Counting Everyday Objects in Everyday Scenes

Prithvijit Chattopadhyay*[,1], Ramakrishna Vedantam*[,2], Ramprasaath R. Selvaraju[2], Dhruv Batra[1], Devi Parikh[1]

[1]Georgia Institute of Technology, [2]Virginia Tech

*equal contribution

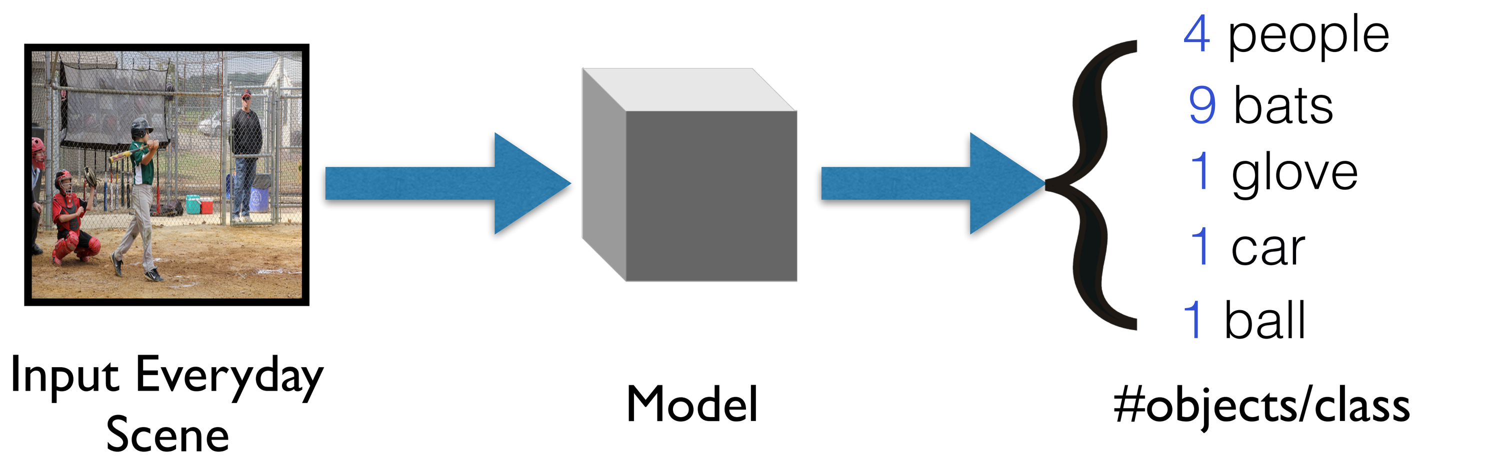
IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR 2017, July 21-26, Honolulu

Highlights

- **Scene Understanding Problem**: Counting instances of object categories in everyday scenes
- **Baseline Approaches**: Detection, Glancing, Associative Subitizing
- **Proposed Approach**: Sequential Subitizing
- **Experimental Results**: PASCAL VOC'07 and COCO
- **Applications**: Counting to improve object detection, and Visual Question Answering **(VQA)**

Task

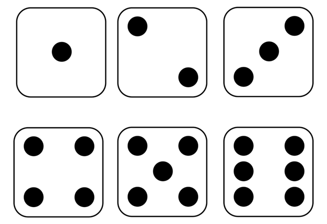


Key Motivations

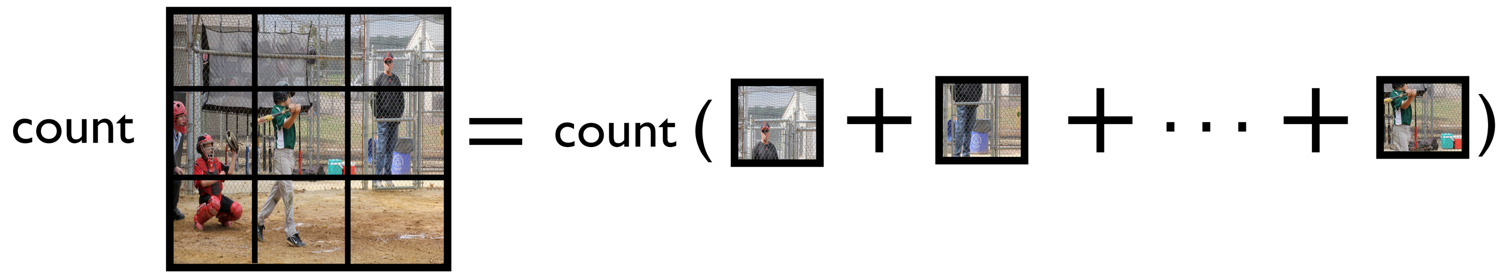
Subitizing

- The ability to see a 'small' number of objects and know how many there are without actually counting

How we count the number of pips on a die



Associative Property of Counts



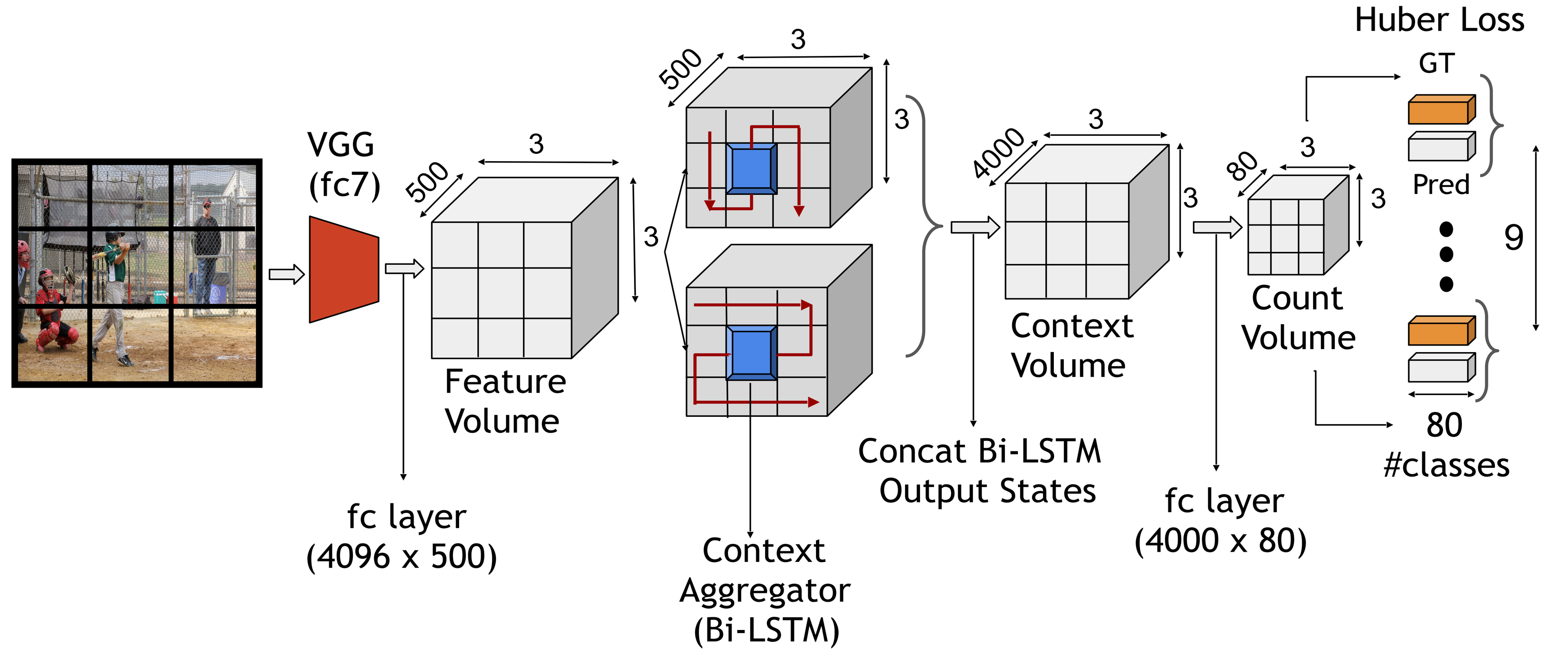
Context

- Model global context across the image while making predictions at one particular cell (partition)

Proposed Approach

Counting By Sequential Subitizing (Seq-sub)

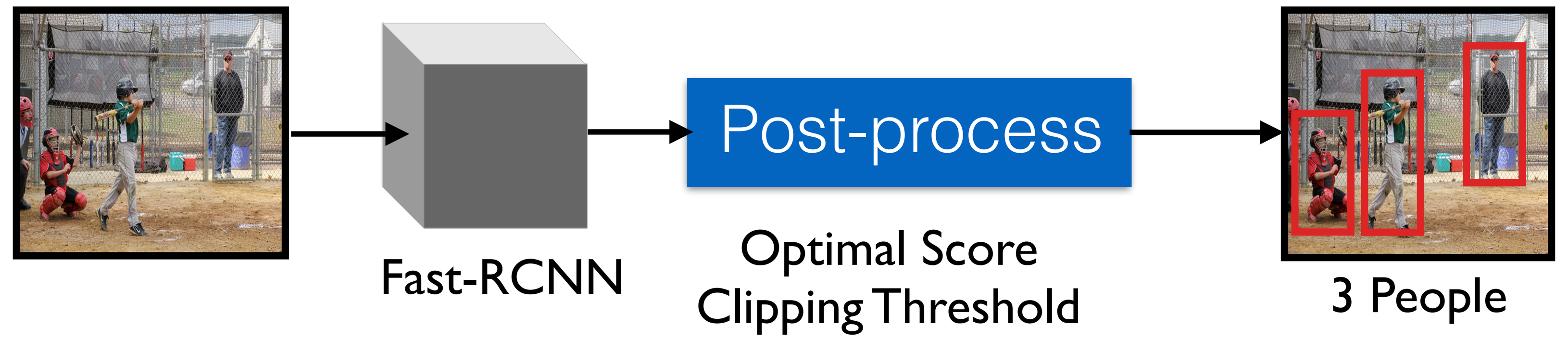
- Associative Nature + Subitizing + Context



Baseline Approaches

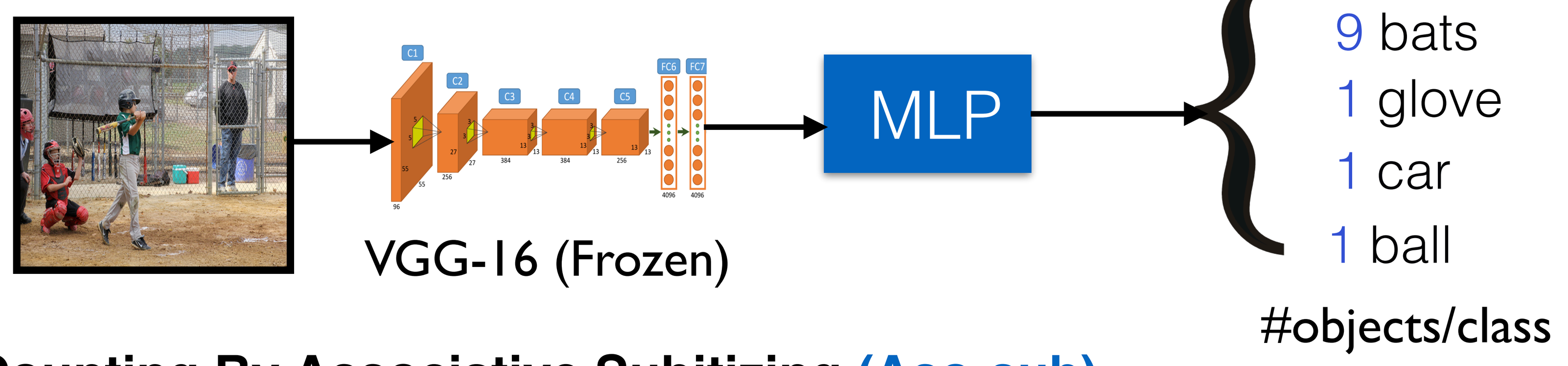
Counting By Detection (Detect)

- Object localization sufficient but not necessary for counting



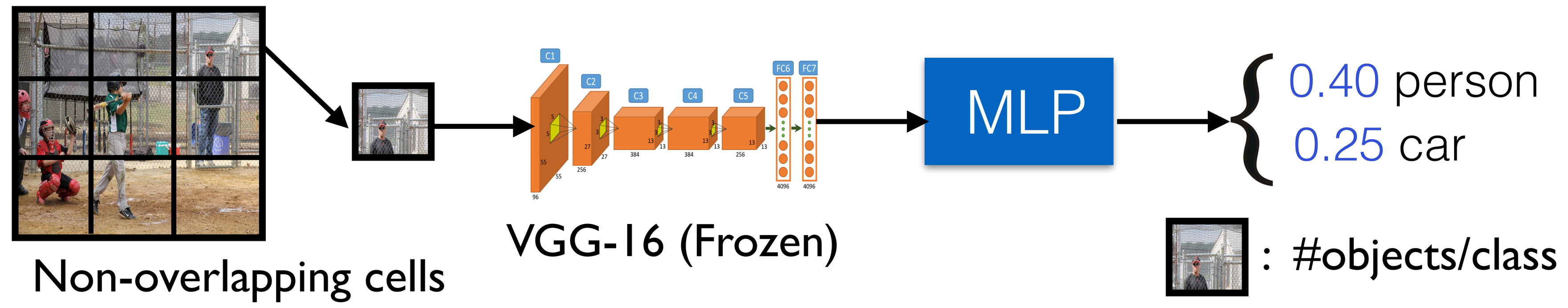
Counting By Glancing (Glance)

- Estimate global count in one-shot or *glance*



Counting By Associative Subitizing (Aso-sub)

- Associative Nature + Subitizing
- Drawback: Unaware of partial presence of objects in other cells



Datasets

Datasets

- PASCAL VOC 2007: 2501 train images, 2510 val images, 4952 test images and 20 object classes
- COCO: 82783 train images, 20252 val images (first half of COCO-val), 20252 test images (second half of COCO-val) and 80 object classes

Results

Metrics

- c_{ik} Ground truth count for class-k and image-i
- $\hat{c_{ik}}$ Predicted count for class-k and image-i

$$RMSE_k = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{c_{ik}} - c_{ik})^2} \qquad relRMSE_k = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{c_{ik}} - c_{ik})^2}{c_{ik}+1}}$$

Root-mean Squared Error — Relative Root-mean Squared Error

Models	mRMSE	mRMSE-nz	mrelRMSE	mrelRMSE-nz
Detection (Baseline)	0.49(0.00)	2.78(0.03)	0.20(0.00)	1.13(0.01)
Glancing (Baseline)	0.42(0.00)	2.25(0.02)	0.23(0.00)	0.91(0.00)
Aso-sub (Baseline)	0.38(0.00)	2.08(0.02)	0.24(0.00)	0.87(0.01)
Seq-sub (Proposed)	**0.35(0.00)**	**1.96(0.02)**	**0.18(0.00)**	**0.82(0.01)**

Counting performance on COCO Count-test split. nz = non-zero counts

Bottle
GT: 8
Detect: 1
Glance: 4
Aso-sub: 10
Seq-sub: 8

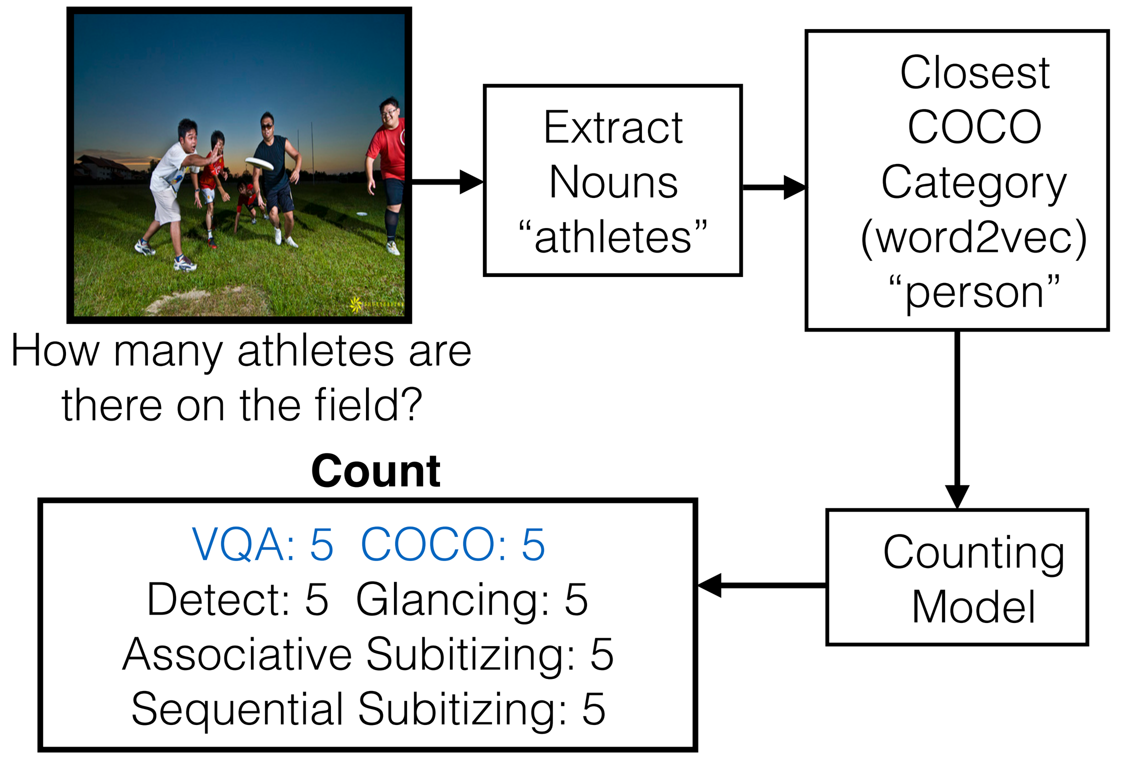
Elephant
GT: 16
Detect: 3
Glance: 9
Aso-sub: 22
Seq-sub: 17

Qualitative Examples

Applications

Visual Question Answering

- 10.28% questions in VQA are counting-Q
- 7.07% questions in COCO-QA are counting-Q
- Count-QA: Subset of counting questions in VQA + COCO-QA



Existing VQA Models	mRMSE
Deeper LSTM + Norm. CNN (Lu et al. 2015)	2.71(0.23)
MCB (Fukui et al. 2016)	3.25(0.94)
Seq-sub (Proposed)	**1.81(0.09)**

Improving Object Detection

- Detectors are typically operated at some threshold which is usually set on a global basis
- Use counting to set per-image thresholds, based on count estimate

Method	mF(%)
Category-wise Threshold	15.26
Ground Truth (oracle)	20.17
Seq-sub (Proposed)	**17.00**

Evaluation Metric: mean F-measure (mF)