HICO: A Benchmark for Recognizing Human-Object Interactions in Images

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng
ICCV 2015

Presented by
Chia-Wen Cheng, Chia-Cheng Hsu

HICO

~47,000 labeled images in 600 human-object interaction (HOI) categories

<u>Object-Verb</u>

sports ball - block X

sports ball - carry V

sports ball - hold V

sports ball - sign X

wine glass - fill ?

apple - peel ?

....

Human-Object Interaction Prediction

Horse-Ride
Horse-Sit on

Evaluate the best proposed model

Pipeline of the DNN Model

Pretrained on ImageNet

AlexNet

feature vector

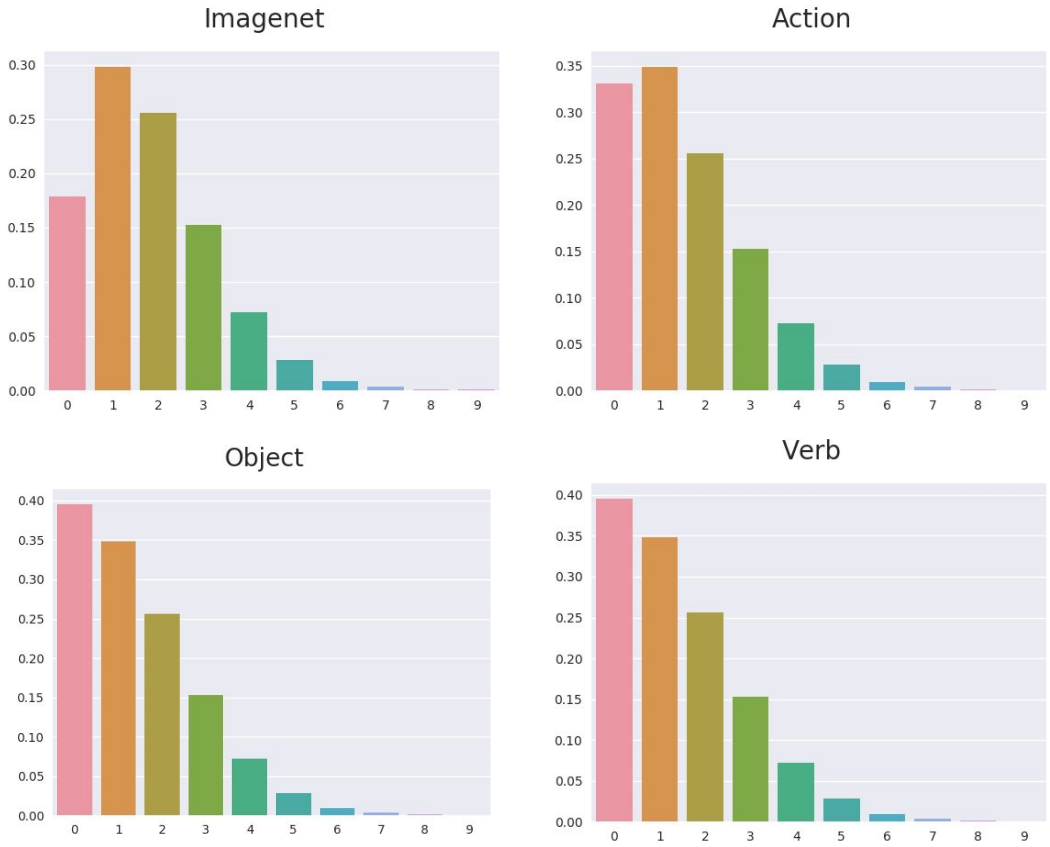
binary SVM per category

SVM

SVM

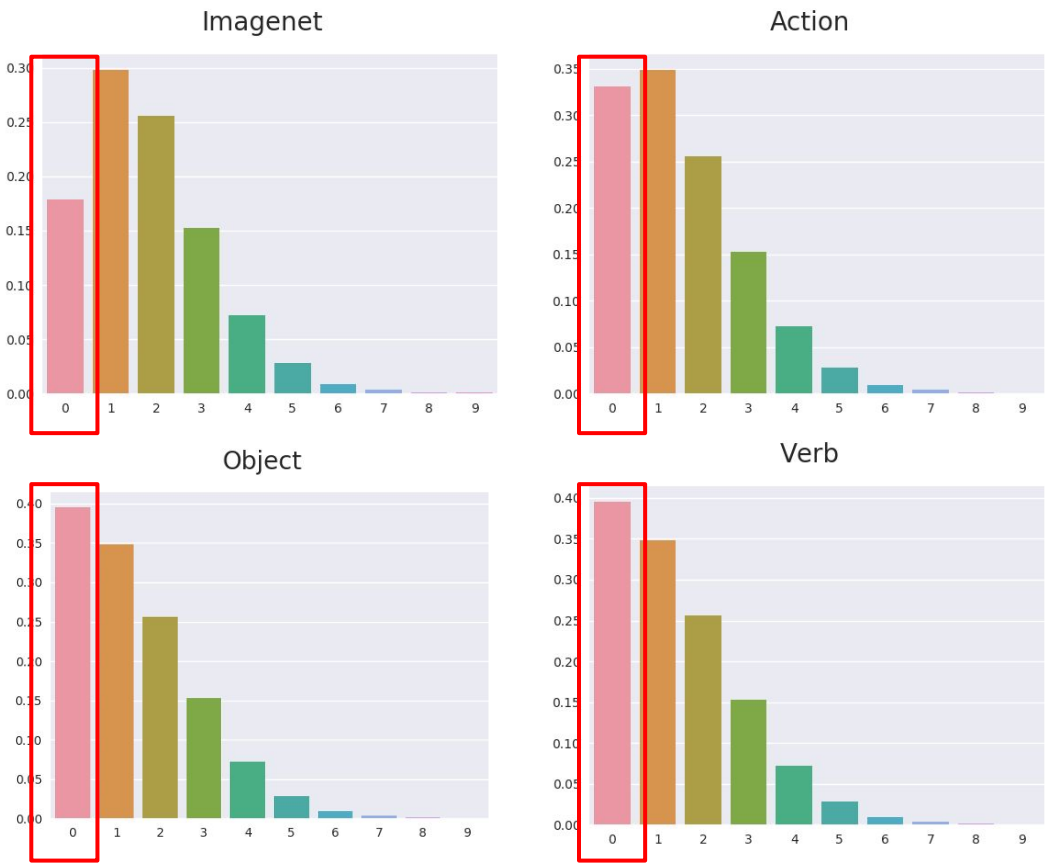
SVM

Weird Output Distribution



x-axis: number of prediction labels
y-axis: % of testing sets

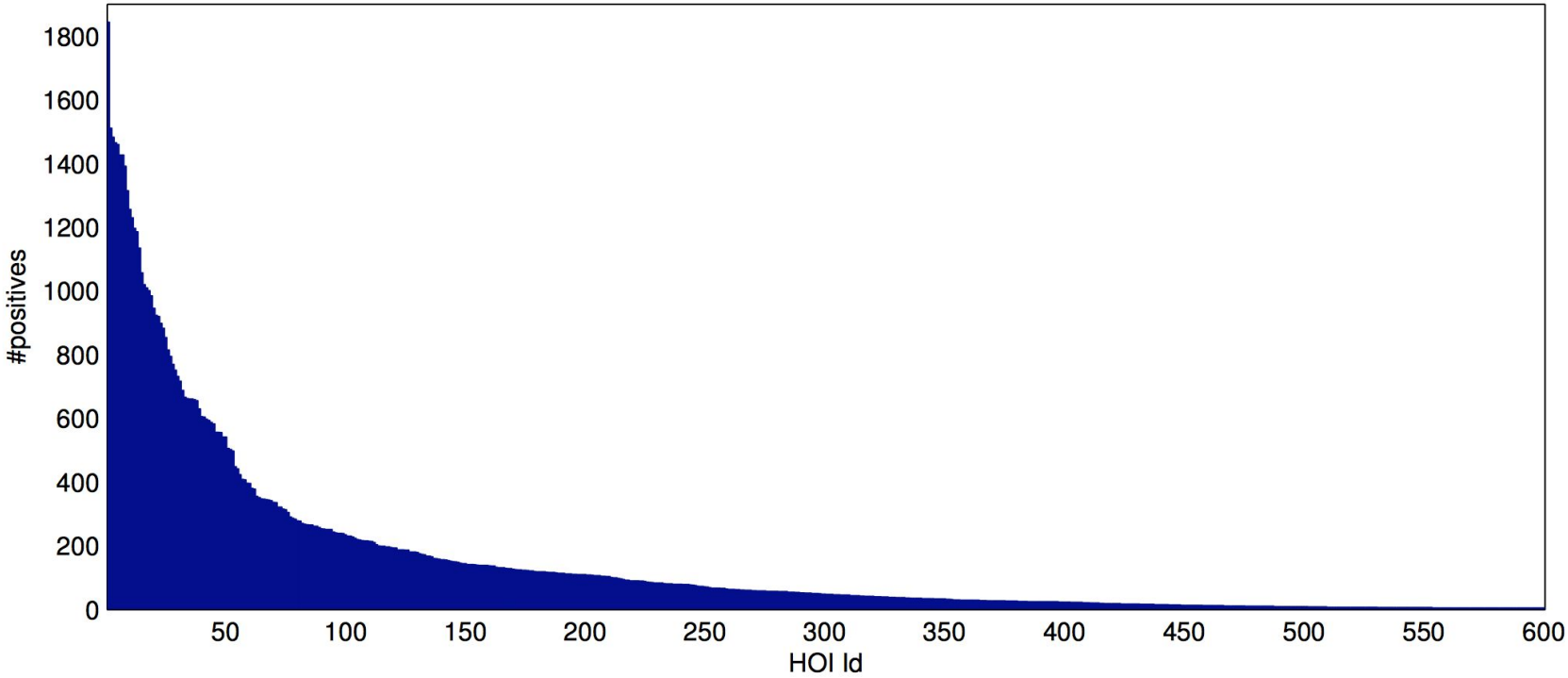
Weird Output Distribution



x-axis: number of prediction labels
y-axis: % of testing sets

A lot of testing images are not predicted as any category.

Long Tail Distribution of Categories


#positives
1800
1600
1400
1200
1000
800
600
400
200
0
50
100
150
200
250
300
350
400
450
500
550
600
HOI Id

Weighted Loss for Unbalanced Dataset

Binary Classifier for Class 1

Positive Sample	Negative Sample
Class 1	Class 2, 3, …,600

Total Loss = w_p * loss on positive samples + w_n * loss on negative samples

Experiments on w_p/w_n

w_p/w_n	**mAP (%)**
1	18.58
3	19.05
10	19.39
30	19.24

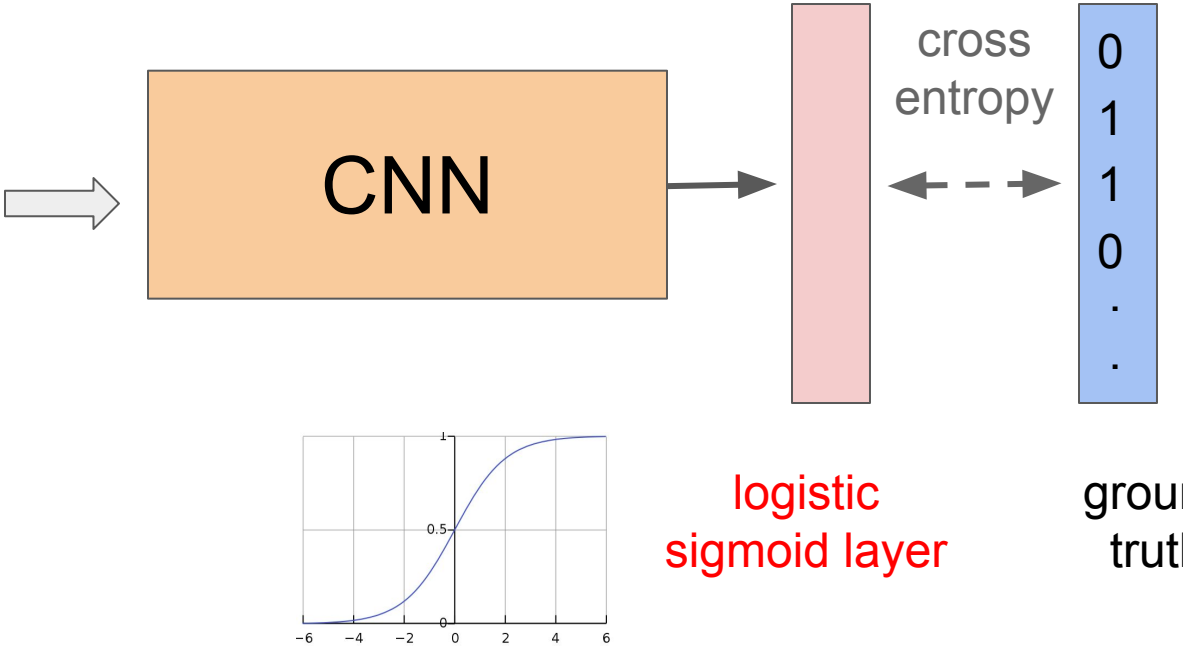
Experiment on w_p/w_n

w_p/w_n	mAP (%)
1	18.58
3	19.05
10	19.39
30	19.24

	mAP
Random	0.57
RF [36]	7.30
FV [29]	4.21
DNN (ImageNet)	18.58
DNN (fine-tune V)	17.65
DNN (fine-tune O)	**19.38**
DNN (fine-tune VO)	18.08
HOCNN	4.90

Our Implementation: End-to-End Network

Multi-Label Classification



Experimental Setting

CNN Model:

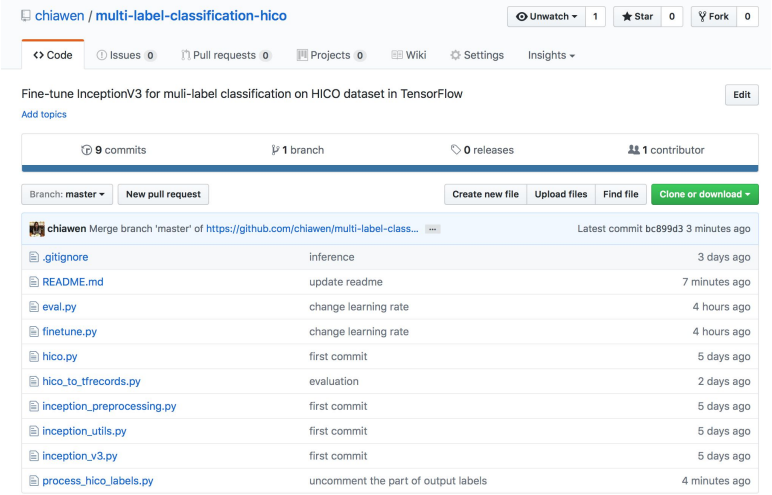
- Inception v3
- softmax layer -> logistic sigmoid layer
- number of classes -> 600

Training:

- Use pretrained model on ImageNet
- Fine-tune only the last layer
- Optimizer: Adam
- Learning rate: 0.001
- Batch size: 64
- Epochs: 10

Source Code

- Implemented in TensorFlow
- TF-Slim Library
- Github: https://github.com/chiawen/multi-label-classification-hico



Usage

First, extract the filenames and labels of the training set and the testing set.

```
$ python process_hico_labels.py
```

Second, convert the image files and annotations to TFRecords.

```
$ python hico_to_tfrecords.py
```

To fine-tune the last layer of Inception v3 for 10 epochs:

```
$ python finetune.py
```

To evaluate mAP scores on the testing set:

```
$ python eval.py
```

Performance

Method	mAP (%)
DNN (fine-tune O)	19.38
DNN (ImageNet) + weighted loss (ours)	19.39
Inception V3 + fine-tune (ours)	**26.31**

Related Work

Performance of HICO Benchmark

Arun Mallya and Svetlana Lazebnik. Learning Models for Actions and Person-Object Interactions with Transfer to Question Answering. In *ECCV*, 2016.

	Method	Full Im.	Bbox	MIL	Wtd. Loss	mAP
a)	AlexNet+SVM [16]	✓				19.4
	VGG-16, full image	✓				29.4
b)	VGG-16, bounding box		✓	✓		14.6
	VGG-16, R*CNN		✓	✓		28.5
	VGG-16, Scene-RCNN	✓	✓	✓		29.0
c)	Fusion-1	✓	✓	✓		33.6
	Fusion-1, weighted loss	✓	✓	✓	✓	36.0
	Fusion-2	✓	✓	✓		33.8
	Fusion-2, weighted loss	✓	✓	✓	✓	**36.1**

Method	mAP (%)
DNN (fine-tune O)	19.38
DNN (ImageNet) + weighted loss (ours)	19.39
Inception V3 + fine-tune (ours)	**26.31**