

Semantic Compositional Networks for Visual Captioning: Supplementary Material

Zhe Gan[†], Chuang Gan^{*}, Xiaodong He[‡], Yunchen Pu[†]
Kenneth Tran[‡], Jianfeng Gao[‡], Lawrence Carin[†], Li Deng[‡]

[†]Duke University, ^{*}Tsinghua University, [‡]Microsoft Research, Redmond, WA 98052, USA

{zhe.gan, yunchen.pu, lcarin}@duke.edu, ganchuang1990@gmail.com

{xiaohe, ktran, jfgao, deng}@microsoft.com

A. More results for Figure 4




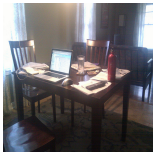

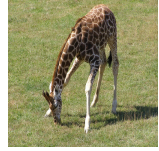
	<p>Tags: outdoor (1), elephant (0.995), animal (0.988), grass (0.962), standing (0.89), rock (0.781), zoo (0.682), enclosure (0.619)</p>		<p>Tags: indoor (0.966), table (0.919), food (0.849), kitchen (0.714), sitting (0.545), counter (0.436), top (0.285), doughnut (0.251)</p>		<p>Tags: outdoor (0.998), building (0.996), man (0.613), front (0.434), standing (0.333), woman (0.255), walking (0.249), next (0.247)</p>
<p>Generated captions: SCN-LSTM-T: a couple of elephants standing next to each other SCN-LSTM: a large elephant standing next to a tree</p>		<p>Generated captions: SCN-LSTM-T: a kitchen with a lot of food on it SCN-LSTM: a bunch of doughnuts sitting on top of a counter</p>		<p>Generated captions: SCN-LSTM-T: a man standing in front of a building SCN-LSTM: a statue of a man standing next to a building</p>	
	<p>Tags: table (0.997), indoor (0.893), chair (0.876), room (0.692), sitting (0.583), window (0.58), wooden (0.542), small (0.344)</p>		<p>Tags: fence (1), giraffe (0.994), animal (0.921), wooden (0.677), fenced (0.66), standing (0.592), next (0.493), zoo (0.442)</p>		<p>Tags: grass (1), outdoor (0.992), giraffe (0.985), mammal (0.98), animal (0.978), field (0.93), eating (0.558), standing (0.508)</p>
<p>Generated captions: SCN-LSTM-T: a dining room with a table and chairs SCN-LSTM: a wooden table with a laptop on it</p>		<p>Generated captions: SCN-LSTM-T: a giraffe standing next to a wooden fence SCN-LSTM: a couple of giraffe standing next to each other</p>		<p>Generated captions: SCN-LSTM-T: a giraffe standing on top of a lush green field SCN-LSTM: a giraffe is eating grass in a field</p>	

Figure 1: Detected tags and sentence generation results on COCO. The output captions are generated by: 1) SCN-LSTM, and 2) SCN-LSTM-T, a SCN-LSTM model without the visual feature inputs, *i.e.*, with only tag inputs.

B. More results on image captioning





	Tags: polar (0.999), rock (0.998), animal (0.997), bear (0.993), mammal (0.988), zoo (0.881), white (0.779), large (0.748)		Tags: refrigerator (0.992), food (0.976), open (0.97), cabinet (0.953), shelf (0.582), filled (0.45), door (0.426), lots (0.329)		Tags: person (0.925), building (0.839), people (0.787), umbrella (0.779), group (0.704), child (0.519), standing (0.369), holding (0.272)
Generated captions: LSTM-R: a polar bear standing on top of a rock LSTM-RT₂: a polar bear standing on a rock SCN-LSTM: a large white polar bear standing on a rock		Generated captions: LSTM-R: a display case filled with lots of food LSTM-RT₂: a shelf full of different kinds of food SCN-LSTM: a refrigerator filled with lots of food and drinks		Generated captions: LSTM-R: a group of people standing next to each other LSTM-RT₂: a group of people standing in front of an umbrella SCN-LSTM: a group of people standing in the rain with umbrellas	
	Tags: bicycle (1), parked (0.923), next (0.889), group (0.829), sidewalk (0.783), many (0.698), lot (0.611), rack (0.596)		Tags: food (0.939), oranges (0.839), fruit (0.836), slice (0.792), sliced (0.783), orange (0.764), plate (0.759), table (0.704)		Tags: clock (1), building (0.999), large (0.902), station (0.876), mounted (0.644), sitting (0.621), tower (0.574), building (0.418)
Generated captions: LSTM-R: a group of motorcycles parked next to each other LSTM-RT₂: a row of bikes parked in a row SCN-LSTM: a bunch of bikes parked in a park lot		Generated captions: LSTM-R: a bowl of fruit sitting on top of a table LSTM-RT₂: a bunch of oranges sitting on a table SCN-LSTM: a bunch of oranges sitting on a plate		Generated captions: LSTM-R: a clock on the wall of a building LSTM-RT₂: a clock on the side of a building SCN-LSTM: a large clock mounted to the side of a building	
	Tags: dog (1), water (0.998), beach (0.805), standing (0.666), walking (0.451), next (0.435), ocean (0.301), white (0.225)		Tags: water (0.985), beach (0.975), ocean (0.655), next (0.493), shore (0.324), sand (0.288), sandy (0.209), bench(0.204)		Tags: bench (0.997), fence (0.98), park (0.974), grass (0.877), sitting (0.771), wooden (0.582), next (0.511), green (0.377)
Generated captions: LSTM-R: a dog that is playing with a frisbee LSTM-RT₂: a couple of dogs standing on a beach SCN-LSTM: a white dog walking on a beach		Generated captions: LSTM-R: a bench that is sitting on the beach LSTM-RT₂: a person sitting on a bench on a beach SCN-LSTM: a wooden bench sitting on top of a sandy beach		Generated captions: LSTM-R: a park bench sitting in the middle of a forest LSTM-RT₂: a park bench sitting on a park bench SCN-LSTM: a wooden bench sitting in the middle of a park	
	Tags: road (0.958), street (0.911), green (0.856), sign (0.601), traffic (0.549), car (0.401), truck (0.382), city (0.374)		Tags: person (0.958), woman (0.728), sitting (0.708), bench (0.394), people (0.381), next (0.371), group (0.361), front (0.311)		Tags: person (0.932), man (0.787), young (0.458), black (0.439), white (0.43), jumping (0.342), riding (0.242), trick (0.156)
Generated captions: LSTM-R: a bus that is driving down the road LSTM-RT₂: a bus parked on the side of a road SCN-LSTM: a green bus driving down a city street		Generated captions: LSTM-R: a couple of women standing next to each other LSTM-RT₂: a couple of people sitting on a toilet SCN-LSTM: a group of people sitting on a bench		Generated captions: LSTM-R: a man sitting on top of a wooden bench LSTM-RT₂: a man riding a skateboard down a street SCN-LSTM: a black and white photo of a skateboarder doing a trick	

Figure 2: Detected tags and sentences generation results on COCO. The output captions are generated by: 1) LSTM-R, 2) LSTM-RT₂, and 3) our SCN-LSTM.

C. More results on video captioning

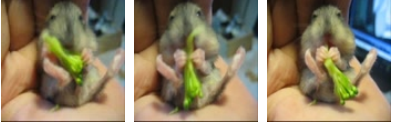
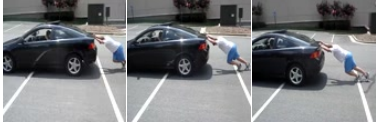

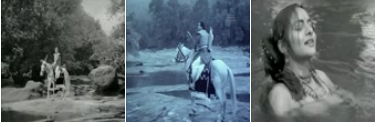



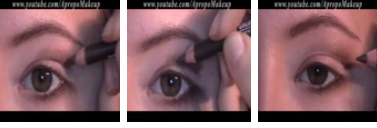
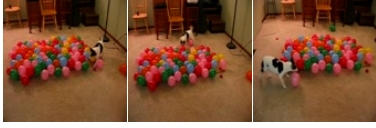
		
<p>Tags: playing (0.694), animal (0.673), baby (0.63), person (0.471), eating (0.419), something (0.333), food (0.329), hand (0.311)</p>	<p>Tags: man (0.807), person (0.733), car (0.442), driving (0.39), playing (0.382), road (0.365), moving (0.189), pushing (0.129)</p>	<p>Tags: woman (0.88), girl (0.732), lady (0.699), making (0.516), something (0.501), water (0.267), glass (0.244), drinking (0.204)</p>
<p>Generated captions: LSTM-CR: a person is eating LSTM-CRT₂: a person is holding a small animal SCN-LSTM: a small animal is eating</p>	<p>Generated captions: LSTM-CR: a man is doing a wheelie LSTM-CRT₂: a man is riding a bike SCN-LSTM: a man is pushing a car</p>	<p>Generated captions: LSTM-CR: a woman is pouring sugar in a glass LSTM-CRT₂: a woman is pouring water SCN-LSTM: a woman is drinking something</p>
		
<p>Tags: man (0.635), woman (0.545), riding (0.541), person (0.465), water (0.465), girl (0.4), doing (0.387), horse (0.132)</p>	<p>Tags: man (0.843), person (0.774), doing (0.393), playing (0.385), open (0.298), gun (0.283), shooting (0.276), field (0.259)</p>	<p>Tags: man (0.958), song (0.869), stage (0.866), singing (0.859), men (0.845), music (0.826), playing (0.762), guitar (0.759)</p>
<p>Generated captions: LSTM-CR: a girl is riding a horse LSTM-CRT₂: a woman is riding a horse SCN-LSTM: a man is riding a horse</p>	<p>Generated captions: LSTM-CR: a girl is firing a gun LSTM-CRT₂: a girl is shooting SCN-LSTM: a man is shooting a gun</p>	<p>Generated captions: LSTM-CR: a group of people are dancing on stage LSTM-CRT₂: a man is dancing on stage SCN-LSTM: a band is performing on stage</p>
		
<p>Tags: doing (0.616), boy (0.557), room (0.554), playing (0.51), floor (0.493), dancing (0.491), dance (0.361), kid (0.281)</p>	<p>Tags: woman (0.829), girl (0.743), doing (0.593), using (0.408), makeup (0.211), applying (0.2), face (0.171), hand (0.139)</p>	<p>Tags: playing (0.776), dog (0.625), floor (0.423), trying (0.399), woman (0.356), running (0.293), puppy (0.202), toy (0.182)</p>
<p>Generated captions: LSTM-CR: a baby is walking LSTM-CRT₂: a baby is dancing SCN-LSTM: a boy is dancing</p>	<p>Generated captions: LSTM-CR: a girl is singing LSTM-CRT₂: a woman is playing SCN-LSTM: a woman is plucking her eyebrow</p>	<p>Generated captions: LSTM-CR: a group of girls are playing with a toy LSTM-CRT₂: the children are playing SCN-LSTM: a dog is playing with a toy</p>

Figure 3: Detected tags and sentence generation results on Youtube2Text. The output captions are generated by: 1) LSTM-CR, 2) LSTM-CRT₂, and 3) our SCN-LSTM.