Privacy Preserving Visual Question Answering

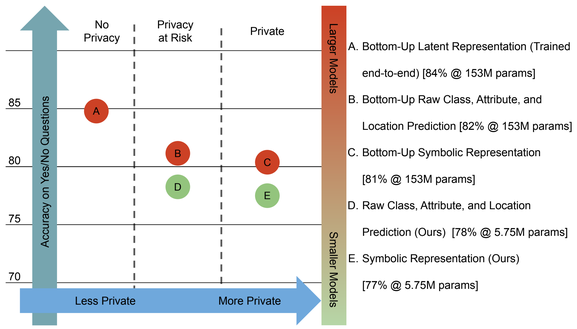
Cristian-Paul Bara, [1] Qing Ping, [2] Abhinav Mathur, [2] Govind Thattai, [2] Rohith MV, [2] Gaurav S. Sukhatme [2,3]

[1]University of Michigan, [2]Amazon, [3]University of Southern California

cpbara@umich.edu, {pingqing,mathuabh,kurohith,thattg,sukhatme}@amazon.com

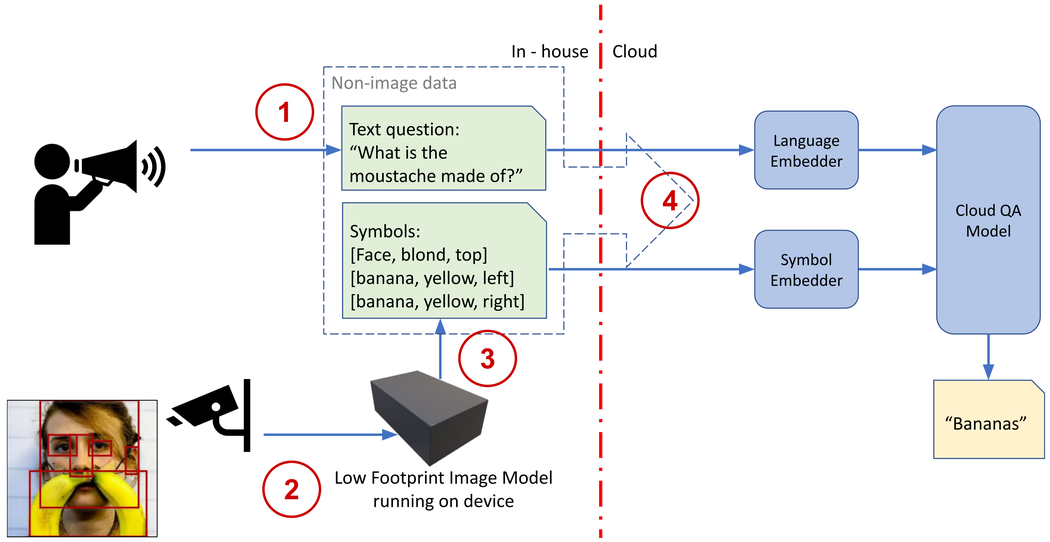
Overview

- Developing machine learning models with high privacy, high accuracy, and low model-size is the holy grail of computer vision systems.
- Current SOTA VQA models are too large for feasible deployment on user devices like virtual assistants.
- We propose a hybrid deployment framework that strengthens privacy while limiting performance degradation to $< 5\%$ on Yes/Yes Question. The model is up to $25\times$ smaller than SOTA on device model and up to $100\times$ smaller than SOTA end-to-end model.

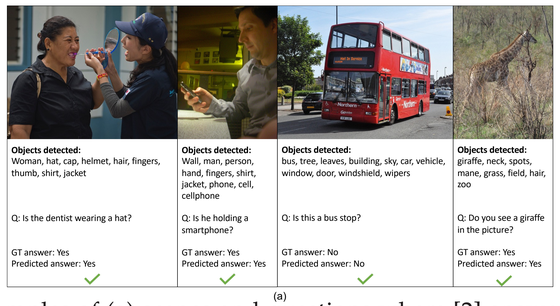


- We start from [1] using [2] as a visual model and [3] as a QA model. We implement a modified [4] trained Visual Genome to detect objects with associated classes and attributes.
- We achieve similar performance to [2] with a $25\times$ reduction in footprint.
- We use the output of our modified [4] and create a non-differentiable symbolic representation in lieu of the intermediate representation of the original paper. We transfer these representations to the cloud to be processed by [3]. We present comparative results below.

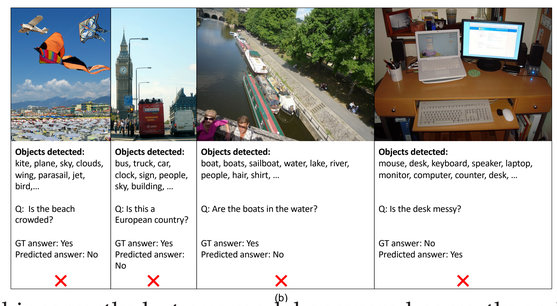
Hybrid Edge-Cloud Framework



Prediction Examples



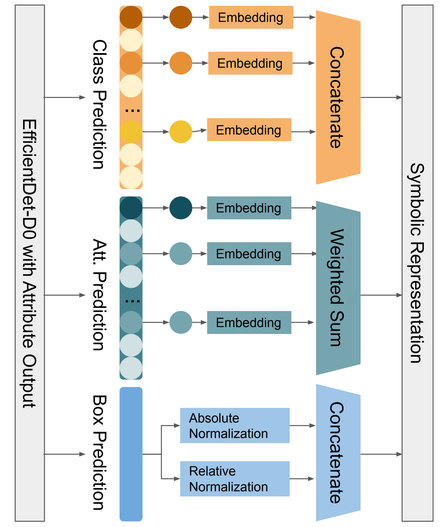
(a)



(b)

Examples of (a) scenes and questions where [2] answered incorrectly, but our model answered correctly and (b) scenes and questions where Bottom-Up answered correctly, but our model answered incorrectly.

Symbolic Representations



- The main contribution is the privacy preserving symbolic embedding.
- We implement a modified [4] that is run on the edge device. This provides objects with associated classes, attributes, and bounding boxes. From this prediction, for each object, we take the names of the top five class predictions, the names of the top five attributes, and associated bounding box.
- We then take the GloVe embeddings of the class and attribute names, concatenate the class embeddings ordered by prediction value, and average the attribute embeddings. The bounding boxes are normalized to the image and to the encompassing bounding box of all bounding boxes.

Results

Visual Model	Privacy Status	Deployability	Visual Model Num. Param.	Feature Type	Performance Overall	Other	Yes/No	Count
Bottom-Up	Not Private	No	153M	Intermediate	67.08	58.41	84.73	49.19
Bottom-Up	At risk	No	153M	Raw	63.31	54.48	81.72	43.81
Bottom-Up	Private	No	153M	Symbolic (GloVe)	62.49	53.78	81.05	42.10
Ours	At Risk	Yes	5.75M	Raw	56.67	44.19	77.98	42.57
Ours	Private	Yes	5.75M	Symbolic (GloVe)	55.41	42.95	77.27	41.89
Ours	Private	Yes	5.75M	Symbolic (BERT)	54.17	72.10	75.25	39.16

Performance of [3] on VQA 2.0 using the output from Bottom-Up or EfficientDet with raw predictions representations or fully symbolic output. We also provide the performance of [3] using the intermediate output provided by Bottom-Up as trained end-to-end as measured by us.

Results with Ground Truth Captions

Visual Model	Privacy Status	Deployability	Visual Model Num. Param.	Feature Type	Performance Overall	Other	Yes/No	Count
Bottom-Up	Not Private	No	153M	Raw	65.16	56.72	82.49	47.25
Bottom-Up	At risk	No	153M	Symbolic (GloVe)	64.58	56.32	81.91	46.01
Caption Only	Private	N/A	N/A	N/A	57.55	47.28	76.53	41.83
Ours	At risk	Yes	5.75M	Raw	60.10	49.42	79.26	45.32
Ours	Private	Yes	5.75M	Symbolic (GloVe)	59.76	49.02	79.14	44.62

Performance of [3] on VQA 2.0 using captions along side the output from Bottom-Up or EfficientDet with raw predictions representations or fully symbolic output. We also provide the performance of [3] using captions alone.

[1] Jiang, Huaizu, et al. "In defense of grid features for visual question answering." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[2] Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[3] Yu, Zhou, et al. "Deep modular co-attention networks for visual question answering." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[4] Tan, Mingxing, Ruoming Pang, and Quoc V. Le. "Efficientdet: Scalable and efficient object detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.