Australian National University [*1]

DATA 61 [*2]

SentiCap: Generating Image Descriptions with Sentiments

Alexander Mathews[*1], Lexing Xie[*1 *2], Xuming He[*2 *1]

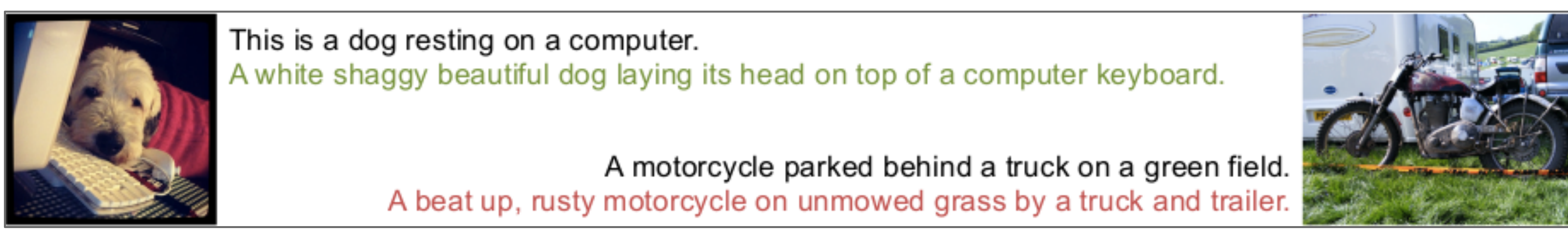


Image Captions and Sentiment

Sentiment is common in everyday language

Sentiment drives decision making

Where to eat for lunch

What to read

Who to vote for

I had a very tasty burger with some crunchy fries.

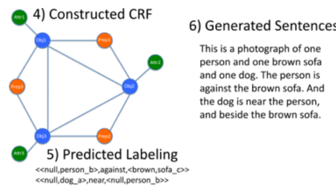
My overcooked burger and soggy fries.

With sentiment we can:

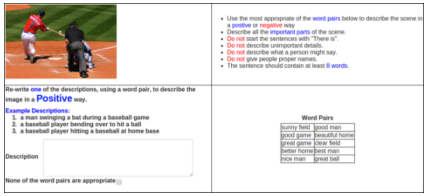
- Make more interesting and more human captions
- change the way people feel about an image

Contents

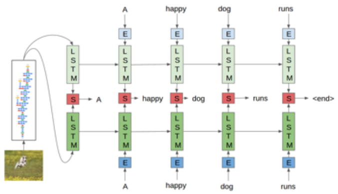
Related work



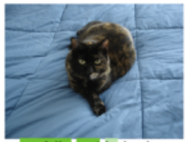
Dataset construction



Switching RNN Model



Evaluation + Results

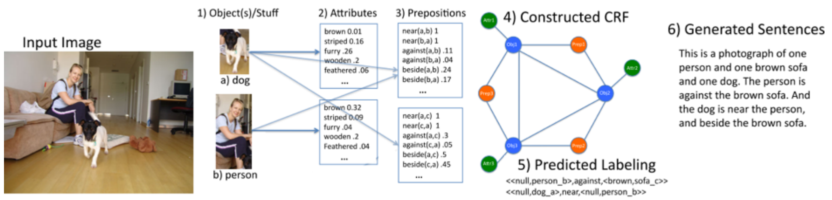


a cuddly cat is laying on a bed

Related work: Image to Sentence

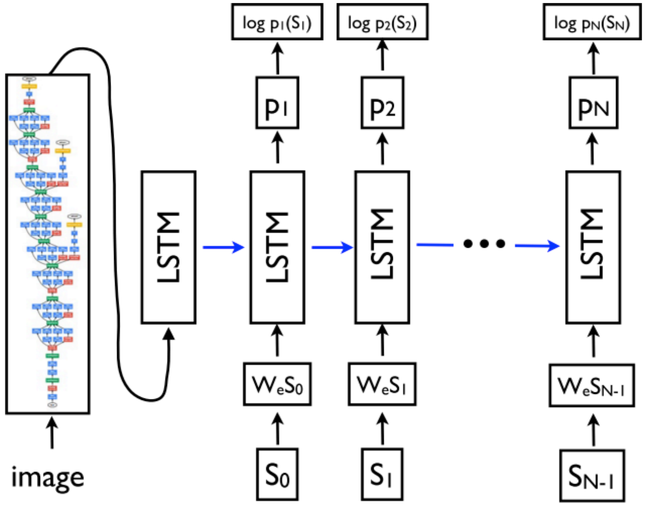
Nearest neighbour images + caption transfer (Farhadi, 2010)

Detectors for nouns, scenes, actions. With template filling and/or language model (Kulkarni, 2011)



Convolution Neural Network + Recurrent neural network

(Vinyals, 2014), (Donahue, 2015), (Karpathy, 2015), (Mao, 2014), (Kiros, 2014)

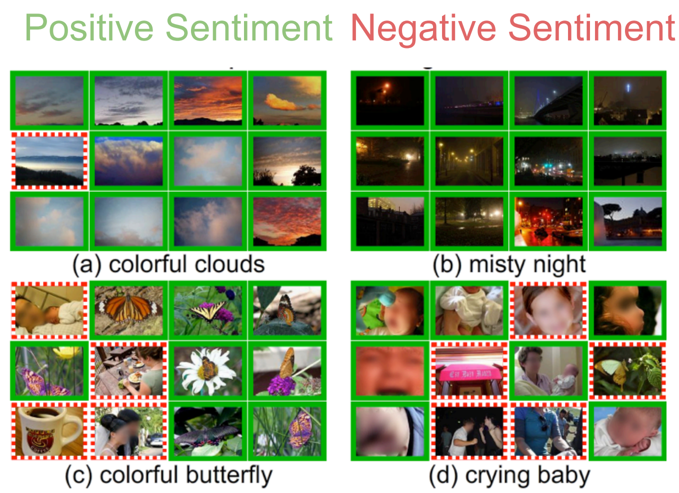


Related Work: Sentiment

Recognising sentiment has been studied extensively

Used in areas such as:

- Predicting movie reviews (Pang, 2005)
- Understanding public opinion (Tumasjan, 2010)
- Exploring large text collections (Mei, 2007)
- Predicting sentiment of images (Borth 2013)
    - Using Adjective Noun Pair (ANP) detectors



Generating image captions with sentiment is still an open problem.

Sentiment Dataset

Existing image-caption datasets focus on descriptiveness (eg MSCOCO)

Captions are short so we need a compact way of incorporating sentiment

Use **Adjective Noun Pairs (ANPs)**

Collect captions from Amazon Mechanical Turk

Task: Re-write a descriptive sentence using an ANP from a list

Word Pairs

sunny field	good man
good game	beautiful home
great game	clear field
better home	best man
nice man	great ball

1. **a man swinging a bat during a baseball game**
2. **a baseball player bending over to hit a ball**
3. **a baseball player hitting a baseball at home base**

Description

Dataset Validation

Validation: Another AMT task asking if the sentiment is appropriate

The painted train drives through a lovely city with country charm.

The abandoned train sits alone in the gloomy countryside.

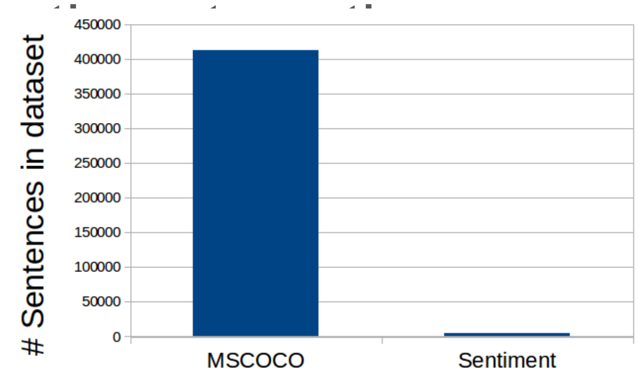
A train on the train tracks.

	#imgs	#sentence	descrip-tiveness	Correct sentiment: #votes			
				3	2	1	0
COCO	124	372	3.42±0.81	355	16	1	0
POS	124	335	3.34±0.79	315	20	0	0
NEG	123	305	2.69±1.11	250	49	6	0

Incorporating Sentiment: Approach

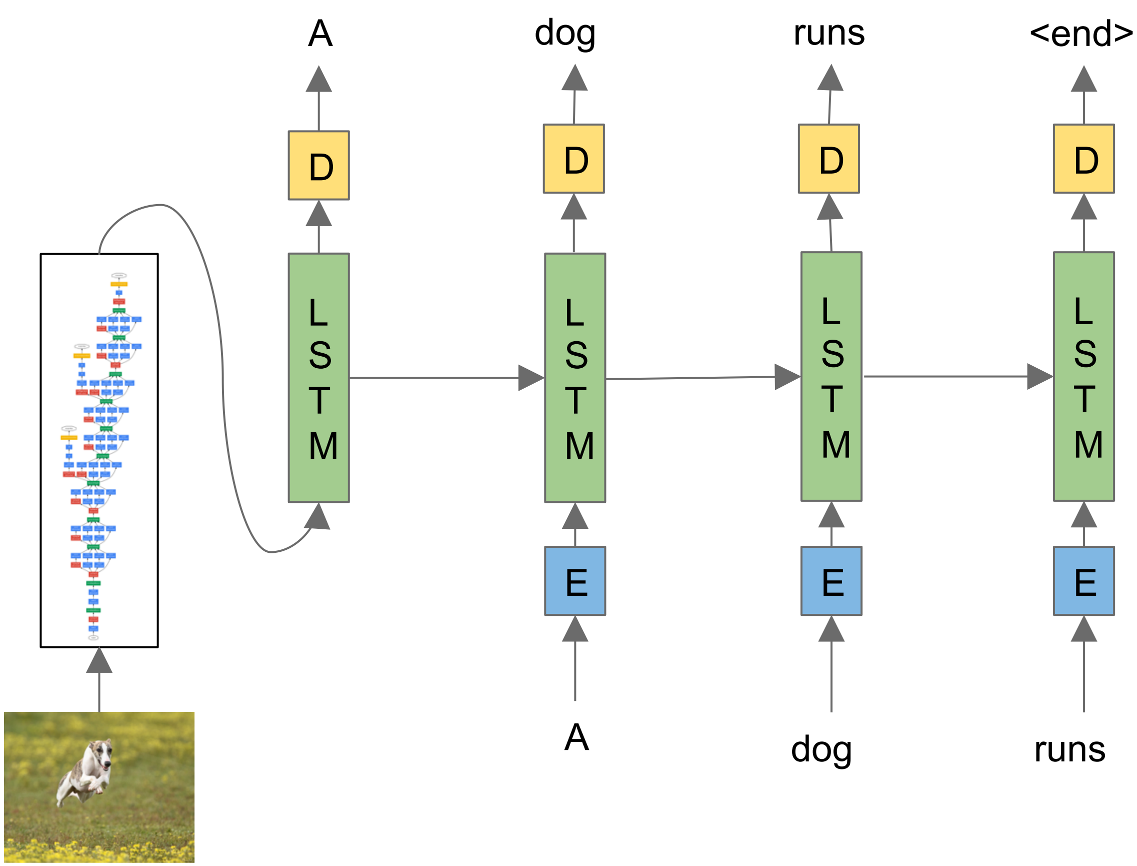
Challenges:

1. Big data + Small data: many descriptive captions, few sentiment captions
2. Generate descriptive captions that **also** have sentiment
3. Identify the important parts of the se

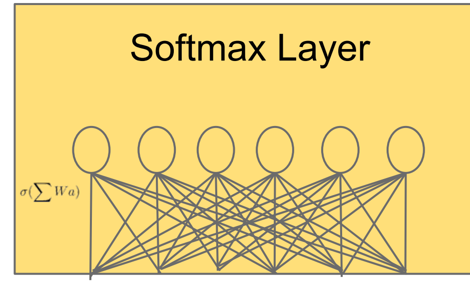


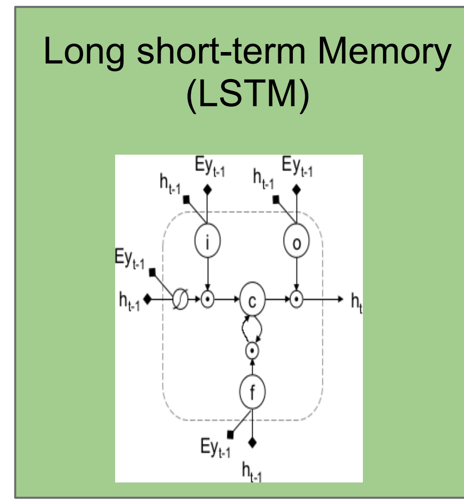
Design a switching RNN that addresses these challenges

Sentence Generation: Recurrent Neural Networks



Architecture of:
O. Vinyals and A. Toshev, 2014.

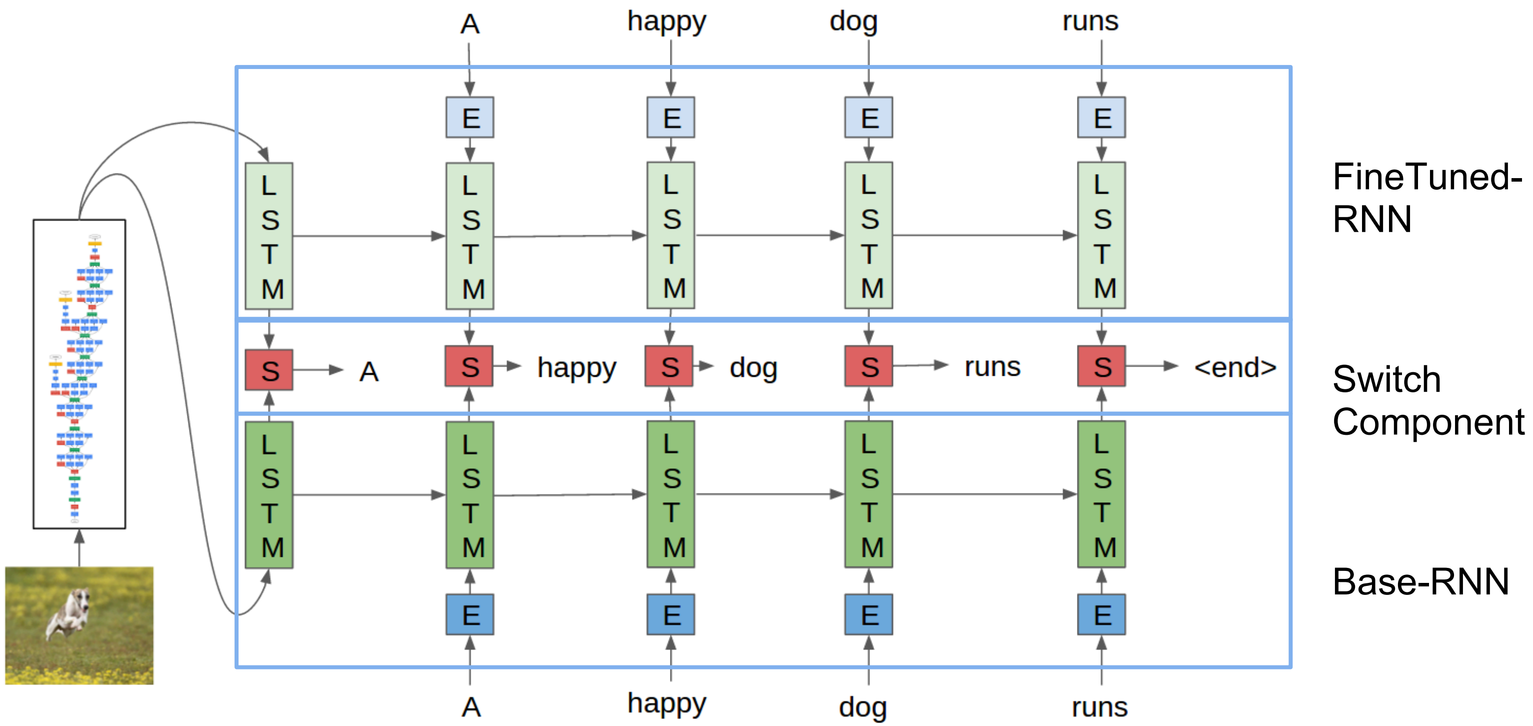




Embedding Layer

A	0.3, 0.1, 0.2, ...
dog	0.5, 0.7, 0.8, ...
runs	0.3, 0.2, 0.9, ...

SentiCap: Our Model



The Base-RNN produces descriptive sentences. (Trained on large data)

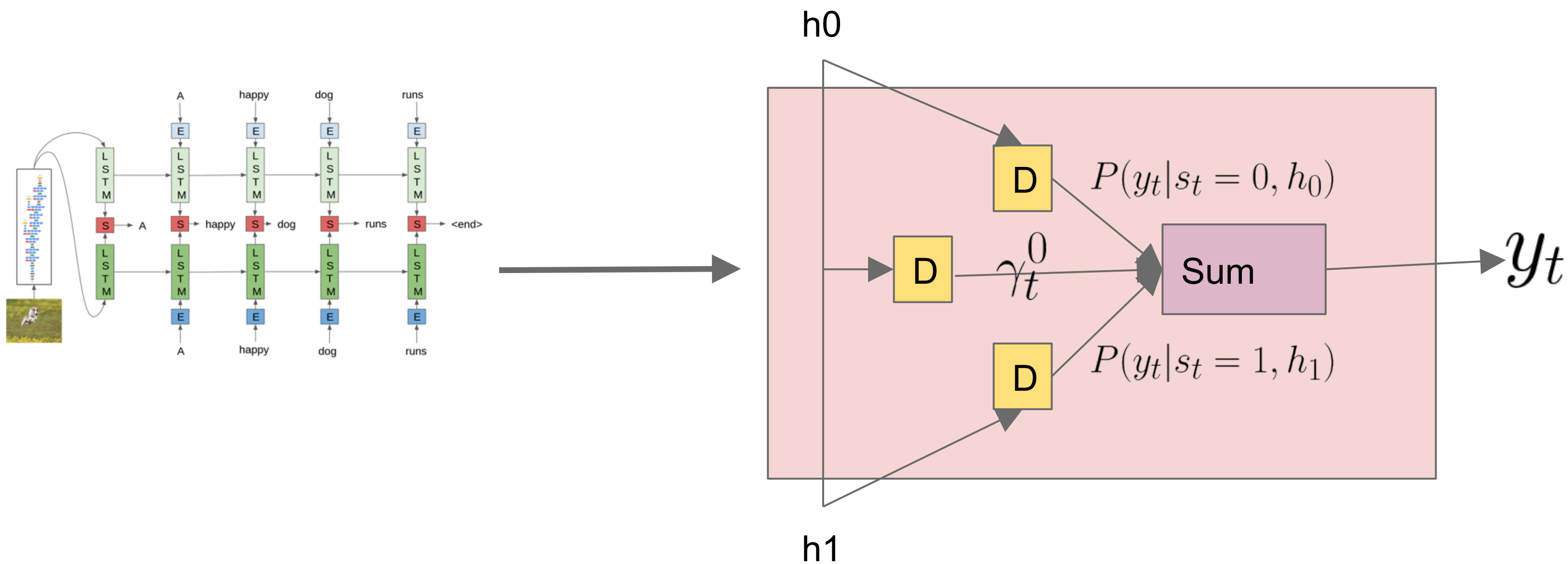
The FineTuned-RNN produces captions with sentiment. (Tuned on our dataset)

Switch component

$$\gamma_t^0 = \sigma(W_s[h_t^0; h_t^1])$$

$$\gamma_t^0 = P(s_t = 0 | x, y_{1:t-1})$$

Switch Component



γ_t^0 Indicates the presence or absence of a sentiment word.

$$P(y_t|h_1, h_0) = \gamma_t^0 P(y_t|s_t = 0, h_0) + \gamma_t^1 P(y_t|s_t = 1, h_1)$$

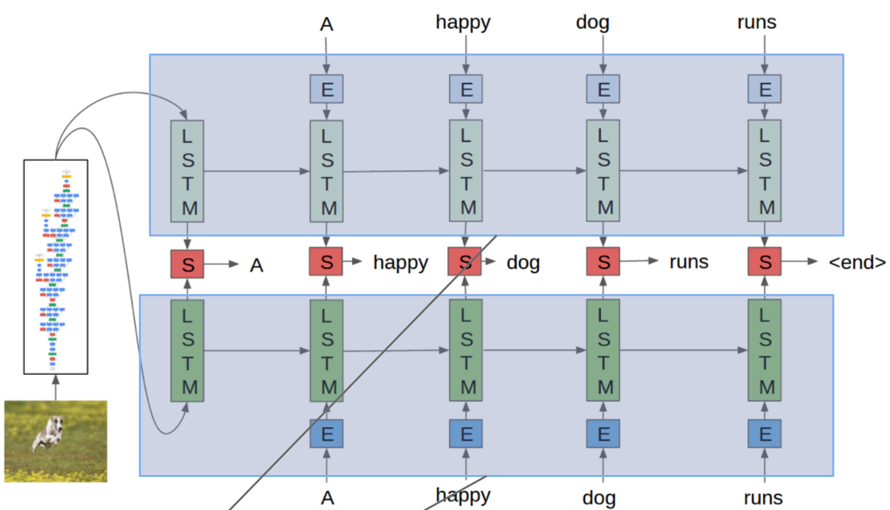
Training Objective

Train the joint model on the sentiment dataset.

- Keep the parameters in the Finetuned RNN “close” to the BaseRNN parameters

- Cross-entropy term ensures:
  - both RNNs are used
  - an increased weight for correctly generating sentiment words

Switching RNN



$$\mathcal{L}(\Theta,\mathcal{D}) = -\sum_i\sum_t(1+\lambda_\eta\eta_t^i)[L_t(\Theta,x^i,y^i)+ \boxed{\lambda_\gamma(\eta_t^i\log\gamma_t^{1,i}+(1-\eta_t^i)\log\gamma_t^{0,i})]} + \boxed{R(\Theta)}$$

$$R(\Theta) = \frac{\lambda_\theta}{2}\left\|\Theta^1 - \Theta^2\right\|^2$$

Results: Examples

a great variety of fresh fruits and vegetables

a cuddly cat is laying on a bed

an ugly car is parked in front of an abandoned building

a lonely train pulling into a train station

a delicious piece of cake sitting on top of a white plate

a clock on the side of a beautiful building

a man in a stupid hat is riding on the back of a crazy horse

a silly cat standing in front of a dirty wall

Evaluating the Result

Automatic:

N-gram based metrics: BLEU, ROUGE, METEOR, CIDEr

Human:

Used Amazon Mechanical Turk

- Most positive caption
- Most interesting caption
- How descriptive is the caption

Avoiding poor quality workers

- Reject using average accuracy on human written captions
- More restrictive worker qualifications

Caption	Most positive	More interesting	Describes the image			
			Correctly	Almost	Barely	Unrelated
a group of people on a boat in a body of water	○	○	○1	○2	○3	○4
a great group of people on a boat in the calm water	○	○	○1	○2	○3	○4

☐ Sentences are identical

Results

		SEN%	B-1	B-2	B-3	B-4	ROUGE$_L$	METEOR	CIDE$_r$	SENTI	DESC	DESCCMP
POS	CNN+RNN	1.0	48.7	28.1	17.0	10.7	36.6	15.3	55.6	–	2.90±0.90	–
	ANP-Replace	90.3	48.2	27.8	16.4	10.1	36.6	16.5	55.2	84.8%	2.89±0.92	95.0%
	ANP-Scoring	90.3	48.3	27.9	16.6	10.1	36.5	16.6	55.4	84.8%	2.86±0.96	95.3%
	RNN-Transfer	86.5	49.3	29.5	17.9	10.9	37.2	17.0	54.1	84.2%	2.73±0.96	76.2%
	SentiCap	93.2	49.1	29.1	17.5	10.8	36.5	16.8	54.4	88.4%	2.86±0.97	84.6%
NEG	CNN+RNN	0.8	47.6	27.5	16.3	9.8	36.1	15.0	54.6	–	2.81±0.94	–
	ANP-Replace	85.5	48.1	28.8	17.7	10.9	36.3	16.0	56.5	61.4%	2.51±0.93	73.7%
	ANP-Scoring	85.5	47.9	28.7	17.7	11.1	36.2	16.0	57.1	64.5%	2.52±0.94	76.0%
	RNN-Transfer	73.4	47.8	29.0	18.7	12.1	36.7	16.2	55.9	68.1%	2.52±0.96	70.3%
	SentiCap	97.4	50.0	31.2	20.3	13.1	37.9	16.8	61.8	72.5%	2.40±0.89	65.0%

Automatic Evaluation:

- sentences are similar to those in the sentiment dataset

Human Evaluation:

- sentences express stronger sentiment according to human evaluators

Summary

1. Introduced the task of generating image captions with sentiment

2. Constructed a dataset of image sentiment caption pairs

3. Designed a switching RNN model which:

   a. Generates image descriptions

   b. Uses a large descriptive dataset and a small sentiment dataset for training

A first step towards more natural and more interesting captions

Future: more fine-grained sentiments

Our dataset is available at:
http://users.cecs.anu.edu.au/~u4534172/senticap.html

Thank You.