Hey, existing datasets have strong scene biases. Maybe we need to collect a even larger dataset?

Why Can’t I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition

Introduction

Hey, existing datasets have strong scene biases. Maybe we need to collect a even larger dataset?

Okay, how are you going to avoid the spurious association between scene and action?

Main Idea

Idea 1: Make a model unable to predict the correct scene category

Idea 2: Make a model unable to predict the correct action class w/o evidence

Experimental Results

Action Classification

Method | Backbone | Top-1 accuracies on UCF-101 | HMDB-51 | DivY48
---|---|---|---|---
Baseline | 3D-ResNet-18 | 83.5 (+1.0) | 53.6 (+3.1) | 18.9 (+2.8)
Ours | 3D-ResNet-18 | 84.5 | 56.7 | 20.5 (+2.5)
C3D | C3D | 82.3 | - | -
Factor-C3D | C3D | 84.5 | - | -
RESOUND-C3D | C3D | - | 16.4 | -
TSN | BN-Inception | 85.1 | 51.0 | 16.8

Method | Inputs | Backbone | Top-1 accuracies on UCFT-101 | HMDB-51 | Diving48
---|---|---|---|---|---
Baseline | RGB | 3D-ResNet-18 | 48.6 | 45.6 | 32.5 (+0)
Ours | RGB | 3D-ResNet-18 | 50.2 | 47.9 | 32.4 (+0.5)
C3D | C3D | 59.8 | 48.3 | 29.8 (+0)
Factor-C3D | C3D | - | 19.9 | 10.7 (+0)
RESOUND-C3D | C3D | - | 16.4 | -
TSN | BN-Inception | 66.0 | 59.4 | 21.9 (+0)

Method | Inputs | Backbone | Top-1 accuracies on UCFT-101 | HMDB-51 | Diving48
---|---|---|---|---|---
Baseline | RGB | 3D-ResNet-18 | 83.5 (+1.0) | 53.6 (+3.1) | 18.9 (+2.8)
Ours | RGB | 3D-ResNet-18 | 84.5 | 56.7 | 20.5 (+2.5)
C3D | C3D | 82.3 | - | -
Factor-C3D | C3D | 84.5 | - | -
RESOUND-C3D | C3D | - | 16.4 | -
TSN | BN-Inception | 85.1 | 51.0 | 16.8

Temporal Action Localization

Method | Input | Backbone | mean AP @ IoU threshold 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | avg
---|---|---|---|---|---|---|---|---|---|---
Baseline | RGB | 3D-ResNet-18 | 48.6 | 45.6 | 40.8 | 32.5 | 25.5 | 15.5 | 36.7 (+0)
Ours | RGB | 3D-ResNet-18 | 50.2 | 47.9 | 42.3 | 32.4 | 26.3 | 16.8 | 38.2 (+1.5)
C3D | C3D | 59.8 | 51.9 | 48.5 | 32.7 | 28.4 | 20.8 | 45.1
Factor-C3D | C3D | - | 19.9 | 10.7 (+0)
RESOUND-C3D | C3D | - | 16.4 | -
TSN | BN-InceptionV3 | 66.0 | 59.4 | 51.9 | 41.0 | 29.8 | 19.6 | 32.3 (+0)

Spatio-Temporal Action Detection

Method | Inputs | Backbone | Pre-train on | mean AP
---|---|---|---|---
Baseline | RGB | VGG | ImageNet+MiniKinetics | 32.5 (+0)
Ours | RGB | VGG | ImageNet | 34.5 (+2.9)
ACT | RGB+Flow | VGG | ImageNet | 65.7
S3D-G | RGB+Flow | Inception (2+1)D | ImageLabel+FullKinetics | 75.2

Ablation Studies

| Method | Split-1 | Split-2 | Split-3 | avg.
---|---|---|---|---
Baseline | None (w/o debiasing) | 52.9 | 55.4 | 52.6 | 53.6 (+0)
Hard | 54.8 | 54.4 | 54.6 | 54.5 (+0.8)
Soft (ours) | 56.4 | 57.3 | 56.4 | 56.2 (+2.6)
Soft (ours) | 56.4 | 57.3 | 56.4 | 56.2 (+2.6)

Visualization

Why Can’t I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition

Jinwoo Choi | Chen Gao | Joseph C. E. Mesou | Jia-Bin Huang
Virginia Tech

Code available at: