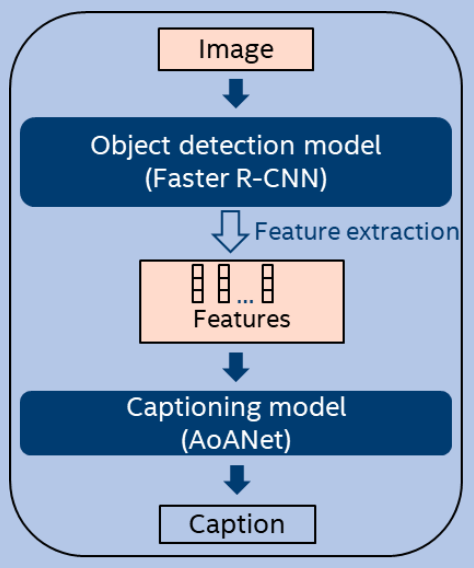
DATA AUGMENTATION TO IMPROVE ROBUSTNESS OF IMAGE CAPTIONING SOLUTIONS

Shashank Bujimalla, Mahesh Subedar, Omesh Tickoo
Intel Corporation, USA

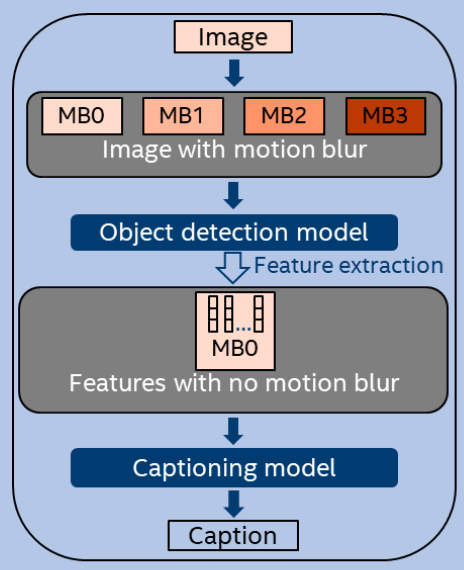
Real world images often have quality flaws which will lead to poor image captioning results.
Our approach: Improve the robustness of the captioning algorithms to image quality flaws by using training data augmentation at each or both stages of the solution.

SOTA Image captioning

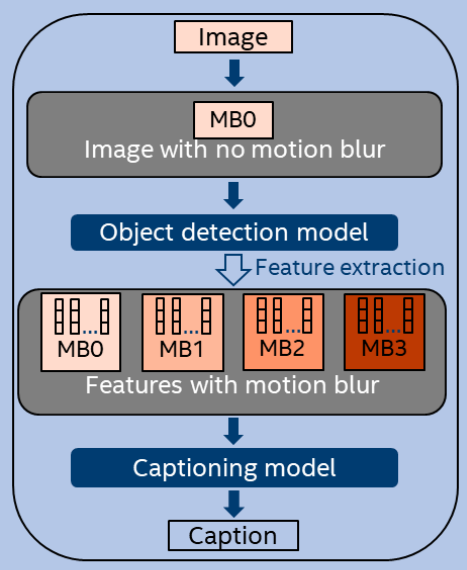


No-Aug

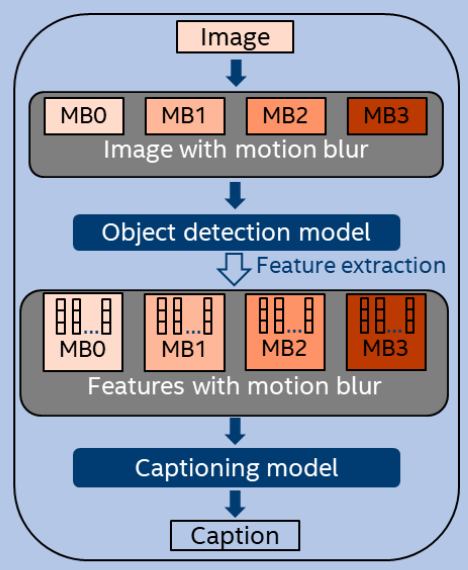
Our Training Approaches



ObjDet-Aug



Cap-Aug



ObjDet-Cap-Aug

Predicted caption

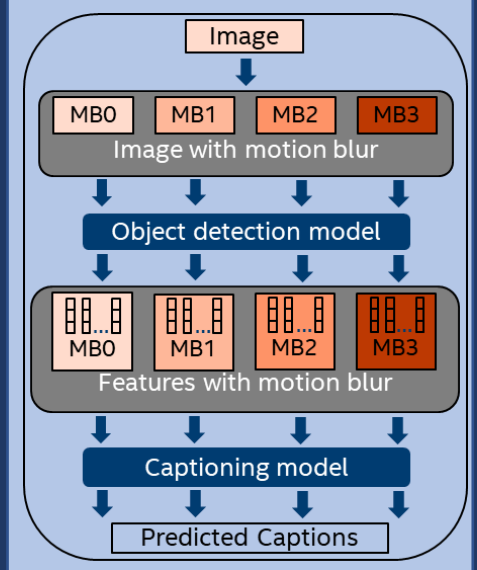
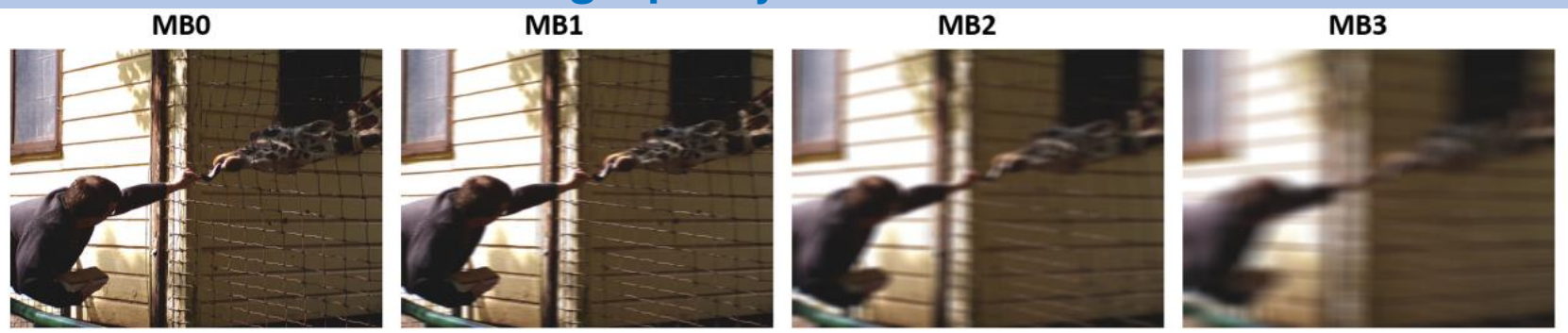


Image quality flaw: Motion Blur



MB0 = No blur ; MB1 = low blur ; MB2 = medium blur ; MB3 = high blur

Ground-truth captions

- a giraffe sticks his nose and tongue out of a wire fence to accept something
- a giraffe eating out of the hands of a man through fence
- a man in black sweater feeding a giraffe through a fence
- a man feeding a giraffe through a metal fence
- a man feeding a giraffe through a fence

Predicted captions (CIDEr-D Scores)

Blur	No-Aug		ObjDet-Cap-Aug	
	Predicted Caption	C	Predicted Caption	C
MB0	a man is feeding a giraffe through a fence	4.028	a man feeding a giraffe through a fence	4.949
MB1	a man feeding a giraffe through a cage	3.993	a man feeding a giraffe through a fence	4.949
MB2	a man is holding a bird in a room	0.121	a man feeding two giraffes through a fence	1.740
MB3	a person is a room with a wooden floor	0.000	a woman feeding two giraffes through a fence	1.284

Training approach	MS COCO Dataset				Vizwiz Dataset					
	MB0	MB1	MB2	MB3	MB0	MB1	MB2	MB3	With blur	No blur
No-Aug	117.1	111.4	95.0	48.4	48.8	47.0	40.9	26.4	47.2	53.0
ObjDet-Aug	116.6	114.6	111.7	100.2	48.9	48.1	45.6	39.5	47.0	53.3
Cap-Aug	116.8	115.0	108.8	85.1	50.0	49.2	46.9	38.2	**49.0**	53.2
ObjDet-Cap-Aug	**117.4**	**116.0**	**113.4**	**105.7**	**50.3**	**49.9**	**48.1**	**43.5**	48.9	**54.1**