What Value Do Explicit High Level Concepts Have in Vision to Language Problems? Supplementary Material

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, Anton van den Hengel
School of Computer Science, The University of Adelaide, Australia
{qi.wu01,chunhua.shen,lingqiao.liu,anthony.dick,anton.vandenhengel}@adelaide.edu.au

1. Image Captioning Results on the Flickr

In this section, we report results on the Flickr8k [3] and Flickr30k [8]. These datasets contain 8,000 and 31,000 images respectively, and each image is annotated with 5 sentences. In our reported results, we use pre-defined splits for Flickr8k, 1000 for validation, 1000 for testing and the rest for training. For Flickr30k, we report results with the widely used publicly available splits in the work of [4], which use 1000 images for validation, 1000 for testing.

Flickr8K

State-of-art-Flickr8k	B-1	B-2	B-3	B-4	$\mathcal{PPL}$
Karpathy & Li (NeuralTalk) [4]	0.58	0.38	0.25	0.16	-
Chen & Zintick (Mind's Eye) [1]	-	-	-	0.14	15.10
Google(NIC)[6]	0.66	0.42	0.27	0.18	-
Mao et al. (m-Rnn-AlexNet) [5]	0.57	0.39	0.26	0.17	24.39
Xu et al. (Hard-Attention) [7]	0.67	0.46	0.31	0.21	-
Baseline - *CNN(I)*					
VggNet+LSTM	0.56	0.37	0.24	0.16	15.71
VggNet-PCA+LSTM	0.56	0.38	0.25	0.16	16.07
GoogLeNet+LSTM	0.56	0.38	0.24	0.16	15.71
VggNet+ft+LSTM	0.64	0.43	0.30	0.20	14.69
Ours - $V_{att}(I)$					
Attributes-GT+LSTM‡	0.76	0.57	0.41	0.29	12.52
Attributes-SVM+LSTM	0.73	0.53	**0.38**	0.26	12.63
Attributes-CNN+LSTM	**0.74**	**0.54**	**0.38**	**0.27**	**12.60**

Flickr30K

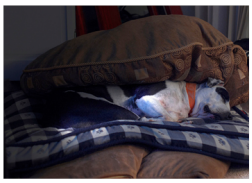
State-of-art-Flickr30k	B-1	B-2	B-3	B-4	$\mathcal{PPL}$
Karpathy & Li (NeuralTalk) [4]	0.57	0.37	0.24	0.16	-
Chen & Zintick (Mind's Eye) [1]	-	-	-	0.13	19.10
Google(NIC) [6]	0.66	-	-	-	-
Donahue et al. (LRCN) [2]	0.59	0.39	0.25	0.17	-
Mao et al. (m-Rnn-AlexNet) [5]	0.54	0.36	0.23	0.15	35.11
Mao et al. (m-Rnn-VggNet) [5]	0.60	0.41	0.28	0.19	20.72
Xu et al. (Hard-Attention) [7]	0.67	0.44	0.30	0.20	-
Baseline - *CNN(I)*					
VggNet+LSTM	0.57	0.38	0.25	0.17	18.83
VggNet-PCA+LSTM	0.59	0.40	0.26	0.17	18.92
GoogLeNet+LSTM	0.58	0.39	0.26	0.17	18.77
VggNet+ft+LSTM	0.67	0.47	0.31	0.21	16.62
Ours - $V_{att}(I)$					
Attributes-GT+LSTM‡	0.78	0.57	0.42	0.30	14.88
Attributes-SVM+LSTM	0.68	0.49	0.33	0.23	16.01
Attributes-CNN+LSTM	**0.73**	**0.55**	**0.40**	**0.28**	**15.96**

Table 1. BLEU-1,2,3,4 and $\mathcal{PPL}$ metrics compared to other state-of-the-art methods and our baseline on Flickr8K and Flickr30K dataset. ‡ indicates ground truth attribute labels are used, which (in gray) will not participate in rankings. Our $\mathcal{PPL}s$ are based on Flickr8k word dictionaries of size 2538 and Flickr30k word dictionaries of size 7414.

Results Table 1 reports image captioning results on Flickr8k, Flickr30k. The **Attributes-GT+LSTM** models perform best over all datasets and all evaluation metrics, because the ground truth attributes are used. Apart from using ground truth attributes, our **Attributes-CNN+LSTM** models generate the best results on both Flickr8k and Flickr30K over all evaluation metrics.

2. Some Example Results

This section shows some qualitative results demonstrating our attribute predictions, image captions and question answering.



Top -5 Attributes
- couch, sleeping , brown, laying, dog

Question Answering
Q1: Does this animal appear to be resting ?
A1: yes (yes)
Q2: What is the color of the cushions ?
A3: brown (brown)

Generated Caption
- a dog laying on top of a couch next to a pillow.

Top -5 Attributes
- group, people, children, cake, grass

Question Answering
Q1: What kind of event does this look like?
A1: birthday party (birthday party)
Q2: Is this woman trying to give the kids an un cooked cake?
A2: no (no)

Generated Caption
- a group of children standing around a cake.

Top -5 Attributes
- cake, wedding, cutting, knife, people

Question Answering
Q1: What is the occasion that this photo depicts?
A1: wedding (wedding)
Q2: What is the woman and man cutting?
A2: cake (cake)

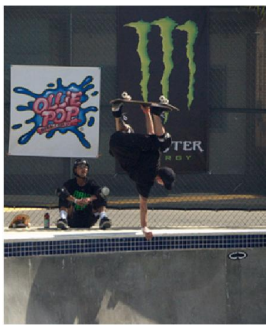
Generated Caption
- a bride and groom cutting their wedding cake.

Figure 1. Some qualitative results demonstrating our attribute predictions, image captions and question answering. Ground truth answers are in parentheses. Blue indicates we give the right answer, red means we are wrong.

Top -5 Attributes
- top, cake, fruits, table, plates

Question Answering
Q1: What are the orange sticks?
A1: carrots (carrots)
Q2: How many carrots are in the bowls?
A2: 2 (over 10)
Q3: Is this set up for a party?
A3: yes (yes)

Generated Caption
- a table topped with plates of food.



Top -5 Attributes
- jumping, watching, skate, board, people

Question Answering
Q1: What is the guy doing?
A1: skateboarding (skateboarding)
Q2: Are both of these skateboarders upside down?
A2: yes (no)
Q3: Why are there signs on the wall?
A3: to keep clean (advertising)

Generated Caption
- a man doing a trick on a skate board.

Generated Caption
- a young man riding a skateboard down the side of a road.

Top -5 Attributes
- skate, road, board, people, hill

Question Answering
Q1: What is the man standing on?
A1: skateboard (skateboard)
Q2: Is the man facing the camera?
A2: no (no)
Q3: Is this man worshiping the local mountains?
A3: no (no)

Generated Caption
- a man is holding a hot dog in his hand.

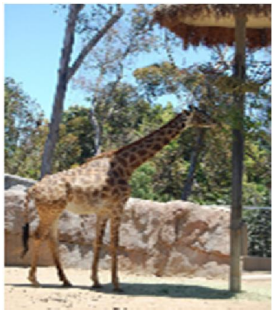
Top -5 Attributes
- hotdog, people, eating, young, red

Question Answering
Q1: What color is the woman's jacket?
A1: red (red)
Q2: Is the food good?
A2: yes (yes)
Q3: What condiments did this woman put on the hot dog?
A3: ketchup (ketchup and mustard)

Top -5 Attributes
- shelf, small, book, room, television

Question Answering
Q1: What pattern is on the curtain?
A1: floral (leaves)
Q2: What sport is being displayed on the television?
A2: football (football)
Q3: Is there a bookcase nearby?
A3: yes (yes)

Generated Caption
- a living room with a couch and a television.



Top -5 Attributes
- giraffe, standing, tree, tall, zoo

Question Answering
Q1: Is it warm or cold in this picture?
A1: warm (warm)
Q2: What type of animal is this?
A2: giraffe (giraffe)
Q3: Do you find a stone wall?
A3: yes (yes)

Generated Caption
- a giraffe standing next to a tree in a zoo enclosure.

Generated Captions
- a bunch of green vegetables on a table along with some literature.

Top -5 Attributes
- table, vegetables, broccoli, carrots, onions

Question Answering
Q1: Do all the vegetables have roots?
A1: no (no)
Q2: Wouldn't you like to participate in a CSA with veggies like these?
A2: yes (yes)

Top -5 Attributes
- room, couch, pillows, table, coffee

Question Answering
Q1: What room is this?
A1: living room (living room)
Q2: What shape is the table without any lamps on it?
A2: round (round)

Generated Captions
- a living room with a couch and a coffee table.

Figure 2. Some qualitative results demonstrating our attribute predictions, image captions and question answering. Ground truth answers are in parentheses. Blue indicates we give the right answer, red means we are wrong.

Generated Caption
- two zebras standing next to each other in a zoo enclosure.

Top -5 Attributes
- zebra, standing, ground, two, zoo

Question Answering
Q1: Where is this picture taken?
A1: zoo (zoo)
Q2: How many zebras?
A2: 2 (2)
Q3: Is the zebra eating cake?
A3: yes (no)

Generated Captions
- a plate of food sitting on a table with a glass of wine.

Top -5 Attributes
- wine, table, meat, white, vegetables

Question Answering
Q1: What is this drink?
A1: wine (wine)
Q2: How many slices of meat is here?
A2: 2 (6)
Q3: What brand of wine is that?
A3: Daisies (Bock)

Generated Caption
- a baseball player is swinging a bat at a ball.

Top -5 Attributes
- baseball, bat, swinging, red, people

Question Answering
Q1: What brand of cleats is the athlete wearing?
A1: Nike (Nike)
Q2: What type of hat is the better wearing?
A2: baseball (helmet)

Generated Caption
- a man is holding a cell phone in his hand.

Top -5 Attributes
- people, holds, cellphone, air, racket

Question Answering
Q1: Is he holding a camera?
A1: yes (yes)
Q2: What is on the man's back?
A2: backpack (backpack)
Q3: Is he bald?
A3: yes (yes)

Generated Caption
- a small bathroom with a toilet and a sink.

Top -5 attributes
- bathroom, door, small, wall, sink

Question Answering
Q1: What type of room is this?
A1: bathroom (bathroom)
Q2: Is there medicine in the medicine cabinet?
A2: no (no)
Q3: What is the wall made of?
A3: brick (stone)

Generated Caption
- a bunch of bananas hanging from a tree.

Top -5 Attributes
- bananas, tree, large, green, ground

Question Answering
Q1: Is this a fruit or vegetable?
A1: fruit (fruit)
Q2: Are these bananas ripe?
A2: no (no)
Q3: Does this tree have large leaves?
A3: no (yes)

Generated Caption
- a herd of sheep standing on top of a lush green field.

Top -5 Attributes
- sheep, field, grass, standing, green

Question Answering
Q1: Which animals are these?
A1: sheep (sheep)
Q2: Will the sheep taste good?
A2: yes (yes)
Q3: What type of ecosystem was this picture taken in?
A3: sheep (farm)

Generated Caption
- a man sitting on a bench in front of a building.

Top -5 Attributes
- bench, park, people, sitting, white

Question Answering
Q1: What color are the slats on the bench?
A1: green (green)
Q2: What's the statue holding?
A2: umbrella (newspaper)
Q3: What color is the statue?
A3: white (white)

Figure 3. Some qualitative results demonstrating our attribute predictions, image captions and question answering. Ground truth answers are in parentheses. Blue indicates we give the right answer, red means we are wrong.

Generated Caption
- a bathroom with a toilet sink and a mirror.

Top -5 Attributes
- bathroom, sink, wall, two, yellow

Question Answering
Q1: What kind of room is this?
A1: bathroom (bathroom)
Q2: How many sinks are there?
A2: 2 (2)
Q3: Is the door across from the sinks?
A3: yes (yes)

Generated Caption
- a person riding a snow board in the air.

Top -5 Attributes
- snow, people, snowboard, air, riding

Question Answering
Q1: What sport is the man engaging in?
A1: snowboarding (snowboarding)
Q2: Is the man touching the ground?
A2: no (no)
Q3: What is on the man's hands?
A3: gloves (gloves)

Generated Caption
- a brown horse standing on top of a grass covered field.

Top -5 Attributes
- horse, field, brown, grass, standing

Question Answering
Q1: What is the animal eating?
A1: grass (grass)
Q2: Is there a fence?
A2: yes (yes)
Q3: Is there a house in the background?
A3: yes (yes)

Generated Caption
- a busy city street filled with lots of cars.

Top -5 Attributes
- car, traffic, road, tree, people

Question Answering
Q1: Is the street crowded?
A1: yes (yes)
Q2: Can you see the body of ocean in the back?
A2: yes (yes)
Q3: How many red trucks are there?
A3: 2 (8)

Generated Caption
- a police officer riding a motorcycle on a city street.

Top -5 Attributes
- motorcycle, riding, people, office, helmet

Question Answering
Q1: Is this a police officer?
A1: yes (yes)
Q2: Is the police officer happy?
A2:yes (yes)
Q3: What type of vehicle is the policeman driving?
A3: motorcycle (motorcycle)

Generated Caption
- a man swinging a tennis racket at a tennis ball.

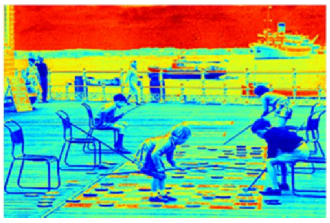
Top -5 Attributes
- people, tennis, racket, ball, hitting

Question Answering
Q1: What is the sport the man is playing?
A1: tennis (tennis)
Q2: Did the man hit the ball?
A2: yes (yes)
Q3: What car advertisement is in the background?
A3: Mercedes-benz (Mercedes-benz)

Generated Caption
- a man riding a skateboard down a sidewalk.

Top -5 Attributes
- skate, board, people, road, riding

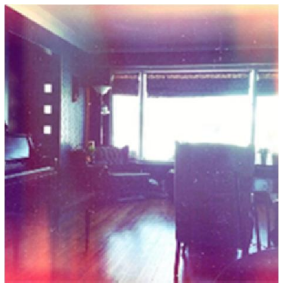
Question Answering
Q1: Is the skateboarder casting a shadow?
A1: yes (yes)
Q2: Is the boy airborne?
A2: yes (yes)



Generated Caption
- a group of young men playing a game of frisbee on a beach.

Top -5 Attributes
- boat, young, playing, water, children

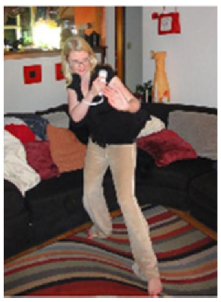
Question Answering
Q1: Is it a chilly day?
A1: yes (yes)
Q2: Are the children fishing?
A2: no (no)
Q3: Is this a recent photo?
A3: no (no)



Generated Caption
- a living room filled with furniture and a large window.

Top -5 Attributes
- room, furniture, large, windows, couch

Question Answering
Q1: Is it a sunny day?
A1: yes (yes)
Q2: Is there a big window?
A2: yes (yes)
Q3: What room is this?
A3: living room (living room)



Generated Caption
- a woman is playing a video game in a living room.

Top -5 Attributes
- people, playing, wii, room, glass

Question Answering
Q1: What gaming system is the woman playing?
A1: Wii (Wii)
Q2: Do you see pillows on the couch?
A2: yes (yes)
Q3: What color is the game controller?
A3: white (white)

Figure 4. Some qualitative results demonstrating our attribute predictions, image captions and question answering. Ground truth answers are in parentheses. Blue indicates we give the right answer, red means we are wrong.

References

[1] X. Chen and C. Lawrence Zitnick. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2015. 1

[2] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015. 1

[3] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, pages 853–899, 2013. 1

[4] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015. 1

[5] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). In *Proc. Int. Conf. Learn. Representations*, 2015. 1

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014. 1

[7] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proc. Int. Conf. Mach. Learn.*, 2015. 1

[8] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Proc. Conf. Association for Computational Linguistics*, 2, 2014. 1