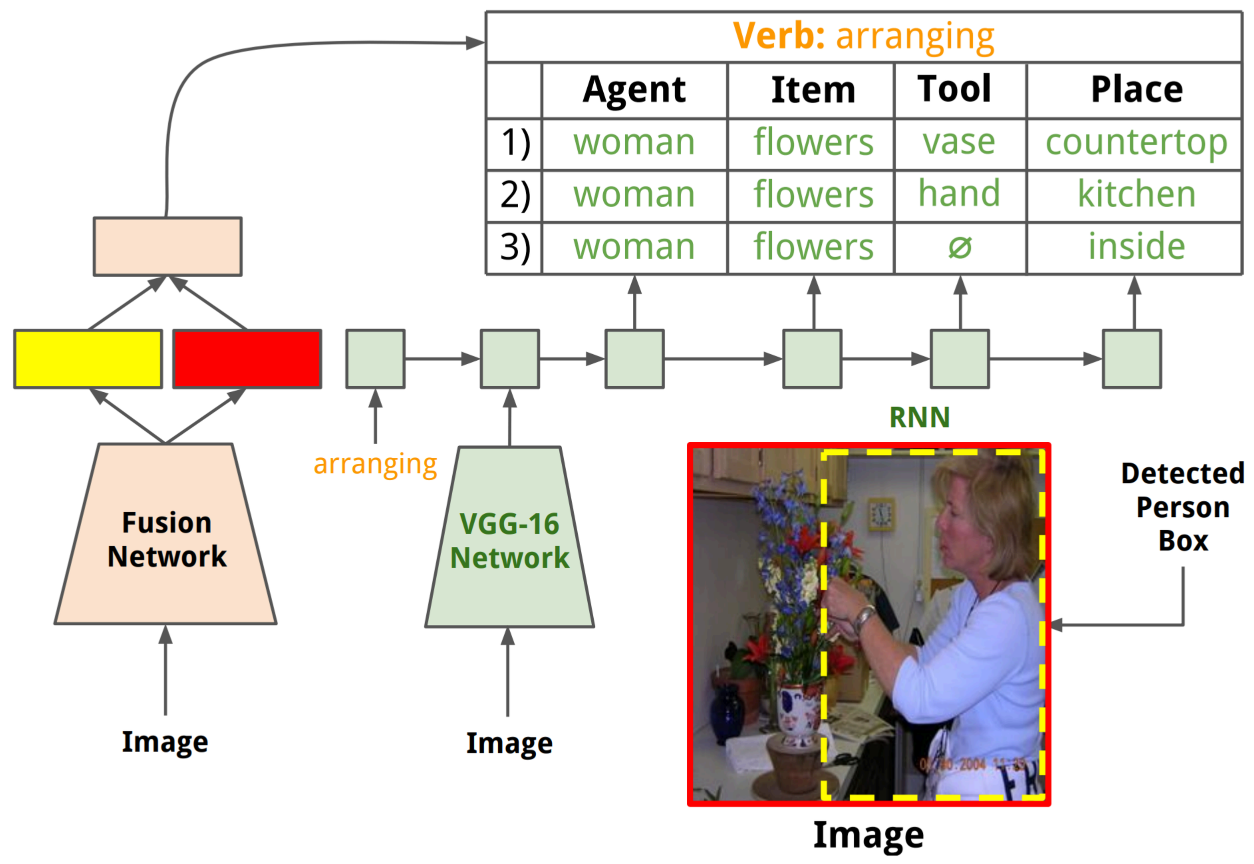
ILLINOIS

Recurrent Models for Situation Recognition

Arun Mallya and Svetlana Lazebnik, University of Illinois at Urbana-Champaign

ICCV 17
International Conference on Computer Vision 2017

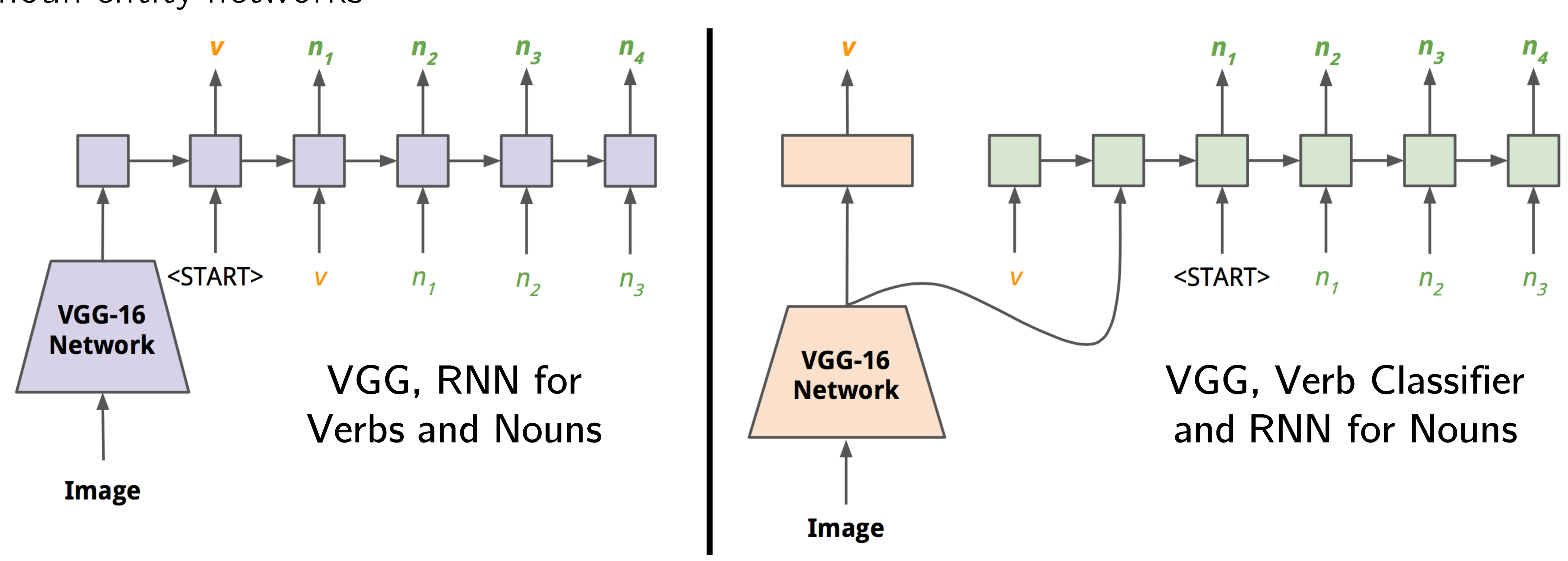
Overview



- Each image in imSitu [1] is labeled with an action verb (out of 504 verbs), and each verb is associated with a unique set of semantic roles (out of 1,700 roles) fulfilled by noun entities in the image (out of 11,000 nouns)
- We pose the structured imSitu prediction as that of sequential noun entity prediction conditioned on the verb
- We use the fusion action prediction network [4] to predict the verb
- Conditioned on the verb, we use a separate network with an RNN to predict the noun entities in an arbitrary but fixed order

Model Evolution and Results

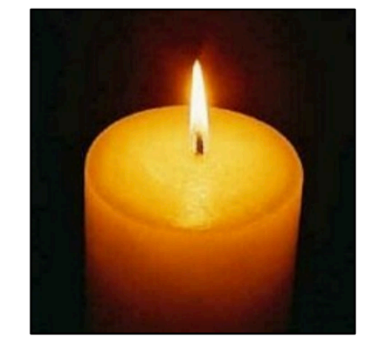
- **No-vision, RNN for nouns:** model that predicts most likely noun entity sequence given verb
- **VGG, RNN for Verbs and Nouns:** An RNN model that takes in visual features, predicts verb at first time step, and then noun entities conditioned on predicted noun
- **VGG, Verb Classifier and RNN for Nouns:** A separate verb classifier with an RNN for noun entity prediction, with shared visual features
- **Fusion for Verbs, VGG+RNN for Nouns:** Final model (in Overview figure), with separate verb and noun entity networks



Models	Top-1 Predicted Verb			Ground Truth Verbs		Mean
	verb	value	value-all	value	value-all	
ImSitu Dev Set (Full)						
Tensor Comp. (TC) + Image Reg. (IR) CRF [2]	32.91	25.39	14.87	69.39	33.17	38.02
Above + Extra 5M Imgs. [2]	34.20	26.56	15.61	**70.80**	34.82	39.57
No-vision, RNN for nouns	—	—	—	52.12	17.62	—
VGG, RNN for Verbs & Nouns	26.52	20.08	11.80	68.27	32.67	33.87
VGG, Verb class. and RNN for Nouns	35.35	26.80	15.77	68.44	32.98	38.74
Fusion for Verbs, VGG+RNN for Nouns	**36.11**	**27.74**	**16.60**	70.48	**35.56**	**40.40**
ImSitu Test Set (Full)						
CRF (IR + TC) + 5M Extra Imgs. [2]	34.12	26.45	15.51	**70.44**	34.38	39.48
Fusion for Verbs, VGG+RNN for Nouns	**35.90**	**27.45**	**16.36**	70.27	**35.25**	**40.16**
ImSitu Test Set (Rare)						
CRF (IR + TC) + 5M Extra Imgs. [2]	20.32	11.87	2.52	55.72	12.28	22.95
Fusion for Verbs, VGG+RNN for Nouns	**22.07**	**12.96**	**3.37**	**56.38**	**13.79**	**23.89**

Qualitative Results and Applications

Example Image Situation Predictions



GT) Verb: glowing	
Agent	Place
candle	∅

Predictions

1) Verb: glowing	
Agent	Place
candle	∅

2) Verb: igniting			
Agent	Item	Tool	Place
person	candle	match	∅

GT) Verb: browsing		
Agent	GoalItem	Place
woman	book	bookshop

Predictions

1) Verb: browsing		
Agent	GoalItem	Place
woman	book	bookshop

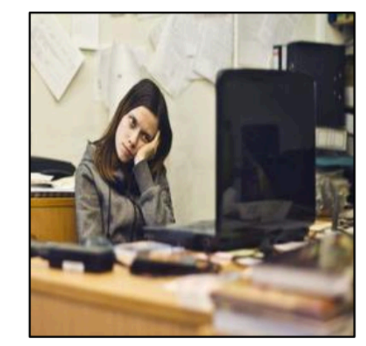
2) Verb: shelving			
Agent	Item	Destination	Place
woman	book	shelf	library

GT) Verb: scoring	
Agent	Place
soccer player	field

Predictions

1) Verb: deflecting			
Agent	Deflec-tedItem	Desti-nation	Place
soccer player	soccer ball	∅	field

2) Verb: scoring	
Agent	Place
soccer player	field

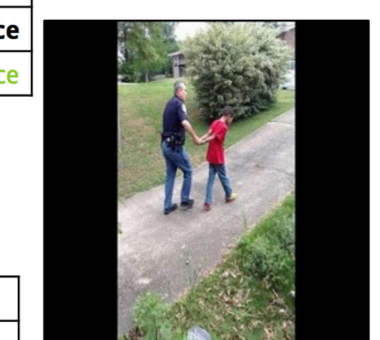


GT) Verb: leaning			
Agent	Item	Against	Place
woman	head	hand	office

Predictions

1) Verb: studying	
Agent	Place
woman	desk

2) Verb: phoning		
Agent	Tool	Place
woman	telephone	office



GT) Verb: misbehaving	
Agent	Place
boy	walkway

Predictions

1) Verb: arresting		
Agent	Suspect	Place
policeman	boy	sidewalk

2) Verb: grieving	
Agent	Place
child	cemetery

GT) Verb: celebrating		
Agent	Occasion	Place
people	parade	river

Predictions

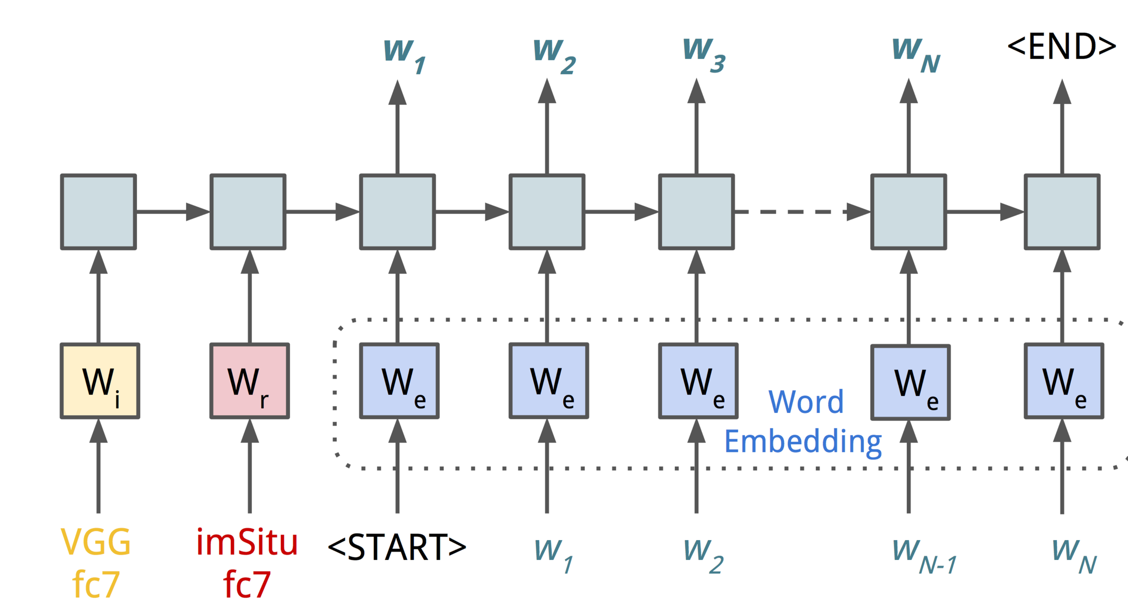
1) Verb: celebrating		
Agent	Occasion	Place
people	∅	outside

2) Verb: parading	
Agent	Place
people	street

Application to Image Captioning

We augment the Neuraltalk2 [3] captioning model with features from the noun entity prediction network (green VGG-16 network in the Overview figure) as input at the second time step

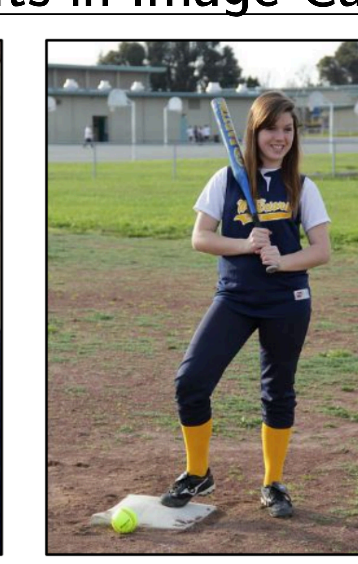
We observe an improvement in semantic content of captions, as shown below



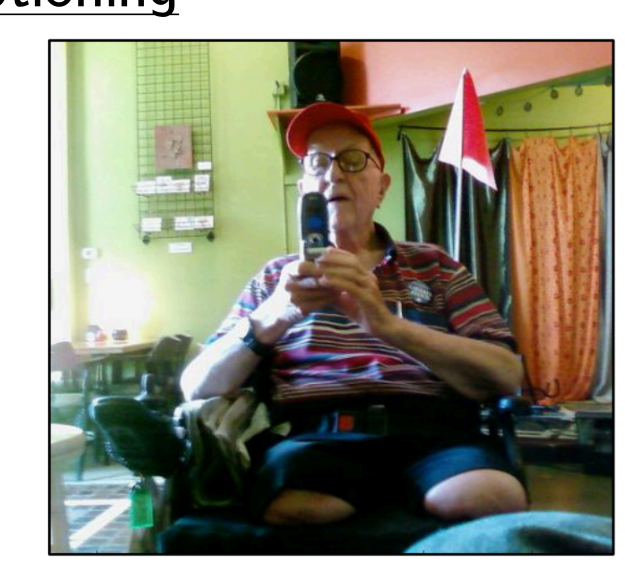
Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
COCO test set of 5000 images (Karpathy split)							
NeuralTalk2 [3]	70.8	53.7	40.1	30.1	24.5	—	93.0
VGG + imSitu	71.5	54.6	41.1	31.1	24.8	—	95.2
COCO test2014 (40 reference captions)							
NeuralTalk2 [3]	87.9	77.8	66.1	54.7	32.4	66.0	89.1
VGG + imSitu	88.7	79.4	68.2	57.2	33.2	67.0	91.8

Qualitative Improvements in Image Captioning

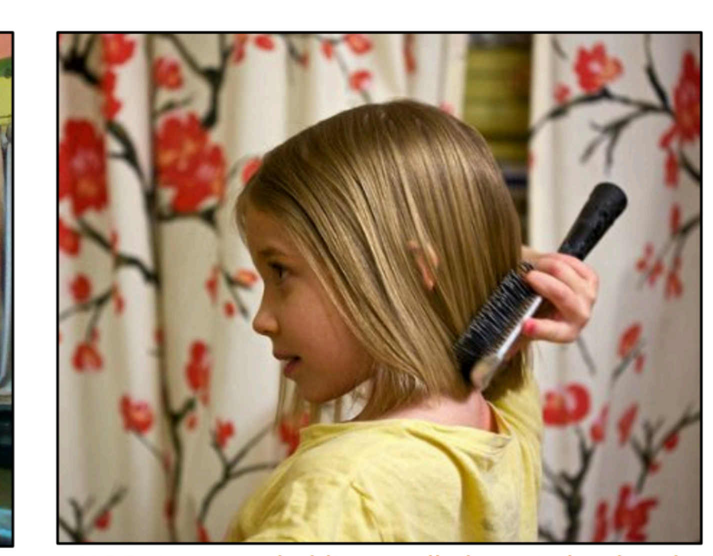
VGG: A man with a beard and a tie
VGG+imSitu: A man is holding a pair of scissors
GT: A person holding a pair of scissors open intently



VGG: A woman is holding a frisbee in a park
VGG+imSitu: A young girl is holding a baseball bat on a field
GT: A girl with a bat standing in a field



VGG: A man sitting on a couch with a cat
VGG+imSitu: A man sitting on a chair with a cell phone
GT: An old man is trying to use his cell phone



VGG: A woman holding a cell phone in her hand
VGG+imSitu: A woman is brushing her hair in a bathroom
GT: A little girl is brushing her hair in a bathroom

References

1. Visual semantic role labeling for image understanding, CVPR 2016
2. Commonly uncommon: Semantic sparsity in situation recognition, CVPR 2017
3. Neuraltalk2, https://github.com/karpathy/neuraltalk2
4. Learning Models for Actions and Person-Object Interactions with Transfer to Question Answering, ECCV 2016