2019 IEEE Winter Conference on Applications of Computer Vision

Improving Image Captioning by Leveraging Knowledge Graphs

Yimin Zhou, Yiwei Sun, Vasant Honavar
Artificial Intelligent Research Laboratory
The Pennsylvania State University
zhoumingdetiger@gmail.com, yus162@psu.edu, vhonavar@ist.psu.edu

Abstract

We explore the use of a knowledge graphs, that capture general or commonsense knowledge, to augment the information extracted from images by the state-of-the-art methods for image captioning. We compare the performance of image captioning systems that as measured by CIDEr-D, a performance measure that is explicitly designed for evaluating image captioning systems, on several benchmark data sets such as MS COCO. The results of our experiments show that the variants of the state-of-the-art methods for image captioning that make use of the information extracted from knowledge graphs can substantially outperform those that rely solely on the information extracted from images.

1. Introduction

Advances in digital technologies have made it possible to acquire and share vast amounts of data of all kinds, including in particular, images. The availability of such data, together with recent advances in machine learning, has resulted in robust and practical machine learning based solutions to object recognition, e.g., Inception[1], vgg16[2], ResNet[3].

Recent years have witnessed a growing interest in describing visual scenes, a task that is remarkably easy for humans yet remains difficult for machines [4]. Of particular interest in this context is the image captioning problem, which requires analyzing the visual content of an image, and generating a caption, i.e., a textual description that summarizes the most salient aspects of the image. Just as question answering presents challenges beyond text processing, image captioning presents several challenges beyond image processing. Effective image captions need to provide information that is not explicit in the image, e.g., "People gathered to watch a volleyball match" when describing a crowd seated around a volleyball court, even if the image shows no players on the field (perhaps because the game is yet to begin), or "An impressionist painting of a garden by Claude Monet", even if the image makes no explicit mention of Monet or impressionism. Generating such captions calls for incorporating background knowledge with information that is available in the image. However, existing methods for image captioning (See [5] for a review) fail to take advantage of readily available general or commonsense knowledge about the world, e.g., in the form of *knowledge graphs*.

Inspired by the success of information retrieval and question answering systems that leverage background knowledge [6], we explore an approach to image captioning that uses information encoded in knowledge graphs. Specifically, we augment the neural image caption (NIC) method introduced in [7, 8] where a convolutional neural network (CNN) [9] trained to encodes an image into a fixed length vector space representation or embedding and uses the embedding to specify the initial state of a recurrent neural network (RNN) that is trained to produce sentences describing the image in two important aspects: In addition to a CNN trained to generate vector space embedding of image features, we use an object recognition module that given an image as input, produces as output, a collection of terms that correspond to objects in the scene We use an external knowledge graph, specifically, ConceptNet [10, 11], a labeled graph which connects words and phrases of natural language connected by edges that denote *commonsense* relationships between them, to infer a set of terms directly or indirectly related to the words that describe the objects found in the scene by the object recognition module. Vector space embeddings of the terms as well as the image features are then used to specify the initial state of an LSTM-based RNN that is trained to produce the caption for the input image. We call the resulting image captioning system ConceptNet enhanced neural image captioning system (CNet-NIC). The results of our experiments on the MS COCO captions benchmark dataset [12] show that CNet-NIC is competitive with or outperforms the state-of-the-art image captioning systems on several of the commonly used performance measures (BLEU [13] , METEOR[14], ROUGE-L[15], all of which are measures designed originally for evaluating machine translation systems as opposed to image

978-1-7281-1975-5/19/$31.00 ©2019 IEEE
DOI 10.1109/WACV.2019.00036

captioning systems). More importantly, CNet-NIC substantially outperforming the competing methods on CIDEr-D, a variant of the CIDEr [16], the only measure that is designed explicitly for evaluating image captioning systems. Because CIDEr-D measures the similarity of a candidate image caption to a collection of human generated reference captions, our results suggest that the incorporation of background knowledge from ConceptNet enables CNet-NIC to produce captions that are more similar to those generated by humans than those produced by methods that do not leverage such background knowledge.

The rest of the paper is organized as follows. Section 2 summarizes the related work on image captioning that sets the stage for our work on CNet-NIC. Section 3 the design and implementation of CNet-NIC. Section 4 describes our experimental setup and the results of our experiments assessing the performance of CNet-NIC on the MS COCO image captioning benchmark dataset along with comparisons with the competing state-of-the-art methods using the standard performance measures (BLEU@N ($N \in 1, 2, 3, 4$), METEOR, ROUGE-L, and CIDEr-D) as well as a qualitative analysis of a representative sample of the captions produced by CNet-NIC. Section 5 concludes with a summary and an outline of some directions for further research.

2. Related Work

Existing image captioning methods can be broadly grouped into the following (not necessarily disjoint) categories: (i) Template-based methods e.g., [17, 18, 19, 20] which rely on (often hand-coded) templates. Such methods typically detect the object types, their attributes, scene types (e.g., indoor versus outdoor), etc., based on a set of visual features, and generate image captions by populating a template with the information extracted from the image. (ii) Retrieval-based methods which can be further subdivided into two groups: (ii.a) Image similarity based methods e.g., [21, 18, 22, 23, 24] which retrieve captioned images that are visually most similar to the target image and transfer their captions to the target image; and (ii.b) Multimodal similarity based methods that use features of images as well as the associated captions to retrieve or synthesize the caption for the target image [25, 26, 27, 28, 29, 30, 7]; (iii) Embedding-based methods, including those that use recurrent, convolutional, or deep neural networks [7, 31, 32, 8, 33, 34, 35, 36] that make use of the learned low-dimensional embeddings of images to train caption generators.

However, none of the existing methods take advantage of the readily available background knowledge about the world (e.g., in the form of *knowledge graphs*. Such background knowledge has been shown to be useful in a broad range of applications ranging from information retrieval to question answering [6], including most recently, visual question answering (VQA) from images [37]. We hypothesize that such background knowledge can address an important drawback of existing image captioning methods, by enriching captions with information that is not explicit in the image.

Unlike the state-of-the-art image captioning systems, CNet-NIC is specifically designed to take advantage of background knowledge to augment the information extracted from the image (image features, objects) to improve machine-produced captions or image descriptions. Unlike VQA [37], which uses a knowledge graph to extract better image features and hence better answer questions about the image, CNet-NIC first detects objects (not just image features) in the image and uses the detected objects to identify related terms or concepts which are then used to produce better image captions.

3. CNet-NIC: ConceptNet-Enhanced Neural Image Captioning

We proceed to describe our design for an image captioning system that takes advantage of background knowledge in the form of a knowledge graph.

3.1. CNet-NIC Architecture

Fig. 1 shows a schematic of the CNet-NIC system. CNet-NIC uses YOLO9000[38], a state-of-the-art general-purpose real-time object recognition module that is trained to recognize 9000 object categories. YOLO9000 takes an image as input and produces as output, a collection of terms that refer to objects in the scene. CNet-NIC use an external knowledge graph, specifically, ConceptNet [10, 11], a labeled graph which connects words and phrases of natural language connected by edges that denote *commonsense* relationships between them, to infer two sets of terms related to the words that describe the objects found in the scene by the object recognition module. The first set of terms are retrieved based on the individual objects in the scene. The second set of terms are retrieved based on the entire collection of objects in the scene. The resulting terms are then provided to a pre-trained RNN to obtain the corresponding vector space embedding of the terms. A CNN is used to obtain vector space embedding of the image features. The two resulting vector space embeddings are used to specify the initial state of an LSTM-based RNN which is trained to produce the caption for the input image.

We proceed to describe each key element of the CNet-NIC system in detail.

3.2. Improving Image Captioning by Incorporating Background Knowledge

To test our hypothesis, we use the ConceptNet[10, 11], a kind of knowledge graph, specifically, one that connects words and phrases of natural language connected by edges

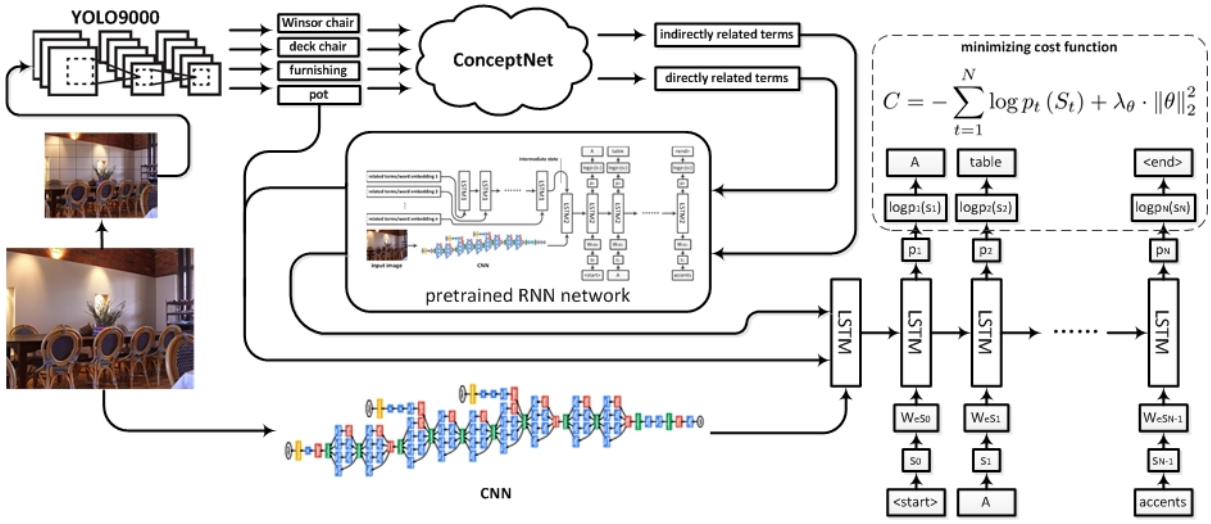

Figure 1. Our model architecture by inviting common sense from external resources

that denote *commonsense* relationships between them. ConceptNet integrates information from resources provided by experts as well as through crowd-sourcing. It encodes general knowledge that is of use in natural language understanding, and has been shown to enrich the semantic information associated with words, beyond that supplied by distributional semantics [11].

3.3. Generating Semantic Representations from ConceptNet

ConceptNet can be used to learn word embeddings using a variant of "retrofitting" [39]. Let $V = \{w_1, \ldots, w_n\}$ be a vocabulary, i.e., the set of word types, and Ω be an ontology encoding semantic relations between words in V. Ω is represented as an undirected graph (V, E) with one vertex for each word type and edges $(w_i,\ w_j) \in E \subseteq V \times V$ indicating a semantic relationship of interest.

Let $\hat{Q}$ be the collection of vectors $\hat{q}_i \in \mathbb{R}^d$ for each $w_i \in V$ that is learned using a standard data-driven method where d is the length of word vectors. The objective is to learn the matrix $Q = (q_1, \ldots, q_n)$ such that the columns are both close to their counterparts in $\hat{Q}$ and to adjacent vertices in Ω where closeness is measured using an appropriate distance measure, e.g., the Euclidean distance. This is achieved by minimizing the following objective function:

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j)\in E} \beta_{ij} \|q_i - q_j\|^2 \right] \quad (1)$$

where α and β are parameters that control the the relative strengths of associations. The procedure is called retrofitting because the word vectors are first trained independent of the information in the semantic lexicons and are then retro-fitted by optimizing the objective function specified above. Because Ψ is convex in Q, the solution of the resulting optimization problem is straightforward. Q can be initialized to $\hat{Q}$ and iteratively updated using the following update equation:

$$q_i = \frac{\sum_{j:(i,j)\in E} \beta_{ij} q_j + \alpha_i \hat{q}_i}{\sum_{j:(i,j)\in E} \beta_{ij} + \alpha_i} \quad (2)$$

3.4. Simple Recurrent Neural Network Image Caption Generator

We use a simple recurrent neural network image caption generator based on LSTM introduced in [7] where a CNN is used to extract image features; and vector space embedding of the extracted features is used by an LSTM-based RNN to generate the caption text. The architecture of this model is shown in Fig. 2.

Let X be an input image and $S = (S_0, \ldots, S_N)$ the corresponding caption sentence. Let

$$x_{-1} = CNN(X) \quad (3)$$

$$x_t = W_e S_t, \quad t \in \{0 \ldots N-1\} \quad (4)$$

where S_t is the one-hot vector representation of the word with a size of the dictionary, S_0 a special start word, and S_N a special end word.

$$p_{t+1} = LSTM(x_t), \quad t \in \{0 \ldots N-1\} \quad (5)$$

The loss function is given by:

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t) \quad (6)$$

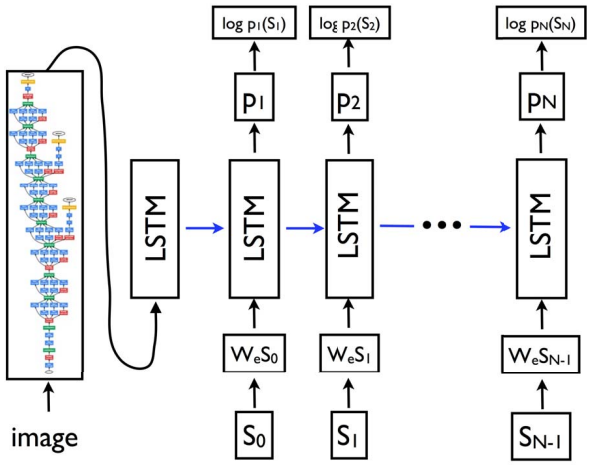

Figure 2. The architecture of the simple recurrent neural network image caption generator

The loss function is minimized with respect to the parameters of the LSTM, CNN and W_e.

3.5. Identifying Semantically Related Words

Given set of input words $W = \{w_1 \cdots w_n\}$, and their associated weights $u_1 \cdots u_n$ (e.g., based on their frequency distribution), a target word w can be scored based on its semantic relatedness to the input words, as measured by the weighted distance (e.g., cosine distance) between the semantic vector representation of the query word with each of the target words. Let s_w and s_{w_i} denote the semantic vector representations of words w and w_i ($i \in \{1, \cdots n\}$; and $d(a, b)$ denote the (cosine) distance between vectors a and b.

$$score(W, w) = \frac{\sum_{i=1}^{n} u_i d(s_w, s_{w_i})}{\sum_{i=1}^{n} u_i} \tag{7}$$

Fig. 3 Identifying Semantically Related Words. Blue rectangular nodes denote the input words (concepts). Red ovals denote the words that are most closely related to the input words, whereas the green ovals the next most closely related, and light blue ovals the next most closely related.

3.6. CNet-NIC

Let X be an input image, and O a set of terms corresponding to the objects detected in the image I by the YOLO9000 object recognition system. Thus, $O = YOLO(X)$. For each $o \in O$, let $r_o = ConceptNet(o)$ be the set of terms related to o in the ConceptNet knowledge graph; and $R_O = ConceptNet(O)$ the set of terms related to the entire set O of terms referring to all of the objects detected in the image X by the YOLO9000 object recognition system. Let $D = \bigcup_{o \in O} r_o \cup \{o\}$ denote the set of terms directly related to individual objects in X. Loosely speaking, R_O provides terms that are descriptive of the scene as a whole, whereas I provides terms that are descriptive of some or all of the objects depicted in the image. Thus, $I = R_O - D$ denote a set of terms that are indirectly related to objects in X. Let $d = RNN_D(D)$ and $i = RNN_I(I)$ denote the vector space embeddings of D and I produced by the pre-trained RNNs RNN_D and RNN_I respectively. Let $a = CNN(X)$ be an embedding of the image features of X produced by a pre-trained CNN. The image captions are produced by an LSTM-based RNN whose state is initialized as follows:

$$x_{-1} = a \parallel d \parallel i \tag{8}$$

where $\parallel$ denotes the concatenation operation.

$$x_t = W_e S_t, \quad t \in \{0 \ldots N-1\} \tag{9}$$

where S_t denotes the one-hot vector representation of the word with a size of the dictionary, and S_0 a special start word, and S_N a special end word.

$$p_{t+1} = LSTM(x_t), \quad t \in \{0 \ldots N-1\} \tag{10}$$

The cost function is given by:

$$C = -\sum_{t=1}^{N} \log p_t(S_t) + \lambda_\theta \cdot \|\theta\|_2^2 \tag{11}$$

where θ represents the model parameters and $\lambda_\theta \cdot \|\theta\|_2^2$ is a regularization term.

The only trainable parameters are within the LSTM and W_e. The pre-trained RNN network is shown in Fig. 4. Let r_i be the ith word embedding.

$$x_i = W_r r_i, \quad i \in \{i, \ldots L-1\} \tag{12}$$

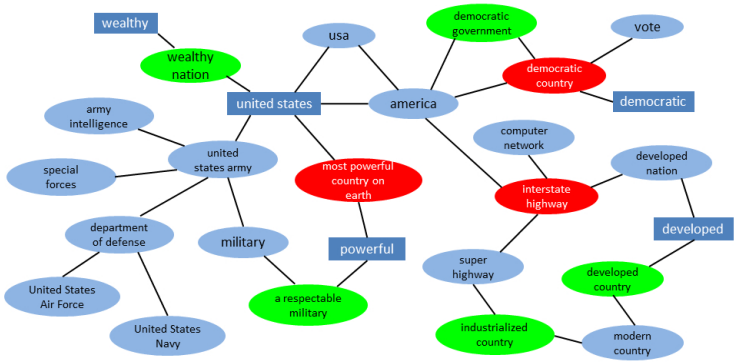

Figure 3. An example of relevant words generated by ConceptNet

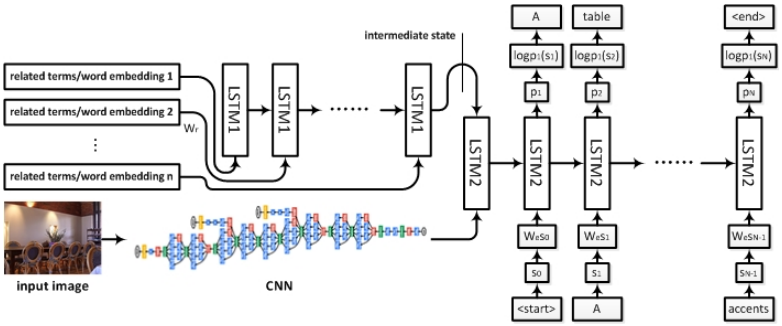

Figure 4. The pre-trained RNN network architecture

$$a = CNN\left(X\right) \tag{13}$$

a represents the image attribute embedding from CNN.

$$y_{-1} = concatenate\left(a,\ x_{L-1}\right) \tag{14}$$

$$a = CNN\left(X\right) \tag{15}$$

$$y_t = W_e S_t, \quad t \in \{0,\ \ldots N-1\} \tag{16}$$

$$p_{t+1} \ = \ LSTM\left(x_t\right), \quad t \in \{0\ \ldots\ N-1\} \tag{17}$$

The cost function to be minimized is given by:

$$C = -\sum_{t=1}^{N} \log p_t\left(S_t\right) \tag{18}$$

Note that the only parameters to be trained are those associated with LSTM1, LSTM2 and W_e. The trained LSTM1 is used in the model in Fig. 3 to extract intermediate state of the related terms.

4. Experiments

4.1. Data

We used the Microsoft COCO captioning data set(COCO)[12], the most widely used image captioning benchmark data set in our evaluations. The data set includes 82,783 images for training and 40,504 images for validation. For each image, the data set includes 5 or 6 descriptions or captions provided by human annotators. In our experiments, from the training and validation set provided, we used 117,211 images for training, 2,026 images for validation and 4050 images for testing.

4.2. Experimental Setup

Details of the experimental setup are summarized below:

- **Data Preprocessing**: Following [28], we convert all of the image captions in training set to lower case and discard rare words which occur less than 4 times, resulting in the final vocabulary with 11,519 unique words in COCO data set. Each word in the sentence is represented as "one-hot" vector.

- **Attribute Extraction**: To extract the image attributes, we use the output of inception v3[40] network image recognition model pre-trained on the ILSVRC-2012-CLS[41] image classification data set.

- **Object Detection**: We use the YOLO9000 object detection network (with 23 layers) and 544×544 resolution. YOLO9000 is able to detect 9419 object classes.

- **Leveraging Background Knowledge**: Because YOLO9000 object detection system is inherently imperfect, in identifying related terms using ConceptNet, we limit ourselves to only the objects detected with high confidence. Based on preliminary experiments, we set 30% as the detection threshold.

- **Training the model**: Our model is implemented on the TensorFlow platform in Python language. The size of LSTM for each embedding(attributes, related terms) is set to 512. Initial learning rate is set to 2.0 with an exponential decay schedule. Batch size is set to 32. Along the training, the learning rate is shrunk by 5 for three or four times. The number of iterations is set to 500,000.

- **Testing the model**: Two approaches can be utilized for sentence generation during the testing stage. One approach is to select the word with maximum probability at each time step and set it as LSTM input for next time step until the end sign word is emitted or the maximum length of sentence is reached. Another approach is to conduct a beam search that selects the top-k best sentences at each time step and use them as the candidates to generate the top-k best sentences at the next time step. We adopt the second approach and set the beam size k empirically to 3.

- **Evaluation Metrics**: To evaluate CNet-NIC, we use 4 metrics: BLEU@N[13], METEOR[14], ROUGE-E[15], and CIDEr-D[16]. All the metrics are computed by using the codes released by [42].

4.3. Performance Comparison

We compare the performance of CNet-NIC with that of several state-of-the-art image captioning methods (as reported in the respective papers):

- **Neural Image Caption (NIC)**[7], which uses a vector space representation of image features produced by a CNN to initialize an LSTM-based RNN trained to generate image captions from vector space representation of image features.

- **Hard and Soft Attention**[31], which combines two attention-based image captioning mechanisms under an encoder-decoder framework: a soft deterministic attention mechanism trainable by standard back-propagation methods and 2) a hard stochastic attention mechanism that is trained using reinforcement learning.

- **LRCN**[36] which combines CNN with LSTMs to perform visual recognition and image captioning.

- **ATT**[8] which combines top-down and bottom-up attention models to extract image features that are used to train an RNN to produce image captions.

- **Sentence-Condition**[43] which uses a *text-conditional attention* mechanism for focusing the caption generator on specific image features that should inform the caption given the already generated caption text.

- **LSTM-A**[44] which extends the basic LSTM model with image attributes model by rearranging image and attributes input in different positions and time to boost the accuracy of image captioning.

Table 1 shows the performance of each method on MS COCO image captioning data set. Bold represents the best in that metric and italic represents the second best. Overall the performance of CNet-NIC is comparable to or better than all other models on all measures, especially with respect to CIDEr-D, the only measure that is explicitly designed for the purpose of evaluating image captions.

4.4. CNet-NIC Ablation Study Results

We report results of an *ablation study* of CNet-NIC, where we examine the relative contributions of the different components of the CNet-NIC architecture.

From the results summarized in Table 2, we see that detected objects and directly related terms contribute to greater improvements in performance as compared to indirectly related terms. We conjecture that the detected objects and directly related terms provide more information about the individual objects in an image whereas the indirectly related terms provide information about the scene as a whole. This perhaps explains why only adding indirectly related terms to image embedding improves performance as measured by METEOR, ROUGE-L and CIDEr-D, albeit at the cost of a slight decrease in BLEU. We further note that the indirectly related terms contribute to increases in CIDEr-D, even when no image features are available. Overall, we find that CNet-NIC which combines the background knowledge (ConceptNet derived terms) related to the detected objects and the scene in generating image captions outperforms all other methods that do not make use of such background knowledge.

4.5. Qualitative Analysis of Captions

Table 3 presents several representative examples of captions produced by CNet-NIC. Here we take a qualitative look at the captions to explore the role played by the commonsense or background knowledge provided by the Con-

Table 1. Performance of our proposed models and other state-of-the-art methods on MS COCO dataset, where B@N, M, R, and C are short for BLEU@N, METEOR, ROUGE-L, and CIDEr-D scores. Except CIDEr-D, all values are reported as percentage(%).

Model	B@1	B@2	B@3	B@4	M	R	C
NIC[7]	66.6	45.1	30.4	20.3	-	-	-
LRCN[36]	62.8	44.2	30.4	21	-	-	-
Soft Attention[31]	70.7	49.2	34.4	24.3	23.9	-	-
Hard Attention[31]	71.8	50.4	35.7	25	23	-	-
ATT[8]	70.9	53.7	40.2	*30.4*	24.3	-	-
Sentence Condition[43]	72	54.6	40.4	29.8	24.5	-	95.9
LSTM-A[44]	*73*	**56.5**	**42.9**	**32.5**	*25.1*	*53.8*	*98.6*
CNet-NIC	**73.1**	*54.9*	*40.5*	29.9	**25.6**	**53.9**	**107.2**

Table 2. Performance of variants of CNet-NIC on MS COCO dataset, where B@N, M, R, and C are short for BLEU@N, METEOR, ROUGE-L, and CIDEr-D scores. Except CIDEr-D, all values are reported as percentage(%).

Input of Model	B@1	B@2	B@3	B@4	M	R	C
none(only seqs input)	48.4	24.7	10.2	3.9	11	34.2	8.6
image embedding	70.3	52.9	38.3	27.5	24.3	51.8	99.5
detected objects and directly related terms	63.3	43.4	29.1	20	19.8	46.1	74.3
indirectly related terms	47.6	27	15.7	10.2	13.7	36.6	31.8
detected objects and directly related terms + image embedding	70.9	53.3	38.7	28	24.8	52.4	103.2
indirectly related terms + image embedding	70.1	52.8	38.2	27.7	24.5	52	100.5
detected objects and directly related terms + indirectly related terms + image embedding	72.1	54.2	38.9	28.5	24.8	52.9	103.6
detected objects and directly related terms + indirectly related terms + image embedding + fine tune CNN	73.1	54.7	40.5	29.9	25.6	53.9	107.2

ceptNet knowledge graph. In the first example, the ConceptNet derived terms such as "upholstered", "found in house", etc. appear to yield more accurate captions. In the third example, the standard model and the model without indirectly related terms completely ignore the large furniture such as tables and chairs while the model that incorporates indirectly related terms such as "item of furniture", "reupholstery", "end table", etc. leads to what appear to be better captions. For the fourth example, the indirectly related terms appear to yield a more accurate caption model, e.g., one that mentions the book rack. For the sixth image listed, only the model with indirectly related terms as "dairy farm", "feed lot", etc. from the knowledge graph correctly recognizes that the scene is occurring in a "barn". In the seventh example, the model with indirectly related terms correctly deduces that most people in the image are travelers and conclude that they are in a baggage claim area. These examples offer further qualitative evidence that shows the utility and effectiveness of background knowledge supplied by knowledge graphs to improve the quality of image captions.

5. Summary and Discussion

The focus of this paper is on the image captioning problem, which requires analyzing the visual content of an image, and generating a caption, i.e., a textual description that summarizes the most salient aspects of the image. Image captioning presents several challenges beyond those addressed by object recognition, e.g., inferring information that is not explicitly depicted in the image. However, existing methods for image captioning (See [5] for a review) fail to take advantage of readily available general or commonsense knowledge about the world.

In this paper, we have presented CNet-NIC, an approach to image captioning that incorporates background knowledge available in the form of knowledge graphs to augment the information extracted from images. We have compared the performance of image captioning systems that as measured by CIDEr-D, a performance measure that is explicitly designed for evaluating image captioning systems, on several benchmark data sets such as MS COCO. The results of our experiments show that the variants of the state-of-the-art methods for image captioning that make use of the information extracted from knowledge graphs can substantially outperform those that rely solely on the information extracted from images.

Some promising directions for future work include: variants and extensions of CNet-NIC, including those that substantially improve the quality of captions, provide justifications for the captions that they produce, tailor captions for visual question answering, tailoring captions to different audiences and contexts, etc. by bringing to bear on such tasks, all available background knowledge.

Table 3. Image example of model showing "Common Sense" from external resource.

Image	Detected	Indirectly Related	Sentences Generated by Model with Indirectly Related	Sentences Generated by Model without Indirectly Related	Standard Model
	Winsor chair, deck chair, furnishing, pot	item of furniture, **upholstered**, **found in house**, **chairs**	0) a dining room with a **table** and **chairs** 1) a dining room with a table, chairs and a table 2) a dining room with a table and chairs and a **fireplace**	0) a table with a vase of flowers on it 1) a dining room with a table and chairs 2) a table with a vase of flowers on it	0) a table with a vase of flowers on it 1) a table with a vase of flowers on it 2) a table with a vase of flowers on it
	fishmonger, cereal bowl, phial, banana, waiter	**food storage jar**, canaree, **storing food**, fruit bowl, food can	0) a **chef** preparing food in a kitchen on a counter 1) a **chef** preparing food in a kitchen on a table 2) a man in a kitchen preparing food for a customer	0) a man and a woman preparing food in a kitchen 1) a man and a woman preparing food in a kitchen 2) a chef preparing food in a kitchen next to a woman	0) a group of people in a kitchen preparing food 1) a group of people standing around a kitchen preparing food 2) a group of people in a kitchen preparing food
	straight chair, furnishing	**item of furniture**, **reupholstery**, end table	0) a kitchen **filled with appliances and lots of clutter** 1) a kitchen **filled with appliances and lots of counter space** 2) a kitchen with a table and chairs	0) a kitchen with a stove a sink and a counter 1) a kitchen with a stove a sink and a window 2) a kitchen with a stove top oven next to a sink	0) a kitchen with a stove a sink and a stove 1) a kitchen with a stove a sink and a refrigerator 2) a kitchen with a stove a sink and a counter
	book(s), toilet seat	**bookrack**, **bookshelving**, **bookrest**	0) a bathroom with a toilet and a **book shelf** 1) a bathroom with a toilet and a book shelf 2) a bathroom with a toilet and a sink	0) a white toilet sitting in a bathroom next to a wall 1) a white toilet sitting next to a book shelf 2) a white toilet sitting in a bathroom next to a shelf	0) a kitchen with a stove a sink and a stove 1) a kitchen with a stove a sink and a refrigerator 2) a kitchen with a stove a sink and a counter
	trolleybus(es), park bench, commuter	**tram stop**, bus rapid transit	0) a couple of buses that are **sitting** in the street 1) a couple of buses that are **parked next to each other** 2) a couple of buses driving down a street next to a tall building	0) a double decker bus driving down a street 1) a double decker bus driving down the street 2) a double decker bus is driving down the street	0) a double decker bus driving down a street 1) a double decker bus driving down a city street 2) a double decker bus driving down the street
	Friesian(s), Brown Swiss, private, settler	**dairy farm**, **cows**, **feed lot**	0) a group of cows standing next to each other 1) a group of cows that are standing in the dirt 2) a group of cows standing in a **barn**	0) a group of cows that are standing in the dirt 1) a group of cows that are standing in the grass 2) a group of cows that are standing in a pen	0) a group of cows are standing in a pen 1) a group of cows standing in a pen 2) a group of cows are standing in a field
	overnighter(s), pilgrim(s), square dancer, general, peddler, hand luggage, backpack	**wayfaring**, **day tripper**, **journeyer**, **excursionist**, **traveller**	0) a group of people standing around with luggage. 1) a group of people standing around a luggage carousel. 2) a group of standing around a **baggage claim area**.	0) a group of people standing around a luggage carousel. 1) a group of people standing around a luggage cart. 2) a group of people standing next to a luggage cart.	0) a group of people standing around a luggage carousel. 1) a group of people standing around a luggage cart. 2) a group of people standing next to each other.
	horse wrangler(s), Tennessee walker, wild horse	**found on ranch**	0) a man standing next to a brown horse. 1) a man is standing next to a horse 2) a man standing next a brown horse in a **stable**.	0) a couple of people standing next to a horse. 1) a woman standing next to a brown horse. 2) a couple of people standing next to a brown horse.	0) a group of people standing next to a horse. 1) a group of people standing next to a brown horse. 2) a group of men standing next to a brown horse.

Acknowledgements

This project was supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health through the grant UL1 TR000127 and TR002014, by the National Science Foundation, through the grants 1518732, 1640834, and 1636795, the Pennsylvania State Universitys Institute for Cyberscience and the Center for Big Data Analytics and Discovery Informatics, the Edward Frymoyer Endowed Professorship in Information Sciences and Technology at Pennsylvania State University and the Sudha Murty Distinguished Visiting Chair in Neurocomputing and Data Science funded by the Pratiksha Trust at the Indian Institute of Science [both held by Vasant Honavar]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

References

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich *et al.*, "Going deeper with convolutions." Cvpr, 2015.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?" *Journal of vision*, vol. 7, no. 1, pp. 10–10, 2007.

[5] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.

[6] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 3156–3164.

[8] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4651–4659.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] H. Liu and P. Singh, "Conceptneta practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.

[11] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge." in *AAAI*, 2017, pp. 4444–4451.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[14] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[15] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[16] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[17] K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1107–1135, 2003.

[18] A. Gupta and P. Mannem, "From image annotation to image description," in *International Conference on Neural Information Processing*. Springer, 2012, pp. 196–204.

[19] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *European conference on computer vision*. Springer, 2010, pp. 15–29.

[20] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[21] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Advances in neural information processing systems*, 2011, pp. 1143–1151.

[22] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 359–368.

[23] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, "Exploring nearest neighbor approaches for image captioning," *arXiv preprint arXiv:1505.04467*, 2015.

[24] P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions," *Transactions of the Association of Computational Linguistics*, vol. 2, no. 1, pp. 351–362, 2014.

[25] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[26] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, 2013, pp. 2121–2129.

[27] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association of Computational Linguistics*, vol. 2, no. 1, pp. 207–218, 2014.

[28] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in neural information processing systems*, 2014, pp. 1889–1897.

[29] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2407–2415.

[30] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2623–2631.

[31] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[32] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 203–212.

[33] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," *CoRR*, vol. abs/1410.1090, 2014. [Online]. Available: http://arxiv.org/abs/1410.1090

[34] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," *OpenReview*, vol. 2, no. 5, p. 8, 2016.

[35] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 652–663, 2017.

[36] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[37] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[38] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 6517–6525.

[39] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, "Retrofitting word vectors to semantic lexicons," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 1606–1615.

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[41] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[42] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *CoRR*, vol. abs/1504.00325, 2015. [Online]. Available: http://arxiv.org/abs/1504.00325

[43] L. Zhou, C. Xu, P. A. Koch, and J. J. Corso, "Image caption generation with text-conditional semantic attention," *CoRR*, vol. abs/1606.04621, 2016. [Online]. Available: http://arxiv.org/abs/1606.04621

[44] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *IEEE International Conference on Computer Vision, ICCV*, 2017, pp. 22–29.