Distinctive-attribute Extraction for Image Captioning

Boeun Kim, Young Han Lee, Hyedong Jung, and Choongsang Cho*

AI Research Center, Korea Electronics Technology Institute, Korea
{kbe36, yhlee, hudson, ideafisher}@keti.re.kr

Abstract. Image captioning has evolved with the progress of deep neural networks. However, generating qualitatively detailed and distinctive captions is still an open issue. In previous works, a caption involving semantic description can be generated by applying additional information into the RNNs. In this approach, we propose a distinctive-attribute extraction (DaE) method that extracts attributes which explicitly encourage RNNs to generate an accurate caption. We evaluate the proposed method with a challenge data and verify that this method improves the performance, describing images in more detail. The method can be plugged into various models to improve their performance.

Keywords: Image captioning, Semantic information, Distinctive-attribute, and Term frequency-inverse document frequency (TF-IDF)

1 Introduction

Image captioning is a potent and useful tool for automatically describing or explaining the overall situation of an image [22, 5, 24]. However, generate qualitatively detailed and distinctive captions is still an open issue. Although in most cases captions with unique expressions are more useful than those with only safe ones, the current evaluation metrics do not adequately reflect this aspect. After the numerical performance of the previous researches increased to some extent, some works are studying how to generate detailed and accurate captions [4].

In this paper, we propose a Distinctive-attribute Extraction (DaE) mehtod that extracts attributes which explicitly encourages RNNs to generate a caption that describes a significant meaning of an image. The main contributions of this paper are as follows: (i) We propose a semantics extraction method by using the TF-IDF caption analysis. (ii) We propose a scheme to infer distinctive-attribute by the model trained with semantic information. (iii) We perform quantitative and qualitative evaluations, demonstrating that the proposed method improves the performance of a base caption generation model by a substantial margin while describing images more distinctively.

2 Related Work

Combinations of CNNs and RNNs are widely used for the image captioning networks [22, 5, 24, 8, 6, 23]. The CNN was used as an image encoder, and an

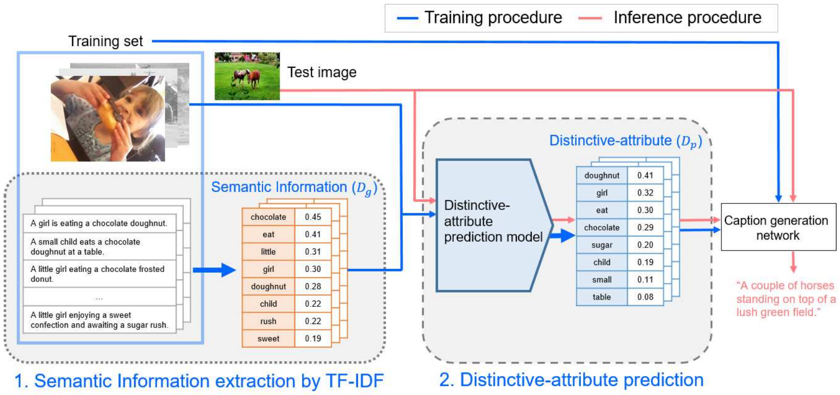

Training procedure
Inference procedure
Training set
Test image
A girl is eating a chocolate doughnut.
A small child eats a chocolate doughnut at a table.
A little girl eating a chocolate frosted donut.
...
A little girl enjoying a sweet confection and awaiting a sugar rush.
Semantic Information (D_g)
chocolate 0.45
eat 0.41
little 0.31
girl 0.30
doughnut 0.28
child 0.22
rush 0.22
sweet 0.19
Distinctive-attribute (D_p)
Distinctive-attribute prediction model
doughnut 0.41
girl 0.32
eat 0.30
chocolate 0.29
sugar 0.20
child 0.19
small 0.11
table 0.08
Caption generation network
"A couple of horses standing on top of a lush green field."
1. Semantic Information extraction by TF-IDF
2. Distinctive-attribute prediction

Fig. 1. An overview of the proposed framework including a semantic information extraction procedure and a Distinctive-attribute prediction model

output of its last hidden layer is fed into the RNN decoder that generates sentences. Recent approaches can be grouped into two paradigms. Top-down includes attention-based mechanisms, and many of the bottom-up methods used semantic concepts. For the latter, Fang *et al.* [6] used multiple instance learning (MIL) to train word detectors with words that commonly occur in captions. The word detector outputs guided a language model to generate descriptions to include the detected words. Wu *et al.* [23] predicted attributes by treating the problem as a multi-label classification. The CNN framework was used and outputs from different proposal sub-regions are aggregated. Gan *et al.* [8] proposed Semantic Concept Network (SCN) integrating semantic concepts to the LSTM network. SCN factorized each weight matrix of the attribute integrated the LSTM model to reduce the number of parameters.

More recently, Dai *et al.* [4] proposed Contrastive Learning(CL) method which encourages the distinctiveness of captions. In addition to true image-caption pairs, this method used mismatched pairs which include captions describing other images for learning.

3 Distinctive-attribute Extraction

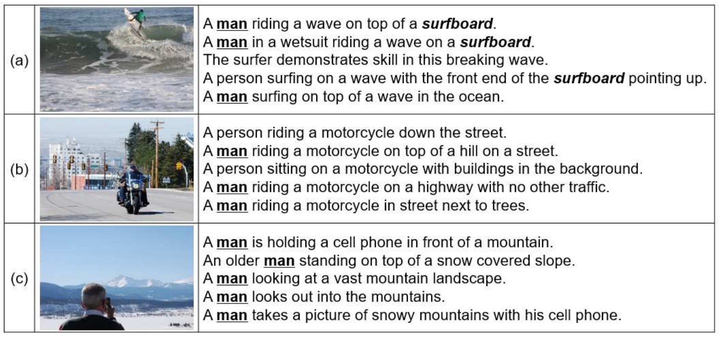
In this paper, we describe a semantic information processing and extraction method, which affects the quality of generated captions. We propose a method to generate captions that can represent the unique situation of the image. Different from CL [4] that improved target method by additional pairs on a training set, our method lies on the bottom-up approaches using semantic attributes. We assign more weights to the attributes that are more informative and distinctive to describe the image. As illustrated in Figure 1, there are two main steps, one is semantic information extraction, and the other is the distinctive-attribute prediction. First, we extract meaningful information from reference captions.

Next, we learn the distinctive-attribute prediction model with image-information (D_g) pairs. After getting distinctive-attribute (D_p) from images, we apply these attributes to a caption generation network to verify their effect. For the network, we used SCN-LSTM [8] which is a tag integrated network.

3.1 Semantic Information Extraction by TF-IDF

Most of the previous methods constituted semantic information that was a ground truth attribute, as a binary form [8, 6, 23, 25]. They first determined vocabulary using K most common words in the training captions. The vocabulary included nouns, verbs, and adjectives. If the word in the vocabulary existed in reference captions, the corresponding element of an attribute vector became 1. Different from previous methods, we weight semantic information according to their significance. Informative and distinctive words are weighted more, and the weight scores are estimated from reference captions by TF-IDF scheme which was widely used in text mining tasks.

Figure 2 represents samples of COCO datasets. In 2(a), there is a common word "surfboard" in 3 out of 5 captions, which is a key-word that characterizes the image. Intuitively, this kind of words should get high scores. To implement this concept, we apply average term frequency $TF_{av}(w, d)$, the number of times word w occurs in document d divided by the number of captions for an image. Another common word "man" appears a lot in other images. Therefore, that is a less meaningful word for distinguishing one image from another. To reflect this, we apply inverse document frequency term weighting $IDF(w) = \log\{(N_d + 1)/(DF(w) + 1)\} + 1$, where N_d is the total number of documents, and $DF(w)$ is the number of documents that contain the word w. "1" is added in denominator and numerator to prevent zero-divisions [19]. Then a semantic information vector is derived by multiplying two metrics as $TF - IDF(w, d) = TF_{av}(w, d) \times IDF(w)$. We apply L2 normalization to TF-IDF vectors of each image for training performance. The normalized value is the



(a)		A **man** riding a wave on top of a ***surfboard***. A **man** in a wetsuit riding a wave on a ***surfboard***. The surfer demonstrates skill in this breaking wave. A person surfing on a wave with the front end of the ***surfboard*** pointing up. A **man** surfing on top of a wave in the ocean.
(b)		A person riding a motorcycle down the street. A **man** riding a motorcycle on top of a hill on a street. A person sitting on a motorcycle with buildings in the background. A **man** riding a motorcycle on a highway with no other traffic. A **man** riding a motorcycle in street next to trees.
(c)		A **man** is holding a cell phone in front of a mountain. An older **man** standing on top of a snow covered slope. A **man** looking at a vast mountain landscape. A **man** looks out into the mountains. A **man** takes a picture of snowy mountains with his cell phone.

Fig. 2. Examples of images and their reference captions brought from MS COCO datasets [15, 2]

ground truth distinctive-attribute vector D_g. We apply stemming using Porter Stemmer [20] before extracting TF-IDF.

The next step is to construct vocabulary with the words in the reference captions. The vocabulary should contain enough characteristic words to represent each image. At the same time, the semantic information should be trained well for prediction accuracy. We determine the words to be included in the vocabulary based on the IDF scores which indicates the uniqueness of the word. The vocabulary contains the words whose IDF is higher than the IDF threshold (th_{IDF}) regardless of the part of speech. We observe the performance of the attribute prediction model and overall captioning model while changing the IDF value threshold in Section 4.3.

3.2 Distinctive-attribute Prediction Model

For the Distinctive-attribute prediction model, convolutional layers are followed by four fully-connected layers (FCs). We use ResNet-152 [10] architecture for CNN layers and the output of the 2048-way pool5 layer is fed into a stack of fully connected layers. Training data for each image consist of input image I and ground truth distinctive-attribute $\mathbf{D}_{g,i} = [D_{g,i1}, D_{g,i2}, \ldots, D_{g,iN_w}]$, where N_w is the number of the words in vocabulary and i is the index of the image. Our goal is to predict attribute scores as similar as possible to D_g. The cost function to be minimized is defined as mean squared error:

$$C = \frac{1}{M}\frac{1}{N_w}\sum_i\sum_w[D_{g,iw} - D_{p,iw}]^2 \tag{1}$$

where $\mathbf{D}_{p,i} = [D_{p,i1}, D_{p,i2}, \ldots, D_{p,iN_w}]$ is predictive attribute score vector for ith image and M denotes the number of training images. The first three FCs have 2048 channels each, the fourth contains N_w channels. We use ReLU [17] as nonlinear activation function for all FCs. We adopt batch normalization [11] right after each FC and before activation. The training is regularized by dropout with ratio 0.3 for the first three FCs. Each FC is initialized with a Xavier initialization [9]. We note that our network does not contain softmax as a final layer, different from other attribute predictors described in previous papers [8, 23]. Hence, we use the output of an activation function of the fourth FC layer as the final predictive score $\mathbf{D}_{p,i}$.

4 Results

4.1 Experiment settings

Our results are evaluated on the popular MS COCO dataset [15, 2]. The dataset contains 82,783 images for training, 40,504 and 40,775 images for validation and testing. The model described in Section 3.2 is implemented in Keras [3] and we used scikit-learn toolkit [19] to implement TF-IDF scheme. We set IDF threshold value to 7 in this experiment. The mini-batch size is fixed at 128 and Adam's optimization [13] with learning rate 3×10^{-3} is used and stopped after 100 epochs. For the prediction model, we train 5 identical models with different initializations, and then ensemble by averaging their outcomes. SCN-LSTM training procedure follows [8] and we use the public implementation [7] of this method opened by Gan who is the author of the published paper [8].

4.2 Evaluation

Table 1. COCO evaluation server results using 5 references and 40 references captions. DaE improves the performance by significant margins across all metrics

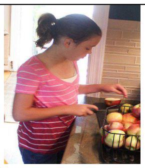
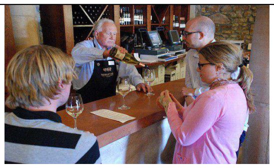
	B-1	B-2	B-3	B-4	M	R	CIDEr
5-refs							
SCN	0.729	0.563	0.426	**0.324**	0.253	0.537	0.967
DaE + SCN-LSTM	**0.734**	**0.568**	**0.429**	**0.324**	**0.255**	**0.538**	**0.981**
40-refs							
SCN	0.910	0.829	0.727	0.619	0.344	0.690	0.971
DaE + SCN-LSTM	**0.916**	**0.836**	**0.734**	**0.625**	**0.348**	**0.694**	**0.990**

Table 2. Results of published image captioning models tested on the COCO evaluation server

	B-1	B-2	B-3	B-4	M	R	CIDEr
5-refs							
Hard-Attention [24]	0.705	0.528	0.383	0.277	0.241	0.516	0.865
Google NIC [22]	0.713	0.542	0.407	0.309	0.254	0.530	0.943
ATT-FCN [25]	0.731	0.565	0.424	0.316	0.250	0.535	0.943
Adaptive Attention [16]	0.735	0.569	0.429	0.323	0.258	0.541	1.001
Adaptive Attention + CL [4]	0.742	0.577	0.436	0.326	0.260	0.544	1.010
DaE + SCN-LSTM	0.734	0.568	0.429	0.324	0.255	0.538	0.981
40-refs							
Hard-Attention [24]	0.881	0.779	0.658	0.537	0.322	0.654	0.893
Google NIC [22]	0.895	0.802	0.694	0.587	0.346	0.682	0.946
ATT-FCN [25]	0.900	0.815	0.709	0.599	0.335	0.682	0.958
Adaptive Attention [16]	0.906	0.823	0.717	0.607	0.347	0.689	1.004
Adaptive Attention + CL [4]	0.910	0.831	0.728	0.617	**0.350**	**0.695**	**1.029**
DaE + SCN-LSTM	**0.916**	**0.836**	**0.734**	**0.625**	0.348	0.694	0.990

Firstly, we compared our method with SCN [7]. We evaluate both results on the online COCO testing server [2] and list them in Table 1. For SCN, we use the pre-trained weights provided by the author. The vocabulary size of the proposed scheme is 938, which is smaller than that of SCN [7] with 999. Results of both methods are derived from ensembling 5 models, respectively. The widely used metrics, BLEU-1,2,3,4 [18], METEOR [1], ROUGL-L [14], CIDEr [21] are selected to evaluate overall captioning performance. DaE improves the performance of SCN-LSTM by significant margins across all metrics. Specifically, DaE improves CIDEr from 0.967 to 0.981 in 5-refs and from 0.971 to 0.990 in 40-refs. The increase is greater at 40-refs which have relatively various expressions. The results for other published models tested on the COCO evaluation server are summarized in Table 2. In 40-refs, our method surpasses the performance of *AdaptiveAttention* + *CL* [4] which is the state-of-the-art in terms of four BLEU scores. The qualitative evaluation is shown in Table 6. We listed the top eight attributes. For DaE, words after stemming with Porter Stemmer [20] are displayed as they are. Scores in the right parentheses of the tags and distinctive-attributes have different meanings, the former is probabilities, and the latter is distinctiveness values of the words. The attributes extracted using DaE include important words to represent the situation in an image; as a result, the caption generated by using them are represented more in detail compared with those of SCN. The result of the proposed method in (a), "A woman cutting a piece of fruit with a knife" explains what the main character does exactly. In the SCN, the general word "food" get a high probability, on the other hand, DaE extracts more distinctive words such as "fruit" and "apple." For verbs, "cut", which is the most specific action that viewers would be interested in, gets high distinctiveness score. In the case of (b), "wine" and "drink" are chosen as the words with the first and the third highest distinctiveness through DaE. Therefore, the characteristic phrase "drinking wine" is added. More examples are in Appendix A.

Table 3. This table illustrates several images with extracted attributes and captions. The captions generated by using DaE+SCN-LSTM are explained more in detail with more distinctive and accurate attributes

	(a)	(b)	(c)
			0 of 12 items found Hints
SCN	Generated captions: **A woman standing in a kitchen preparing food** Tags: person (0.99), food (0.91), indoor (0.85), table (0.58), woman (0.51), preparing (0.50), kitchen (0.42), small (0.35)	Generated captions: **A group of people sitting at a table** Tags: person (1.00), table (0.99), indoor (0.90), sitting (0.80), woman (0.76), man (0.57), front (0.46), group (0.36)	Generated captions: **A group of people standing in front of a table** Tags: indoor (0.83), table (0.63), standing (0.51), photo (0.49), computer (0.34), front (0.31), man (0.31), next (0.26)
DaE + SCN-LSTM	Generated captions: **A woman cutting a piece of fruit with a knife** Distinctive-attribute: **cut** (0.41), woman (0.28), **knife** (0.27), cake (0.18), **fruit** (0.14), food (0.13), kitchen (0.42), **appl** (0.11)	Generated captions: **A group of people sitting at a table drinking wine** Distinctive-attribute: **wine** (0.41), peopl (0.16), **drink** (0.13), tabl (0.12), woman (0.09), man (0.07), girl (0.07), group (0.06)	Generated captions: **A room filled with lots of colorful decorations** Distinctive-attribute: **color** (0.15), room (0.12), **decor** (0.11), hang (0.10), display (0.10), of (0.09), with (0.08), and (0.07)

4.3 Vocabulary construction

To analyze DaE in more detail, we conduct experiments with differently constructed vocabularies. We set seven different IDF threshold values, th_{IDF}, from 5 to 11.

$$Vocab_i = \{w \mid IDF(w) > i, i = th_{IDF}\}. \tag{2}$$

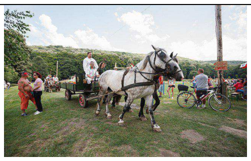
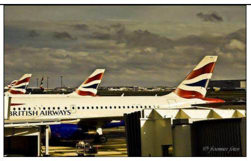
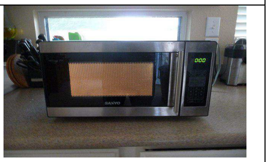
The vocabulary contains only the words whose IDF is bigger than th_{IDF}. The number of vocabulary words is shown in the second row of Table 4(a) and (b). Semantic information of the images are extracted corresponding to this vocabulary, and we use them to learn the proposed prediction model. Widely used splits [12] of COCO datasets are applied for the evaluation. We evaluate the prediction considering it as a multi-label and multi-class classification problem. The distinctiveness score between 0 and 1 are divided into four classes; $(0.0, 0.25]$, $(0.25, 0.5]$, $(0.5, 0.75]$, and $(0.75, 1.0]$ and the macro-averaged F1 score is computed globally. The performance, of the prediction model is shown in the third row. Each extracted distinctive-attribute is fed into SCN-LSTM to generate a caption, and the evaluation result, CIDEr, is shown in the fourth row. The CIDErs increase from $Vocab_5$ to $Vocab_7$, and then monotonically decrease in the rest. In other words, the maximum performance is derived from $Vocab_7$ to 0.996. The vocabulary size and the prediction performance are in a trade-off in this experiment. With the high th_{IDF} value, captions can be generated with various vocabularies, but the captioning performance is not maximized because the performance of distinctive-attribute prediction is relatively low. $Vocab_6$ and $Vocab_9$ have almost the same CIDEr. In this case, If the vocabulary contains more words, it is possible to represent the captions more diversely and accurately for some images. Table 5 shows examples corresponding to this case. For the case of (a), the $Vocab_6$ does not include the stemmed word "carriag", but the $Vocab_9$ contains the word and is extracted as the word having the seventh highest value through DaE. The word led the phrase "pulling a carriage" to be included the caption, well describing the situation. "Tamac" in (b), and "microwav" in (c) plays a similar role.

Table 4. Results of experiments with differently constructed vocabularies

	Vocab$_5$	Vocab$_6$	Vocab$_7$	Vocab$_8$	Vocab$_9$	Vocab$_{10}$	Vocab$_{11}$
(a) With stemming							
# of vocabulary	276	546	938	1660	2656	4009	**5530**
F1(DaE)	**0.432**	0.401	0.389	0.379	0.378	0.373	0.374
CIDEr(caption)	0.978	0.991	**0.996**	0.994	0.991	0.984	0.981
(b) Without stemming							
# of vocabulary	241	582	1121	2039	3572	5900	**8609**
F1(DaE)	**0.437**	0.399	0.383	0.374	0.366	0.362	0.358
CIDEr(caption)	0.955	0.989	**0.991**	0.986	0.990	0.988	0.979

Table 4 (b) presents experimental results without stemming. The maximum value was 0.911, which is lower than the maximum value of the experiments applying stemming. When stemming is applied, the distinctiveness and significance of a word can be better expressed because it is mapped to the same word even if the tense and form are different. In addition, the size of vocabulary required to achieve the same performance is less when stemming is applied.

Table 5. Several cases that more diverse and accurate captions are generated using $Vocab_9$ than using $Vocab_6$, although their CIDErs are similar

	(a)	(b)	(c)
			
Vocab$_6$	Generated captions: **A couple of people standing next to a horse** Distinctive-attribute: hors (0.58), pull (0.11), peopl (0.10), two (0.10), stand (0.07), field (0.07), of (0.07), in (0.06)	Generated captions: **A large air plane on a run way** Distinctive-attribute: airport (0.28), plane (0.26), airplan (0.25), jet (0.22), park (0.13), runway (0.12), an (0.12), on (0.09)	Generated captions: **A toaster oven sitting on top of a counter** Distinctive-attribute: oven (0.51), counter (0.18), kitchen (0.13), on (0.06), of (0.06), top (0.06), an (0.05), in (0.05)
Vocab$_9$	Generated captions: **A couple of horses pulling a carriage in a field** Distinctive-attribute: horse (0.58), pull (0.17), peopl (0.10), two (0.08), of (0.07), in (0.07), **carriag** (0.06), stand (0.06)	Generated captions: **A large jetliner sitting on top of an airport tarmac** Distinctive-attribute: airport (0.30), airplan (0.28), plane (0.25), jet (0.18), runway (0.16), an (0.12), **tarmac** (0.12), park (0.09)	Generated captions: **A microwave oven sitting on top of a counter** Distinctive-attribute: oven (0.46), **microwav** (0.40), counter (0.14), kitchen (0.07), on (0.06), of (0.06), top (0.06), an (0.05)

5 Conclusion

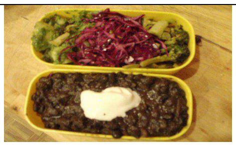
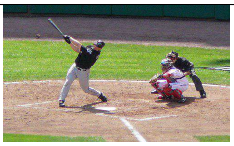
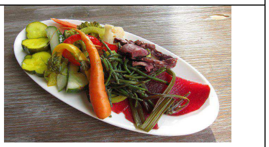
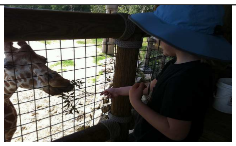
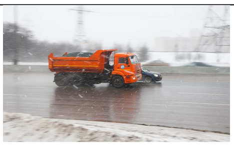
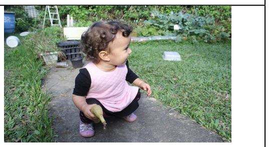
In this study, we propose a Distinctive-attribute Extraction (DaE) method for image captioning. In particular, the TF-IDF scheme is used to extract meaningful information from the reference captions. Then the attribute prediction model is trained by the extracted information and used to infer the semantic-attribute for generating a description. DaE improves the performance of SCN-LSTM scheme by significant margins across all metrics; moreover, detailed and unique captions are generated. The proposed method can be plugged into various models to improve their performance.

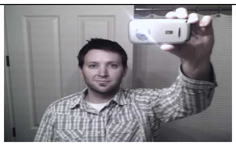
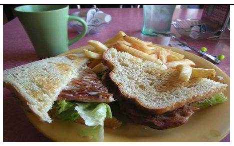
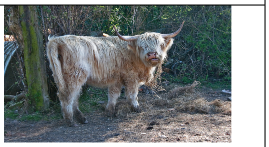
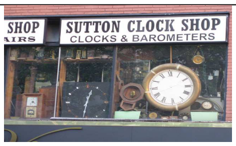
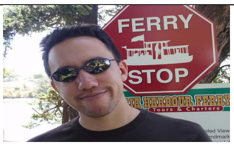
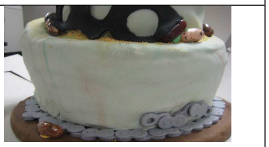
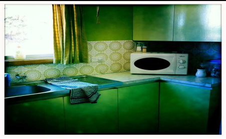
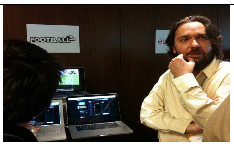
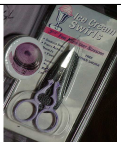
Acknowledgement

This work was supported by IITP/MSIT [2017-0-00255, Autonomous digital companion framework and application].

A Qualitative evaluation of DaE

Table 6. This figure is an expansion in Table 3 which is the qualitative evaluation of the proposed method

	(a)	(b)	(c)
			
SCN	Generated captions: **A close up of a bowl of food** Tags: food (1.00), table (0.97), indoor (0.92), container (0.71), sitting (0.67), wooden (0.61), sauce (0.53), plate (0.53)	Generated captions: **A baseball player swinging a bat at a ball** Tags: grass (1.00), baseball (1.00), player (0.99), bat (0.97), person (0.95), game (0.95), sport (0.95), swinging (0.93)	Generated captions: **A close up of a plate of food on a table** Tags: food (1.00), plate (0.99), table (0.98), hot (0.43), sitting (0.35), small (0.29), fruit (0.24), filled (0.23)
DaE + SCN-LSTM	Generated captions: **Two plastic containers filled with different types of food** Distinctive-attribute: contain (0.34), food (0.22), **veget** (0.16), **and** (0.12), **brocoli** (0.11), dish (0.09), **meat (0.08)**, of (0.09)	Generated captions: **A batter catcher and umpire during a baseball game** Distinctive-attribute: basebal (0.49), bat (0.32), player (0.18), swing (0.18), **catcher** (0.11), **umpir** (0.11), ball (0.10), **batter** (0.10)	Generated captions: **A white plate topped with a variety of vegetables** Distinctive-attribute: plate (0.48), **veget** (0.33), **carrot** (0.16), **salad** (0.16), **and** (0.13), food (0.10), on (0.09), with (0.09)
	(d)	(e)	(f)
			
SCN	Generated captions: **A dog is looking out of a fence** Tags: person (0.99), fence (0.87), building (0.65), window (0.61), looking (0.52), dog (0.47), standing (0.45), small (0.35)	Generated captions: **A fire hydrant spraying water from a fire hydrant** Tags: outdoor (0.99), orange (0.97), fire (0.83), water (0.76), hydrant (0.55), car (0.52), yellow (0.46), truck (0.44)	Generated captions: **A little boy is playing with a frisbee'** Tags: outdoor (1.00), grass (1.00), person (0.99), child (0.98), little (0.97), young (0.94), boy (0.93), small (0.85)
DaE + SCN-LSTM	Generated captions: **A person feeding a giraffe through a fence** Distinctive-attribute: **giraff** (0.40), fenc (0.25), **feed** (0.12), dog (0.10), out (0.07), look (0.06), in (0.05), is (0.05)	Generated captions: **A red truck driving down a snow covered road** Distinctive-attribute: **truck** (0.40), **snow** (0.19), orang (0.12), **drive** (0.11), car (0.09), the (0.09), toy (0.08), **red** (0.07)	Generated captions: **A small child sitting on the ground holding a banana** Distinctive-attribute: **banana** (0.35), boy (0.22), child (0.18), little (0.15), **hold** (0.12), young (0.11), skateboard (0.09), on (0.08)

	(g)	(h)	(i)
			
SCN	Generated captions: **A man holding a nintendo wii game controller** Tags: person (1.0), indoor (0.99), holding (0.99), man (0.96), controller (0.91), remote (0.89), video (0.87)	Generated captions: **A close up of a sandwich on a plate** Tags: food (1.00), sandwich (1.00), cup (0.98), plate (0.94), dish (0.90), indoor (0.87), sitting (0.84), coffee (0.80)	Generated captions: **A close up of a cow in a field** Tags: outdoor (1.00), grass (0.97), cow (0.97), animal (0.95), mammal (0.93), standing (0.87), hay (0.79), brown (0.64)
DaE + SCN-LSTM	Generated captions: **A man is taking a picture of himself** Distinctive-attribute: **take** (0.35), man (0.27), **phone** (0.24), hold (0.20), **hi** (0.19), pictur (0.17), **camera** (0.15), **cell** (0.14)	Generated captions: **A sandwich cut in half on a plate** Distinctive-attribute: sandwich (0.70), plate (0.28), **cut** (0.16), **half** (0.13), and (0.11), on (0.10), with (0.09), fri (0.09)	Generated captions: **A bull is standing next to a tree** Distinctive-attribute: cow (0.27), stand (0.19), tree (0.13), in (0.09), **bull** (0.08), brown (0.08), the (0.06), field (0.06)
	(j)	(k)	(l)
			
SCN	Generated captions: **A large clock on the side of a building** Tags: building (0.99), outdoor (0.93), clock (0.85), front (0.70), sign (0.49), large (0.44), sitting (0.27), next (0.24)	Generated captions: **A man in a blue shirt is holding a sign** Tags: person (1.00), outdoor (1.00), man (0.99), sign (0.65), front (0.61), eating (0.55), holding (0.55), food (0.45)	Generated captions: **A close up of a cake on a plate** Tags: cake (1.00), food (0.96), plate (0.92), table (0.91), chocolate (0.86), indoor (0.86), decorated (0.85), top (0.83)
DaE + SCN-LSTM	Generated captions: **A store window with a clock on display** Distinctive-attribute: **store** (0.33), **window** (0.32), clock (0.31), **display** (0.30), **shop** (0.16), sign (0.10), of (0.09), front (0.07)	Generated captions: **A man wearing sunglasses standing next to a stop sign** Distinctive-attribute: sign (0.39), **stop** (0.23), man (0.21), wear (0.13), **sunglass** (0.12), stand (0.09), smile (0.08), in (0.06)	Generated captions: **A chocolate cake with white frosting on top** Distinctive-attribute: cake (0.42), chocol (0.41), plate (0.12), decor (0.12), on (0.11), **frost** (0.10), with (0.08), top (0.08)
	(m)	(n)	(o)
			
SCN	Generated captions: **A kitchen with green walls and green walls** Tags: green (1.00), indoor (1.00), window (0.89), sitting (0.70), small (0.69), room (0.43), table (0.42), painted (0.41)	Generated captions: **A man holding a cell phone in his hand** Tags: person (1.00), man (0.99), indoor (0.85), front (0.66), looking (0.58), photo (0.38), standing (0.35), holding (0.31)	Generated captions: **A cell phone sitting on top of a book** Tags: indoor (0.94), sitting (0.53), small (0.39), book (0.27), case (0.27), next (0.25), table (0.20), top (0.20)
DaE + SCN-LSTM	Generated captions: **A kitchen with a sink and a microwave** Distinctive-attribute: **microwav** (0.44), **kitchen** (0.43), counter (0.23), and (0.11), green (0.09), with (0.09), **sink** (0.09), oven (0.09)	Generated captions: **A man sitting in front of a computer monitor** Distinctive-attribute: **comput** (0.36), man (0.24), phone (0.17), desk (0.13), hi (0.12), at (0.12), **laptop** (0.09), sit (0.08)	Generated captions: **A close up of a pair of scissors** Distinctive-attribute: **scissor** (0.32), **pair** (0.13), phone (0.10), of (0.10), cell (0.07), and (0.06), on (0.06), book (0.05)

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: proc. of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
3. Chollet, F., et al.: Keras. https://github.com/keras-team/keras (2015)
4. Dai, B., Lin, D.: Contrastive learning for image captioning. In: Advances in Neural Information Processing Systems. pp. 898–907 (2017)
5. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: proc. of IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 2625–2634 (2015)
6. Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., et al.: From captions to visual concepts and back (2015)
7. Gan, Z.: Semantic compositional nets. https://github.com/zhegan27/Semantic_Compositional_Nets (2017)
8. Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., Deng, L.: Semantic compositional networks for visual captioning. In: proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2 (2017)
9. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: proc. of 13t International Conference on Artificial Intelligence and Statistics. pp. 249–256 (2010)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456 (2015)
12. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: proc. of IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 3128–3137 (2015)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
16. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 6 (2017)
17. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: proc. of 27th international conference on machine learning (ICML). pp. 807–814 (2010)

18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: proc. of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
20. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
21. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: proc. of IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575 (2015)
22. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3156–3164. IEEE (2015)
23. Wu, Q., Shen, C., Liu, L., Dick, A., van den Hengel, A.: What value do explicit high level concepts have in vision to language problems? In: proc. of IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 203–212 (2016)
24. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057 (2015)
25. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4651–4659 (2016)