Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels

Ishan Misra C. Lawrence Zitnick Margaret Mitchell Ross Girshick

Carnegie Mellon University

Facebook AI Research

Microsoft Research

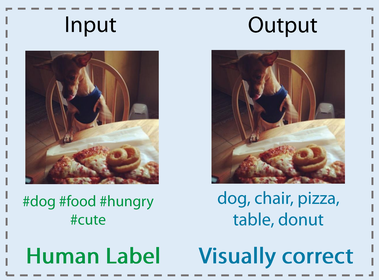
Overview

Goal
- Use noisy human labels
- Learn correct classifiers

Key Ideas
- Estimate noise from data
- Factor predictions

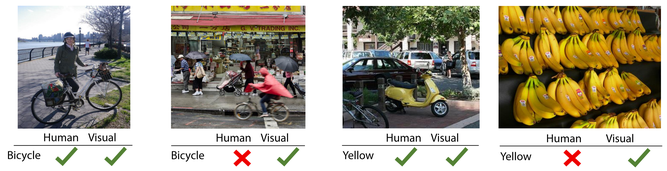
Properties
- No "clean" labels
- Only human labels

Outcome
- Presence: What is visually present
- Relevance: What to say and when



What is "human-centric"* noise?

*Berg et al. 2012

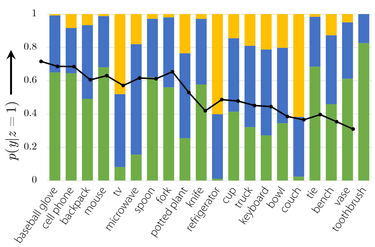


How noisy are these labels?

MSCOCO Dataset
- Caption Label y
- Visual Label z
- Label noise $p(y|z=1)$ shown by the black line

Small Medium Large

Train Labels ⟶	y	z
mAP on z	63.7	76.3



Factoring human-centric predictions

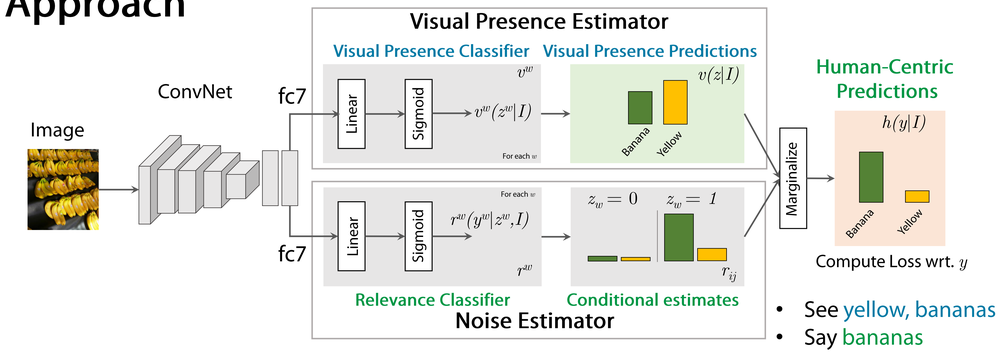
Visual Presence: v
- Is it "visually present"?
- Noise-free
- Depends on the image

Banana ✓ Yellow ✓

Relevance: r
- Is it "relevant" for a human?
- Models Noise/Human bias
- Depends on concept & image

Banana ✓ Yellow ✗

Approach



- See yellow, bananas
- Say bananas

Marginalize: $h(y|I) = \sum_{j \in \{0,1\}} r(y|z=j,I)\, v(z=j|I)$

Notation

	Banana	Yellow	Label	Model output
Visually correct ground truth (Unknown)	✓	✓	z	v
Available ground truth (human-centric)	✓	✗	y	h

Relevance

Per concept: $r_{ij} = r(y=i|z=j)$

Visually present, irrelevant: r_{01}

Visually present, relevant: r_{11}

Factors Predictions
- $v(z=j|I)$: Is the object present?
- $r(y|z=1,I)$: Is the object relevant?

Handles mislabeled concepts
- For an unlabeled concept, assign high visual prediction & low relevance
- Produce both "clean" and "human-centric" output: relevant & irrelevant

Exploits structure in data
- Trained using only human labels

Evaluation using noisy labels

Method	Prob	NN	VB	JJ	DT	PRP	IN	Others	All
COCO Dataset. 1000 Visual concepts from Captions									
MILVC	-	41.6	20.7	23.9	33.4	20.4	22.5	16.3	34.0
MILVC + Multiple fc8	-	41.1	20.9	23.7	33.6	21.1	22.8	16.8	33.8
MILVC + Latent (Ours)	v	42.9	21.7	24.9	33.1	19.6	23.0	16.2	35.1
MILVC + Latent (Ours)	h	**44.3**	**22.3**	**25.8**	**34.4**	**21.8**	**23.6**	**17.3**	**36.3**
Classif.	-	34.9	18.1	20.5	32.8	19.2	21.8	16.3	29.0
Classif. + Multiple fc8	-	34.2	17.7	19.9	32.6	19.0	21.5	15.9	28.4
Classif. + Latent (Ours)	v	37.7	19.6	22.0	32.6	20.2	22.0	16.3	31.2
Classif. + Latent (Ours)	h	**38.7**	**20.1**	**22.6**	**33.8**	**21.2**	**23.0**	**17.5**	**32.0**
YFCC100M: Flickr images with tags (90k images, 1k tags)									
MILVC	-	5.7	9.2	5.2	-	3.8	8.8	6.1	5.7
MILVC + Multiple fc8	-	4.6	6.2	3.8	-	2.7	7.3	3.1	4.5
MILVC + Latent (Ours)	v	9.8	15.1	8.9	-	8.3	12.4	12.4	9.8
MILVC + Latent (Ours)	h	**11.2**	**15.4**	**9.9**	-	**8.2**	**16.3**	**12.5**	**11.2**

All methods use VGG16. Trained using binary cross-entropy loss.
MILVC: Fang et al., 2015; Classif.: Simple classification baseline;
Multiple-fc8: Same # parameters as our model.

Evaluation using clean labels

Baseline	Ours v	Ours h
63.7	**66.8**	66.5

Train on noisy labels (1000 classes) from Captions
Test on fully labeled data (73 classes) from COCO

Noise model w/o image conditioning

Baseline	w/o Image	w/ Image
34.0	34.3	**36.3**

Train and test on 1000 visual concepts from COCO Captions

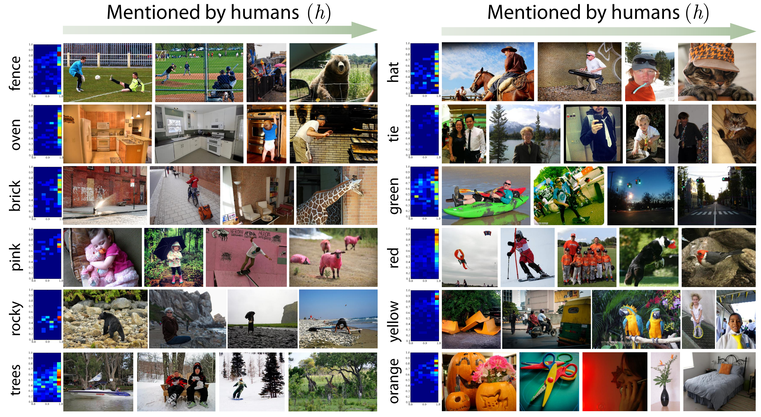
Human-like Image captioning

	BLEU-4	ROUGE	CIDEr
Baseline	27.7	51.8	89.7
Ours	**29.2**	**52.4**	**92.8**

Train and test on COCO Captions. We use the model output as image features for an LSTM trained for image caption generation.

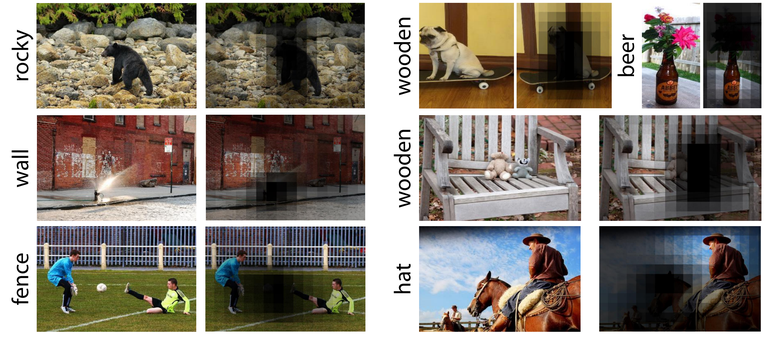
Qualitative Results

What to mention?



When to mention it?

When would you mention something not worth mentioning?



Correcting error modes