An Empirical Evaluation of Visual Question Answering for Novel Objects

Supplementary Material

Santhosh K. Ramakrishnan[1,2] Ambar Pal[1] Gaurav Sharma[1] Anurag Mittal[2]

[1]IIT Kanpur* [2]IIT Madras†

1. Possible extensions

Here, we tackle the problem (P1) of answering *known* (e.g. similar to those in train set) questions containing *novel* objects and having *known* answers, at test time. More challenging cases include (P2) answering novel questions about novel objects (as suggested by AR2) and (P3) generating answers containing novel objects (as suggested by AR5). While problems P2 and P3 are more difficult problems than P1, we highlight that P1 is itself a very challenging subproblem which has not been addressed so far. In the proposed setting, questions that come under P1 account for a significant fraction of the test questions (71.79%, $83,508$ out of $116,323$). If a perfect model were available for P1, then the overall accuracy would be 71.79%. However, current methods obtain accuracies of around 40%, highlighting that P1 itself is very hard and arguably should be the first stepping stone in this direction.

2. Image sharing

Image sharing takes place in our proposed split, statistics are shown in Tab. 2. We claim that overfitting does not happen and justify the claim with the performances of the system computed separately on the common and exclusive parts of test set; Tab. 1 gives these performances and we see that the differences in overall performance are very small (≤ 0.5) in all cases. Also, the improvements obtained by various models over the baseline model are similar in the common and exclusive parts of the test set.

We would like to also highlight that sharing images does not make the task easier or the split prone to overfitting (as already demonstrated by the results above). Even if the same image is present in train and test set, the object being queried for is different at train and test time (by design of the split). Hence, the system can not memorize or overfit on the train set *and* give good performance on the test set.

# Images	Common to train and test		Train only	Test only
	73487		43583	6216
# Corres. Questions	Train	Test	Train	Test
	108857	97675	115847	18648

Table 2. Statistics of images in train/test splits

*The project started when Santhosh Ramakrishnan and Ambar Pal were summer interns at IIT Kanpur. Ambar Pal is a student at IIIT Delhi. ambar14012@iiitd.ac.in, grv@cse.iitk.ac.in

†{ee12b101@ee, amittal@cse}.iitm.ac.in

						Open Ended Questions				Multiple Choice Questions			
arch	feat	model	aux	vocab	set	ov	oth	num	y/n	ov	oth	num	y/n
Architecture 1	VGG	1	none	oracle	s1	39.64	21.4	28.94	73.22	46.39	33.14	30.64	73.27
					s2	39.33	23.37	27.24	74.19	46.57	35.23	29.15	74.27
		2	text	train	s1	40.39	21.82	29.49	74.61	47.20	33.66	31.12	74.68
					s2	40.04	23.75	28.73	75.25	47.23	35.62	30.21	75.33
		3	text	oracle	s1	40.89	21.91	28.81	76.15	47.91	34.10	30.66	76.21
					s2	40.35	23.69	28.13	76.60	47.60	35.62	29.73	76.68
	INCEP	4	none	oracle	s1	40.71	23.00	29.88	73.47	46.72	33.52	31.01	73.52
					s2	40.19	24.82	27.65	74.04	46.42	35.07	29.09	74.11
		5	text	train	s1	40.40	22.61	28.77	73.53	46.79	33.70	30.47	73.61
					s2	40.14	24.39	28.15	74.55	47.05	35.74	29.80	74.64
		6	text	oracle	s1	41.20	23.20	28.61	74.99	47.76	34.63	29.97	75.11
					s2	41.19	25.30	28.41	76.12	47.89	36.24	30.29	76.23
Architecture 2	VGG	1	none	oracle	s1	35.45	15.53	28.86	70.55	43.02	28.90	29.79	70.57
					s2	34.87	17.24	28.16	71.17	42.80	30.39	29.35	71.24
		2	text	train	s1	37.52	17.37	27.37	74.12	44.45	29.63	27.85	74.20
					s2	37.26	19.88	26.02	74.56	44.27	31.54	26.95	74.62
		3	text	oracle	s1	37.99	17.66	28.16	74.78	45.04	30.00	29.47	74.83
					s2	37.62	19.83	28.30	75.11	45.13	32.25	29.67	75.17
	INCEP	4	none	oracle	s1	37.95	18.50	28.64	73.10	44.59	30.20	29.52	73.17
					s2	37.61	20.49	28.25	73.81	44.59	32.05	29.28	73.89
		5	text	train	s1	37.73	17.98	27.63	73.64	44.48	29.93	28.07	73.73
					s2	37.30	20.36	25.56	73.94	44.39	32.17	26.30	74.00
		6	text	oracle	s1	38.71	18.87	28.26	74.87	45.80	31.31	29.44	74.95
					s2	38.49	21.14	28.04	75.50	45.85	33.28	29.35	75.60

Table 1. Arch. 1 (top) and Arch. 2 (bottom) with VGG and INCEP features: s1 is images exclusive to test and s2 is common images between train and test; in both cases questions are test only.