SAP
KTH VETENSKAP OCH KONST
UNIVERSITY OF TRENTO - Italy
POLITECNICO MILANO 1863

When the Distribution Is the Answer

VizWiz Challenge


Denis Dushi

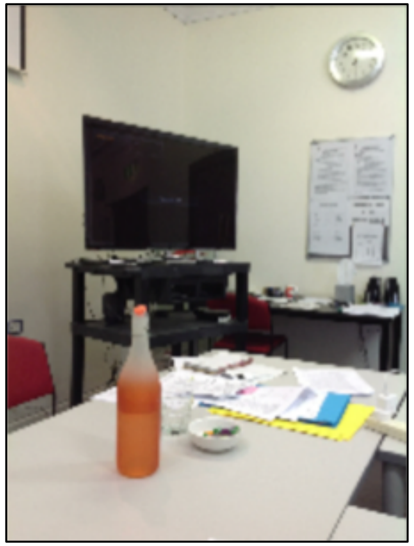

Sandro Pezzelle


Tassilo Klein


Moin Nabi

VQA Task

Input



Q: “What is this?”

Annotations

A1	bottle
A2	bottle
A3	tv
A4	office
A5	bottle
A6	tv
A7	bottle
A8	room
A9	office
A10	bottle

answer	**count**
bottle	5
tv	2
office	2
room	1

Ground Truth

“bottle”

VQA Evaluation metric[1]

$$accuracy = min(\frac{\#\ \text{Annotators providing that answer}}{3}, 1)$$

Annotations

answer	count
bottle	5
tv	2
office	2
room	1

Training Loss

$$H(p,q) = -\sum_x p(x)\,\log q(x)$$

Ground Truth

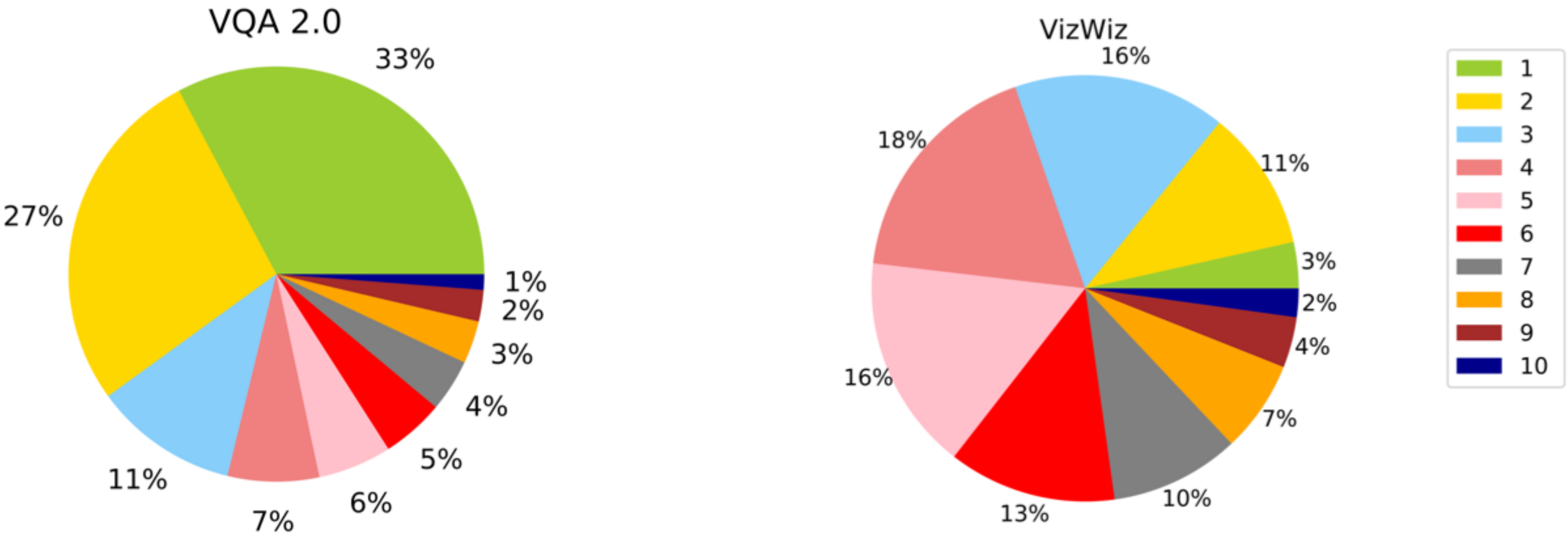
"bottle"

Evaluation Accuracy

prediction	accuracy
bottle	100%
tv	~ 67%
office	~ 67%
room	~ 33%

[1] Antol et al. (2015). VQA: Visual Question Answering. Proceedings of the IEEE international 076 conference on Computer Vision: 2425–2433

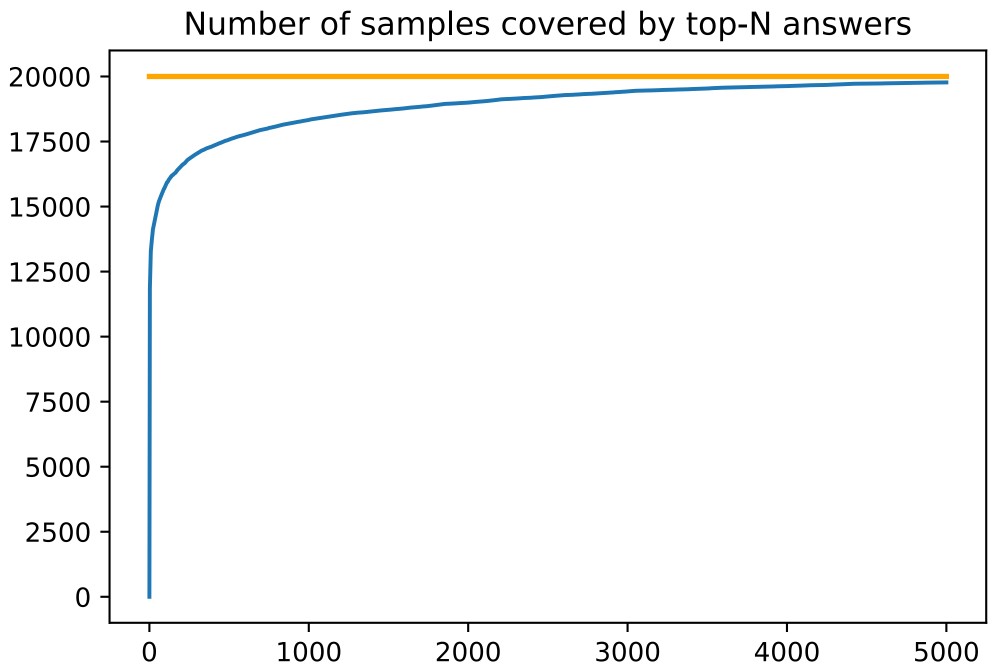
Subjectivity



[2] Jolly, Pezzelle et al. (2018). The Wisdom of MaSSeS: Majority, Subjectivity, and Semantic Similarity in the Evaluation of VQA

Coverage analysis

- Coverage of samples considering all the annotations



num answers/classes	**1**	**2**	**5**	**50**	**300**	**3000**	**40271**
num samples (train)	9541	11570	12531	14963	17046	19425	20K
% samples (train)	47.70	57.85	62.65	74.81	85.23	97.12	100

Table 1: Number and percentage of samples covered by using the top-N answers (row 1).

Most frequent answer : *unanswerable*

count	covered samples	% covered samples
1	3059	32%
2	1878	20%
$\geq$ *3*	4604	48%

Uncertainty-aware training

- Methods that use only the most-frequent answer ignore :

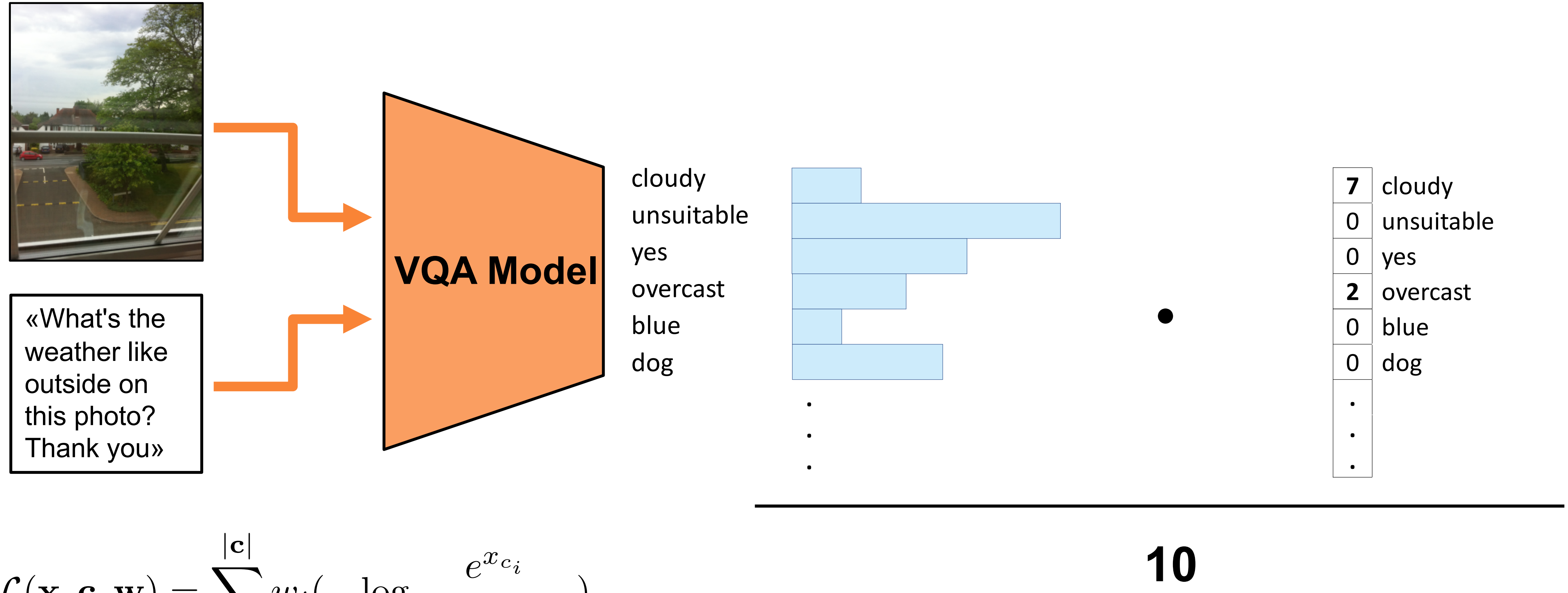
  1. Contribution of other answers

  2. Uncertainty of each answer

Uncertainty-aware training ➡ Uncertainty modeled as **agreement over humans**

Soft cross-entropy loss[3]

- Standard VQA model [4]



$$\mathcal{L}(\mathbf{x}, \mathbf{c}, \mathbf{w}) = \sum_{i=1}^{|\mathbf{c}|} w_i \left(-\log \frac{e^{x_{c_i}}}{\sum_{j=1}^{|\mathbf{x}|} e^{x_j}}\right)$$

[3] Ilievski et al. (2017). A simple loss function for improving the convergence and accuracy of visual question answering models.

[4] Kazemi et al. (2017). Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering.

Results

- Accuracy on validation split

num answers/classes	**1**	**2**	**5**	**50**	**300**	**3000**	**40271**
soft-loss model acc. (val)	**0.349**	0.402	0.424	0.481	0.504	0.516	0.512

Table 2 Accuracy of soft-loss model using N classes in prediction.

- Accuracy on **test-challenge** split

method	acc
SoA[5]	0.475
Ours	**0.512**

[5] Gurari et al. (2018). VizWiz Grand Challenge: Answering Visual Questions from Blind People.

Preprocessing

1. Smartly stripping punctuation

    e.g. “can’t” → “cant”

2. Filtering conversational words

    e.g. “hello”, “please”, “thank you”, “goodbye” ...

- Accuracy on **test-challenge**

method	acc
SoA[5]	0.4750
Ours	0.5120
Ours + prepro	**0.5163**

[5] Gurari et al. (2018). VizWiz Grand Challenge: Answering Visual Questions from Blind People.

Answerability task

1. Change output layer of multi-class model

   Label : **0/1** (*unanswerable/answerable*)

2. Balance dataset

   Imbalanced dataset (71.3 % answerable) →

   - Up-sampling
   - Down-sampling

- Accuracy on test-dev

method	F1	AP
Ours	65.02	74.71
Ours + Up	68.84	74.73

- Accuracy on **test-challenge**

method	F1	AP
SoA[5]	-	71.7
Ours + Up	67.71	**73.11**

[5] Gurari et al. (2018). VizWiz Grand Challenge: Answering Visual Questions from Blind People.

Conclusion

1. **Multi-class task**
    - Soft cross-entropy
    - Smart preprocessing

2. **Answerability task**

    Binary classifier with up-sampling of unanswerable samples

SAP

KTH VETENSKAP OCH KONST

UNIVERSITY OF TRENTO - Italy

POLITECNICO MILANO 1863

Thank you.

(Answerable) Questions?

Denis Dushi

Sandro Pezzelle

Tassilo Klein

Moin Nabi