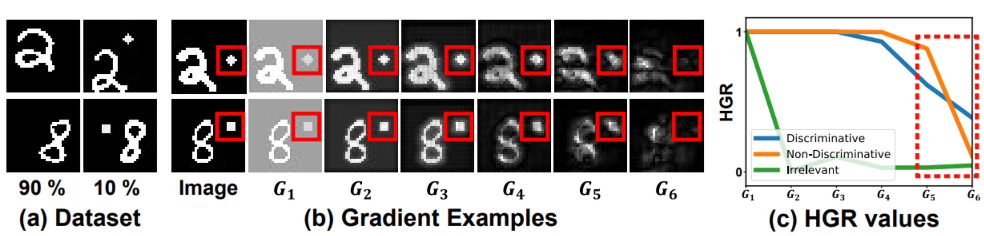

Figure 1: (a) Examples of toy images. (b) Examples of gradient maps G_k. (c) Plot of HGR values of $\mathcal{R}_{\mathrm{D}}$, $\mathcal{R}_{\mathrm{ND}}$, and $\mathcal{R}_{\mathrm{BG}}$ for each layer, averaged over 100 images.

2.2 Reducing Information Bottleneck

Authors deal with this problem by simply removing the sigmoid or softmax function used in the final layer of the DNN. Instead of BCE loss, this proposed method use Loss(RIB):

$$\mathcal{L}_{\text{BCE}} = -\sum_{c=1}^{\mathcal{C}} t_c \log \texttt{sigmoid}(y^c) + (1 - t_c)\log(1 - \texttt{sigmoid}(y^c)),\ \mathcal{L}_{\text{RIB}} = -\sum_{c=1}^{\mathcal{C}} t_c \min(m, y^c),$$

where m is a margin.
The procedure is described as follows:
First train a classifier with $\mathcal{L}_{BCE}$ whose trained weights are denoted by θ_0, and for a given image x, adapt the weights toward a bottleneck-free model of x. Then fine-tune the initial model using $\mathcal{L}_{RIB}$. And to further deal with overfitting, establish a batch of size B for RIB sampling random $B-1$ samples other than x at each RIB iteration. Note that for each iteration, $B-1$ samples are randomly selected, while x is fixed.
To prove that, authors committed a experiment shows that RIB is actually really effective.

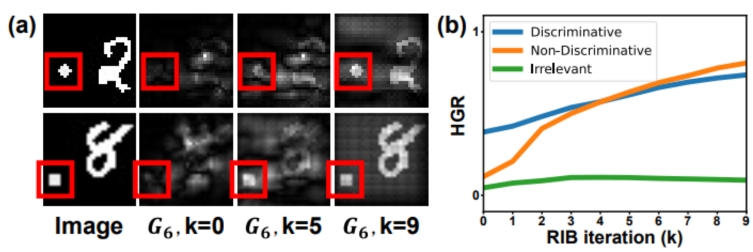

Figure 2: Analysis of G_6 for $\mathcal{R}_{\text{D}}$, $\mathcal{R}_{\text{ND}}$, and $\mathcal{R}_{\text{BG}}$ at each RIB iteration.

Limiting the transmission of information from discriminative regions
Authors also introduce e a global non-discriminative region pooling (GNDRP). Contrary to GAP which aggregates all the values of the spatial location in the feature map, GNDRP selectively aggregates the values of spatial locations whose CAM scores are below a threshold τ , as follows:

$$\text{GAP}(T_l) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} T_l(u),\quad \text{GNDRP}(T_l) = \frac{1}{|\mathcal{U}_\tau|}\sum_{u\in\mathcal{U}_\tau} T_l(u),\quad \mathcal{U}_\tau = \{u \in \mathcal{U}\,|\,\texttt{CAM}(u) \le \tau\},$$

Obtaining a final localization map Obtain the final localization map M by aggregating all the CAMs obtained from the classifier at each RIB iteration k: $\mathcal{M} = \sum_{0\le k\le K} CAM(x;\theta_k)$.

3 Results

The boundary information provided by saliency supervision allows this method to produce a more precise boundary (yellow boxes). However, the non-salient objects in an image are often ignored when using saliency supervision, while RIB successfully identifies them.

4 Ablation study

1, influence of the total number of RIB iteration is limited, and even slightly decrease after $k = 5$. But the proposed GNDRP alleviated that.
2, Fine-tuning with $\mathcal{L}_{RIB}$ actually helps a lot.
3, RIB is robust against hyperparameters if m, λ is relatively large.

5 Summary

First this paper points out the bottleneck hindered the representative ability of CAMs. And proposed a new method to reduce the information bottleneck through two simple modifications to the existing method.

References