



# Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking

Rajarshi Biswas<sup>1</sup> · Michael Barz<sup>2</sup> · Daniel Sonntag<sup>1</sup>

Received: 30 March 2020 / Accepted: 22 June 2020 / Published online: 8 July 2020  
© The Author(s) 2020

## Abstract

Image captioning is a challenging multimodal task. Significant improvements could be obtained by deep learning. Yet, captions generated by humans are still considered better, which makes it an interesting application for interactive machine learning and explainable artificial intelligence methods. In this work, we aim at improving the performance and explainability of the state-of-the-art method Show, Attend and Tell by augmenting their attention mechanism using additional bottom-up features. We compute visual attention on the joint embedding space formed by the union of high-level features and the low-level features obtained from the object specific salient regions of the input image. We embed the content of bounding boxes from a pre-trained Mask R-CNN model. This delivers state-of-the-art performance, while it provides explanatory features. Further, we discuss how interactive model improvement can be realized through re-ranking caption candidates using beam search decoders and explanatory features. We show that interactive re-ranking of beam search candidates has the potential to outperform the state-of-the-art in image captioning.

**Keywords** Image captioning · Deep learning · Explainable artificial intelligence (XAI) · Visual explanations · Interactive machine learning (IML) · Beam search · Re-ranking

## 1 Introduction

The goal of image captioning is to automatically generate descriptions for a given image, i.e., to capture the relationship between the objects present in the image, generate natural language expressions (see an example in Fig. 1), and judge the quality of the generated descriptions. The problem, therefore, is seemingly more difficult than popular computer vision tasks, e.g., object detection or segmentation, where the emphasis is solely on identifying the different entities present in the image. With recent advancements

in training neural networks [26], the availability of GPU computing power, and large datasets [31], neural network driven approaches are the most popular choice for handling the caption generation problem. However, humans are still better at interpreting images and constructing useful and meaningful captions, with or without a particular application context, which renders it an interesting applications for IML [10, 43] and explainable artificial intelligence (XAI) [11]. Promising technologies include active learning [41], which was already applied for automating the assessment of image captioning [4, 5], IML methods to incrementally train, e.g., re-ranking models for selecting the best caption candidate similar to [3, 39], and XAI methods that can improve the user's understanding of a model and, eventually, enable it to provide better feedback for a second IML process.

In this work, we adopt and extend the architecture proposed in [49] since it is the most cited seminal work in the area of image captioning. It introduced the encoder-decoder architecture and the visual attention mechanism for image captioning in a simple yet powerful approach. Compared to [49] other captioning approaches are task specific, more complex and derivative in nature. Moreover, we believe the simplicity of the Show, Attend and

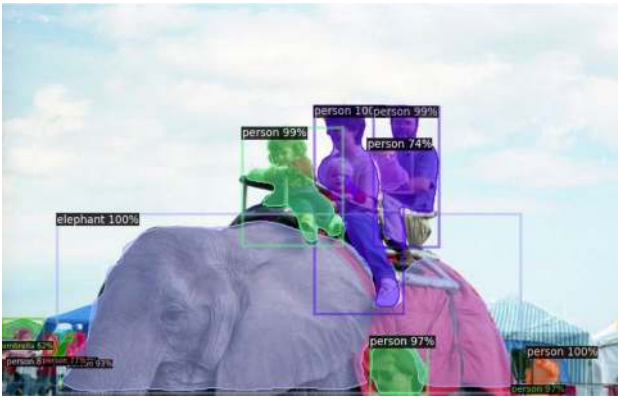
✉ Rajarshi Biswas  
rajarshi.biswas@dfki.de

Michael Barz  
michael.barz@dfki.de

Daniel Sonntag  
daniel.sonntag@dfki.de

<sup>1</sup> German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus, 66123 Saarbrücken, Germany

<sup>2</sup> German Research Center for Artificial Intelligence (DFKI), Saarbrücken Graduate School of Computer Science, Saarland Informatics Campus, 66123 Saarbrücken, Germany



**Fig. 1** Image with caption generated using our approach: “a group of people riding on top of an elephant”

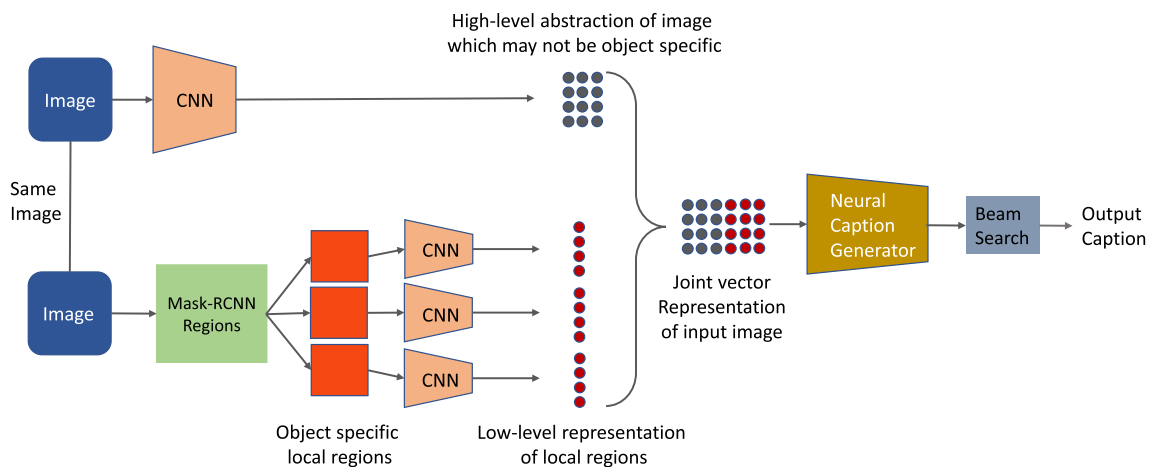
Tell model, compared to its other counterparts, would help to add explainability into the captioning task. This approach uses a transparent attention mechanism which, in the domain of sequence-to-sequence tasks, translates to the ability to dynamically select important features in the image instead of maintaining one feature representation for the image at all times. However, the selected image features on which the attention mechanism works are obtained from a deep convolutional encoder which mostly capture high-level image abstractions and not the low-level object specific details. These high-level features are top-down in nature since their primary purpose is to provide context for the decoder in producing the next word based on the partially generated caption. In doing so, these features often fail to attend and provide direct visual cues, i.e., specific object details, to the decoder. We explore the possibilities to make this attention mechanism more effective, in particular towards novel IML and XAI approaches,

by implementing a combination of *top-down* image features and *bottom-up* object specific details.

We discuss a novel augmentation of the attention mechanism in [49] with bottom-up features, in terms of localization maps encoded in the feature space obtained by the deep convolutional encoder (see Fig. 2). For each input image, we embed the content of a constant number of localized bounding boxes from a pre-trained Mask R-CNN model [13] for augmenting the attention mechanism from [49]. We use Resnet-101 [14] pre-trained on the Imagenet dataset [7] for extracting fixed size feature vectors per bounding box. The resulting set of vectors represents object specific salient regions of the input image.

Further, we compute visual attention on the joint embedding space formed by the union of high-level features obtained from the encoder of the caption generator and the low-level features obtained from our object specific encoding of salient regions of the input image. We show that with our approach we obtain better Bleu scores compared to the original scores in [49], specifically, we obtain higher scores in Bleu-2, Bleu-3, Bleu-4 metrics. In a separate experiment, we use beam search to expand the search space of the associated natural language generation problem [32]. Beam search is a greedy tree search algorithm which sorts possible language generations based on a heuristic and keeps the best  $k$  options with  $k$  being the beam width. We show that effective re-ranking of caption candidates from a beam search decoder has a huge potential for improving results. Further, we discuss how interactive model improvements and explainability can be obtained.

To summarize our contributions: first, we implement a novel image captioning architecture that augments the visual attention mechanism introduced by [49]. We show that our approach achieves comparable or better results on the image captioning task compared to Show, Attend and Tell, while it



**Fig. 2** Caption generation with augmented visual attention

offers more explanatory cues for XAI at the same time. Second, we show the potential of an implemented beam search generation process and how it improves the resulting captions by interactively re-ranking the candidates. Third, we discuss how our architecture can be used for novel IML and XAI approaches. We describe several directions of future work, concerning the bottom-up features and beam search results and the challenges towards explanatory interactive image captioning.

## 2 Related Work

Recently there has been renewed interest in the problem of image captioning in spite of considerable focus in the recent past on language grounding in perceptual data [12, 33, 38]. This is due to a wider push to investigate the intersection between vision and language. In this work, the caption generation method employs the neural framework proposed in [6] where instead of translating text from one language to another, an image is translated into a caption or sentence that describes it. In general, the image caption generator is a neural architecture consisting of a deep convolutional network [17] and a recurrent network [16]. Kiros et al. [24, 25] is credited with the first attempt in this direction where the authors develop a joint multimodal embedding space and provide a natural way of performing both ranking and generation. As a slight modification, the works of [9, 47] employ LSTMs (*Long Short Term Memory*) instead of regular recurrent neural networks. Karpathy et al. [22], on the other hand, advocates learning a joint embedding space for both ranking and generation. As a matter of fact, their model learns to score sentence and image similarity as a function of convolutional network object detection with outputs of a bidirectional RNN (*Recurrent Neural Network*).

The caption generation problem also is a structured learning problem since both the input and output of this problem have a rich structure. That is, the image of a natural scene is made up of multiple random variables, such as, the position of objects and their inter relationships and all of them have a rich joint distribution. Moreover, there needs to be an alignment between the output words of a caption with the spatial regions of the input image. So, to properly address the structured nature of this problem, we make use of an attention mechanism in our work. Hence, we have adopted the Show, Attend and Tell architecture by Xu et al. [49] which uses attention to generate the captions for images. The attention mechanism tries to learn the latent alignments between the objects in the image and output words of the caption or sentence from scratch. Thus, they learn to attend to the higher level dependencies between different entities present in the image. It is worthwhile to note that the use of an attention mechanism with neural networks is not entirely

new. In fact, in the computer vision community there exists some works, such as, [8, 27] which employed attention with neural networks to handle different vision tasks.

In general, the attention mechanism operates on a grid of image features obtained from a layer of a convolutional neural network, where each feature represents a high-level abstraction of a region in the image, and provides a weighting for each spatial region. There by, a higher weight would translate to more importance for the corresponding image region. However, often times it is difficult to find the optimal number of image regions which should capture all the relevant details in the image. Additionally, the high-level image features may fail to capture the finer object specific details or low-level salient regions in the image. So in our approach, we try to augment the attention mechanism by combining low-level fine details in the image with the high-level image abstractions. Previously, only a couple of works [19, 36] have tried to use salient image regions. The work in [19] utilizes a search technique proposed in [45] to identify salient image regions which are subsequently used in image captioning. Pedersoli et al. [36], on the other hand, uses spatial transformer networks [18] or edge-based boxes [50] for generating image features which are processed using a model based on three bilinear pairwise interactions. In our work, for the purpose of generating object specific localized maps or salient regions, we utilize the Mask R-CNN [13], a close variant of the Faster R-CNN [37] technique. We extract the image regions inside the bounding boxes and embed them into the feature space learned by a pre-trained deep convolutional network.

Image caption generation, in addition to being an important task in computer vision, is also a major problem in the area of natural language generation which requires proper evaluation. The common criteria here include readability or fluency, which refer to the linguistic quality of the text, and also accuracy or relevance relative to the input which shows the natural language generating system's ability to satisfactorily reproduce content. In our evaluation we use standard metrics, such as Bleu [34], ROUGE [30], METEOR [2], CIDEr [46] and SPICE [1] which try to emulate human judgement.

## 3 Implementation

In this section, we describe our implementation of the neural encoder-decoder architecture for generating image captions based on [49] and our extension of the visual attention mechanism: we use bounding boxes from Mask R-CNN [13] to encode object specific bottom-up features which complement the currently used top-down representation. Further, we describe our beam search decoder and two heuristic approaches for re-ranking its generated caption candidates.

### 3.1 Image Caption Generation

For generating the image captions, we use an own implementation of the Show, Attend and Tell method [49] as depicted in Fig. 3, with several modifications for extensions. Xu et al. [49] suggested to use a set of fixed dimensional vectors from a lower convolution layer of the CNN (*Convolutional Neural Network*) architecture instead of using a single fixed dimensional vector to represent the image. This helps to maintain a fine grained correspondence between the different portions of a 2D image represented through the corresponding vectors. With this the decoder becomes more powerful as it can focus selectively on different parts of an image during the generation process by selecting a subset of the feature vectors. The detailed operations of the LSTM based decoder, used in [49] for generating the captions, are described through the following equations:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ g_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} T_{D+m+n,n} \begin{bmatrix} E^{y_{(t-1)}} \\ h_{t-1} \\ \hat{z}_t \end{bmatrix} \tag{1}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{2}$$

$$h_t = o_t \odot \tanh(c_t). \tag{3}$$

The variables  $i_t, f_t, c_t, o_t, h_t$  denote input, forget, memory, output gates and the hidden state respectively.  $T$  represents a mapping of the form  $f_{s,t} : \mathbb{R}^s \rightarrow \mathbb{R}^t$ . Thus,  $T_{D+m+n,n}$  is a mapping from  $\mathbb{R}^{(D+m+n)}$  to  $\mathbb{R}^n$ .  $\hat{z} \in \mathbb{R}^D$  denotes the context vector responsible for capturing the visual information related to a specific location in the input image.  $E$  denotes the embedding matrix and has the dimension  $m \times k$ . The dimension of the embedding vector is given by  $m$  while the dimension of the LSTM hidden state is denoted by  $n$ . Furthermore,  $\sigma$  and

$\odot$  represent the logistic sigmoid and element-wise multiplication respectively. The model implementation and training details are as follows:

- We use the MSCOCO dataset for training the model [31], adopting the data splits proposed in [21]: the training set contains 113,287 images with 5 corresponding captions, the validation and test sets contain 5000 images each with 5 groundtruth captions per image.
- We build our vocabulary by dropping word types with a frequency  $< 5$ , resulting in a vocabulary of  $size = 10,000$  words.
- The dimensions for the LSTM hidden state, image, word and attention embeddings are set to 512 for our model.
- We train our model under the cross entropy objective, using the ADAM [23] optimizer.
- We use the PyTorch [35] framework for our implementation and the RESNET-101 [15] neural architecture with 101 layers for extracting image features.
- We do not perform any pre-processing on the images by way of cropping or rescaling them. Instead we utilize the final convolution layer of Resnet for encoding the full image.
- We apply spatially adaptive max-pooling which results in a fixed size output of  $14 \times 14 \times 2048$ .

### 3.2 Augmented Attention Mechanism

The visual attention mechanism used in image captioning models can be described as the expectation of an annotation function. In general, this expectation is computed over a set of image features and the previous history of the generation process. This form of attention works primarily with high-level abstractions captured by the convolution network which may or may not include specific objects and salient regions in the image. We propose a strategy to enrich the

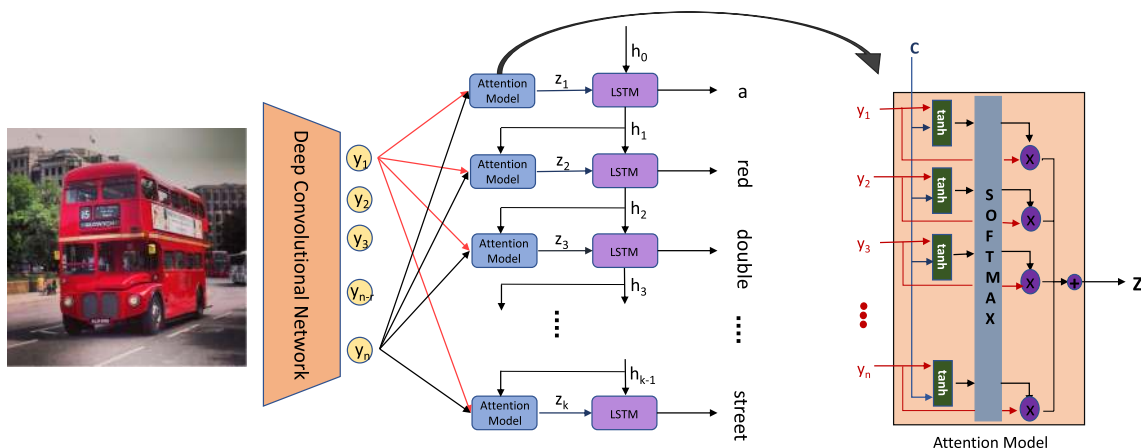


Fig. 3 Neural caption generation mechanism based on [49] depicting the processing of a red double decker bus

present attention mechanism by incorporating object-specific localized maps from the region proposal network Mask R-CNN [13] (see Fig. 4 for examples).

We represent an input image  $I$  as a set which includes a constant number of fixed size feature vectors:  $I = \{f_1, f_2, \dots, f_n\}$ ,  $f_i \in \mathbb{R}^d$ . Each feature vector represents the encoding of a bounding box detected by Mask R-CNN that is encoded using the Resnet-101 model. At every spatial location of an image, Mask R-CNN predicts an objectiveness score accompanied by refinement of anchor boxes of varying scales and aspect ratios which result in tighter bounding boxes. These bounding boxes are further refined using non-maximum suppression:

1. We extract the image regions inside the final bounding boxes and embed them into the feature space learned by Resnet-101 pre-trained on the Imagenet dataset.
2. We re-train Resnet-101 on MSCOCO images and set a high threshold for the classification probability for the regions to be selected.
3. In contrast to the original architecture in [49], we compute visual attention on the joint embedding space formed by the union of high-level features obtained from the encoder of the caption generator and the low-level features obtained from the object specific salient regions of the input image, i.e., the embedded bounding boxes.

The augmented attention mechanism is shown in Fig. 2. For every image we use 10 additional feature vectors of dimension 2048 to represent the salient regions. Thus, at every time-step, our attention model produces a mask over 206 spatial locations. This mask is applied to a set of image features and then the result is spatially averaged to produce a 2048 dimensional representation of the attended portion of the image. Most hyper-parameters of the training procedure stay the same. The initial learning rate for this model is  $4 \times 10^{-4}$  which is annealed by a factor of 0.8 every three

epochs. Further, we use batch size of 32. We evaluate the model at each epoch on the development set.

### 3.3 Beam Search and Re-ranking

Beam search [32] as a decoding technique allows for the generation of a more diverse set of caption candidates. A previous investigation [4, 5] has shown that beam search is to be preferred over other techniques such as [28, 29] for generating diverse captions. We use a beam width of  $k = 20$  to generate caption candidates that can be re-ranked in an additional step. To estimate the potential improvement for our caption generation method through re-ranking, we compute the upper bound of Bleu scores using the scores of all generated candidates. Our long-term goal is to leverage the objects, or their respective embeddings, detected by Mask R-CNN for such a re-ranking. In this work, we implement and test two heuristic re-ranking methods that rely on the similarity between the generated captions and the corresponding object classes: we estimate the similarity using the Euclidean distance with (1) bag-of-words and (2) TF-IDF based text representations.

## 4 Evaluation

In this section, we evaluate the image caption generation process, the extended attention mechanism, and the beam search and re-ranking approach. We compare the performance of our approach with and without beam search to the scores reported in [49]. Further, we investigate the potential improvement that can be achieved by the re-ranking of caption candidates of our beam search decoder.

We compute a set of common metrics as dependent variables: Bleu, METEOR, ROUGE-L, CIDEr and SPICE, which primarily focus on the n-gram overlap between the



Fig. 4 Object-specific salient regions highlighted with corresponding bounding boxes as bottom-up features

generated and ground truth captions. To be more specific, we provide short descriptions for each metric.

Bleu is an automatic metric for evaluating the quality of a machine generated text. Bleu scores are computed from individual machine generated sentences by direct comparison between them and with a set of good quality references or ground truth references. It is always between 0 and 1 and indicates the similarity between the generated captions and the ground truth. So, a score of 0 indicates no overlap whereas 1 indicates complete overlap. Depending on the size of the n-grams we want to match between the candidate caption and the ground truth captions we have different BLEU scores, i.e., Bleu-1, Bleu-2, Bleu-3, Bleu-4.

METEOR is a metric for evaluating outputs from a machine translation system. The metric is based on the harmonic mean of unigram precision and recall, where, recall is weighted higher than precision. METEOR uses features, such as, stemming and synonymy matching along with the standard exact word matching. ROUGE-L measures the longest matching sequence of words. An advantage of it is that it does not require consecutive matches but in-sequence matches that reflects sentence level order. One does not require a predefined n-gram length since it automatically includes longest in-sequence common n-grams. CIDEr denotes Consensus based Image Description Evaluation. It measures the similarity of a generated sentence against a set of ground truth sentences composed by humans and shows high agreement with consensus as assessed by humans. While SPICE stands for semantic propositional image caption evaluation.

The upper bounds for our architecture are estimated by sorting the generated caption candidates from the beam search by their Bleu-1 to Bleu-4 scores, i.e., assuming we had access to a perfect re-ranking. For all tests, we use the MSCOCO dataset [31] using the data splits as described above.

We hypothesize that our approach improves the caption generation process and, hence, outperforms the scores reported in Xu et al. [49].

Another hypothesis is that re-ranking of beam search candidates has a high potential for improving image captions and that our heuristic approaches supports this conjecture.

Finally, we expect that our approach paves the way for novel IML and XAI methods that can be used to further

improve the image captioning results. We qualitatively discuss this topic based on the results of this experiment.

## 4.1 Results

Table 1 shows the scores of the three approaches we evaluated. Our approach without beam search obtains higher scores Bleu-2, Bleu-3, Bleu-4, measuring bi-gram, tri-gram and tetra-gram overlaps than the baseline approach [49]. This is a significant improvement because the Bleu metric computation does not remove stop words and so higher scores should lead to more natural and pertinent generations. The Bleu-1 and METEOR scores are on par with the baseline approach. Additionally, we obtain high scores in ROUGE-L and CIDEr; unfortunately we cannot compare these results with the those from the baseline approach [49] since the original baseline does not report on these metrics.

Naturally, the results for the top-1 captions (beam search approach) are worse than the scores for the version without beam search. In particular, the scores for Bleu-3 and Bleu-4 are significantly worse (see Sect. 5 for an extensive discussion about this).

In addition to these quantitative results, we visually inspect generated captions from our approach without beam search and our baseline implementation based on [49] (see Fig. 5). A third caption is shown, which was selected from the 20 beam search candidates which have a zero Bleu-4 score (beam candidate). Obviously, the Beam metric, which the international leader board uses, does not work properly on their own gold standard. To summarize, all generated captions distinctly describe the objects and their inter relationships in the corresponding images in natural language text. Our approach correctly aligns the image concepts, i.e., the objects with the output words in the generated captions. A more detailed qualitative analysis of generated captions can be found in the discussion Sect. 5. Further examples from the beam search decoder are shown in the Appendix Table 3, where we show all 20 generations for randomly selected images from the test set, along with the 5 ground truth captions.

The upper bounds for Bleu-1 to Bleu-4 metrics and the results from our heuristic re-ranking methods are reported in Table 2. The upper bounds are reported for top- $i$  candidates from our beam search decoder with  $i \in 3, 5, 10, 20$ . The re-ranking methods perform slightly better compared

**Table 1** Different metrics computed for the captions generated on the MSCOCO test set

Classifier	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L	CIDEr	SPICE
Xu et al. [49]	<b>0.707</b>	0.492	0.344	0.243	<b>0.239</b>	–	–	–
our approach (no beam search)	0.706	<b>0.528</b>	<b>0.382</b>	<b>0.275</b>	0.238	0.515	0.872	0.167
our approach (beam search top-1)	0.685	0.493	0.307	0.17	–	–	–	–

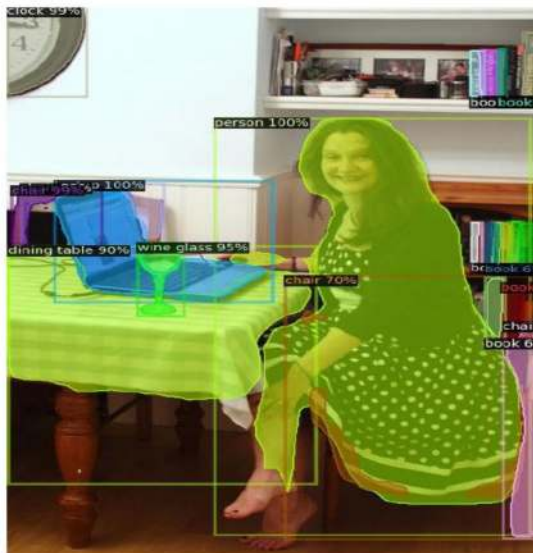
The values in bold represent higher scores



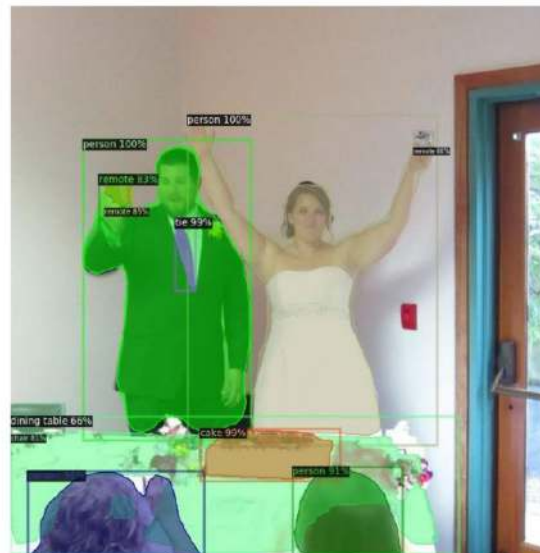
(a) **our approach:** a red double decker bus driving down a street; **Xu et al.:** a double deckered bus on a city street; **beam candidate:** a double decker bus drives down the street



(b) **our approach:** a man sitting at a table in a restaurant; **Xu et al.:** a man is holding a glass of wine; **beam candidate:** a man is sitting at a table with a glass of wine



(c) **our approach:** a woman sitting in front of a laptop computer; **Xu et al.:** a woman sitting in front of a laptop; **beam candidate:** a woman sitting on a chair using a laptop computer



(d) **our approach:** a newly married couple cutting their wedding cake; **Xu et al.:** a couple of people standing in a room; **beam candidate:** a man standing next to a woman in a white dress

**Fig. 5** Generated image captions with our approach (without beam search), the Show, Attend and Tell method and one of the beam candidates with zero Bleu-4 score

to the top-1 result, but with the same deterioration for Bleu-3 and Bleu-4 compared to our approach without beam search. We observe much better Bleu scores for the upper bounds compared to any other method, in particular for  $i = 20$ : a plus of 0.153 for Bleu-1, 0.205 for Bleu-2, 0.225 for Bleu-3, and 0.199 for Bleu-4 when compared to Table 1.

## 5 Discussion

### 5.1 Qualitative Analysis

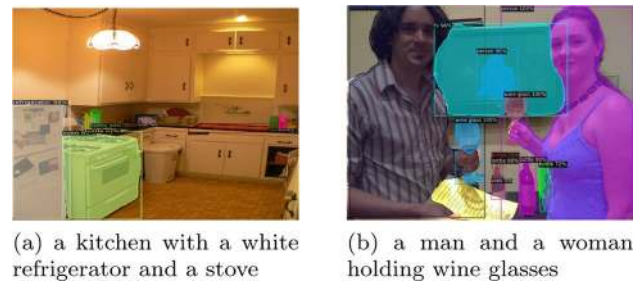
Our approach without beam search outperforms the state-of-the-art method from [49] for several metrics and

**Table 2** Bleu scores for the two re-ranking approaches based on bag-of-words (BoW) and TF-IDF and the estimated upper bounds for Bleu scores, i.e., top- $i$  are ranked according to Bleu-1 to Bleu-4 scores

$i$	Metric	Re-ranking		Upper bounds			
		BoW	TF-IDF	Bleu-1	Bleu-2	Bleu-3	Bleu-4
20	Bleu-1	0.701	0.694	<b>0.860</b>	0.845	0.830	0.823
	Bleu-2	0.509	0.502	0.691	<b>0.733</b>	0.725	0.716
	Bleu-3	0.321	0.315	0.513	0.587	<b>0.607</b>	0.603
	Bleu-4	0.177	0.173	0.349	0.427	0.462	<b>0.474</b>
10	Bleu-1	0.701	0.695	0.839	0.826	0.816	0.810
	Bleu-2	0.509	0.501	0.666	0.704	0.698	0.692
	Bleu-3	0.319	0.313	0.482	0.548	0.565	0.562
	Bleu-4	0.177	0.172	0.318	0.386	0.412	0.422
5	Bleu-1	0.697	0.691	0.801	0.792	0.784	0.781
	Bleu-2	0.506	0.498	0.626	0.654	0.649	0.646
	Bleu-3	0.317	0.310	0.438	0.485	0.500	0.498
	Bleu-4	0.174	0.169	0.277	0.323	0.341	0.347
3	Bleu-1	0.695	0.690	0.768	0.762	0.756	0.755
	Bleu-2	0.504	0.498	0.591	0.611	0.608	0.606
	Bleu-3	0.316	0.312	0.402	0.434	0.445	0.444
	Bleu-4	0.175	0.173	0.246	0.274	0.287	0.291

The values in bold represent higher scores

achieves similar scores for the remaining metrics (see Table 1). In particular, we achieve better Bleu scores with long  $n$ -grams, i.e., 3-grams and 4-grams, showing a better alignment to formulations in the ground truth captions. This latent alignment is important because neural caption generation is often regarded as translating an image into a natural language description. Together with the qualitative analysis (visual inspection) of generations, this shows that our architecture can effectively produce meaningful image captions. Results, as shown in Table 1 and compared to [49], suggest that particular localized information in conjunction with the high-level features obtained from deep convolutional encoder improves the correspondence resolution problem (i.e., image and word entity alignment) at the heart of this multimodal task. The results underline the positive influence of bottom-up features (or object specific localized maps) for the image captioning task; also they deliver explainable features. We note that the application of Mask R-CNN [13] in obtaining the localized maps or salient regions in our work puts in an important step towards better exposition of the object features involved in the caption generation process compared to previous approaches including [49] which only uses high level image abstractions, i.e., the top-down features. The specific object masks in addition to the bounding boxes provide important explanatory cues for the generated text describing the corresponding image as is shown in Fig. 6 where bounding boxes help localize the white refrigerator and stove present in the generated caption. Similarly, for the other image in Fig. 6 the context for the generated caption is provided by bounding boxes localizing the



**Fig. 6** Explanatory cues provided by localized object masks for one of the generated beam candidate captions

man, woman and wine glasses. We believe the proposed approach is a good step in the direction of infusing image caption generation with explainable AI.

In addition, we use our beam search decoder for generating a more diverse set of caption candidates. Averaging over all test images, we computed the upper bounds for all Bleu metrics. For  $i = 20$ , i.e., including all caption candidates from the beam search with beam width  $k = 20$ , all Bleu scores potentially outperform the state-of-the-art method and our method without beam search by 0.196 (average over all Bleu metrics). We find that the scores increase with higher values of  $k$  which is probably caused by a higher recall due to more captions from which the Bleu score can be selected. Of course, this gain in Bleu scores motivates an optimal method for selecting from the 20 candidates and indicates a high potential of re-ranking methods. The results from our evaluation show that simple heuristics-based re-ranking methods do not improve the caption selection process considerably. This leaves the challenge to future research, i.e.,



to develop re-ranking methods that optimize the caption selection over the top- $k$  results and, hence, approximate the upper bound of the caption generation pipeline. Examples include interactive training of a re-ranking model using the bottom-up features from Mask R-CNN as additional input (see Sect. 5.2.1).

Without a selection or re-ranking, the top-1 candidate from beam search yields worse results compared to all other methods. This phenomenon is well understood, beam search expands the search space for the natural language generation which does not guarantee that the first generation has the best overlap with the ground truth captions. This, however, is measured by the Bleu metric and can affect the corresponding scores. The examples in Figs. 5 and 7 show semantically meaningful and fluent generations originating from the beam search with zero Bleu-4 score, which further demonstrates some shortcoming of the Bleu metric in this regards. More examples can be seen in the Appendix in Table 3. Compared to greedy decoding which is locally optimal, candidates obtained from beam search may contain different words compared to corresponding ground truths which can dramatically harm Bleu scores since they measure only the overlap with the ground truths without taking into account the semantic meaning of the generations.

However, beam search alone and re-ranking with two heuristic methods (based on bag-of-words and tf-idf) turn out to fail in improving the overall scores. But we believe our approach with its use of Mask R-CNN producing bottom-up features provides new opportunity for making image-captioning an IML task apart from only generating captions which achieve higher metric scores.

## 5.2 Towards Interactive and Explanatory Captioning

Based on our findings, we believe that our image captioning system with its augmented attention mechanism and the beam search decoder has the potential to facilitate interactive improvement of the captioning system and to improve the explainability of the caption generation process. In the



(a) a market with a variety of fruits and vegetables



(b) a view of a city street with traffic lights

**Fig. 7** Correctly generated one of the beam candidate captions with zero Bleu-4 score

following, we describe opportunities and challenges for future research in this direction.

### 5.2.1 Interactive Re-ranking

Effective re-ranking can leverage the inherent potential of the beam search decoder to improve generated image captions. Besides the output of diverse image captions, our architecture yields several opportunities for user interaction such as interactive training of a caption re-ranking model: The additional bottom-up features can be used as input to a re-ranking that learns from continuous user feedback to score the 20 generated caption candidates. Corrective feedback to the model can be realized by selecting relevant areas of the image, that are important for generating the caption, based on the Mask R-CNN bounding boxes. This enables users to easily change the focus for the generation process, e.g., if the model wrongly puts emphasis on an irrelevant object. The challenge lies in the development of interactive machine learning (IML) mechanisms that facilitate efficient and effective model training, i.e., that model training requires low annotation effort, is scalable and, yet, converges to a model that improves image captioning. Active learning can be used to reduce the annotation effort for the humans involved in that process or, due to a better selection of training samples from an unlabelled pool, improving the overall quality of the model [41]. Including active learning techniques was shown to be effective for different natural language processing tasks, e.g., for reducing the number of training samples in machine translation without a loss in quality [20] and for training quality assessment models for image captioning [5]. The latter model for caption quality assessment can also be used as a baseline for a future re-ranking system. Crowdsourcing can scale up the annotation process as shown for, e.g., dialogue systems [3, 39, 48] and in the context of image captioning [4, 5]. Promising techniques for improving caption generation can also be found in coactive learning [42].

### 5.2.2 XAI Methods for Image Caption Generation

We discuss different future extensions of our work pertaining to domain of XAI methods, particularly through deep explanations. The field of deep explanations subsumes methods that introduce more transparency in how black box models, in particular neural network models, work. A prominent approach is to generate visual explanations which describe how the objective is achieved by the neural model. Our approach provides a relation between image captions, bounding boxes, and pixel-wise segmentations from Mask R-CNN that localize regions that are important to the generation process. An interesting direction of future work is

to develop segmentation-based visual explanation methods and to compare them with state-of-the-art approaches like Grad-Cam [40].

The specific object masks in addition to the bounding boxes provide important explanatory cues for the generated text describing the corresponding image as is shown in Fig. 6 where bounding boxes help to localize the white refrigerator and stove present in the generated caption. Similarly, for the other image in Fig. 6 context for the generated caption is provided by bounding boxes localizing the man, woman and wine glasses. This can also be used as an extension to interactive re-ranking, e.g., as a part of explanatory interactive machine learning interfaces [44].

## 6 Conclusion

In this work, we presented a new architecture for image captioning that incorporates a top-down attention mechanism with bottom-up features of a scene: we encoded the object specific bounding boxes provided by the Mask R-CNN model [13] using the Resnet-101 architecture [14]. We show that our approach achieves scores on par with the state-of-the-art, Show, Attend and Tell [49], for the Bleu-1 and METEOR metrics, and better scores for the Bleu-2, Bleu-3 and Bleu-4 metrics using the MSCOCO dataset, while at the same time, providing explanatory features. In addition,

we showed that using our beam search decoder has great potential for further improvements of the image captioning process. We discussed opportunities in interactive machine learning for leveraging this potential, in particular by interactively training re-ranking models that effectively select the best options from the generated caption candidates. Further, we discussed how XAI method can be developed based on our image captioning system to better understand the image captioning process, which in turn delivers valuable feedback to users of such intelligent user interfaces for incremental model improvements.

**Acknowledgements** Open Access funding provided by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## Appendix

See Table 3.

**Table 3** Image caption candidates from our the beam search decoder




Image	Generated	Bleu-4	Bleu-3		
 <p>people skiing and standing on a ski slope multiple skiing down hill skiing in the snow a group of skiers skiing down a mountain on a sunny day a skier is skiing down a bunny hill slope people skiing down a mountain made of artificial snow</p>	a person riding skis on a snowy surface a person riding skis on a snowy slope a man skiing down a snow covered slope a person skiing down a snow covered slope a man is skiing down a snowy hill a man riding skis down a snow covered slope a person riding skis down a snow covered slope a couple of people on skis in the snow a group of people on skis in the snow a group of people skiing down a snowy slope a group of people skiing down a snow covered slope a man riding skis down a snow covered ski slope a couple of people skiing down a snow covered slope a couple of people that are skiing in the snow a group of people riding skis on a snowy surface a group of people riding skis on a snowy slope a group of people skiing down a snow covered hill a man riding skis down a snow covered slope a group of people riding skis down a snow covered slope a couple of people riding skis down a snow covered slope	0.000 0.000 0.000 0.000 0.382 0.000 0.000 0.000 0.000 0.000 0.446 0.404 0.000 0.000 0.312 0.302 0.000 0.000 0.404 0.000 0.000 0.000	0.629 0.000 0.329 0.329 0.474 0.000 0.000 0.302 0.502 0.619 0.572 0.000 0.000 0.387 0.292 0.307 0.572 0.000 0.264 0.000		
	a bathroom with a sink and a mirror a bathroom with two sinks and a mirror a bathroom with a sink sink and mirror a bathroom with a sink toilet and mirror a bathroom with a toilet sink and mirror a bathroom with a sink mirror and mirror a bathroom with a sink mirror and shower a bath room with a sink and a mirror a bathroom with two sinks and a large mirror a bathroom with a sink and a mirror in it a bathroom with a large mirror and a sink a bathroom with a white sink and a mirror a bathroom that has a sink and a mirror a bathroom with a sink mirror and a mirror a bathroom with a sink a mirror and a mirror a bathroom with a sink and a mirror in it a bathroom with two sinks a mirror and a mirror a bathroom with a sink a mirror and a window a bathroom with a sink a mirror and a shower a bathroom with a sink a mirror and a sink	0.740 0.000 0.594 0.562 0.000 0.562 0.562 0.354 0.000 0.510 0.368 0.422 0.000 0.513 0.451 0.526 0.562 0.000 0.467 0.451 0.451	0.793 0.376 0.678 0.529 0.550 0.629 0.629 0.456 0.287 0.652 0.480 0.575 0.381 0.592 0.526 0.615 0.292 0.550 0.526 0.526		
	 <p>a variety of fruit is displayed in a market fruits and vegetables for sale in a farmers market many fruits displayed in a produce store including bananas, oranges, apples and lemons the produce section of a grocery store showing many various fruits a market has many fruits and vegetables out for display</p>	a store filled with lots of fresh produce a market filled with lots of fresh produce a grocery store filled with lots of produce a market with lots of fruits and vegetables a grocery store filled with lots of fresh produce a farmers market filled with lots of fresh produce a market filled with lots of fruits and vegetables a store filled with lots of fruits and vegetables a grocery store filled with lots of fresh fruit a market filled with lots of fruit and vegetables a market with a variety of fruits and vegetables a grocery store filled with lots of fresh fruits a grocery store filled with lots of different fruits a farmers market filled with lots of fresh fruits a grocery store filled with lots of fruits and vegetables a store filled with lots of fresh fruits and vegetables a grocery store filled with lots of fruit and vegetables a store filled with lots of ripe and unripe bananas a store filled with lots of different types of fruits a store filled with lots of different types of fruit	0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000	0.000 0.000 0.273 0.332 0.270 0.270 0.329 0.287 0.270 0.000 0.541 0.270 0.270 0.270 0.426 0.255 0.338 0.000 0.000 0.000	
		a man and a woman holding wine glasses a couple of people that are drinking wine a man and woman pose for a picture a couple of people that are standing together a group of people standing around a table a man and a woman are drinking wine a couple of people standing next to each other a couple of women standing next to each other a group of people standing next to each other a man and woman standing next to each other a couple of people that are holding wine glasses a man and a woman posing for a picture a man and a woman are holding wine glasses a man and a woman pose for a picture a man and woman pose for a picture together a man and a woman posing for a photo a man and a woman standing next to each other a group of people sitting at a table with wine glasses a group of people standing around a table with wine glasses a couple of women standing next to each other holding glasses a woman is pouring a glass of wine a woman is holding a glass of wine a woman is pouring wine into a glass a woman sitting at a table with wine glasses a man sitting at a table with wine glasses a woman sitting at a table with a wine glass a man sitting at a table with a wine glass a woman sitting at a table with glasses of wine a man sitting at a table with glasses of wine a woman pouring a glass of wine at a table a woman is sitting at a table with wine glasses a woman sitting at a table with a glass of wine a man sitting at a table with a glass of wine a woman sitting at a table with a bottle of wine a man sitting at a table with a bottle of wine a woman is sitting at a table with a wine glass a man is sitting at a table with a wine glass a woman is sitting at a table with a glass of wine a woman sitting at a table in front of a wine glass a man is sitting at a table with a glass of wine	0.634 0.000 0.000 0.000 0.000 0.477 0.000 0.000 0.000 0.000 0.000 0.467 0.596 0.467 0.390 0.467 0.525 0.000 0.000 0.000 0.320 0.320 0.000 0.718 0.671 0.503 0.467 0.516 0.481 0.411 0.707 0.722 0.683 0.467 0.436 0.449 0.417 0.652 0.359 0.617	0.770 0.000 0.394 0.000 0.000 0.528 0.287 0.270 0.287 0.592 0.287 0.724 0.522 0.390 0.522 0.643 0.000 0.000 0.000 0.407 0.000 0.000 0.698 0.608 0.550 0.629 0.572 0.584 0.759 0.817 0.758 0.575 0.525 0.546 0.494 0.742 0.531 0.689	
		 <p>the dinner plate has asparagus carrots and some kind of meat a plate containing a piece of meat, baby carrots and asparagus a plate at a table with a piece of meat, carrots and asparagus some food is laying on a white plate</p>	a plate of food on a table a plate of food on a wooden table a plate of food sitting on a table a plate of food is on a table a plate of food sits on a table a plate of food that is on a table a couple of plates of food on a table a plate of food is sitting on a table a plate of food and a glass of wine a plate of food that is sitting on a table a close up of a plate of food with broccoli a close up of a plate of food with food a close up of a plate of food with vegetables a close up of a plate of food with a a close up of a plate of food with meat a plate of food on a table with a fork a close up of a plate of food on a table a close up of a plate of food with a fork a close up of a plate of food with a sandwich a close up of a plate of food with a table	0.000 0.000	0.000 0.000

Table 3 (continued)

Image	Generated	Bleu-4	Bleu-3					
 <p>a bus stop next to a curvy road surrounded by traffic lights a street with some street lights and lamps empty city street with multiple street signs on it there is a glass awning over a sidewalk a car driving down a quiet road during a sunny day</p>	<p>a view of a city street at night a city street filled with lots of traffic a city street with lots of traffic lights a city street filled with lots of traffic lights a city street filled with traffic and traffic lights a traffic light on a city street at night a city street at night with a traffic light a view of a street with a traffic light a view of a city street with traffic lights a city street at night with lots of traffic a city street at night with lots of traffic lights a view of a city street with a traffic light a city street filled with traffic lights and traffic lights a city street filled with lots of traffic and lights a city street filled with lots of traffic and traffic a city street filled with lots of traffic and traffic lights a city street with lots of traffic lights and traffic lights a view of a city street at night with traffic lights a city street filled with lots of traffic lights and buildings a city street filled with lots of traffic and traffic signals</p>	<p>0.000 0.000</p>	<p>0.000 0.000 0.376 0.000 0.000 0.000 0.000 0.287 0.346 0.000 0.000 0.268 0.000 0.000 0.000 0.304 0.000 0.000 0.000 0.000</p>					
	 <p>a picture of a very nice kitchen that is white a small residential kitchen with white cabinets and appliances a kitchen filled with a white stove top oven a kitchen scene complete with a fridge and an oven a dated kitchen reveals a wall hung faucet under a closed off window</p>	<p>a kitchen with white cabinets and white appliances a kitchen with a stove and a refrigerator a kitchen with a refrigerator and a sink a kitchen with a stove sink and refrigerator a kitchen with a sink and a refrigerator a kitchen with a stove and a sink a kitchen with a refrigerator and a stove a kitchen with a refrigerator and a microwave a kitchen with white cabinets and a white refrigerator a kitchen with a white refrigerator and a sink a kitchen that has a refrigerator and a stove a kitchen that has a refrigerator and a sink a kitchen with a white refrigerator and a stove a kitchen with a refrigerator and a stove top a kitchen with a refrigerator and a stove top oven a kitchen with a refrigerator a sink and a refrigerator a kitchen with a refrigerator a sink and a stove a kitchen with a refrigerator a sink and a microwave a kitchen with a refrigerator a sink and a window a kitchen with a stove top oven and a refrigerator</p>	<p>0.524 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.537 0.381 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000</p>	<p>0.598 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.629 0.381 0.000 0.000 0.398 0.000 0.396 0.000 0.000 0.000 0.000 0.396</p>				
		 <p>a man taking a swing at a ball on the court a tennis game is being played on a brown court a woman on a tennis court swinging her racquet a man on a tennis court hitting the ball a guy playing tennis while being photographed</p>	<p>a man playing tennis on a tennis court a woman playing tennis on a tennis court a man is playing tennis on a court a man standing on top of a tennis court a group of people playing a game of tennis a man that is standing on a tennis court a man on a court with a tennis racket a tennis player getting ready to hit a ball a man hitting a tennis ball with a racquet a man swinging a tennis racquet at a ball a man standing on a tennis court holding a racquet a man swinging a tennis racquet on a tennis court a man hitting a tennis ball on a tennis court a man swinging a tennis racket on a tennis court a group of people playing tennis on a tennis court a man swinging a tennis racquet at a tennis ball a man hitting a tennis ball with a tennis racquet a man standing on a tennis court with a racquet a man on a tennis court with a tennis racket a man standing on a tennis court holding a tennis racquet</p>	<p>0.451 0.451 0.000 0.000 0.000 0.368 0.368 0.000 0.000 0.000 0.335 0.345 0.365 0.335 0.312 0.000 0.335 0.000 0.537 0.299</p>	<p>0.592 0.000 0.329 0.000 0.000 0.480 0.480 0.000 0.000 0.414 0.446 0.464 0.499 0.446 0.405 0.000 0.446 0.000 0.579 0.401</p>			
			 <p>a man standing holding a game controller and two people sitting young adult using video game controller with other looking on people sitting around watching a man playing a video game the group of friends is enjoying playing the video games a man standing on a piece of equipment for a video game while holding the remote as four other people look on</p>	<p>a group of people playing a video game a man and woman playing a video game a man and a woman playing a video game a group of people are playing a video game a man and woman are playing a video game a man and a woman are playing video games a group of people standing around a living room a group of people sitting in a living room a group of people sitting around a living room a couple of people that are playing a video game a man and a woman are playing a video game two people sitting on a couch playing a video game a group of people that are playing a video game a group of people in a living room playing wii a group of people playing video games in a living room a group of people sitting on a couch playing video games a group of people playing a game with nintendo wii controllers a man standing on a piece of equipment for a video game while holding the remote as four other people look on a group of people in a living room playing wii a group of people in a living room playing a game</p>	<p>0.389 0.377 0.381 0.381 0.368 0.000 0.000 0.000 0.000 0.302 0.335 0.423 0.335 0.000 0.000 0.000 0.000 0.000 0.000 0.000</p>	<p>0.507 0.485 0.502 0.502 0.480 0.000 0.000 0.000 0.346 0.387 0.446 0.608 0.446 0.000 0.000 0.000 0.000 0.000 0.000 0.000</p>		
				 <p>a clock on a pole next to a city street an outdoor clock across the street from a building a clock on a sidewalk with cars passing by on a street a single pillar with a clock is near approaching traffic, including a police car a clock stands on the sidewalk while a police car turns the intersection in front of an old looking building</p>	<p>a clock on the side of a street a clock on a pole on a city street a clock tower in the middle of a street a clock on the side of a city street a clock tower on the side of a street a clock is on the side of a street a black and white photo of a clock tower a clock tower in the middle of a city a clock tower on the side of a road a black and white photo of a city street a clock is on the side of the road a clock is on the side of a road a clock that is on the side of a building a clock tower in the middle of a city street a clock that is on the side of a street a clock that is on the side of a road there is a clock on the side of the road a clock that is on the side of the road there is a clock on the side of the street a clock on the side of a street next to buildings</p>	<p>0.000 0.638 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000</p>	<p>0.793 0.000 0.541 0.000 0.398 0.000 0.000 0.000 0.000 0.270 0.346 0.346 0.000 0.000 0.321 0.000 0.000 0.307 0.000 0.354 0.356</p>	
					 <p>a tennis player winds up to hit a serve to her opponent this is an aerial view of a tennis match a scene of something that is quite attractive a woman that is standing on a blue tennis court a tennis match is being played on a blue court</p>	<p>a crowd of people watching a tennis match a group of people watching a tennis match a crowd of people watching a tennis match a group of people on a tennis court a group of people playing a game of tennis a crowd of people playing a game of tennis a group of people on a court playing tennis a crowd of people watching a game of tennis a group of people playing tennis on a court a group of people standing on a tennis court a tennis match with a crowd of people watching a man on a tennis court with a racket a group of people playing tennis on a tennis court a group of people are playing a game of tennis a group of people on a tennis court playing tennis a crowd of people on a tennis court playing tennis a group of people are playing tennis on a court a man on a tennis court with a tennis racket a group of people on a tennis court playing a a group of people on a tennis court playing a game</p>	<p>0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000</p>	<p>0.309 0.309 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.381 0.270 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000</p>

Table 3 (continued)

Image	Generated	Bleu-4	Bleu-3	
 <p>there is a small bus with several people standing next to it people standing besides a bus taking to each other people eating from food trucks near a commemorative archway several people walking on a sidewalk near a large arch with figures on it a very big pretty arch way with a bunch of people near it</p>	a group of people standing around a bus a group of people standing next to a bus a group of people standing outside of a bus a group of people standing on a city street a group of people walking on a city street a group of people standing in a parking lot a group of people standing outside of a truck a group of people standing in front of a bus a group of people standing in front of a truck a couple of people that are standing in the street a group of people that are standing in the street a group of people standing on the side of a road a group of people standing on the side of a street a group of people standing in front of a food truck a group of people standing on the side of the road a group of people standing in front of a large truck a group of people standing on the side of the street a group of people standing in front of a white truck a group of people standing in front of a red bus a group of people standing in front of a red truck	0.000 0.403 0.000 0.000 0.354 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000	0.000 0.541 0.000 0.000 0.456 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000	
	a group of children playing soccer on a field a group of kids playing soccer on a field a group of children playing a game of soccer a group of kids playing a game of soccer a group of kids playing soccer in a field a group of people playing soccer on a field a group of children playing soccer in a field a group of boys playing soccer on a field a group of young children playing a game of soccer a group of young men playing a game of soccer a group of young boys playing a game of soccer a group of young people playing a game of soccer a group of young children playing soccer on a field a group of young girls playing a game of soccer a group of young boys playing soccer on a field a group of young men kicking around a soccer ball a group of young people playing soccer on a field a group of children are playing soccer on a field a group of kids are playing soccer on a field a group of young men playing soccer on a field	0.614 0.747 0.000 0.000 0.000 0.596 0.000 0.999 0.000 0.000 0.000 0.000 0.577 0.000 0.675 0.000 0.541 0.541 0.541 0.541	0.733 0.854 0.302 0.346 0.563 0.724 0.414 0.999 0.321 0.268 0.321 0.268 0.729 0.268 0.786 0.268 0.669 0.669 0.669 0.669	
	 <p>kids playing a game of baseball while people watch parents watching young boys playing baseball in the sun a young boy is at home plate in a baseball game a baseball game taking place at a baseball field two young baseball players and one adult</p>	a boy swinging a bat at a ball a boy swinging a bat at a baseball a man swinging a bat at a baseball a man swinging a bat at a ball a baseball player swinging a bat at a ball a young boy swinging a bat at a ball a young boy swinging a bat at a baseball a man swinging a baseball bat at a ball a boy swinging a baseball bat at a ball a baseball player swinging a bat on a field a baseball player swinging a bat during a game a man swinging a baseball bat at a baseball a baseball player swinging a bat at a baseball a boy swinging a baseball bat at a baseball a young boy swinging a baseball bat at a ball a group of young men playing a game of baseball a young boy swinging a baseball bat at a baseball a young boy swinging a baseball bat on a field a young boy holding a baseball bat on a field a young boy swinging a baseball bat during a game	0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.427 0.000 0.000 0.000 0.000	0.000 0.309 0.287 0.000 0.000 0.309 0.456 0.000 0.000 0.000 0.000 0.309 0.309 0.329 0.321 0.488 0.459 0.292 0.292 0.321

References

- Anderson P, Fernando B, Johnson M, Gould S (2016) Spice: semantic propositional image caption evaluation. In: ECCV
- Banerjee S, Lavie A (2005) 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, pp 65–72
- Barz M, Sonntag D (2019) Incremental improvement of a question answering system by re-ranking answer candidates using machine learning. CoRR. arXiv:1908.10149
- Biswas R (2019) Diverse image caption generation and automated human judgement through active learning. Master’s thesis, Saarland University
- Biswas R, Mogadala A, Barz M, Sonntag D, Klakow D (2019) Automatic judgement of neural network-generated image captions. In: Statistical language and speech processing—7th international conference, SLSP 2019, Ljubljana, Slovenia, October 14–16, 2019, Proceedings, pp 261–272
- Cho K, Merriënboer B, Gulcehre C, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. EMNLP
- Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. CVPR
- Denil M, Bazzani L, Larochelle H, Freitas N (2012) Learning where to attend with deep architectures for image tracking. Neural Comput
- Donahue J, Hendriks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. CVPR
- Dudley JJ, Kristensson PO (2018) A review of user interface design for interactive machine learning. ACM Trans Interact Intell Syst 8(2):1–37. <https://doi.org/10.1145/3185517>. <http://dl.acm.org/citation.cfm?doid=3232718.3185517>
- Gunning D, Aha D (2019) DARPA’s explainable artificial intelligence (XAI) program. AI Mag 40(2):44–58. <https://doi.org/10.1609/aimag.v40i2.2850>
- Harnad S (1990) The symbol grounding problem. Physica 335–346
- He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: IEEE international conference on computer vision (ICCV), pp 2980–2988
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, pp 770–778
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. CVPR
- Hochreiter S, Schmidhuber J (1997) Long short term memory. Neural Comput 1735–1780
- Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167
- Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. Adv Neural Inf Process Syst 28:2017–2025
- Jin J, Fu K, Cui R, Sha F, Zhang C (2015) Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv:1506.06272

20. Kalimuthu M, Barz M, Sonntag D (2019) Incremental domain adaptation for neural machine translation in low-resource settings. In: Proceedings of the fourth arabic natural language processing workshop, pp 1–10. Association for Computational Linguistics (ACL). <https://doi.org/10.18653/v1/w19-4601>. <https://www.aclweb.org/anthology/W19-4601>
21. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. CVPR
22. Karpathy A, Joulin A, Fei-Fei L (2014) Deep fragment embeddings for bidirectional image sentence mapping. NIPS
23. Kingma D, Ba J (2015) Adam: a method for stochastic optimization. ICLR
24. Kiros R, Salakhutdinov R, Zemel R (2014) Multimodal neural language models. ICLR, pp 595–603
25. Kiros R, Salakhutdinov R, Zemel R (2014) Unifying visual-semantic embeddings with multimodal neural language models. [arXiv:1411.2539](https://arxiv.org/abs/1411.2539)
26. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. NIPS
27. Larochelle H, Hinton G (2010) Learning to combine foveal glimpses with a third-order Boltzmann machine. NIPS
28. Li J, Galley M, Brockett C, Gao J, Dolan B (2015) A diversity-promoting objective function for neural conversation models. [arXiv:1506.01057](https://arxiv.org/abs/1506.01057)
29. Li J, Monroe W, Jurafsky D (2016) A simple, fast diverse decoding algorithm for neural generation. [arXiv:1611.08562](https://arxiv.org/abs/1611.08562)
30. Lin C (2004) Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out
31. Lin T, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C (2014) Microsoft coco: common objects in context. ECCV, pp 740–755
32. Lowerre B, Reddy R (1990) The harpy speech understanding system. In: Readings in speech recognition, pp 576–586
33. Oviatt S, Schuller B, Cohen P, Sonntag D, Potamianos G (2017) The handbook of multimodal-multisensor interfaces In: Foundations, user modeling, and common modality combinations, vol 1. ACM, New York
34. Papineni K, Roukos S, Ward T, Zhu W (2002) Bleu: a method for automatic evaluation of machine translation. In: Association for computational linguistics, pp 311–318
35. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32, pp 8026–8037. Curran Associates, Inc., New York. <https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
36. Pedersoli M, Lucas T, Schmid C, Verbeek J (2017) Areas of attention for image captioning. In: 2017 IEEE international conference on computer vision (ICCV), pp 1251–1259
37. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28:91–99
38. Roy D, Reiter E (2005) Connecting language to the world. Artif Intell 1–12
39. Savenkov D, Agichtein E (2016) CRQA: crowd-powered real-time automatic question answering system. In: Fourth AAAI conference on human computation and crowdsourcing
40. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: International conference on computer vision, ICCV 2017, Venice, Italy, October 22–29, 2017, pp 618–626
41. Settles B (2010) Active learning literature survey. University of Wisconsin, Madison, vol 52, no 55–66, p 11
42. Shivaswamy P, Joachims T (2015) Coactive learning. J Artif Intell Res 53:1–40
43. Simard P, Amershi S, Chickering M, Edelman Pelton A, Ghosh S, Meek C, Ramos G, Suh J, Verwey J, Wang M, Wernsing J (2017) Machine teaching: a new paradigm for building machine learning systems. Technical report. <https://www.microsoft.com/en-us/research/publication/machine-teaching-new-paradigm-building-machine-learning-systems/>
44. Teso S, Kersting K (2019) Explanatory interactive machine learning. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, pp 239–245
45. Uijlings J, van de Sande K, Gevers T, Smeulders A (2013) Selective search for object recognition. Int J Comput Vis
46. Vedantam R, Zitnick C, Parikh D (2015) Cider: consensus-based image description evaluation. Comput Vis Pattern Recognit 4566–4575
47. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. CVPR
48. Williams JD, Niraula NB, Dasigi P, Lakshmiratan A, Suarez CGJ, Reddy M, Zweig G (2015) Rapidly scaling dialog systems with interactive learning. In: Natural language dialog systems and intelligent assistants, pp 1–13. Springer International Publishing, Cham. [https://doi.org/10.1007/978-3-319-19291-8\\_1](https://doi.org/10.1007/978-3-319-19291-8_1)
49. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. ICML
50. Zitnick L, Dollár P (2014) Edge boxes: Locating object proposals from edges. Comput Vis ECCV 2014:391–405