Show, Infer and Tell: Contextual Inference for Creative Captioning

Ankit Khare
ankit.khare@mavs.uta.edu
The University of Texas at Arlington
Texas, USA

Manfred Huber
huber@cse.uta.edu

1 Supplementary Material

1.1 Qualitative Results and Analysis

Code has been made public at: https://ankit1khare.github.io/Show_Infer_and_Tell-CIC. We utilize the open source pytorch implementation [1] of the Up-Down Captioner to compare and contrast the qualitative results showing improvements in our final CIC-RCNN model in reference to up-down captioner.

In these supplemental examples, we demonstrate unique features of our model in terms of the quality of captions it produces. More specifically, we present cases where the model describes: (i) more than two items present in an image (Fig. 1), (ii) "out of context" scenarios like the presence of black and white photo frames (Fig. 2), (iii) images with the captions comprised of novel semantic constellations (Fig. 3), and (iv) images with the captions showing better utilization of visual features getting reflected in sophisticated language containing collocations (Fig. 4 and 5). Cases where the model fails to generalize well are shown in Fig. 6. Fig. 7 includes instances relating to the skateboard context where our model successfully understands the context and associates the interactions between the objects to learn collocations and contextual inference. Furthermore, we present additional examples (Fig. 8) demonstrating the collocations in captions produced by our model in various scenarios. Finally, we present examples of qualitative improvements comparing Up-Down Captioner, our experimental Mean-Pooled-CIC model, and our final CIC-RCNN model (Fig. 9). Note that the abbreviation GT stands for the randomly selected ground truth for an image.
Figure 1: Descriptive and detailed captioning.

**Up-Down:** A bathroom with a toilet and a sink.

**CIC-RCNN:** A bathroom with a sink, a toilet, and a bathtub.

**GT:** A bathroom with a bathtub next to a white toilet and a sink.

Our model produces descriptive captions for a lot of scenarios. It often describes more than two objects whereas other models only pick up one or two of them.

Figure 2: Detection of photo frames.

**Up-Down:** A cow and a sheep are standing in a field.

**CIC-RCNN:** Black and white photo of a cow and sheep standing on a field.

**GT:** Black and white photograph of a dog standing on a cow’s back.

A lot of images in the MS-COCO dataset contain a series of photos or a photo frame. Our model succeeds in identifying them as shown in the example considered here.
There are many instances in the dataset where the model sees a cement block and skateboarders doing tricks, but mostly in different contexts. Still, it learns to associate them well. Additionally, we observe collocations like "on top of" associated with the "skateboard" in the generated captions.

The model describes the position of the woman more precisely in relation to the presence of pigeons within the image. We also see that the collocation "large flock of" associated with birds is learned by the model.
A failure case where the model makes an intelligent guess that is better than the Up-Down model

**Up-Down:** A dog wearing a tie and a sweater.
**CIC-RCNN:** A dog wearing a pair of shoes.
**GT:** Animal with hoofs made to wear colorful pair of shoes.

The model describes the color of the jersey along with the girl’s current action

**Up-Down:** A woman kicking a soccer ball on a field.
**CIC-RCNN:** A girl in a white uniform kicking a soccer ball.
**GT:** A young lady kicking a soccer ball on a field.

Figure 5: Two examples are shown where the model utilizes the image visual features to a better extent and is able to reflect it in the captions.

**Up-Down:** Two uncooked pizzas sitting on a counter top.
**CIC-RCNN:** A pizza sitting on top of a cutting board.
**GT:** The halved melon is on the counter next to the remote.

**Up-Down:** A woman is looking at her phone in front of a train.
**CIC-RCNN:** A woman is standing in front of a train.
**GT:** A woman looking out the window of a train.

**Up-Down:** A woman wearing a helmet and holding a cup.
**CIC-RCNN:** A woman wearing a helmet and glasses holding a knife.
**GT:** A man dressed in a helmet and goggles indoors has a goofy smile on and his hand raised.

Figure 6: Examples of failure cases.
A blurry photo of a person riding a skateboard.

A white dog sitting on top of a skateboard.

A man riding a skateboard on a street at night.

A man doing a trick on a skateboard in the water.

A person standing next to a skateboard on the street.

A man doing a trick on a skateboard in the air.

A woman riding a skateboard on a street.

A man sitting on a skateboard on a street.

A man doing a trick on a skateboard on a bench.

A young boy holding a skateboard in the street.

A man sitting in front of a building with a skateboard.

A group of people riding skateboards down a street.

A young man riding a skateboard in front of a wall.

A man holding a skateboard in front of a store.

A man riding a skateboard with a dog on a street.

A little girl is standing on the grass with a skateboard.

Figure 7: A series of images highlighting our final model’s ability to associate objects, object’s attributes and object’s surroundings to learn collocation and contextual inference. We specifically chose the skateboard related context with diverse surroundings and a variety of different interactions between objects. All images are taken from the MS-COCO 2014 test-set and thus, they are never seen by the model during training.
A series of photos showing different types of food.

A cat wearing a pink hat sitting on top of a carpet.

A stone patio with a stone walkway and a lot of chairs.

A group of men standing around a sheet cake holding a sword.

A bathroom with a walk in shower next to a toilet.

A classic car parked in the grass near a group of people.

A man flying through the air while riding a kiteboard.

A man in a yellow shirt catching a white frisbee.

A school bus is reflected in a rear view mirror.

Figure 8: Additional examples showing the CIC-RCNN model’s ability to utilize the rich image features from the encoder, showing creativity, a higher level of detail and more human-like inferences.
GT: Two girls under an umbrella with health information on it.

Up-Down: A woman holding an umbrella in the rain.
Mean-Pooled-CIC: A couple of people standing under a blue umbrella.
CIC-RCNN: Two young girls are under a blue umbrella in the rain.

GT: A large clock tower on a hill overlooking a city.

Up-Down: A clock tower on top of a building.
Mean-Pooled-CIC: A clock tower with a weather vane on top.
CIC-RCNN: A clock tower with a weather vane on top and a clock on each of its sides.

GT: A long buffet table of food with people standing around it.

Up-Down: A group of people standing under a table with food.
Mean-Pooled-CIC: A group of people standing under a table filled with food.
CIC-RCNN: A group of people standing around a table with plates of food.

GT: A very large elephant standing next to a baby elephant.

Up-Down: A baby elephant standing next to a tree.
Mean-Pooled-CIC: A baby elephant standing next to an adult elephant.
CIC-RCNN: A baby elephant standing next to a larger elephant.

Figure 9: Examples showing the qualitative comparison between Up-Down Captioner, our experimental Mean-Pooled-CIC model and our final CIC-RCNN model.
References