Areas of Attention for Image Captioning
— Supplementary Material —

Marco Pedersoli[1] Thomas Lucas[2] Cordelia Schmid[2] Jakob Verbeek[2]
[1] École de technologie supérieure, Montréal, Canada
[2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[1]Marco.Pedersoli@etsmtl.ca [2]firstname.lastname@inria.fr

In this supplementary material we report additional detailed performance measures in Section 1, we describe the gated recurrent unit (GRU) used in our work with more detail in Section 2, and provide additional visualizations of our attention model in Section 3.

1. Additional evaluation results

For sake of brevity we reported only the three metrics that are most commonly used in the recent captioning literature in the main paper, the same three as in *e.g.* [1, 8, 11]. While the evaluation of caption quality remains a challenging issue, the CIDEr-D metric [7] is generally considered to be correlating the best to human judgement. The BLEU metrics [6] are based on N-gram matching statistics. In particular, the BLEU1 metric completely disregards word ordering, and is thus of little interest to measure sentence quality. From the BLEU measures, BLEU4 (based on 4-grams) is most commonly used [7].

In tables 1, 2, and 3 provide the evaluation results including the BLEU 1–3 metrics. The conclusion of the comparisons among the variants of our model and to the state of the art remain unchanged. The tables here correspond to those with the same numbers in the main paper. We refer to the main paper for a full description of the experimental setup.

	BLEU1	BLEU2	BLEU3	BLEU4	Meteor	CIDEr
Baseline: θ_{wh}	66.3	48.5	35.5	26.4	22.2	78.9
Ours: θ_{wh}, θ_{wr}	68.0	50.8	37.5	28.0	22.9	83.6
Ours: $\theta_{wh}, \theta_{wr}, \theta_{rh}$	68.2	51.2	37.9	28.4	23.3	85.5
Ours: conditional feedback	68.3	51.1	38.1	28.7	**23.7**	86.8
Ours: full model	**69.1**	**51.9**	**38.5**	**28.8**	**23.7**	**87.4**

Table 1. Evaluation of the baseline and our attention model using activation grid regions, including variants with certain components omitted, and word-conditional instead of marginal feedback.

2. Details on gated recurrent units (GRUs)

To alleviate the problem of vanishing or exploding gradients encountered in deep and recurrent neural networks (RNNs), gated units have been proposed. The most two well known ones are LSTMs [4] and GRUs [2]. Such units use a gating mechanism to control the flow of information depending on the input, enabling better learning of long-term dependencies. In our work we used GRUs, based on better results in initial experiments with the baseline model. We provide here a brief description of the GRU mechanism, see [2] for more details.

We use $h_t \in \mathbb{R}^{d_h}$ to denote the RNN state, and x_t as the input to the RNN at time t. To compute the state evolution, a gated recurrent unit (GRU) makes use of two gates: a "forget gate" $z_t \in [0,1]^{d_h}$ and a "read gate" $r_t \in [0,1]^{d_h}$. Both gates are computed as a sigmoid of a linear function of the input and previous state:

$$z_t = \sigma\left(\omega_{zx} x_t + \omega_{zh} h_{t-1}\right), \tag{1}$$

$$r_t = \sigma\left(\omega_{rx} x_t + \omega_{rh} h_{t-1}\right). \tag{2}$$

	BLEU1	BLEU2	BLEU3	BLEU4	Meteor	CIDEr
			RNN training only			
Baseline	66.3	48.5	35.5	26.4	22.2	78.9
Activation grid	69.1	51.9	38.5	28.8	23.6	87.4
Object proposals	69.4	52.2	38.7	28.9	23.7	89.0
Spatial transformers	**70.2**	**53.3**	**40.0**	**30.2**	**24.2**	**91.1**
			CNN-RNN fine-tuning			
Baseline	68.6	51.3	38.1	28.7	23.5	87.1
Activation grid	70.4	53.4	40.1	30.3	**24.5**	92.6
Object proposals	**71.0**	**54.1**	40.3	30.1	**24.5**	93.7
Spatial transformers	70.8	53.9	**40.6**	**30.7**	**24.5**	**93.8**

Table 2. Captioning performance of the baseline and our model using different attention regions, with and without fine tuning.

	CNN	FT	B1	B2	B3	B4	Meteor	CIDEr
Xu *et al*. [10], soft	V19	N	70.7	49.2	34.4	24.3	23.9	—
Xu *et al*. [10], hard	V19	N	71.8	50.4	35.7	25.0	23.0	—
Yang *et al*. [11]	V16	N	—	—	—	29.0	23.7	88.6
Jin *et al*. [5]	AN+V16	N	69.7	51.9	38.1	28.2	23.5	83.8
Donahue *et al*. [3]	AN	N?	66.9	48.9	34.9	24.9	—	—
Bengio *et al*. [1]	GN	N?	—	—	—	30.6	24.3	92.1
Wu *et al*. [9]	V?	Y	74	56	42	31	26	94
Areas of Attention	V16	Y	70.8	53.9	40.6	30.7	24.5	93.8
			Ensemble methods					
Vinyals *et al*., ens. [8]	GN	N	—	—	—	27.7	23.7	85.5
You *et al*., ens. [12]	GN	N?	70.9	53.7	40.2	30.4	24.3	—
Bengio *et al*., ens. [1]	GN	N?	—	—	—	32.3	25.4	98.7
Areas of Attention, ens.	V16	Y	73.2	56.1	42.5	31.9	25.2	98.1

Table 3. Comparison of our results to the state of the art on the COCO dataset. The key for CNN column is: GN=GoogLeNet, VXX=VGG-XX, AN=AlexNet. The FT column indicates if the CNN component of the model was finetuned or not.

The read gate r_t is used, together with the previous hidden state h_{t-1} and the input x_t, to compute a "tentative" state $\tilde{h}_t$:

$$\tilde{h}_t = \tanh\left(\omega_{hr}(r_t \odot h_{t-1}) + \omega_{hx}x_t\right), \tag{3}$$

where $\odot$ denotes the element-wise product of two vectors.

Finally, the forget gate controls to what extent the previous state h_{t-1} is maintained, or replaced by the tentative state $\tilde{h}_t$:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t. \tag{4}$$

These updates together constitute the state update function $h_{t+1} = g(h_t, w_t)$ used in the main paper.

In the baseline model, presented in Section 3.1 of the main paper, the input x_t used to compute h_{t+1} is the previously generated word w_t. In our attention model the input to compute h_{t+1} is a concatenation of the previously generated word w_t and the visual feedback vector v_t as defined in Equation 6 of the main paper.

3. Additional visualizations

We present two sets of additional visualizations. First we provide visualizations similar to those of Figure 5 of the main paper. We visualize the focus of our attention model during sequential word generation for the three different region types: activation grids, object proposals, and spatial transformers. The attention areas are drawn with line widths directly proportional to weights $p(r_t|h_t)$, see Section 3.2 of the main paper. The images displayed for the object proposals differ

slightly from the others, since the high-resolution network used in that case uses a different cropping and scaling scheme. To visualize the attention regions, we show the areas from which the convolutional features are pooled. For the spatial transformers, we show the transformed anchor boxes. For the activation grid regions, we show the back-projection of a 3×3 activation block, which allows for direct comparison with the spatial transformers. For object proposals we show the edge-boxes. Note that in all cases the underlying receptive fields are significantly larger than the depicted areas.

In a second set of visualizations, starting at page 7, we show for a set of random images not seen during training the sentence generated by our ensemble model, together with the five ground-truth sentences provided with these images.

References

[1] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, 2015.

[2] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS Deep Learning Workshop*, 2014.

[3] J. Donahue, L. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[5] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv:1506.06272, 2015.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002.

[7] R. Vedantam, C. Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015.

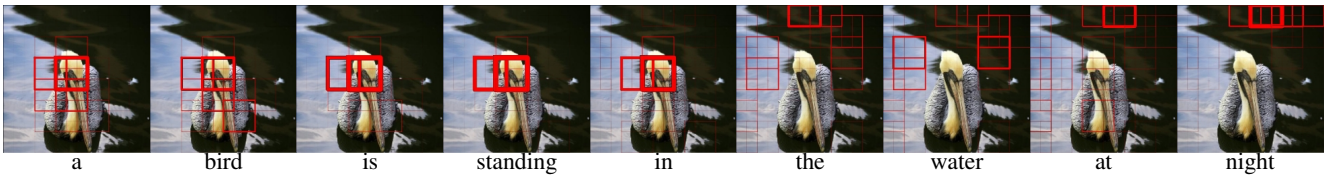
[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

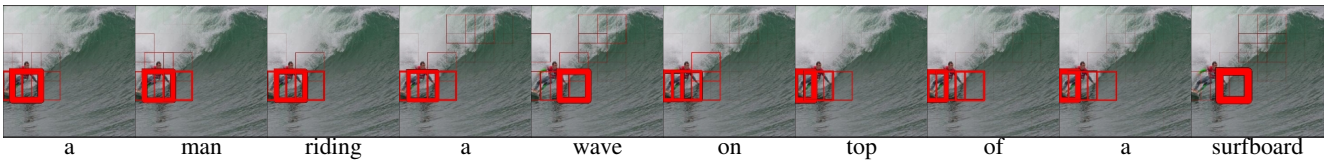
[9] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel. What value do explicit high level concepts have in vision to language problems? In *CVPR*, 2016.

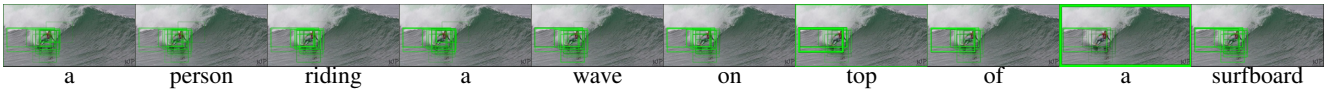
[10] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

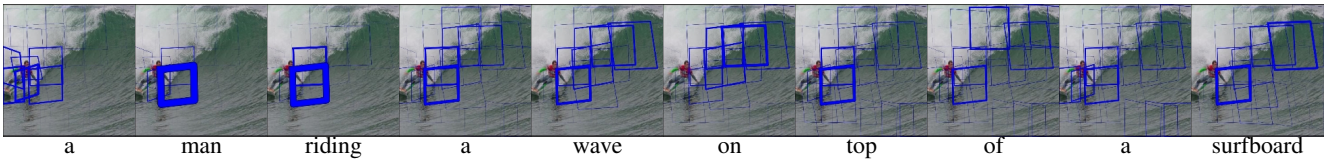
[11] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. Cohen. Encode, review, and decode: Reviewer module for caption generation. In *NIPS*, 2016.

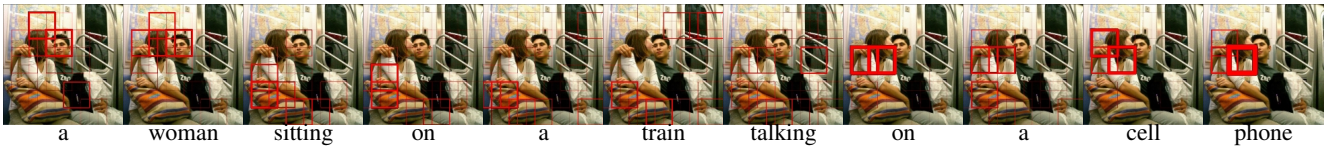
[12] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, 2016.


a bird is standing in the water at night
a bird is standing on a rock in the water
a bird perched on top of a bird feeder
a group of children sitting around a table with food
a group of people sitting around a table eating food
a little girl sitting at a table with a plate of food
a large jetliner sitting on top of an airport tarmac
a large jetliner sitting on top of an airport tarmac
a large jetliner sitting on top of an airport runway

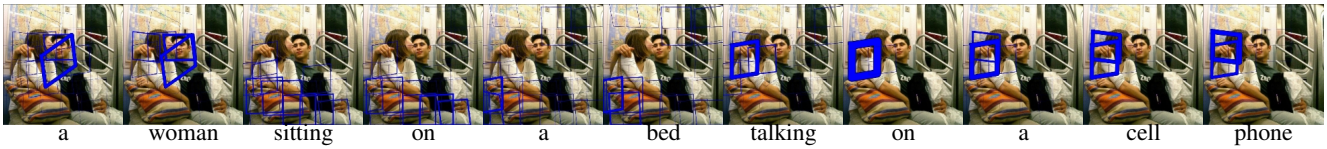

a man riding a wave on top of a surfboard

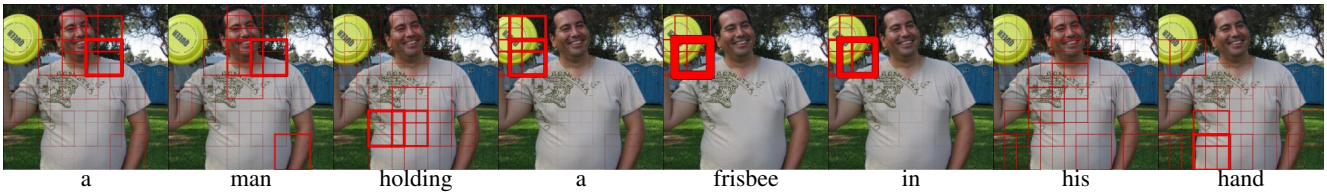

a person riding a wave on top of a surfboard

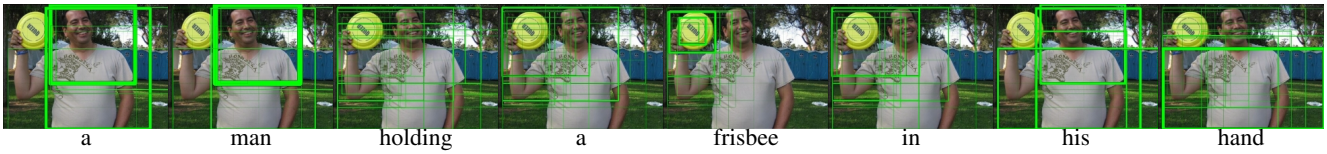

a man riding a wave on top of a surfboard


a woman sitting on a train talking on a cell phone

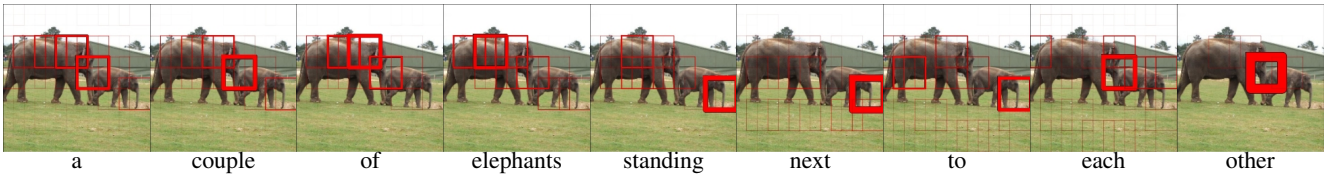

a woman sitting on a bed with a bag of luggage

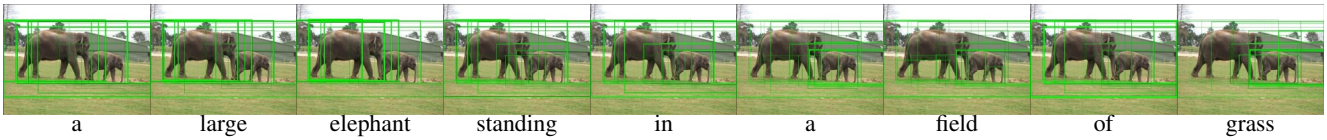

a woman sitting on a bed talking on a cell phone

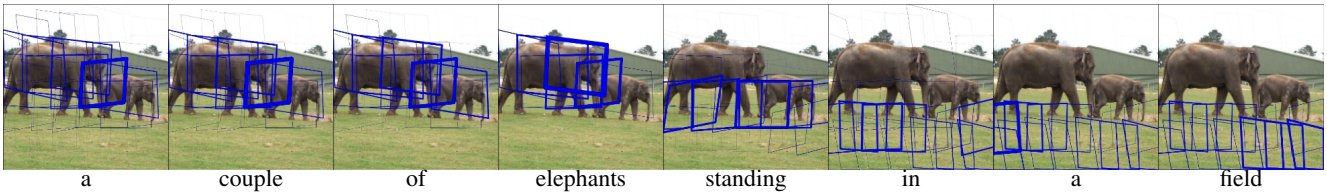

a man holding a frisbee in his hand

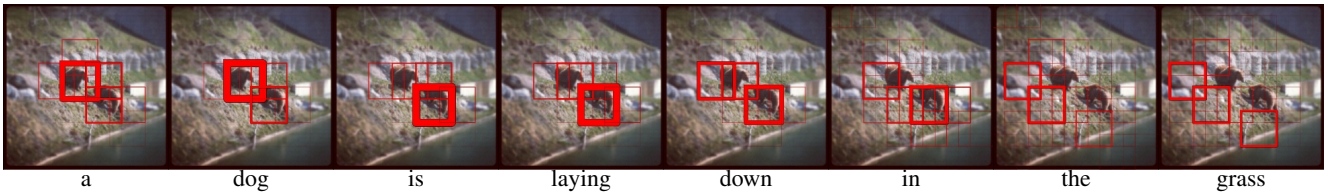

a man holding a frisbee in his hand


a man holding a frisbee in his hand

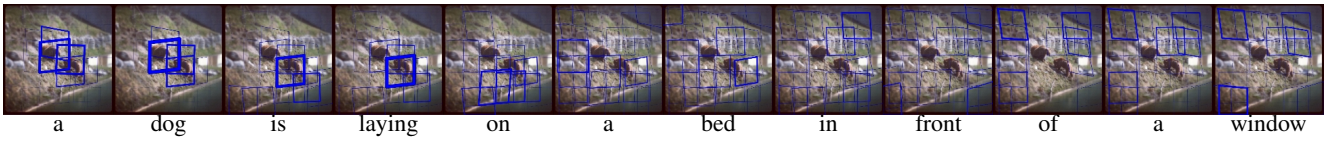

a couple of elephants standing next to each other

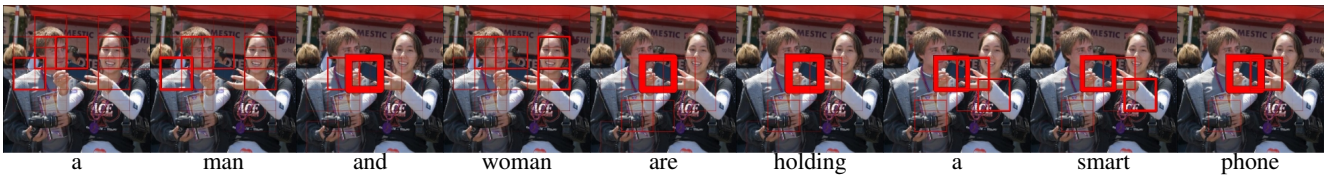

a large elephant standing in a field of grass


a couple of elephants standing in a field

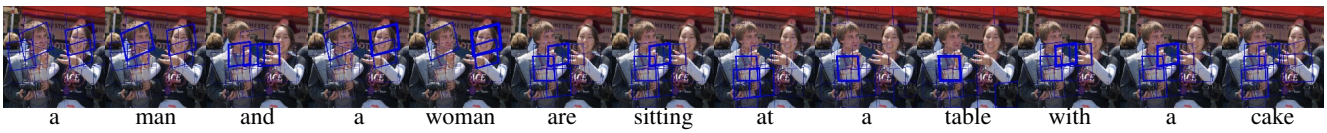

a dog is laying down in the grass


a bear is standing in the middle of a field


a dog is laying on a bed in front of a window


a man and woman are holding a smart phone


a group of people sitting around a table with a cell phone


a man and a woman are sitting at a table with a cake

Generated caption:
- A man is doing a trick on a skateboard
Ground-truth captions:
- A man doing a trick on a skateboard
- The man is jumping over the skate board
- Man in a skateboard being handed another skateboard in the air.
- A man that is holding a skateboard in the air.
- A man flying through the air on top of a skateboard.

Generated caption:
- A little girl sitting at a table with a plate of food
Ground-truth captions:
- A group of children standing around a table together.
- Four children are sitting at a table eating desserts.
- A group of kids that are sitting around a table.
- Four small children that are enjoying a small snack.
- Four young children eat cake off of disposable plates around a low plastic table on a porch.

Generated caption:
- A fire hydrant sitting on the side of a street
Ground-truth captions:
- A dumpster sitting in front of a building covered in graffiti.
- A small garbage can on a sidewalk with graffiti
- A graffiti covered dumpster sits on a sidewalk.
- Spray painted garbage dumpster sitting in front of a black fire hydrant.
- Street scene of black fire hydrant in front a dumpster with graffiti on its side.

Generated caption:
- A man riding a wave on top of a surfboard
Ground-truth captions:
- A man riding a wave on top of a surfboard.
- a man is enjoying surfing the the raging waves.
- A man on a surfboard riding a large wave with one hand skimming against the water.
- a surfer in a red shirt is surfing on a white board
- A surfer is riding a big wave in the ocean.

Generated caption:
- A man is doing a trick on a skateboard
Ground-truth captions:
- a man flying through the air while riding a skateboard.
- A skateboarder is high in the air separated from his board in the break of a ramp.
- A boy in a blue shirt at a skate park doing tricks on his skateboard.
- Man riding skateboard over up a ramp and over a gap.
- a skateboarder is wearing a blue shirt and doing a trick

Generated caption:
- A man standing in a living room playing a video game
Ground-truth captions:
- A man standing in a living room holding a Nintendo Wii game controller.
- Man standing in a family room with a game controller in his hand.
- A man is standing and playing video games.
- A man is standing in a living room with a Wii controller.
- A middle aged man is playing with a WII game.

Generated caption:
- A baseball player swinging a bat at a ball

Ground-truth captions:
- A batter holds the bat behind his head for a powerful swing.
- A man holding a bat at a professional baseball game.
- a baseball player in the batters box at a game
- A man is at bat during a professional baseball game.
- There are spectators watching a baseball game going on.

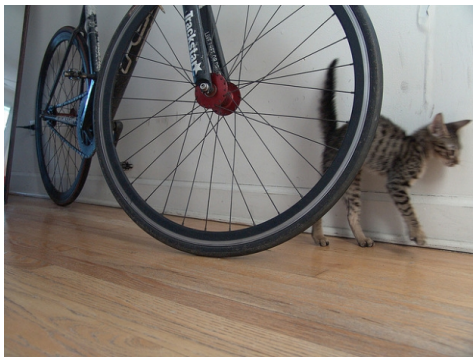
Generated caption:
- A vase filled with flowers on a table

Ground-truth captions:
- A counter containing two silver vases with colorful flowers.
- A base with yellow pink and orange daisies in it.
- There are some flowers in decorative silver vases
- Colorful flowers in a metal vase sitting on a mirror ledge.
- Flowers in a paper that is silver and standing.

Generated caption:
- A bathroom with a toilet sink and bathtub

Ground-truth captions:
- A tiled bathroom containing a vanity sink, toilet and bathtub.
- a bath room with a toilet a bath tub and a sink
- There is a bathroom with a toilet and tub.
- A white toilet and bath in a room.
- A white bathtub sitting next to a toilet.

Generated caption:
- A man in a suit and tie wearing a hat
Ground-truth captions:
- A man wearing a hat and a gray jacket.
- A man with a suit and tie and a fedora on.
- A man wears a suit with a blue shirt and a multicolored tie.
- A man in a suit and tie wearing a hat.
- a man that is in a suite and a tie



Generated caption:
- A cat is standing next to a bicycle
Ground-truth captions:
- A cat standing next to a bike parked against a wall.
- A ca walking with its tail straight up .
- A kitten is walking next to a parked bike inside.
- A small cat is walking behind a bike.
- A baby tabby cat walking behind a bicycle leaning against a wall

Generated caption:
- A large airplane flying through a blue sky
Ground-truth captions:
- A picture of a plane is flying in the air.
- An airplane with two propellor engines flying in the sky.
- A propeller plane flying through a blue sky.
- A small airplane is flying through a clear blue sky.
- A propeller plane flying through a blue sky.

Generated caption:
- A couple of kids laying on a bed
Ground-truth captions:
- A custom cake featuring a fisherman for a man's 65th birthday.
- A birthday cake made to look like a man on a pier fishing.
- a fathers birthday cake with a pond and sheep on the hill
- The birthday cake is in the shape of a hill with a fisherman sitting on it.
- The birthday cake for a 65 year old.

Generated caption:
- A large body of water with boats in the background
Ground-truth captions:
- Two boats that are sitting in the water.
- Those boats are waiting by the pier in the water.
- A large city is on the water with boats.
- City skyline as seen beyond waterway docking area.
- A large body of water filled with boats next to a tall building.

Generated caption:
- A man holding a piece of cake on a plate
Ground-truth captions:
- A man is sitting down with a piece of chocolate cake in front of him with a fork in his hand.
- a person at a table with a large piece of cake
- This man is holding a fork to eat a piece of chocolate cake.
- The man sitting at a table with a large slice of cake.
- a man sitting at a table with a piece of cake and holding a fork

Generated caption:
- A man standing in a kitchen preparing food
Ground-truth captions:
- A cook in a restaurant kitchen putting chopped vegetables in a bow.
- At a restaurant's kitchen, a gentleman wearing sanitary gloves prepares a salad.
- A man in a blue shirt is putting food into a bowl.
- The man is in the kitchen preparing a meal.
- A man in blue shirt preparing food in a kitchen.

Generated caption:
- A herd of sheep grazing on a lush green field
Ground-truth captions:
- A herd of animals traveling down a country road surrounded by a lush green landscape.
- Jogger running past a flock of sheep in a rural area
- A man walking beside sheep on a country road.
- Jogger running past a flock of white sheep.
- A runner wearing spandex has come across a large herd traveling down the dirt road.

Generated caption:
- A woman is eating a piece of pizza
Ground-truth captions:
- A woman wearing a hat bites into a pastry.
- A girl eating a donut out of a bag
- a woman eating a powdered sugar pastry out of a bag
- A woman eating a pastry at a coffee shop.
- The woman in the hat is taking a bite of a pastry.