DeepSTLD: Deep Neural Network and Focal Regression Loss for Small Traffic Lights Detector

박경민, 김대진 *포항공과대학교 컴퓨터공학과*

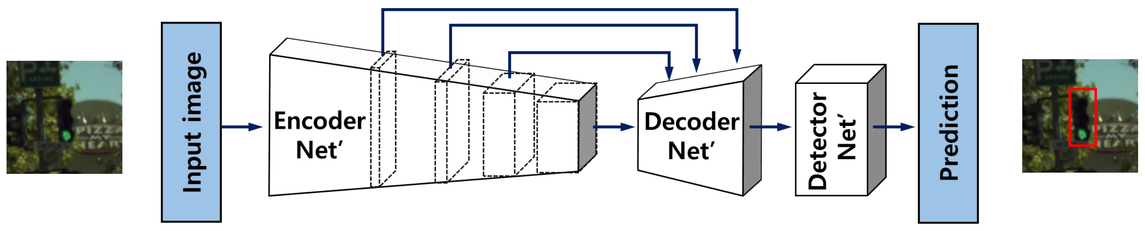
한국정보과학회 KCC 2018

Abstract

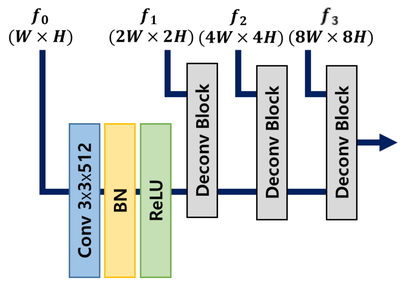
- **Main Objection**
  - Vision based traffic light detection
- **Problem**
  - Traffic lights are too small
    - A traffic light occupies only 12×4 px in 1280×720 image
  - (When using one-stage detector such as YOLO and SSD) Too large amount of background dominate training process
- **Our Approaches**
  - DeepSTLD
    - Deep neural network for small traffic light detector
    - YOLOv2 based, encoder-decoder hourglass structure
  - Focal Regression Loss
    - Focal loss based loss function for regression
    - We substitute L2 in original YOLOv2 loss with focal regression loss

DeepSTLD

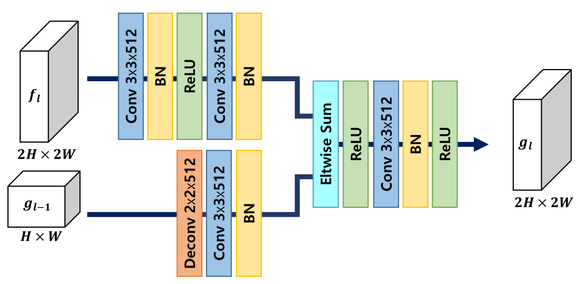
- **Overview**



- **Encoder Network**
  - Encode an input image to feature maps
  - We used *ResNet-101* as the encoder network
- **Decoder Network**

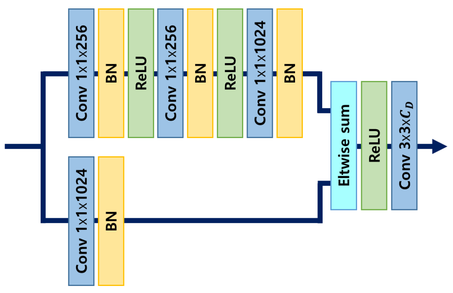


Decoder network

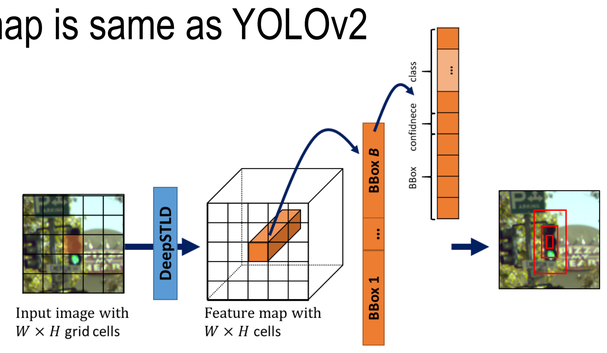


Deconvolutional block

  - Decode feature maps from the proceeding encoder network
    - Upsample late feature maps of the encoder network by deconvolutional process
    - Combine upsampled feature maps with early feature maps of the encoder network
  - The *result feature map* has *detailed* information in the early feature maps as well as *contextually strong* information in the late feature maps
- **Detector Network**
  - Predict *bounding boxes, confidences, class probabilities* from the result of the proceeding decoder network
  - Configuration of final feature map is same as YOLOv2



Detector network



Configuration of final feature map

Focal Regression Loss

- **Focal Regression Loss**
  - Reduce loss of easy examples
  - Most of backgrounds are easy example
  - By reducing loss of easy examples, backgrounds do not dominate training process
  - $\mathcal{L}^{FR}(\boldsymbol{p}, \boldsymbol{q}) = -|p - q|^{\gamma} \log(1 - |p - q|)$
    - $\boldsymbol{p} \in [\mathbf{0}, \mathbf{1}]$: regressed value
    - $\boldsymbol{q} \in [\mathbf{0}, \mathbf{1}]$: regression target
    - $\boldsymbol{\gamma} \geq \mathbf{0}$: focusing parameter
    - $|\boldsymbol{p} - \boldsymbol{q}|^{\boldsymbol{\gamma}}$: modulating factor
- **Training DeepSTLD with focal regression loss**
  - We *substitute L2 loss for confidence regression* in YOLOv2 with focal regression loss.
  - Loss of DeepSTLD for confidence regression $\mathcal{L}_{obj}$

$$\mathcal{L}_{obj} = \lambda_{obj} \sum_i \sum_j \mathrm{I}_{ij} \mathcal{L}^{FR}\left(\sigma\left(p_{ij}^{conf}\right), t_{ij}^{conf}\right)$$

$$+\lambda_{noobj} \sum_i \sum_j (1 - \mathrm{I}_{ij}) \mathcal{L}^{FR}\left(\sigma\left(p_{ij}^{conf}\right), 0\right)$$

  - $\boldsymbol{\lambda_{obj}, \lambda_{noobj}}$: weights for foreground and background respectively
  - $\boldsymbol{p_{ij}^{conf}}$: confidence of the bounding box which is predicted by the j-th anchor box at the i-th grid cell
  - $\boldsymbol{t_{ij}^{conf}} = IOU(predicted\ bbox, target\ bbox)$: regression target when foreground
  - $\boldsymbol{I_{ij}}$: indication function for foreground

Experimental Results

- **Dataset**
  - Bosch Small Traffic Lights Dataset
    - 5093 training images, 8334 test images
    - *Median width of traffic lights : 8.5px*
- **Experimental model**
  - $deconv$: 3 deconvolutional blocks, original YOLOv2 loss (L2)
  - $deconv + frl$: 3 deconvolutional blocks, focal regression loss
- **Result**

  - mAP

model	$IOU \geq 0.5$	$IOU \geq 0.3$
$deconv$	0.5021	0.6850
$deconv + frl$	0.5641	0.7871