

Reflective Decoding Network for Image Captioning: Supplementary Material

Lei Ke¹, Wenjie Pei², Ruiyu Li², Xiaoyong Shen², Yu-Wing Tai²

¹The Hong Kong University of Science and Technology, ²Tencent

We provide more qualitative and quantitative evaluation of our Reflective Decoding Network (RDN) in this document.

1. Additional qualitative results

In Figure 1, we present more examples for our RDN, particularly, including two cases where RDN has no obvious improvements on the baseline model. To comprehensively show the caption decoding process, textual visualization cases for multiple key words in a single sentence are shown in Figure 2. Additional example captions from Up-Down [1] and our RDN, along with their corresponding ground truth captions, are shown in Figure 3 and Figure 4.


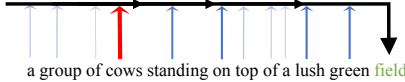

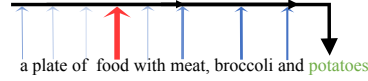

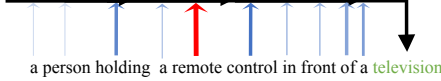



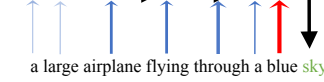


	Generated sentence	Reflective weight visualization
	Baseline: a herd of cows standing next to each other Ours: a group of cows standing on top of a lush green field	
	Baseline: a white plate topped with meat and broccoli Ours: a plate of food with meat, broccoli and potatoes	
	Baseline: a person that is holding a remote control Ours: a person holding a remote control in front of a television	
	Baseline: a dog that is standing in a bathroom Ours: a dog standing in a bathroom next to a sink	
	Baseline: an airplane is flying in the blue sky Ours: a large airplane flying through a blue sky	
	Baseline: a man on a surfboard riding a wave Ours: a man on a surfboard riding on wave	

Figure 1. More examples of captions generated by our RDN compared to the baseline (Up-Down [1]) and their reflective attention weight distribution over the past generated hidden states when predicting the key word highlighted in green. The thicker line indicates a relatively larger weight and the red line means the largest contribution. The bottom two examples show the situation that the baseline model has sufficient capability to cope with and our RDN model has no obvious improvement.


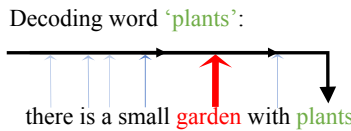
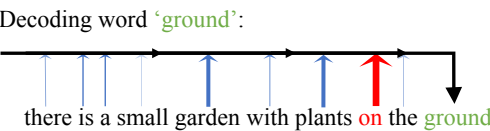

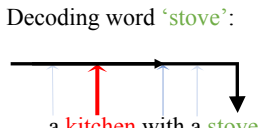
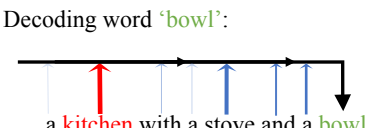
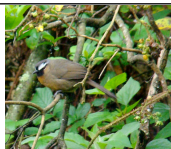
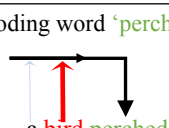
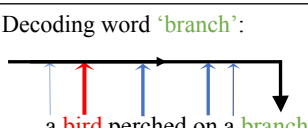
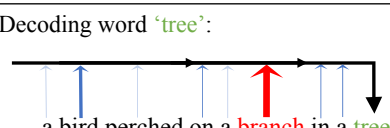

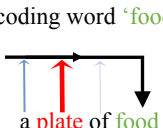
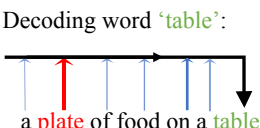
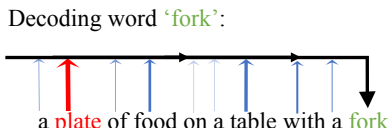

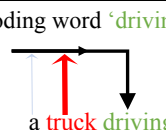
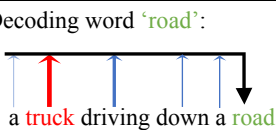
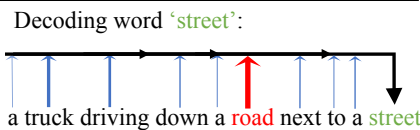

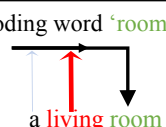
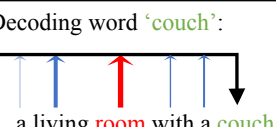
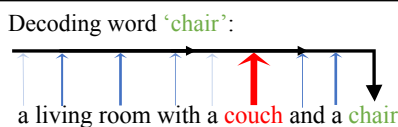

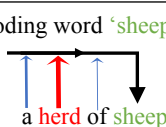
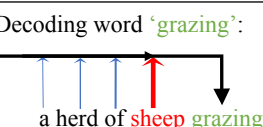
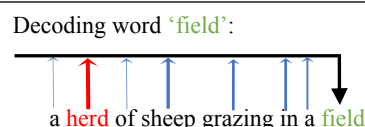

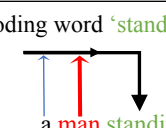
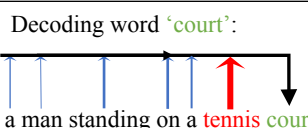
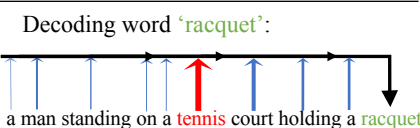
Complete sentence	RDN: there is a small garden with plants on the ground		
	Decoding word 'plants': 	Decoding word 'ground': 	
Complete sentence:	RDN: a kitchen with a stove and a bowl of fruit		
	Decoding word 'stove': 	Decoding word 'bowl': 	
Complete sentence:	RDN: a bird perched on a branch in a tree		
	Decoding word 'perched': 	Decoding word 'branch': 	Decoding word 'tree': 
Complete sentence:	RDN: a plate of food on a table with a fork		
	Decoding word 'food': 	Decoding word 'table': 	Decoding word 'fork': 
Complete sentence:	RDN: a truck driving down a road next to a street sign		
	Decoding word 'driving': 	Decoding word 'road': 	Decoding word 'street': 
Complete sentence:	RDN: a living room with a couch and a chair		
	Decoding word 'room': 	Decoding word 'couch': 	Decoding word 'chair': 
Complete sentence:	RDN: a herd of sheep grazing in a field		
	Decoding word 'sheep': 	Decoding word 'grazing': 	Decoding word 'field': 
Complete sentence:	RDN: a man standing on a tennis court holding a racquet		
	Decoding word 'standing': 	Decoding word 'court': 	Decoding word 'racquet': 

Figure 2. Further visualization examples show reflective attention weight distribution over the past generated hidden states for multiple key words in a single sentence during caption generation process for our RDN. The words highlighted in green are usually strong related with the words highlighted in red in vocabulary coherence, such as the correlations between words “garden” and “plants”, “kitchen” and “stove”.

Image	Generated sentence	Ground truth
	<p>Baseline: a boat that is in the water</p> <p>Ours: a boat floating on the water in front of a city</p>	<ol style="list-style-type: none"> 1. a ship in the water sailing past the city in the background 2. a shoreline is empty except for a single boat with two red flags 3. a boat moving along the water with city buildings in the backdrop 4. a single boat travels through the water outside a large city 5. an old looking ship sailing in water by a modern city
	<p>Baseline: a black and white dog running in the grass</p> <p>Ours: a black and white dog in a field of sheep</p>	<ol style="list-style-type: none"> 1. a group of sheep being herded by a black dog 2. a flock of sheep being herded by a black and white sheep dog 3. a herd of sheep and their sheep dog run in a pasture 4. a black and white dog herding sheep across a field 5. a sheep dog herding a flock in the field
	<p>Baseline: a boat in the middle of the river</p> <p>Ours: a small boat on a river in front of some red trees</p>	<ol style="list-style-type: none"> 1. a small boat is going down the river in front of colorful trees 2. a terraced hill in fall colors going down to the water with a boat on it 3. a red blue and yellow boats and some red trees 4. boat on water above trees with fall foliage 5. a series of steep stairs lay next to a lake
	<p>Baseline: a man with a hat and a dog</p> <p>Ours: a man wearing a hat and sunglasses with a horse</p>	<ol style="list-style-type: none"> 1. a man with a beard is close to a farm animal 2. a man with a beard and hat is petting a horse 3. a man with a large red beard petting a brown horse 4. the man is petting his horse while he is outside 5. a man with a rough red beard pets a horse
	<p>Baseline: a couple of sheep standing next to each other</p> <p>Ours: a row of sheep standing next to a fence fed by a man</p>	<ol style="list-style-type: none"> 1. a person feeding sheep behind a white picket fence 2. several sheep with their front hooves on a fence 3. several lambs leaning up and over a wooden post towards a red cup held by a man 4. a row of sheep by a wooden fence with a guy holding a bowl 5. four sheep standing against a fence looking at a man with a red bowl
	<p>Baseline: a man playing tennis on a blue court</p> <p>Ours: a man swinging a tennis racquet to hit a ball on a court</p>	<ol style="list-style-type: none"> 1. a man hitting a tennis ball on a court 2. a man leaping to hit a tennis ball with his tennis racket 3. a male tennis player jumping up to hit ball 4. a man holds his racket out while on the tennis court 5. a tennis player playing a game of tennis
	<p>Baseline: a collage of pictures of different images of photos</p> <p>Ours: a collage of different pictures of a dog doing different things</p>	<ol style="list-style-type: none"> 1. there is a photo of a collage of photos of a dog 2. 25 images of dogs are formed into a square grid 3. twenty various pictures of a small dog doing different things 4. the dog enjoys playing outside and laying down in the house 5. many shots grouped together of the same herding breed dog, some images of it playing, sleeping or running through obstacles

Figure 3. Examples of captions generated by our RDN compared to the baseline (Up-Down [1]), as well as the corresponding ground truths.

Image	Generated sentence	Ground truth
	<p>Baseline: a small child eating a piece of food</p> <p>Ours: a baby eating a piece of broccoli in his mouth</p>	<ol style="list-style-type: none"> 1. a small baby girl eating a piece of broccoli 2. a baby is eating a piece of broccoli from a plate 3. a baby in a high chair holds a piece of broccoli 4. a baby eats a piece of broccoli from a plate 5. a little girl sits in her high chair chomping on a plate of broccoli
	<p>Baseline: a cat laying on top of a computer mouse</p> <p>Ours: a cat laying on the floor next to a pair of shoes</p>	<ol style="list-style-type: none"> 1. an orange and white cat laying on top of black shoes 2. a white and gray cat sniffing a woman's heeled shoe 3. a cat is smelling a lady's shoe heel 4. a cat sniffing the heel of a shoe 5. a cat sits on top of a shoe with its noes on the heel
	<p>Baseline: a couple of boats parked next to each other</p> <p>Ours: a couple of boats docked at a bridge</p>	<ol style="list-style-type: none"> 1. houseboats are docked on the water's edge near a long bridge 2. the colorful boats are in the blue water 3. a picture of house boats and a long bridge over water on a cloudy day 4. a dock is near a bridge over a river 5. a few boats that are out on a river
	<p>Baseline: airplane is flying in the air on the beach:</p> <p>Ours: an airplane is flying over the beach on a cloudy day</p>	<ol style="list-style-type: none"> 1. an airplane flying through a cloudy sky flying over the ocean 2. an airplane is flying over the beach crowded with people 3. there is an airplane flying over a crowded beach 4. an airplane flying over a crowded beach 5. people on the beach look at the plane flying overhead
	<p>Baseline: a close up of a plate of food</p> <p>Ours: a white plate topped with pasta and broccoli</p>	<ol style="list-style-type: none"> 1. pasta with sauce and broccoli on white plate 2. a macaroni and broccoli dish on a white plate 3. a pasta dish with mushrooms and broccoli in a cream sauce 4. a plate of pasta is shown with broccoli and ham 5. a plate holds a good size portion of a cooked, mixed dish that includes broccoli and pasta
	<p>Baseline: a bunch of flowers that are in a room</p> <p>Ours: a cake decorated with lots of colorful flowers</p>	<ol style="list-style-type: none"> 1. a large multi layered cake with candles sticking out of it. 2. a cake is decorated with flowers and flags 3. a party decoration containing flowers, flags and candles 4. a cake decorated with flowers and flags 5. a layered cake with many decorations on a table
	<p>Baseline: a cow that is walking down the street</p> <p>Ours: a cow walking down a street in front of a store</p>	<ol style="list-style-type: none"> 1. a cow standing near a curb in front of a store 2. there is a cow on the sidewalk standing in front of a store 3. a cow on the sidewalk on a corner in front of a store 4. cow standing on sidewalk in city area near shops 5. a cow on a city sidewalk in front of a business

Figure 4. More examples.

2. Additional quantitative results

To investigate our RDN’s effectiveness on spatial convolutional feature, we conduct experiments to directly apply RDN to utilize spatial ConvNet features as shown in Table 1. The comparison of our RDN and Att2in [3] on hard image captioning is provided in Figure 5.

We trained a transformer-based network for a more comprehensive analysis. Its performance and model complexity are reported in Table 2. We suspect the worse performance of transformer-based network is due to the limited training data in COCO. Note that the original transformer network requires millions of samples in language translation tasks. Thus, although the high level idea of our works share some similarity with the transformer, our model performs much better and with much lower model complexity.

We also provide more ablation study results: 1) The ablation study on beam search is in Table 3. 2) The comparison between teacher forcing and student forcing is presented in Table 4. 3) Performance comparison on CIDEr optimization is in Table 5.

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr
Baseline (ConvNet)	74.1	33.3	26.1	54.4	105.3
RDN (ConvNet)	74.6	33.9	26.4	54.8	107.8

Table 1. Experiment result for applying RDN using spatial convolutional feature ($2048 \times 7 \times 7$, extracted from the last layer of ResNet) without object dection module on COCO ‘Karpathy’ split test set.

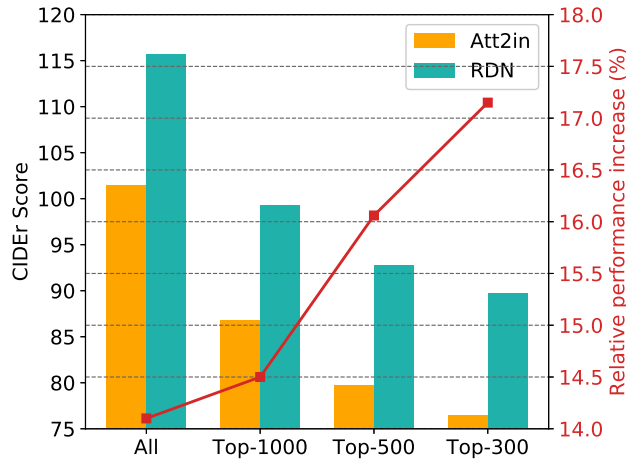


Figure 5. Performance comparison between our RDN and Att2in [3] on hard Image Captioning as a function of average length of annotations. We rank the ‘Karpathy’ test set according to their average length of annotations (ground truth captions) in descending order and extract four different size of subsets. Smaller subset corresponds to averagely longer annotations and harder captioning. It reveals that our model exhibits more superiority over Att2in [3] in harder cases.

Model	CIDEr	BLEU-4	Params(M)	Flops(G)
Transformer-based	111.2	33.8	54.91	3.45
Baseline	113.2	36.1	19.69	1.20
RDN	115.3	36.8	19.85	1.23

Table 2. Comparison between our RDN and the Transformer (adapted directly to image captioning) on COCO ‘Karpathy’ test split using the same object-level visual feature.

Model	CIDEr	BLEU-4
RDN + Greedy Decoding	111.7	35.5
RDN + Beam Search (Default)	115.3	36.8

Table 3. Comparison on using greedy decoding and beam search.

Model	CIDEr	BLEU-4
RDN + Teacher forcing (Default)	115.3	36.8
RDN + Schedule sampling	115.1	37.0

Table 4. Comparison between using teacher forcing and student forcing (with schedule sampling introduced in [2]).

Model	CIDEr	BLEU-4
Up-Down ^Ψ	120.1	36.3
RFNet ^Ψ	121.9	36.5
RDN ^Ψ	123.2	37.2

Table 5. Performance comparison on COCO ‘Karpathy’ test split on single model with CIDEr optimization (marked with Ψ).

References

- [1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and vqa. In *CVPR*, 2018. 1, 3
- [2] X. Chen, L. Ma, W. Jiang, J. Yao, and W. Liu. Regularizing rnns for caption generation by reconstructing the past with the present. In *CVPR*, 2018. 6
- [3] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 5