Integrating Image Captioning with Rule-based Entity Masking

Anonymous authors

Abstract

Given an image, generating its natural language description (i.e., caption) is a well studied problem. Approaches proposed to address this problem usually rely on image features that are difficult to interpret. Particularly, these image features are subdivided into global and local features, where global features are extracted from the global representation of the image, while local features are extracted from the objects detected locally in an image. Although, local features extract rich visual information from the image, existing models generate captions in a blackbox manner and humans have difficulty interpreting which local objects the caption is aimed to represent. Hence in this paper, we propose a novel framework for the image captioning with an explicit object (e.g., knowledge graph entity) selection process while still maintaining its end-to-end training ability. The model first explicitly selects which local entities to include in the caption according to a human-interpretable mask, then generate proper captions by attending to selected entities. Experiments conducted on the MSCOCO dataset demonstrate that our method achieves good performance in terms of the caption quality and diversity with a more interpretable generating process than previous counterparts.

1 Introduction

Over the past few years, the task of generating descriptions for images (i.e., image captioning) [Vinyals *et al.*, 2015; Anderson *et al.*, 2017] has become popular as it effectively brings together vision and natural language to serve various real-world applications. Most of the existing approaches are efficient in learning a correspondence between image and sequence of words with different techniques that either improve how visual information is captured with attention [Xu *et al.*, 2015; Lu *et al.*, 2016; Anderson *et al.*, 2017] or language model interactions [Shen *et al.*, 2017].

Careful analysis of methods that aim to effectively capture visual information reveal that either utilize global image features or attend to regions for local image features to generate captions. However, this makes it hard to interpret, as they do not select or control objects in an image which may be prominent for caption generation. It is especially important for easy understanding of the caption generation process in case of failures in those systems that cater real-world applications such as autonomous driving, medical imaging and surveillance. Also, observed previously [Wang *et al.*, 2018] that rich entities and their interactions in some kind of a layout can help to better understand image captioning.

Therefore, in this paper, we introduce our interpretable image caption generation model (henceforth, Interpret-IC) to address the limitations of previous approaches as shown in the Figure 1. Our proposed approach work with a *human-interpretable mask* which selects the set of local objects observed in an image based on human proposed rules. These rules ensure that only those desirable objects are selected which human wants to observe in the caption. For this to work, the local objects need to be represented with semantically enriched labels so that humans can comprehend. As none of the current approaches provide such local object information. We leveraged relational knowledge provided by the knowledge graph entities to attain semantic labels by building a multi-label image classifier and replace local object visual features with entity distributed representations [Bordes *et al.*, 2011]. We show that these entity labels and its features are superior detected local object features in terms of interpreting knowledge from the image. Very close to our approach by [Cornia *et al.*, 2019], who considers the decomposition of a sentence into noun chunks and models the relationship between image regions and textual chunks. However, we dynamically select the number of objects prior learning the model. Our main contributions are as follows:

- We proposed a novel end-to-end caption model for intepretable image captioning.
- We used knowledge graph entities as image labels for grounding visual and factual knowledge.
- We show that interpretable image captioning can attain diversity in the captions generated with simple visual object masking.

2 Related Work

In the related work, we explore deep neural network based approaches which generate sentence-level natural language description for images.

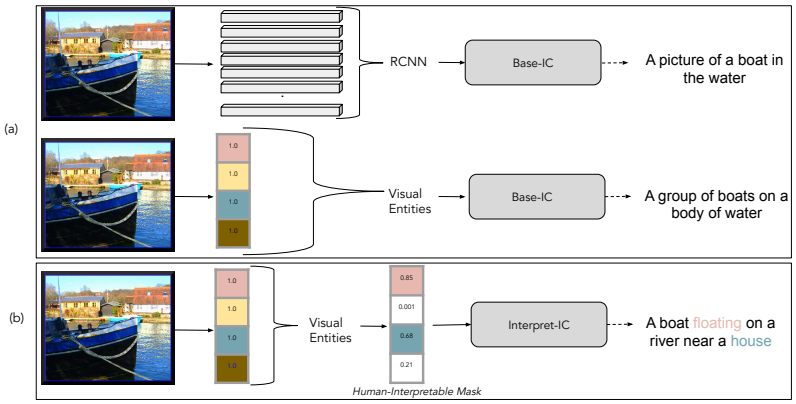


Figure 1: Comparison between two proposed models with different visual features (a) Base-IC (Section 3.1) and (b) Interpret-IC (Section 3.2). Interpret-IC has an extra process of highlighting which objects to cover in the generated caption from a shortlist of all detected objects in the image.

Diverse Image Captioning In the recent years, monolingual image caption generation is explored to incorporate diversity in the generated captions. Approaches [Li *et al.*, 2018] has leveraged adversarial training using either generative adversarial networks [Shetty *et al.*, 2017] or variational autoencoder [Shen *et al.*, 2019]. While, [Vijayakumar *et al.*, 2016] used diverse beam search to decode diverse image captions in English. Approaches were also proposed to describe images from cross-domain [Chen *et al.*, 2017]. However, our goal in this research is to provide better selection procedure for identifying preferable objects in images. Nevertheless, we show that interpretability can also assist diversity.

Controllable Image Captioning Approach that is closer to intepretable image captioning is a procedure to control local objects in images. [Cornia *et al.*, 2019] used either a sequence or a set of local objects by explicitly grounding them with noun chunks observed in the captions to generate diverse captions. Further, instead of making captions only diverse, [Deshpande *et al.*, 2019] made the captioning more accurate. Our work falls into this space, however understanding the important entities that represent the image and controlling them is what we aim to achieve.

3 Interpretable Image Captioning

3.1 Base-IC Model

The base image caption model (Base-IC) is built *without* masking. Given an image I, its global representation $\boldsymbol{I}_v \in \mathbb{R}^V$ denote the encoding of the full image, while the spatial objects $a_v = \{a_{v_1}, \ldots, a_{v_L}\}$ encode local regions of the image provided as $\boldsymbol{a}_{v_j} \in \mathbb{R}^D$. Similar to previous works [Lu *et al.*, 2016; Anderson *et al.*, 2017], our proposed image description model also leverages soft attention mechanism to weigh spatial objects during description generation using the partial output sequence as context. Figure 2 illustrates the architecture.

Initially, L-1 of the model receives input from the global visual context provided by $\boldsymbol{I}_v$ and textual sequence, where each word ($\boldsymbol{w}_t \in \mathbb{R}^T$) at time step t in the textual sequence is initialized with the pretrained word embeddings to produce

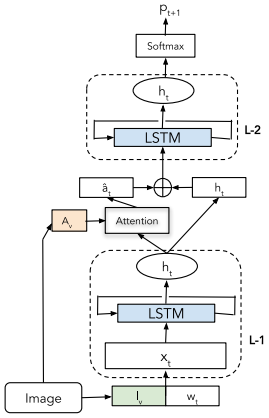


Figure 2: Illustration of Base-IC Model

hidden vectors $\boldsymbol{h}_t^1 \in \mathbb{R}^{H_1}$. Furthermore, $\boldsymbol{h}_t^1$ is used in combination with $\boldsymbol{a}_v$ to compute soft attention. Later, $\boldsymbol{h}_t^1$ and attended spatial features are added and provided as input to L-2 for attaining $\boldsymbol{h}_t^2 \in \mathcal{R}^{H_2}$. For convenience and to reduce many parameter names, we use Θ as the reference for the parameters of the LSTM.

To calculate attended spatial features ($\hat{\boldsymbol{a}}_t$) we leverage $\boldsymbol{a}_v$. Hidden sequences $\boldsymbol{h}_t^1$ at each time step t is used to generate a normalized attention weight $\boldsymbol{\alpha}_t$ for each of the spatial object features ($\boldsymbol{a}_{v_j}$) given by Equation 1 and Equation 2.

$$\boldsymbol{\alpha}_{tj} = \frac{exp(\boldsymbol{e}_{tj})}{\sum_{k=1}^{L} exp(\boldsymbol{e}_{tk})} \tag{1}$$

$$\boldsymbol{e}_{tj} = \tanh(W_{ae}\boldsymbol{a}_{v_j} + W_{he}\boldsymbol{h}_t^1) \tag{2}$$

where L represent the cardinality of set a_v. $W_{ae} \in \mathbb{R}^{M \times D}$ and $W_{he} \in \mathbb{R}^{M \times H_1}$ are learnable parameters. Further, $\hat{\boldsymbol{a}}_t$ is calculated with Equation 3 and is used as input along with $\boldsymbol{h}_t^1$ to the L-2 at every time step t.

$$\hat{\boldsymbol{a}}_t = \sum_{j=1}^{L} \boldsymbol{\alpha}_{tj} a_{v_j} \tag{3}$$

The final Base-IC using $\boldsymbol{w}_t$ and $\boldsymbol{I}_v$ as input to L-1 is given by Equation 4 and $\boldsymbol{h}_t^1$ is given by Equation 5. Further, $\hat{\boldsymbol{a}}_t$ and $\boldsymbol{h}_t^1$ are added using Equation 6 to provide as input for L-2 for generating $\boldsymbol{h}_t^2$ as given by Equation 7. It is then used to predict next words in the sequence as given in the Equation 8.

$$\boldsymbol{x}_t = \boldsymbol{I}_v \oplus \boldsymbol{w}_t \tag{4}$$

$$\boldsymbol{h}_t^1 = \text{L-1}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}^1; \Theta) \tag{5}$$

$$\boldsymbol{x}_t' = \hat{\boldsymbol{a}}_t + \boldsymbol{h}_t^1 \tag{6}$$

$$\boldsymbol{h}_t^2 = \text{L-2}(\boldsymbol{x}_t', \boldsymbol{h}_{t-1}^2; \Theta) \tag{7}$$

$$p_{t+1} = \text{softmax}(W_{vocab}\boldsymbol{h}_t^2) \tag{8}$$

where $W_{vocab} \in \mathbb{R}^{vocab \times (V+H_2)}$, $\oplus$ represents concatenation and $vocab$ refers to vocabulary of the caption dataset.

3.2 Interpret-IC Model

Main aim of the *Interpret-IC* model is to select objects present in the spatial objects set a_v with human-interpretable masking. This is in contrast with earlier approaches [Xu *et al.*, 2015; Anderson *et al.*, 2017], who decoded the caption by attending to spatial objects only by ranking them according to their importance at each time step. Also, these approaches provide no control for humans to select their desirable objects. It clearly sets expectation from *Interpret-IC* model that the selected objects should provide more prominence in caption generation by discarding those objects that are not selected.

Hence, we introduce *masked attention* to select those objects that human wants to see in the generated captions. To achieve it, we leverage ground truth *mask* i.e., mask_{gt} where each object in the a_v is masked with a binary parameter $\beta_1, \beta_2, \ldots, \beta_n$. We set $\beta_i = 1$ if selected and 0 otherwise. Also, β_i is assumed to be independent from each other and is sampled from a bernoulli distribution. Prediction mask i.e., mask_{pred} is estimated during training with a multi-layer perceptron (MLP).

Further, attention weights computed in the Equation 1 is modified with the estimated mask_{pred} as shown in Equation 9.

$$\boldsymbol{\alpha}_{tj}^{mask} = \frac{exp(\boldsymbol{e}_{tj})\text{mask}_{pred}}{\sum_{k=1}^{L} exp(\boldsymbol{e}_{tk})\text{mask}_{pred}} \tag{9}$$

It is then used to calculate $\hat{\boldsymbol{a}}_t^{mask}$ given by Equation 10, which is further used as input along with $\boldsymbol{h}_t^1$ to the L-2 at every time step t. Figure 3 illustrates the overall architecture.

$$\hat{\boldsymbol{a}}_t^{mask} = \sum_{j=1}^{L} \boldsymbol{\alpha}_{tj}^{mask} a_{v_j} \tag{10}$$

Note that our selection strategy is very different from [Cornia *et al.*, 2019], who control spatial objects using the fixed noun-chunks extracted from captions which are not available during testing phase. While, we use human designed rules to change our mask, so that we control the mask as we aim to use it.

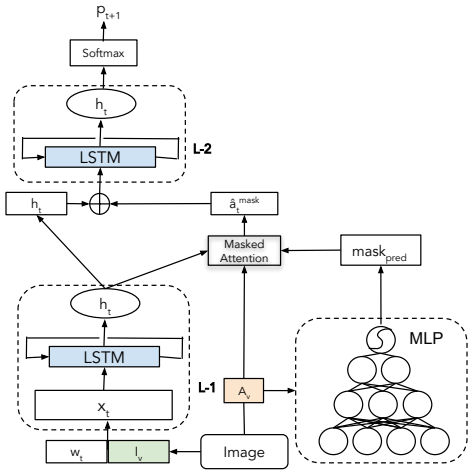


Figure 3: Illustration of Interpret-IC Model

3.3 Ground Truth Mask Selection

In the *Interpret-IC* model, mask_{pred} needs to be optimized during training phase closer to the ground truth binary mask i.e., mask_{gt} such that it can be utilized during the testing phase. However, first we need to create such mask_{gt} based on *human-interpretable rules* to influence the caption generation process.

There can be several ways to create mask_{gt} by changing the rules. In this paper, we apply *visual entities* to caption *noun* matching approach to build the mask_{gt}. Our *rule* here states that for each noun identified[1] in the caption, we need to find the closest visual entity by computing cosine distance between the noun and visual entity vectors attained using pretrained fastText[2] vectors. For all nouns identified, closest visual entities are set to 1, while rest are set to 0. This rule ensures that the nouns observed in the caption representing some kind of objects present in images have to be given higher preference during caption generation. While, rest of the visual entities (e.g., actions) are put on back burner. Algorithm 1 presents the overview of selection process.

4 Training and Inference

Base-IC The parameters (θ) of the *Base-IC* model are trained for optimizing the cost function ($\mathcal{C}$) to minimize the sentence-level categorical cross-entropy loss by finding negative log likelihood of the appropriate ground truth word (y_t^*) at each time step t as shown in Equation 11. Here, we leverage teacher forcing [Sutskever *et al.*, 2014], where ground truth (y_t^*) is fed to next step in the layer L-1, instead of the predicted word in previous step.

$$\mathcal{C}(\theta) = -\sum_{t=0}^{T^{(n)}} \log p_\theta(y_t^*) \tag{11}$$

The $T^{(n)}$ represents the length of the sentence at n-th training sample. During inference, we leverage beam search with beam size is set to 5 in our experiments.

[1] https://spacy.io/

[2] https://fasttext.cc/

Nouns (N), Visual Entities (VE), fastText Embeddings (FTE) mask_{gt} for each caption Initialize N_{emb} = FTE(N) ;
Initialize VE_{emb} = FTE(VE) ;
Initialize Image_{velist} as I_{velist};
Initialize Caption_{list} as C_{list};
Function *mask_{gt} Selection*
  for *C,VE in* C_{list},I_{velist} **do**
    Extract N from caption ;
    Initialize mask_{gt} = zeros[len(VE)] ;
    for *n in N* **do**
      if *n not EMPTY* **then**
        dist = CosineDistance(n_{emb},VE_{emb});
        close_{index}= (dist);
        mask_{gt}[close_{index}] = 1;
      end
    end
    return mask_{gt} ;
  end
end

Algorithm 1: mask_{gt} selection process

Interpret-IC Similar to *Base-IC* model, parameters ($\theta^{'}$) of the *Interpret-IC* are trained for optimizing the cost function ($\mathcal{C}^{'}$) which minimizes both the sentence-level categorical cross-entropy loss along with binary cross-entropy loss that approximate (mask_{pred}) closer to the ground truth mask (mask_{gt}) as shown in Equation 12.

$$\mathcal{C}^{'}(\theta^{'}) = -\left(\left(\sum_{t=0}^{T^{(n)}} \log p_\theta(y_t^*)\right) + \text{mask}_{gt} log(\text{mask}_{pred}) + (1-\text{mask}_{gt}) log(1-\text{mask}_{pred})\right) \tag{12}$$

During inference, similar to *Base-IC* model, we leverage beam search by setting beam size to 5 in our experiments.

5 Evaluation Setup

Datasets For experimental evaluation, we use MSCOCO dataset with splits of [Karpathy and Fei-Fei, 2015].

Local and Global Image Features Spatial object ($\boldsymbol{a}_v$) features are extracted in two different ways.

- Faster R-CNN [Ren *et al.*, 2015] in conjunction with the ResNet-101 [He *et al.*, 2016] trained on visual genome data by [Anderson *et al.*, 2017] is used to extract top 36 local object features ($\boldsymbol{a}_{v_j}$) of dimension 2048. There are pure visual features and we refer to this set as Obj→RCNN.
- Since, Obj→RCNN represent pure visual features without label information. Following [Mogadala *et al.*, 2018], we extracted semantically enriched labels denoting entities from captions aligned to an image in training set of MSCOCO with a knowledge graph annotation tool such as DBpedia spotlight[3]. In total, 812 unique human-interpretable already disambiguated labels are extracted. Further, a multi-label image classifier is trained with sigmoid cross-entropy loss by fine-tuning VGG-16 [Simonyan and Zisserman, 2014] pre-trained on the training part of the ILSVRC12 with training images in MSCOCO. After training, we use the classifier to acquire Top-15 entity labels for each image present in the training, validation and testing set of MSCOCO. Now, to use entity labels similar to Obj→RCNN features. We use knowledge graph embeddings [Ristoski and Paulheim, 2016] and generate 500 dimensional vectors[4] for each entity-label. We refer to this set as Obj→VisualEntity.
- The global visual features ($\boldsymbol{I}_v$) of dimension 2048 is extracted using the average pooling of Obj→RCNN features.

Caption Model Both *Base-IC* and *Interpret-IC* models are built by initializing the model with input ($\boldsymbol{w}_t$) word embeddings pretrained using Glove [Pennington *et al.*, 2014] on the MSCOCO training captions corpora. The dimensions of the hidden units $\boldsymbol{h}_t^1, \boldsymbol{h}_t^2$ in **L-1** and **L-2** of models are set to 512. Also, the hidden units of shared layer $\boldsymbol{h}_t^{(s)}$ are set to 512. All models are then trained with Adam optimizer with gradient clipping having maximum norm of 1.0 and mini-batch size of 50 for 25 epochs. Initially, the learning is set to 0.001 and is reduced by a factor of 10 if there is no improvement in the validation loss for 3 continuous epochs.

Evaluation Measures We first evaluate the generated captions based on correctness which guarantee the generation quality based on standard captioning metrics. Further, we check if our proposed model with human-interpretable masking can generate diverse and interesting captions. For this, we leverage earlier proposed [Shetty *et al.*, 2017; Deshpande *et al.*, 2019] metrics such as vocabulary size and novel caption with best (i.e., Top-1) generated caption. Vocabulary Size (VS) find unique words in generated captions and Novel captions (NC) identify the percentage of generated captions that are not seen in the training set.

6 Results

6.1 Quantitative Results

We compared our proposed *Base-IC* and *Interpret-IC* along with other recent baselines. Table 1 shows the results obtained. It can be observed that the *Interpret-IC* model was able improve over recent approaches by allowing better control over the caption generation process.

6.2 Qualitative Results

To understand the contribution made by *human-interpretable mask* to caption generation. We explored qualitatively the

[3]https://github.com/dbpedia-spotlight/

[4]Please note that these embeddings are different from fastText Vectors used to build mask_{gt}. These embeddings are analogous to pure visual features, however, learned from knowledge graph structure.

Model	Cross-Entropy Loss				
	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Adv-bs [Shetty *et al.*, 2017]	-	23.9	-	-	16.7
CNN+CNN [Wang and Chan, 2018]	26.7	23.4	51.0	84.4	-
Convolutional-IC [Aneja *et al.*, 2018]	31.6	**25.0**	53.1	95.2	17.9
POS+Joint [Deshpande *et al.*, 2019]	-	24.7	-	-	18.0
Base-IC					
+Obj→RCNN	31.8	24.9	52.9	96.7	**18.2**
+Obj→VisualEntity	32.1	24.8	53.6	96.9	18.0
Interpret-IC					
+Obj→VisualEntity	**32.4**	24.9	**53.7**	**97.8**	18.1

Table 1: Results achieved with our models in comparison with baseline approaches.

Figure 4: **Caption Coverage Example** (Entities with $\text{mask}_{pred} > 0.5$ are highlighted in blue): (a) Missing local object (Dog) in the caption generated by *Base-IC*, while "White Dog" is included by *Interpret-IC* providing better coverage. (b) Missing details about the birthday cake, *Interpret-IC* generated better and interesting caption by highlighting objects that need to be focused on.

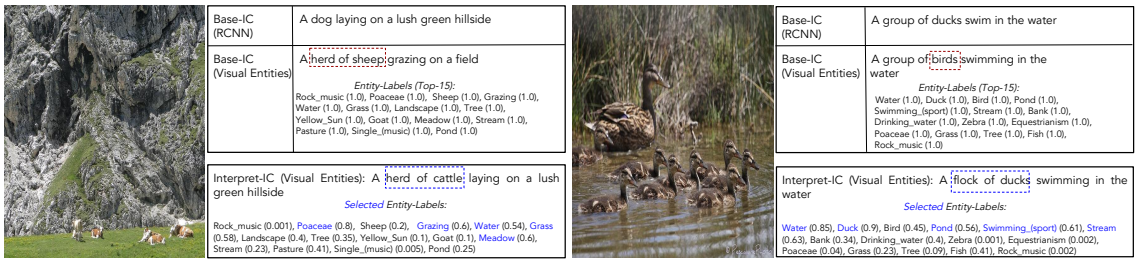
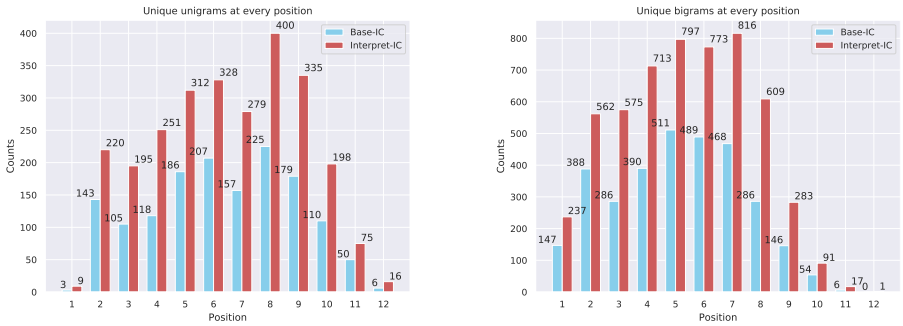


Figure 5: **Caption Correction Example** (Entities with $\text{mask}_{pred} > 0.5$ are highlighted in blue): (a) *Base-IC* generate a caption by including wrong objects i.e., sheep, while "cattle" is included by *Interpret-IC* because a lower weight (0.2) is assigned to "sheep" hence filters out the wrongly detection. (b) Although *Base-IC* covers the correct object (Birds), it is too general and fails to provide more informative caption. *Interpret-IC* replaces it with the exact object by giving a large weight to emphasize the detected entity "duck".



(a) Unigrams (b) Bigrams

Figure 6: Plot of starting unique unigrams and bigrams observed in the generated caption.

captions generated by both *Base-IC* and *Interpret-IC* models with visual entities from two different perspectives. First, we observed the quality of the predicted mask in selecting required visual entities for better coverage. Second, we checked if *Interpret-IC* model could overcome or correct mistakes made by the *Base-IC* model. In the following, we discuss each of these cases briefly by showing some examples.

Caption Coverage We use visual entities such that they represent local objects in images to be incorporated them in the caption. However, this cannot be simply achieved with a *Base-IC* model. As seen in Figure 4, the *Interpret-IC* model which weighs each of these objects differently based on the predicted mask, when compared with the *Base-IC* model giving equal importance to each of them. Although the Base-IC model generated partially relevant caption, masking has shown to improve coverage of local objects in the image. The selector is able to assign higher scores to prominent objects in the image which increases the probability of covering them in the generated caption.

Caption Correction We also observe that, apart from providing better coverage of visual entities in the generated captions. Masking also plays a prominent role in the caption correction. That is, as seen in the Figure 5, although the *Base-IC* model generated a partially relevant caption, *Interpret-IC* generated the most accurate caption with effective selection of relevant visual entities. The selector is expected to assign lower scores to inappropriate (bird in Figure 5b) or wrongly detected objects (sheep in Figure 5a) thus encouraging the decoder to attend to more plausible entities.

6.3 Diversity

Although our aim is not to achieve diverse captions, to comprehend whether our proposed *Base-IC* and *Interpret-IC* models generate best (i.e., Top-1) diverse and interesting caption. We compared our models with other diverse caption generation baselines that compare best generated caption using diversity measures described earlier. Table 2 shows the results attained. We observe that, our *Interpret-IC* model cannot exceed scores of the baseline trained to generate diverse captions in an adversarial setting (i.e., Adv-bs). However, with less effort and simple masking we could see a significant jump on the standard caption model (i.e., Base-bs).

Metrics	Base-bs	Adv-bs	Base-IC	Interpret-IC
VS	756	**1508**	443	862
NC	34.18	**68.62**	36.23	51.54

Table 2: Diversity: Comparison of vocab size (VS) and novel captions (NC) using Top-1 generated caption with Base-bs [Shetty *et al.*, 2017] and Adv-bs [Shetty *et al.*, 2017]. Base-IC and Interpret-IC use +Obj→VisualEntity features.

Also, in Figure 6, we plot unique unigrams and bigrams predicted at every word position. The plot shows that the *Interpret-IC* have higher unique unigrams at different word positions and is consistently higher for the bigrams when compared against *Base-IC* with visual entities as features. This supports our hypothesis that *Interpret-IC* can produce more diverse captions as it can alter caption generation process.

7 Conclusion and Future Work

In this paper, we aimed to address the problem of interpretable image captioning by leveraging knowledge graph entity features. Initially, we obtained local objects as visual entities in the image by grounding knowledge graph entities. Further, the human-interpretable masking rules are developed to select those visual entities for generating desirable captions. Experimental results show that interpretability in caption generation can help to alter caption generation process hence allowing control and selection. In Future, we aim to improve caption generation process by trying different masks and better sampling.

References

[Anderson *et al.*, 2017] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2017.

[Aneja *et al.*, 2018] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5561–5570, 2018.

[Bordes *et al.*, 2011] Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning structured embeddings of knowledge bases. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[Chen *et al.*, 2017] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. *arXiv preprint arXiv:1705.00930*, 2017.

[Cornia *et al.*, 2019] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: a framework for generating controllable and grounded captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2019.

[Deshpande *et al.*, 2019] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10695–10704, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[Li *et al.*, 2018] Dianqi Li, Qiuyuan Huang, Xiaodong He, Lei Zhang, and Ming-Ting Sun. Generating diverse and accurate visual captions by comparative adversarial learning. *arXiv preprint arXiv:1804.00861*, 2018.

[Lu *et al.*, 2016] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv preprint arXiv:1612.01887*, 2016.

[Mogadala *et al.*, 2018] Aditya Mogadala, Umanga Bista, Lexing Xie, and Achim Rettinger. Knowledge guided attention and inference for describing images containing unseen objects. In *European Semantic Web Conference*, pages 415–429. Springer, 2018.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[Ristoski and Paulheim, 2016] Petar Ristoski and Heiko Paulheim. Rdf2vec: Rdf graph embeddings for data mining. In *International Semantic Web Conference*, pages 498–514. Springer, 2016.

[Shen *et al.*, 2017] Xiaoyu Shen, Youssef Oualil, Clayton Greenberg, Mittul Singh, and Dietrich Klakow. Estimation of gap between current language models and human performance. *Proc. Interspeech 2017*, pages 553–557, 2017.

[Shen *et al.*, 2019] Xiaoyu Shen, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine. Select and attend: Towards controllable content selection in text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 579–590, 2019.

[Shetty *et al.*, 2017] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. *arXiv preprint arXiv:1703.10476*, 2017.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[Vijayakumar *et al.*, 2016] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[Wang and Chan, 2018] Qingzhong Wang and Antoni B Chan. Cnn+ cnn: Convolutional decoders for image captioning. *arXiv preprint arXiv:1805.09019*, 2018.

[Wang *et al.*, 2018] Josiah Wang, Pranava Swaroop Madhyastha, and Lucia Specia. Object counts! bringing explicit detections back into image captioning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2180–2193, 2018.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.