Knowledge Guided Attention and Inference for Describing Images Containing Unseen Objects

Aditya Mogadala[1(✉)], Umanga Bista[2], Lexing Xie[2], and Achim Rettinger[1]

[1] Institute of Applied Informatics and Formal Description Methods (AIFB), Karlsruhe Institute for Technology (KIT), Karlsruhe, Germany
{aditya.mogadala,rettinger}@kit.edu
[2] Computational Media Lab, Australian National University (ANU), Canberra, Australia
{umanga.bista,lexing.xie}@anu.edu.au

Abstract. Images on the Web encapsulate diverse knowledge about varied abstract concepts. They cannot be sufficiently described with models learned from image-caption pairs that mention only a small number of visual object categories. In contrast, large-scale knowledge graphs contain many more concepts that can be detected by image recognition models. Hence, to assist description generation for those images which contain visual objects unseen in image-caption pairs, we propose a two-step process by leveraging large-scale knowledge graphs. In the first step, a multi-entity recognition model is built to annotate images with concepts not mentioned in any caption. In the second step, those annotations are leveraged as external semantic attention and constrained inference in the image description generation model. Evaluations show that our models outperform most of the prior work on out-of-domain MSCOCO image description generation and also scales better to broad domains with more unseen objects.

1 Introduction

Content on the Web is highly heterogeneous and consists mostly of visual and textual information. In most cases, these different modalities complement each other, which complicates the capturing of the full meaning by automated knowledge extraction techniques. An approach for making information in all modalities accessible to automated processing is linking the information represented in the different modalities (e.g., images and text) into a shared conceptualization, like entities in a Knowledge Graph (KG). However, obtaining an expressive formal representation of textual and visual content has remained a research challenge for many years.

Recently, a different approach has shown impressive results, namely the transformation of one unstructured representation into another. Specifically, the task of generating natural language descriptions of images or videos [16] has gained

A. Gangemi et al. (Eds.): ESWC 2018, LNCS 10843, pp. 415–429, 2018.
https://doi.org/10.1007/978-3-319-93417-4_27

much attention. While such approaches are not relying on formal conceptualizations of the domain to cover, the systems that have been proposed so far are limited by a very small number of objects that they can describe (less than 100). Obviously, such methods – as they need to be trained on manually crafted image-caption parallel data – do not scale to real-world applications, and can't be applied to cross-domain web-scale content.

In contrast, visual object classification techniques have improved considerably and they are now scaling to thousands of objects more than the ones covered by caption training data [3]. Also, KGs have grown to cover all of those objects plus millions more accompanied by billions of facts describing relations between those objects. Thus, it appears that those information sources are the missing link to make existing image captioning models scale to a larger number of objects without having to create additional image-caption training pairs with those missing objects. In this paper, we investigate the hypothesis, that conceptual relations of entities – as represented in KGs – can provide information to enable caption generation models to generalize to objects that they haven't seen during training in the image-caption parallel data. While there are existing methods that are tackling this task, none of them has exploited any form of conceptual knowledge so far. In our model, we use KG entity embeddings to guide the attention of the caption generator to the correct (unseen) object that is depicted in the image. Our main contributions presented in this paper are summarized as follows:

- We designed a novel approach, called Knowledge Guided Attention (KGA), to improve the task of generating captions for images which contain objects that are not in the training data.
- To achieve it, we created a multi-entity-label image classifier for linking the depicted visual objects to KG entities. Based on that, we introduce the first mechanism that exploits the relational structure of entities in KGs for guiding the attention of a caption generator towards picking the correct KG entity to mention in its descriptions.
- We conducted an extensive experimental evaluation showing the effectiveness of our KGA method. Both, in terms of generating effectual captions and also scaling it to more than 600 visual objects.

The contribution of this work on a broader scope is its progress towards the integration of the visual and textual information available on the Web with KGs.

2 Previous Work on Describing Images with Unseen Objects

Existing methods such as Deep Compositional Captioning (DCC) [4], Novel object Captioner (NOC) [15], Constrained Beam Search (CBS) [2] and LSTM-C [17] address the challenge by transferring information between seen and unseen objects either before inference (i.e. before testing) or by keeping constraints on

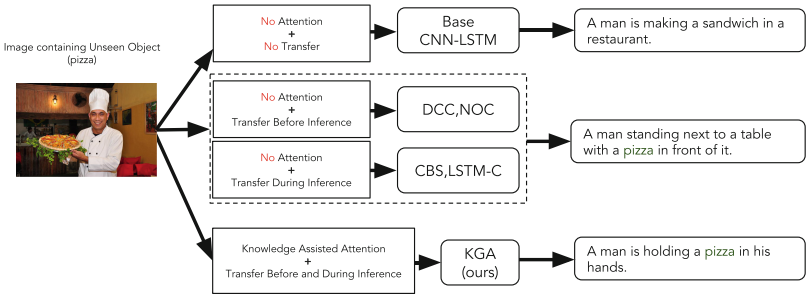

Fig. 1. KGA goal is to describe images containing unseen objects by building on the existing methods i.e. DCC [4], NOC [15], CBS [2] and LSTM-C [17] and going beyond them by adding semantic knowledge assistance. Base refers to our base description generation model built with CNN [13] - LSTM [5].

the generation of caption words during inference (i.e. during testing). Figure 1 provides a broad overview of those approaches.

In DCC, an approach which performs information transfer only before inference, the training of the caption generation model is solely dependent on the corpus constituting words which may appear in the similar context as of unseen objects. Hence, explicit transfer of learned parameters is required between seen and unseen object categories before inference which limits DCC from scaling to a wide variety of unseen objects. NOC tries to overcame such issues by adopting a end-to-end trainable framework which incorporates auxiliary training objectives during training and detaching the need for explicit transfer of parameters between seen and unseen objects before inference. However, NOC training can result in sub-optimal solutions as the additional training attempts to optimize three different loss functions simultaneously. CBS, leverages an approximate search algorithm to guarantee the inclusion of selected words during inference of a caption generation model. These words are however only constrained on the image tags produced by a image classifier. And the vocabulary used to find similar words as candidates for replacement during inference is usually kept very large, hence adding extra computational complexity. LSTM-C avoids the limitation of finding similar words during inference by adding a copying mechanism into caption training. This assists the model during inference to decide whether a word is to be generated or copied from a dictionary. However, LSTM-C suffers from confusion problems since probabilities during word generation tend to get very low.

In general, aforementioned approaches also have the following limitations: (1) The image classifiers used cannot predict abstract meaning, like “hope”, as observed in many web images. (2) Visual features extracted from images are confined to the probability of occurrence of a fixed set of labels (i.e. nouns, verbs and adjectives) observed in a restricted dataset and cannot be easily extended to varied categories for large-scale experiments. (3) Since an attention mechanism is missing, important regions in an image are never attended. While, the attention

mechanism in our model helps to scale down all possible identified concepts to the relevant concepts during caption generation. For large-scale applications, this plays a crucial role.

We introduce a new model called Knowledge Guided Assistance (KGA) that exploits conceptual knowledge provided by a knowledge graph (KG) [6] as external semantic attention throughout training and also to aid as a dynamic constraint before and during inference. Hence, it augments an auxiliary view as done in multi-view learning scenarios. Usage of KGs has already shown improvements in other tasks, such as in question answering over structured data, language modeling [1], and generation of factoid questions [12].

3 Describing Images with Unseen Objects Using Knowledge Guided Assistance (KGA)

In this section, we present our caption generation model to generate captions for unseen visual object categories with knowledge assistance. KGAs core goal is to introduce external semantic attention (ESA) into the learning and also work as a constraint before and during inference for transferring information between seen words and unseen visual object categories.

3.1 Caption Generation Model

Our image caption generation model (henceforth, KGA-CGM) combines three important components: a language model pre-trained on unpaired textual corpora, external semantic attention (ESA) and image features with a textual (T), semantic (S) and visual (V) layer (i.e. TSV layer) for predicting the next word in the sequence when learned using image-caption pairs. In the following, we present each of these components separately while Fig. 2 presents the overall architecture of KGA-CGM.

Language Model. This component is crucial to transfer the sentence structure for unseen visual object categories. Language model is implemented with two long short-term memory (LSTM) [5] layers to predict the next word given previous words in a sentence. If $\overrightarrow{\boldsymbol{w}_{1:L}}$ represent the input to the forward LSTM of layer-1 for capturing forward input sequences into hidden sequence vectors ($\overrightarrow{\boldsymbol{h}^1_{1:L}} \in \mathcal{R}^H$), where L is the final time step. Then encoding of input word sequences into hidden layer-1 and then into layer-2 at each time step t is achieved as follows:

$$\overrightarrow{\boldsymbol{h}^1_t} = \text{L1-F}(\overrightarrow{\boldsymbol{w}_t}; \Theta) \tag{1}$$

$$\overrightarrow{\boldsymbol{h}^2_t} = \text{L2-F}(\overrightarrow{\boldsymbol{h}^1_t}; \Theta) \tag{2}$$

where Θ represent hidden layer parameters. The encoded final hidden sequence ($\overrightarrow{\boldsymbol{h}^2_t} \in \mathcal{R}^H$) at time step t is then used for predicting the probability distribution

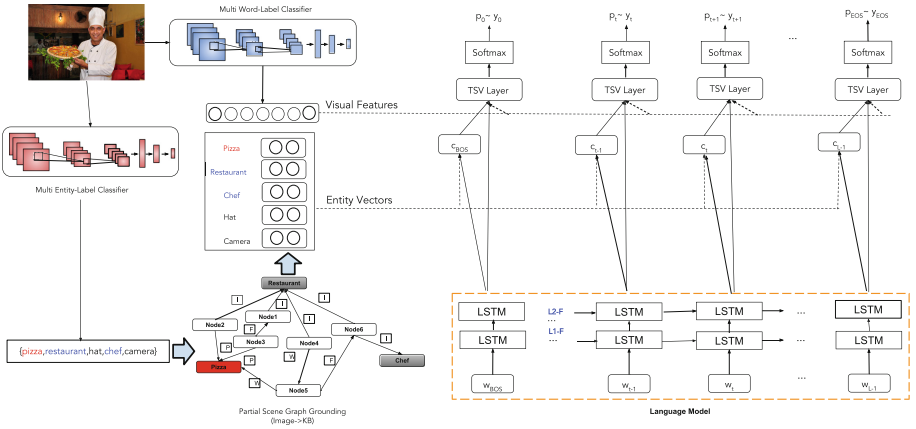

Fig. 2. KGA-CGM is built with three components. A language model built with two-layer forward LSTM (L1-F and L2-F), a multi-word-label classifier to generate image visual features and a multi-entity-label classifier that generates entity-labels linked to a KG serving as a partial image specific scene graph. This information is further leveraged to acquire entity vectors for supporting ESA. w_t represents the input caption word, c_t the semantic attention, p_t the output of probability distribution over all words and y_t the predicted word at each time step t. BOS and EOS represent the special tokens.

of the next word given by $\boldsymbol{p}_{t+1} = softmax(\boldsymbol{h}_t^2)$. The softmax layer is only used while training with unpaired textual corpora and not used when learned with image captions.

External Semantic Attention (ESA). Our objective in ESA is to extract semantic attention from an image by leveraging semantic knowledge in KG as entity-labels obtained using a multi-entity-label image classifier (discussed in the Sect. 4.2). Here, entity-labels are analogous to patches or attributes of an image. In formal terms, if $\boldsymbol{ea}_i$ is an entity-label and $\boldsymbol{e}_i \in \mathcal{R}^E$ the entity-label vector among set of entity-label vectors $(i = 1, .., L)$ and $\boldsymbol{\beta}_i$ the attention weight of $\boldsymbol{e}_i$ then β_i is calculated at each time step t using Eq. 3.

$$\boldsymbol{\beta}_{ti} = \frac{exp(\boldsymbol{O}_{ti})}{\sum_{j=1}^{L} exp(\boldsymbol{O}_{tj})} \tag{3}$$

where $\boldsymbol{O}_{ti} = f(\boldsymbol{e}_i, \boldsymbol{h}_t^2)$ represent scoring function which conditions on the layer-2 hidden state $(\boldsymbol{h}_t^2)$ of a caption language model. It can be observed that the scoring function $f(\boldsymbol{e}_i, \boldsymbol{h}_t^2)$ is crucial for deciding attention weights. Also, relevance of the hidden state with each entity-label is calculated using Eq. 4.

$$f(\boldsymbol{e}_i, \boldsymbol{h}_t^2) = tanh((\boldsymbol{h}_t^2)^T W_{he} \boldsymbol{e}_i) \tag{4}$$

where $W_{he} \in \mathcal{R}^{H \times E}$ is a bilinear parameter matrix. Once the attention weights are calculated, the soft attention weighted vector of the context $\boldsymbol{c}$, which is a dynamic representation of the caption at time step t is given by Eq. 5.

$$\boldsymbol{c}_t = \sum_{i=1}^{L} \beta_{ti} \boldsymbol{e}_i \tag{5}$$

Here, $\boldsymbol{c}_t \in \mathcal{R}^E$ and L represent the cardinality of entity-labels per image-caption pair instance.

Image Features and TSV Layer and Next Word Prediction. Visual features for an image are extracted using multi-word-label image classifier (discussed in the Sect. 4.2). To be consistent with other approaches [4,15] and for a fair comparison, our visual features (I) also have objects that we aim to describe outside of the caption datasets besides having word-labels observed in paired image-caption data.

Once the output from all components is acquired, the TSV layer is employed to integrate their features i.e. textual (T), semantic (S) and visual (V) yielded by language model, ESA and images respectively. Thus, TSV acts as a transformation layer for molding three different feature spaces into a single common space for prediction of next word in the sequence.

If $\boldsymbol{h}_t^2 \in \mathcal{R}^H$, $\boldsymbol{c}_t \in \mathcal{R}^E$ and $\boldsymbol{I}_t \in \mathcal{R}^I$ represent vectors acquired at each time step t from language model, ESA and images respectively. Then the integration at TSV layer of KGA-CGM is provided by Eq. 6.

$$\boldsymbol{TSV}_t = W_{\boldsymbol{h}_t^2} \boldsymbol{h}_t^2 + W_{\boldsymbol{c}_t} \boldsymbol{c}_t + W_{\boldsymbol{I}_t} \boldsymbol{I}_t \tag{6}$$

where $W_{\boldsymbol{h}_t^2} \in \mathcal{R}^{vs \times H}, W_{\boldsymbol{c}_t} \in \mathcal{R}^{vs \times E}$ and $W_{\boldsymbol{I}_t} \in \mathcal{R}^{vs \times I}$ are linear conversion matrices and $\boldsymbol{vs}$ is the image-caption pair training dataset vocabulary size.

The output from the TSV layer at each time step t is further used for predicting the next word in the sequence using a softmax layer given by $\boldsymbol{p}_{t+1} = softmax(\boldsymbol{TSV}_t)$.

3.2 KGA-CGM Training

To learn parameters of KGA-CGM, first we freeze the parameters of the language model trained using unpaired textual corpora. Thus, enabling only those parameters to be learned with image-caption pairs emerging from ESA and TSV layer such as $W_{he}, W_{\boldsymbol{h}_t^2}, W_{\boldsymbol{c}_t}$ and $W_{\boldsymbol{I}_t}$. KGA-CGM is now trained to optimize the cost function that minimizes the sum of the negative log likelihood of the appropriate word at each time step given by Eq. 7.

$$\min_{\theta} -\frac{1}{N} \sum_{n=1}^{N} \sum_{t=0}^{L^{(n)}} log(\boldsymbol{p}(y_t^{(n)})) \tag{7}$$

where $L^{(n)}$ represent the length of sentence (i.e. caption) with beginning of sentence (BOS), end of sentence (EOS) tokens at n-th training sample and N as a number of samples used for training.

3.3 KGA-CGM Constrained Inference

Inference in KGA-CGM refer to the generation of descriptions for test images. Here, inference is not straightforward as in the standard image caption generation approaches [16] because unseen visual object categories have no parallel captions throughout training. Hence they will never be generated in a caption. Thus, unseen visual object categories require guidance either before or during inference from similar seen words that appear in the paired image-caption dataset and likely also from image labels. In our case, we achieve the guidance both before and during inference with varied techniques.

Guidance Before Inference. We first identify the seen words in the paired image-caption dataset similar to the visual object categories unseen in image-caption dataset by estimating the semantic similarity using their Glove embeddings [9] learned using unpaired textual corpora (more details in Sect. 4.1). Furthermore, we utilize this information to perform dynamic transfer between seen words visual features ($W_{\boldsymbol{I}}$), language model ($W_{\boldsymbol{h}_t^2}$) and external semantic attention ($W_{\boldsymbol{c}_t}$) weights and unseen visual object categories. To illustrate, if (v_{unseen}, i_{unseen}) and ($v_{closest}, i_{closest}$) denote the indexes of unseen visual object category "zebra" and its semantically similar known word "giraffe" in a vocabulary (v_s) and visual features (i_s) respectively. Then to describe images with "zebra" in the similar manner as of "giraffe", the transfer of weights is performed between them by assigning $W_{\boldsymbol{c}_t}[v_{unseen},:]$, $W_{\boldsymbol{h}_t^2}[v_{unseen},:]$ and $W_{\boldsymbol{I}_t}[v_{unseen},:]$ to $W_{\boldsymbol{c}_t}[v_{closest},:]$, $W_{\boldsymbol{h}_t^2}[v_{closest},:]$ and $W_{\boldsymbol{I}_t}[v_{closest},:]$ respectively.

Input: M=$\{W_{he}, W_{\boldsymbol{h}_t^2}, W_{\boldsymbol{c}_t}, W_{\boldsymbol{I}_t}\}$
Output: M_{new}
Initialize List(closest) = cosine_distance(List(unseen),vocabulary) ;
Initialize $W_{\boldsymbol{c}_t}[v_{unseen},:]$, $W_{\boldsymbol{h}_t^2}[v_{unseen},:]$, $W_{\boldsymbol{I}_t}[v_{unseen},:] = 0$;
Function ***Before Inference***
  forall *items T in closest and Z in unseen* **do**
    if *T and Z is vocabulary* **then**
      $W_{\boldsymbol{c}_t}[v_Z,:] = W_{\boldsymbol{c}_t}[v_T,:]$;
      $W_{\boldsymbol{h}_t^2}[v_Z,:] = W_{\boldsymbol{h}_t^2}[v_T,:]$;
      $W_{\boldsymbol{I}_t}[v_Z,:] = W_{\boldsymbol{I}_t}[v_T,:]$;
    end
    if i_T *and* i_Z *in visual features* **then**
      $W_{\boldsymbol{I}_t}[i_Z,i_T]=0$;
      $W_{\boldsymbol{I}_t}[i_T,i_Z]=0$;
    end
  end
  M_{new} = M ;
  return M_{new} ;
end

Algorithm 1. Constrained Inference Overview (Before)

Input: M_{new}, Im_{labels}, beam-size k, word w
Output: best k successors
Initialize Im_{labels} = Top-5 (ea) ;
Initialize beam-size k ;
Initialize word w=null ;
Function ***During Inference***
  forall *State st of k words* **do**
    w=st ;
    if *closest[w] in ea* **then**
      st = closest[w];
    end
    else
      $st = w$;
    end
  end
  return best k successors ;
end

Algorithm 2. Constrained Inference Overview (During)

Furthermore, $W_{\boldsymbol{I}_t}[i_{unseen},i_{closest}]$, $W_{\boldsymbol{I}_t}[i_{closest},i_{unseen}]$ is set to zero for removing mutual dependencies of seen and unseen words presence in an image. Hence, aforementioned procedure will update the KGA-CGM trained model before inference to assist the generation of unseen visual object categories during inference as given by Algorithm 1.

Guidance During Inference. The updated KGA-CGM model is used for generating descriptions of unseen visual object categories. However, in the before-inference procedure, the closest words to unseen visual object categories are identified using embeddings that are learned only using textual corpora and are never constrained on images. This obstructs the view from an image leading to spurious results. We resolve such nuances during inference by constraining the beam search used for description generation with image entity-labels ($\boldsymbol{ea}$). In general, beam search is used to consider the best k sentences at time t to identify the sentence at the next time step. Our modification to beam search is achieved by adding a extra constraint to check if a generated unseen visual object category is part of the entity-labels. If it's not, unseen visual object categories are never replaced with their closest seen words. Algorithm 2 presents the overview of KGA-CGM guidance during inference.

4 Experimental Setup

4.1 Resources and Datasets

Our approach is dependent on several resources and datasets.

Knowledge Graphs (KGs) and Unpaired Textual Corpora. There are several openly available KGs such as DBpedia, Wikidata, and YAGO which provide semantic knowledge encapsulated in entities and their relationships. We choose DBpedia as our KG for entity annotation, as it is one of the extensively used resource for semantic annotation and disambiguation [6].

For learning weights of the language model and also Glove word embeddings, we have explored different unpaired textual corpora from out-of-domain sources (i.e. out of image-caption parallel corpora) such as the British National Corpus (BNC)[1], Wikipedia (Wiki) and subset of SBU1M[2]caption text containing 947 categories of ILSVRC12 dataset [11]. NLTK[3] sentence tokenizer is used to extract tokenizations and around 70k+ words vocabulary is extracted with Glove embeddings.

Unseen Objects Description (Out-of-Domain MSCOCO and ImageNet). To evaluate KGA-CGM, we use the subset of MSCOCO dataset [7] proposed by Hendricks et al. [4]. The dataset is obtained by clustering 80 image object category labels into 8 clusters and then selecting one object from each cluster to be held out from the training set. Now the training set does not contain the images and sentences of those 8 objects represented by bottle, bus, couch, microwave, pizza, racket, suitcase and zebra. Thus making the MSCOCO training dataset to constitute 70,194 image-caption pairs. While validation set of 40504 image-caption pairs are again divided into 20252 each for testing and validation. Now, the goal of KGA-CGM is to generate caption for those test images which contain these 8 unseen object categories. Henceforth, we refer this dataset as "out-of-domain MSCOCO".

To evaluate KGA-CGM on more challenging task, we attempt to describe images that contain wide variety of objects as observed on the web. To imitate such a scenario, we collected images from collections containing images with wide variety of objects. First, we used same set of images as earlier approaches [15,17] which are subset of ImageNet [3] constituting 642 object categories used in Hendricks et al. [4] who do not occur in MSCOCO. However, 120 out of those 642 object categories are part of ILSVRC12.

4.2 Multi-label Image Classifiers

The important constituents that influence KGA-CGM are the image entity-labels and visual features. Identified objects/actions etc. in an image are embodied in visual features, while entity-labels capture the semantic knowledge in an image grounded in KG. In this section, we present the approach to extract both visual features and entity-labels.

[1] http://www.natcorp.ox.ac.uk/.
[2] http://vision.cs.stonybrook.edu/~vicente/sbucaptions/.
[3] http://www.nltk.org/.

Multi-word-Label Image Classifier. To extract visual features of out-of-domain MSCOCO images, emulating Hendricks et al. [4] a multi-word-label classifier is built using the captions aligned to an image by extracting part-of-speech (POS) tags such as nouns, verbs and adjectives attained for each word in the entire MSCOCO dataset. For example, the caption "A young child brushes his teeth at the sink" contains word-labels such as "young (JJ)", "child (NN)", "teeth (NN)" etc., that represent concepts in an image. An image classifier is trained now with 471 word-labels using a sigmoid cross-entropy loss by fine-tuning VGG-16 [13] pre-trained on the training part of the ILSVRC12. The visual features extracted for a new image represent the probabilities of 471 image labels observed in that image. For extracting visual features from ImageNet images, we replace the multi-word-label classifier with the lexical classifier [4] learned with 642 ImageNet object categories.

Multi-entity-Label Image Classifier. To extract semantic knowledge for out-of-domain MSCOCO images analogous to the word-labels, a multi-entity-label classifier is build with entity-labels attained from a knowledge graph annotation tool such as DBpedia spotlight[4] on training set of MSCOCO constituting 82,783 training image-caption pairs. In total around 812 unique labels are extracted with an average of 3.2 labels annotated per image. To illustrate, considering the caption presented in the aforementioned section, entity labels extracted are "Brush[5]" and "Tooth[6]". An image classifier is now trained with multiple entity-labels using sigmoid cross-entropy loss by fine-tuning VGG-16 [13] pre-trained on the training part of the ILSVRC12. For extracting entity-labels from ImageNet images, we again leveraged lexical classifier [4] learned with 642 ImageNet object categories. However, as all 642 categories denote WordNet synsets, we build a connection between these categories and DBpedia by leveraging BabelNet [8] for multi-entity-label classifier. To illustrate, for visual object category "wombat" (wordnetid: *n1883070*) in ImageNet can be linked to DBpedia Wombat[7]. Hence, this makes our method very modular for building new image classifiers to incorporate semantic knowledge.

4.3 Entity-Label Embeddings

We presented earlier that the acquisition of entity-labels for training multi-entity-label classifiers were obtained using DBpedia spotlight entity annotation and disambiguation tool. Hence, entity-labels are expected to encapsulate semantic knowledge grounded in KG. Further, entities in a KG can be represented with embeddings by capturing their relational information. In our work, we see the efficacy of these embeddings for caption generation. Thus, we leverage entity-label embeddings for computing semantic attention observed in an image

[4] https://github.com/dbpedia-spotlight/.
[5] http://dbpedia.org/resource/Brush.
[6] http://dbpedia.org/resource/Tooth.
[7] http://dbpedia.org/page/Wombat.

with respect to the caption as observed from KG. To obtain entity-label embeddings, we adopted the RDF2Vec [10] approach and generated 500 dimensional vector representations for 812 and 642 entity-labels to describe out-of-domain MSCOCO and ImageNet images respectively.

4.4 Evaluation Measures

To evaluate generated descriptions for the unseen MSCOCO visual object categories, we use similar evaluation metrics as earlier approaches [4,15,17] such as METEOR and also SPICE [2]. However, CIDEr [14] metric is not used as it is required to calculate the inverse document frequency used by this metric across the entire test set and not just unseen object subsets. F1 score is also calculated to measure the presence of unseen objects in the generated captions when compared against reference captions. Furthermore, to evaluate ImageNet object categories description generation: we leveraged F1 and also other metrics such as Unseen and Accuracy scores [15,17]. The Unseen score measures the percentage of all novel objects mentioned in generated descriptions, while accuracy measure percentage of image descriptions correctly addressed the unseen objects.

5 Experiments

The experiments are conducted to evaluate the efficacy of KGA-CGM model for describing out-of-domain MSCOCO and ImageNet images.

5.1 Implementation

KGA-CGM model constitutes three important components i.e. language model, visual features and entity-labels. Before learning KGA-CGM model with image-caption pairs, we first learn the weights of language model and keep it fixed during the training of KGA-CGM model. To learn language model, we leverage unpaired textual corpora (e.g. entire MSCOCO set, Wiki, BNC etc.) and provide input word embeddings representing 256 dimensions pre-trained with Glove [9] default settings on the same unpaired textual corpora. Hidden layer dimensions of language model are set to 512. KGM-CGM model is then trained using image-caption pairs with Adam optimizer with gradient clipping having maximum norm of 1.0 for about 15–50 epochs. Validation data is used for model selection and experiments are implemented with Keras + Theano backend[8].

5.2 Describing Out-of-Domain MSCOCO Images

In this section, we evaluate KGA-CGM using out-of-domain MSCOCO dataset.

[8] https://github.com/adityamogadala/KGA.

Quantitative Analysis. We compared our complete KGA-CGM model with the other existing models that generate image descriptions on out-of-domain MSCOCO. To have a fair comparison, only those results are compared which used VGG-16 to generate image features. Table 1 shows the comparison of individual and average scores based on METEOR, SPICE and F1 on all 8 unseen visual object categories with beam size 1.

Table 1. Measures for all 8 unseen objects. Underline shows the second best.

F1										
Model	Beam	Microwave	Racket	Bottle	Zebra	Pizza	Couch	Bus	Suitcase	Average
DCC [4]	1	28.1	52.2	4.6	79.9	64.6	45.9	29.8	13.2	39.7
NOC [15]	>1	24.7	55.3	17.7	89.0	69.3	25.5	68.7	39.8	48.8
CBS(T4) [2]	>1	29.7	57.1	16.3	85.7	**77.2**	**48.2**	67.8	**49.9**	54.0
LSTM-C [17]	>1	27.8	70.2	29.6	91.4	68.1	38.7	**74.4**	44.7	**55.6**
KGA-CGM	1	**50.0**	**75.3**	**29.9**	**92.1**	70.6	42.1	54.2	25.6	55.0
METEOR										
DCC [4]	1	22.1	20.3	18.1	22.3	**22.2**	**23.1**	**21.6**	18.3	21.0
NOC [15]	>1	21.5	24.6	21.2	21.8	21.8	21.4	20.4	18.0	21.3
LSTM-C [17]	>1	-	-	-	-	-	-	-	-	23.0
CBS(T4) [2]	>1	-	-	-	-	-	-	-	-	**23.3**
KGA-CGM	1	**22.6**	**25.1**	**21.5**	**22.8**	21.4	23.0	20.3	**18.7**	22.0
SPICE										
DCC [4]	>1	-	-	-	-	-	-	-	-	13.4
CBS(T4) [2]	>1	-	-	-	-	-	-	-	-	**15.9**
KGA-CGM	1	13.3	16.8	13.1	19.6	13.2	14.9	12.6	10.6	14.3

It can be noticed that KGA-CGM with beam size 1 was comparable to other approaches even though it used fixed vocabulary from image-caption pairs. For example, CBS [2] used expanded vocabulary of 21,689 when compared to 8802 by us. Also, our word-labels per image are fixed, while CBS uses a varying size of predicted image tags (T1-4). This makes it non-deterministic and can increase uncertainty, as varying tags will either increase or decrease the performance. Furthermore, we also evaluated KGA-CGM for the rest of seen visual object categories in the Table 2. It can be observed that our KGA-CGM outperforms existing approaches as it did not undermine the in-domain description generation, although it was tuned for out-of-domain description generation.

Qualitative Analysis. In Fig. 3, sample predictions of our best KGA-CGM model is presented. It can be observed that entity-labels has shown an influence for caption generation. Since, entities as image labels are already disambiguated, it attained high similarity in the prediction of a word thus adding useful semantics. Figure 3 presents the example unseen visual objects descriptions.

Table 2. Average measures of MSCOCO seen objects.

Seen objects				
Model	Beam	METEOR	SPICE	F1-score
DCC [4]	1	23.0	15.9	-
CBS(T4) [2]	>1	24.5	18.0	-
KGA-CGM	1	24.1	17.2	-
KGA-CGM	>1	**25.1**	**18.2**	-

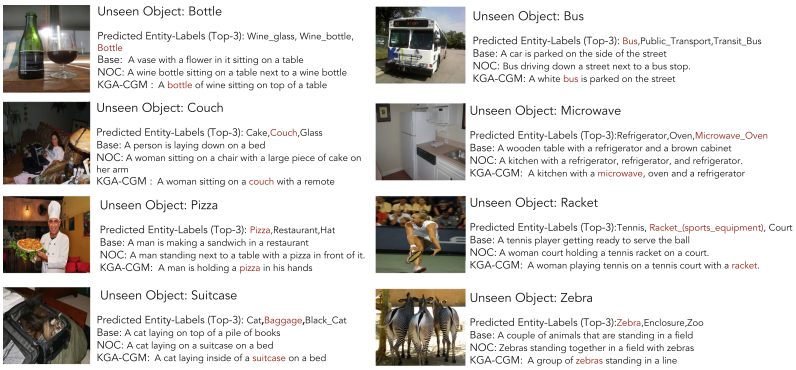

Fig. 3. Sample predictions of KGA-CGM on out-of-domain MSCOCO images with beam size 1 when compared against base model and NOC [15]

5.3 Describing ImageNet Images

ImageNet images do not contain any ground-truth captions and contain exactly one unseen visual object category per image. Initially, we first retrain different language models using unpaired textual data (Sect. 4.1) and also the entire MSCOCO training set. Furthermore, the KGA-CGM model is rebuilt for each one of them separately. To describe ImageNet images, image classifiers presented in the Sect. 4.2 are leveraged. Table 3 summarizes the experimental results attained on 634 categories (i.e. not all 642) to have fair comparison with other approaches. By adopting only MSCOCO training data for language model, our KGA-CGM makes the relative improvement over NOC and LSTM-C in all categories i.e. unseen, F1 and accuracy. Figure 4 shows few sample descriptions.

6 Key Findings

The key observations of our research are: (1) The ablation study conducted to understand the influence of different components in KGA-CGM has shown that using external semantic attention and constrained inference has superior performance when compared to using only either of them. Also, increasing the beam size during inference has shown a drop in all measures. This is basically

Table 3. Describing ImageNet images with beam size 1. Results of NOC and LSTM-C (with Glove) are adopted from Yao et al. [17]

Model	Unpaired-text	Unseen	F1	Accuracy
NOC [15]	MSCOCO	69.1	15.6	10.0
	BNC & Wiki	87.7	31.2	22.0
LSTM-C [17]	MSCOCO	72.1	16.4	11.8
	BNC & Wiki	89.1	33.6	31.1
KGA-CGM	MSCOCO	74.1	17.4	12.2
	BNC & Wiki	90.2	34.4	33.1
	BNC & Wiki & SBU1M	**90.8**	**35.8**	**34.2**

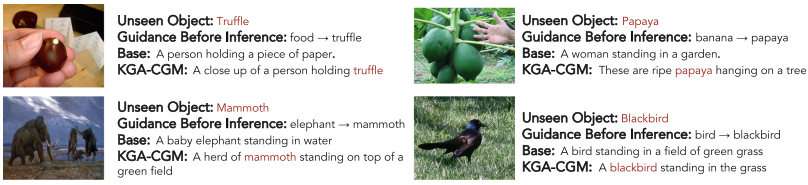

Fig. 4. ImageNet images with best KGA-CGM model from Table 3. Guided before inference shows which words are used for transfer between seen and unseen. (Color figure online)

adhered to the influence of multiple words on unseen objects. (2) The performance advantage becomes clearer if the domain of unseen objects is broadened. In other words: KGA-CGM specifically improves over the state-of-the-art in settings that are larger and less controlled. Hereby, KGA-CGM scales to one order of magnitude more unseen objects with moderate performance decreases. (3) The influence of the closest seen words (i.e. observed in image-caption pairs) and the unseen visual object categories played a prominent role for generating descriptions. For example in out-of-domain MSCOCO, words such as "suitcase"/"bag", "bottle"/"glass" and "bus/truck" are semantically similar and are also used in the similar manner in a sentence added excellent value. However, some words usually cooccur such as "racket"/"court" and "pizza"/"plate" played different roles in sentences and lead to few grammatical errors. (4) The decrease in performance have a high correlation with the discrepancy between the domain where seen and unseen objects come from.

7 Conclusion and Future Work

In this paper, we presented an approach to generate captions for images that lack parallel captions during training with the assistance from semantic knowledge encapsulated in KGs. In the future, we plan to expand our models to build multimedia knowledge graphs along with image descriptions which can be used for finding related images or can be searched with long textual queries.

Acknowledgements. First author is grateful to KHYS at KIT for their research travel grant and Computational Media Lab at ANU for providing access to their K40x GPUs.

References

1. Ahn, S., Choi, H., Pärnamaa, T., Bengio, Y.: A neural knowledge language model. arXiv preprint arXiv:1608.00318 (2016)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Guided open vocabulary image captioning with constrained beam search. In: EMNLP (2017)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)
4. Hendricks, L.A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., Darrell, T.: Deep compositional captioning: describing novel object categories without paired training data. In: CVPR, pp. 1–10 (2016)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
6. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., et al.: DBpedia-a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web **6**, 167–195 (2015)
7. 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: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
8. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)
9. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: EMNLP, pp. 1532–1543 (2014)
10. Ristoski, P., Paulheim, H.: RDF2Vec: RDF graph embeddings for data mining. In: Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y. (eds.) ISWC 2016. LNCS, vol. 9981, pp. 498–514. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46523-4_30
11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
12. Serban, I.V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., Bengio, Y.: Generating factoid questions with recurrent neural networks: the 30M factoid question-answer corpus. arXiv preprint arXiv:1603.06807 (2016)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Vedantam, R., Zitnick, L.C., Parikh, D.: CIDEr: consensus-based image description evaluation. In: CVPR, pp. 4566–4575 (2015)
15. Venugopalan, S., Hendricks, L.A., Rohrbach, M., Mooney, R., Darrell, T., Saenko, K.: Captioning images with diverse objects. In: CVPR (2017)
16. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 652–663 (2017)
17. Yao, T., Yingwei, P., Yehao, L., Mei, T.: Incorporating copying mechanism in image captioning for learning novel objects. In: CVPR (2017)