Resolving Semantic Confusions for Improved Zero-Shot Detection

Sandipan Sarma, Sushil Kumar, Arijit Sur
Department of CSE, Indian Institute of Technology Guwahati, India

Motivation

Impossible to collect images of all possible objects, especially rare, for training supervised object detection models.

Imagery for fine-grained annotations is time-consuming. Supervised detection models cannot adapt to evolving object appearances, like futuristic cars.

Fine-grained annotations require expert analysis.

Semantic Confusion: Knowledge transfer in existing ZSD models not discriminatory enough to differentiate between objects with similar attributes (semantics) like car and train.

Low average precisions can misguide applications like marine debris detection.

Proposed Framework

Conditional WGAN

- Trained mapper function used to ensure visual-semantic consistency
- Cyclic-consistency loss reduces bias of generated features towards seen classes

Visual feature of kite

Mapper function

Semantic feature of kite

ResNet-101 (Feature extractor)

Random noise

FGAN

Bounding box regressor

RPN

Faster-RCNN

Classifer

Update classifier for unseen classes

Diversification

Triplet module

Training time

Classification loss encouraging discriminative feature generation

Mode-seeking loss prevents GAN from falling into a mode collapse

Incorporating inter-class dissimilarity within the generated object features

Margin of dissimilarity (Δ) computed on the basis of mean and variance of class semantics

Zero-shot Detection (ZSD) models can alleviate all these issues.

Experiments

Datasets

- 80 object categories with bounding box annotations
- Seen/Unseen classes = 65/15
- Train/test images = 62,300/10,098

- 20 object categories with bounding-box annotations
- Seen/Unseen classes = 16/4
- Train/test images = 5,981/1,402 + 4,836

Source of inputs

- Attribute (Semantic) source
- 300 dimensional semantic vector per object class

Visual feature source

- Pre-trained ResNet-101 trained on ImageNet, excluding classes common with unseen objects of MS-COCO and PASCAL VOC, as per zero-shot criterion

Results

Mean Average Precision (in %) with MS-COCO

Method | Year | ZSD | Seen | GZSD | Unseen | HM
--- | --- | --- | --- | --- | --- | ---
PL | 2018 | 12.40 | 34.07 | 12.40 | 18.18
BLC | 2020 | 14.70 | 36.00 | 13.10 | 19.20
ACS-ZSD | 2020 | 15.34 | - | - | -
SUZOD | 2020 | 17.30 | 37.40 | 17.30 | 23.65
ZSDTR | 2021 | 13.20 | 40.55 | 13.22 | 20.16
ContrastZSD | 2022 | 18.60 | 40.20 | 16.50 | 23.40
Ours | 2022 | 20.10 | 37.40 | 20.10 | 26.15

Recall@100 with MS-COCO

1

Effect of number of generated features on mAP per class on mAP with MS-COCO

1

Takeaways

- Visual-feature based triplet strategy for addressing inter-class dissimilarities, along with cycle-consistency between generated visual features of an object and its semantic vector, encourages a generative network to synthesize discriminative features, reducing semantic confusion

- False-positive rate and misclassification of localized objects (especially unseen objects) can be addressed as a by-product of resolving semantic confusion

- With conditional feature generation, performance usually stays low for unseen objects that are semantically very different from every seen object class