Batch construction and multitask learning in visual relationship recognition

Shane Josias
Stellenbosch University, CAIR
josias@sun.ac.za

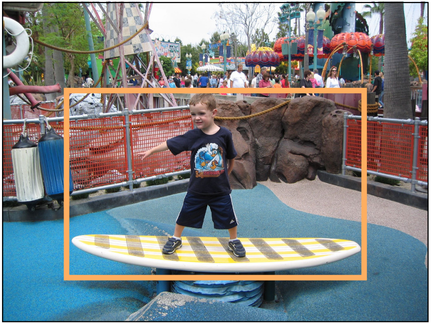
Willie Brink
Stellenbosch University
wbrink@sun.ac.za

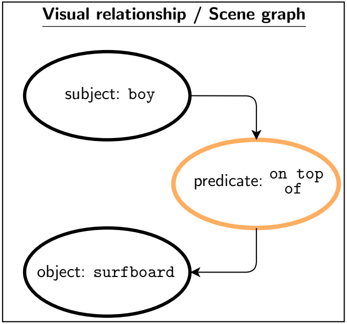
30 January 2020

Visual relationship recognition

Task: produce a (subject, predicate, object) triplet given an image.

Example:

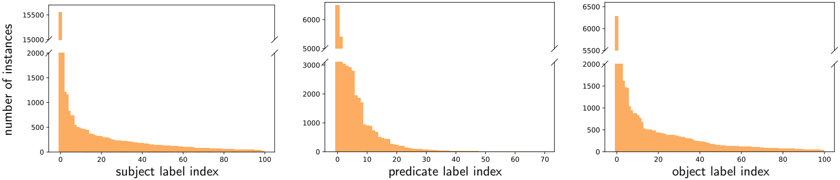




Challenges

Combinatorial: with 100 subject, 70 predicate, and 100 object labels we have 700,000 possible relationships.

Data distribution: is typically long-tailed, making it difficult to learn rare relationships.



Our approach

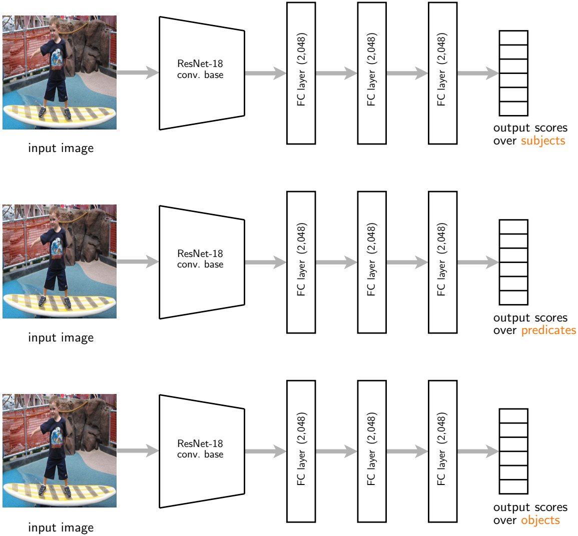
Treat VRR as a classification problem.

Input: image, cropped around a pair of objects.
Output: (subject, predicate, object) triplet.

Three tasks: predict the subject, predict the predicate and predict the object. Avoid predicting over 700,000 classes.

Obtain normalised scores over classes in each task. Combine scores through multiplication.

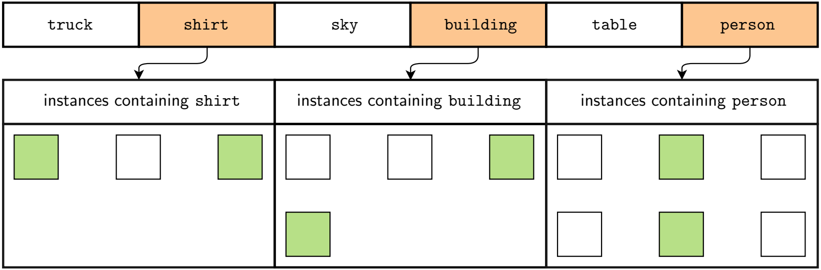
Single task learning with standard batching



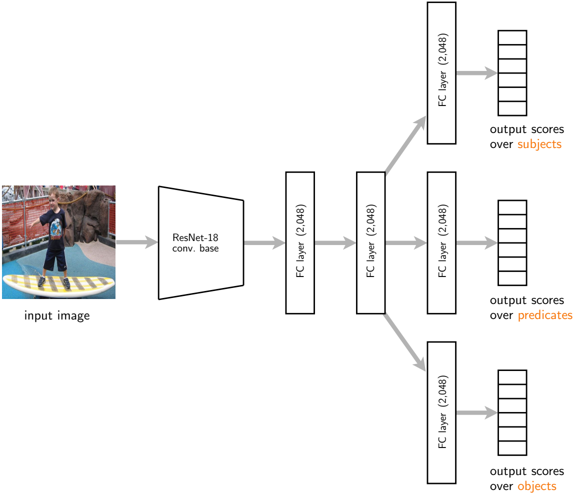
Class-selective batch construction

Select *n* classes from a vocabulary of *N* classes, uniformly at random.

Sample *m* instances from each selected class, uniformly at random.



Multitask learning



VRD dataset (Lu et al. ECCV 2016)

5,000 images, 37,987 visual relationships but only 15,448 unique relationships.

100 labels for both subject and objects, 70 predicate labels in five categories.

				
action verb	spatial	preposition	comparative	non-action verb

`person`
`kick`
`ball`
	`person`
	`on top of`
	`ramp`
		`motorcycle`
		`with`
		`wheel`
			`elephant`
			`taller than`
			`person`
				`person`
				`wear`
				`shirt`

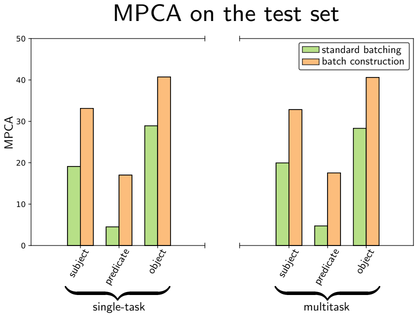
Metrics

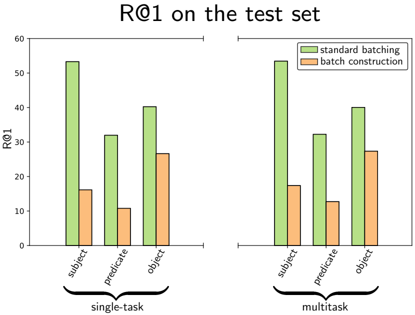
MPCA: mean per-class accuracy; used to measure performance on rare classes in the individual tasks.

R@k: recall-at-k; percentage of times the correct label occurs in the top k predictions (if ordered by output scores).

Tail R@k: R@k measured on visual relationship classes that have fewer than 1,000 samples for subject, predicate, and object labels.

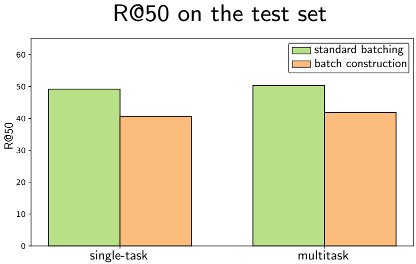
Quantitative results: individual tasks

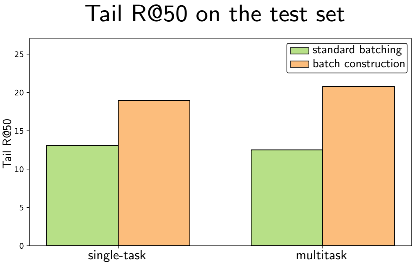




Batch construction is performed with respect to label on x-axis (same as the task being predicted).

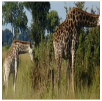
Quantitative results: visual relationship recognition





Batch construction is performed with respect to the object labels since it performed better overall.

Qualitative results

Models	person, on, horse 		giraffe, taller than, giraffe 		person, on, skateboard 		person, feed, elephant 	
ST-SB	person, on, horse	12.0	giraffe, taller than, giraffe	25.1	person, wear, person	11.8	person, above, street	4.3
	person, ride, horse	7.0	giraffe, in front of, giraffe	20.8	person, wear, shirt	10.5	person, on, street	4.1
	person, wear, horse	5.3	giraffe, next to, giraffe	9.5	person, wear, skateboard	10.0	person, under, street	3.0
	person, has, horse	5.2	giraffe, above, giraffe	7.6	person, wear, shoes	5.4	sky, above, street	1.7
	person, on, person	3.1	giraffe, behind, giraffe	7.2	person, wear, pants	4.4	sky, on, street	1.6
ST-BC-O	person, on, horse	18.7	giraffe, in front of, giraffe	98.6	person, wear, skateboard	25.6	person, under, elephant	16.4
	person, has, horse	11.8	giraffe, taller than, giraffe	0.4	person, on, skateboard	10.0	person, in front of, elephant	16.0
	person, wear, horse	7.7	giraffe, behind, giraffe	0.4	person, has, skateboard	9.6	person, above, elephant	10.0
	person, in front of, horse	4.3	giraffe, next to, giraffe	0.1	person, ride, skateboard	5.2	person, near, elephant	4.7
	person, next to, person	3.7	giraffe, beside, giraffe	0.1	person, wear, shoes	3.5	person, behind, elephant	4.1
MT-SB	person, wear, horse	9.3	giraffe, taller than, giraffe	45.4	person, wear, shirt	15.5	person, on, street	4.7
	person, on, horse	6.8	giraffe, in front of, giraffe	18.9	person, wear, person	9.6	person, under, street	3.9
	person, wear, person	3.4	giraffe, next to, giraffe	8.6	person, wear, skateboard	6.9	person, above, street	3.4
	person, behind, horse	3.1	giraffe, behind, giraffe	7.3	person, wear, shoes	6.1	person, on, person	2.4
	person, has, horse	2.6	giraffe, under, giraffe	2.6	person, wear, pants	4.1	person, under, person	1.9
MT-BC-O	person, on, horse	13.2	giraffe, in front of, giraffe	92.5	person, wear, skateboard	20.0	person, in front of, elephant	7.4
	person, above, horse	12.0	giraffe, taller than, giraffe	6.0	person, wear, shoes	14.0	person, near, elephant	6.9
	person, behind, horse	6.3	giraffe, behind, giraffe	0.9	person, wear, helmet	12.0	person, under, elephant	5.1
	person, ride, horse	5.3	giraffe, next to, giraffe	0.3	person, has, skateboard	3.8	person, on, elephant	3.4
	person, has, horse	4.8	giraffe, beside, giraffe	0.07	person, wear, pants	3.7	person, above, elephant	2.4

ST-SB	single-task, standard batching	MT-SB	multitask, standard batching
ST-BC-O	single-task, batch construction from object labels	MT-SB-O	multitask, batch construction from object labels

Conclusion

Class-selective batch construction improves performance on the tail of the distribution, at the cost of performance on the small number of dominating classes.

Conclusion

Class-selective batch construction improves performance on the tail of the distribution, at the cost of performance on the small number of dominating classes.

Multitask learning neither improves nor impedes performance. Reduced capacity can be beneficial.

Conclusion

Class-selective batch construction improves performance on the tail of the distribution, at the cost of performance on the small number of dominating classes.

Multitask learning neither improves nor impedes performance. Reduced capacity can be beneficial.

Predicates are difficult to model. Limitation of pretrained models?

Conclusion

Class-selective batch construction improves performance on the tail of the distribution, at the cost of performance on the small number of dominating classes.

Multitask learning neither improves nor impedes performance. Reduced capacity can be beneficial.

Predicates are difficult to model. Limitation of pretrained models?

Misclassifications are often semantically similar to groundtruth. We could use a language model to incorporate semantics.