

This is the peer reviewed version of the following article:

Learning to Select: A Fully Attentive Approach for Novel Object Captioning / Cagrandi, Marco; Cornia, Marcella; Stefanini, Matteo; Baraldi, Lorenzo; Cucchiara, Rita. - (2021), pp. 437-441. ( 11th ACM International Conference on Multimedia Retrieval, ICMR 2021 Taipei, Taiwan August 21-24, 2021) [10.1145/3460426.3463587].

Association for Computing Machinery, Inc  
*Terms of use:*

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

18/12/2025 18:15

(Article begins on next page)

# Learning to Select: A Fully Attentive Approach for Novel Object Captioning

Marco Cagrandi, Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, Rita Cucchiara

{name.surname}@unimore.it

University of Modena and Reggio Emilia

Modena, Italy

## ABSTRACT

Image captioning models have lately shown impressive results when applied to standard datasets. Switching to real-life scenarios, however, constitutes a challenge due to the larger variety of visual concepts which are not covered in existing training sets. For this reason, novel object captioning (NOC) has recently emerged as a paradigm to test captioning models on objects which are unseen during the training phase. In this paper, we present a novel approach for NOC that learns to select the most relevant objects of an image, regardless of their adherence to the training set, and to constrain the generative process of a language model accordingly. Our architecture is fully-attentive and end-to-end trainable, also when incorporating constraints. We perform experiments on the *held-out* COCO dataset, where we demonstrate improvements over the state of the art, both in terms of adaptability to novel objects and caption quality.

## CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Natural language generation**.

## KEYWORDS

novel object captioning; region selector; constrained beam search.

### ACM Reference Format:

Marco Cagrandi, Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, Rita Cucchiara. 2021. *Learning to Select: A Fully Attentive Approach for Novel Object Captioning*. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21)*, August 21–24, 2021, Taipei, Taiwan. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3460426.3463587>

## 1 INTRODUCTION

Describing images has recently emerged as an important task at the intersection of computer vision, natural language processing, and multimedia, thanks to the key role it can have to empower both retrieval and multimedia systems [4, 6, 8–10, 17, 31]. Recent advances in image captioning, indeed, have demonstrated that fully-attentive architectures can provide high-quality image descriptions when tested on the same data distribution they are trained [11, 13, 20, 27]. As the existing datasets for image captioning [23, 40] are limited in terms of the number of visual concepts they contain, though, the application of such systems in real-life scenarios is still

challenging. For this reason, the task of Novel Object Captioning (NOC) has recently gained a lot of attention due to its affinity towards real-world applications [1, 12, 15]. This setting, indeed, requires a model to describe images containing objects unseen in the training image-text data, also referred to as out-of-domain visual concepts.

Since the language model behind a NOC algorithm can not be trained to predict out-domain words, proper incorporation of such novel words during the generation phase is one of the most relevant issues in this task. Early NOC approaches [12, 35] tried to transfer knowledge from out-domain images by conditioning the model at training time on external unpaired visual and textual data. Further works [21, 38] proposed to integrate coping mechanisms in the language model to select words corresponding to the predictions of a tagger. However, these frameworks do not include a proper and explainable method to identify which objects on the scene are more relevant to be described, and consequently, lack on leveraging all the available information provided by visual inputs. On a different line, Anderson *et al.* [3] devised a Constrained Beam Search algorithm to force the inclusion of selected tag words in the output caption, following the predictions of a tagger.

Inspired by this last line of research, we combine the ability to constrain the predictions from a language model with the usage of object regions and of fully-attentive architectures, which is dominant in traditional image captioning. Precisely, we devise a model with a specific ability to select objects in the scene to be described, with a class-independent module that can work on both in-domain and out-of-domain objects. Further, we combine this with a variant of the Beam Search algorithm which can include constraints produced by the region selector, while assuring end-to-end differentiability. We provide extensive experiments to validate the proposed approach: when tested on the *held-out* portion of the COCO dataset, our model provides state-of-the-art results in terms of caption quality and adaptability to describe objects unseen in the training set. Given its simplicity and effectiveness, our approach can also be thought of as a powerful new baseline for NOC, which can foster future works in the same area.

## 2 PROPOSED METHOD

Our NOC approach can be conceptually divided into two modules: an *image captioner* and a *region selector*. While the image captioning model is conditioned on the input image and is in charge of modeling a sequence of output words, the region selector is in charge of choosing the most relevant objects which need to be described, regardless of their adherence to the training set. The objects picked by the selector are used as constraints during the generation process, so that the output caption is forced to contain their labels

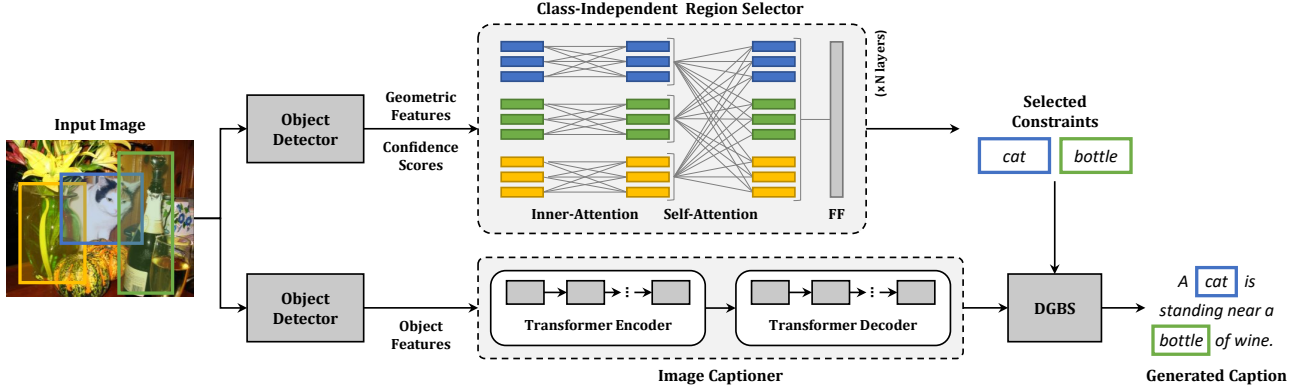


Figure 1: Summary of our approach.

as predicted by an object detector. All the components of our architecture are based on fully-attentive structure, and end-to-end training is allowed also when adding constraints to the language model. Fig. 1 shows an outline of the approach.

## 2.1 Class-Independent Region Selector

The role of the region selector is to identify objects which must be described in the output sentence. Since the object selector will need to work on classes that are unseen in the training set, we adopt a class-independent strategy in which no information about the object class is employed in the feature extraction process. Instead, we model intra-class relationships between objects of the same class, to handle the case in which multiple objects of the same class are present on the scene.

Given a set of regions  $X = \{x_i\}_i$  extracted from the input image, along with their classes  $\{c_i\}_i$ , we extract central coordinates, width, height and, additionally, we compute the object area. We also consider as an extra feature the confidence score  $s_i$  of the object, to obtain a class-independent feature vector:

$$x_i = \left[ \left( \frac{x_c}{W} \right), \left( \frac{y_c}{H} \right), \left( \frac{w_i}{W} \right), \left( \frac{h_i}{H} \right), \left( \frac{w_i \cdot h_i}{W \cdot H} \right), s_i \right] \quad (1)$$

where  $x_c$  and  $y_c$  are the coordinates of the center of the region,  $w_i$  and  $h_i$  its width and height, and  $W$  and  $H$  the image dimensions.

The set of feature vectors obtained for an image is then fed to a sequence of Transformer-like [33] layers, each of them composed by an *inner-attention* operator and a *self-attention* operator. The inner-attention operator is devised to connect together regions belonging to the same class, while the self-attention operator provides complete connectivity between elements in  $X$ . The combination of these two operators allows the region selector to independently focus on specific clusters of objects, in order to exchange semantically related information and learn intra-class dependencies, and then, to model long-range and diverse dependencies.

Given a partition of  $X$  computed according to the class each region belongs to, i.e.  $\{r_c \subseteq X \mid \forall x_i, x_j \in r_c, c_i = c_j\}_c$ , the result of the inner-attention operator applied over an element of the partition is a new set of elements  $\mathcal{I}(r_c)$ , with the same cardinality as  $r_c$ , in which each element is replaced with a weighted sum of values computed from regions of the same class. Formally, it can

be defined as:

$$\mathcal{I}(r_c) = \text{Attention}(W_q r_c, W_k r_c, W_v r_c), \quad (2)$$

where  $r_c$  is the set of all elements of  $X$  belonging to class  $c$ ,  $W_*$  are learnable projection matrices, and  $\text{Attention}$  indicates the standard dot-product attention [33].

The inner attention layer is applied independently over each element of the above-defined partition so that the overall encoding of  $X$  is a new sequence of elements defined as follows:

$$\mathcal{I}(X) = (\mathcal{I}(r_1), \mathcal{I}(r_2), \dots, \mathcal{I}(r_C)), \quad (3)$$

where  $C$  indicates the number of classes. After each inner-attention layer, a self-attention layer is employed to connect elements of different classes together. Formally, it is defined as:

$$\mathcal{S}(X) = \text{Attention}(W_q X, W_k X, W_v X), \quad (4)$$

where  $W_*$  are, again, learnable projection matrices.

After a sequence of inner- and self-attention layers, in which each pair of operators is followed by a position-wise feed-forward network [33], the region selector outputs a selection score  $Y_i$  for each object proposal. To do so, we apply an affine transformation and a non-linear activation to the output of the last layer:

$$Y_i = \sigma(\text{RegionSelector}(X_i) W_o), \quad (5)$$

where  $W_o \in \mathbb{R}^{d \times 1}$  are learnable weights and  $\sigma$  is a sigmoid.

**Training.** The region selector is trained using a binary cross-entropy loss. To build ground-truth data, for each image we collect the object classes identified by the object detector and construct a binary ground-truth vector indicating whether a class name is contained in at least one of the ground-truth captions associated with the image. We also consider as positives synonyms and plural forms of the object class names. At inference time, we extract the selected objects for each image adopting 0.5 as threshold.

## 2.2 Image Captioner

After object selection, our image captioning model is responsible for generating a caption using the chosen class names as constraints. Inspired by recent works which employ fully-attentive models in image captioning [11, 13, 25], we create a captioning model with an encoder-decoder structure, where the encoder refines image region features and the decoder generates captions auto-regressively.

**Table 1: Evaluation on the *held-out* COCO test set, when using different constraint selection approaches.**

	Cross-Entropy Loss							CIDEr Optimization							CIDEr Optimization with DGBS						
	In-Domain			Out-Domain				In-Domain			Out-Domain				In-Domain			Out-Domain			
	M	C	S	M	C	S	F1	M	C	S	M	C	S	F1	M	C	S	M	C	S	F1
No Constraints	27.2	108.9	20.2	22.4	68.5	14.7	0.0	28.4	122.3	22.3	23.5	76.8	16.3	0.0	28.1	120.9	21.9	23.4	76.5	16.1	0.0
Top-1	26.2	97.4	19.2	24.1	75.9	17.6	60.1	27.6	110.5	21.3	25.4	84.6	18.8	60.2	27.9	115.9	21.0	25.3	84.7	18.7	60.2
Top-2	24.4	81.9	16.4	23.8	68.7	16.1	68.1	26.2	95.4	18.4	25.1	77.6	17.3	68.1	27.1	102.9	18.4	25.6	80.0	17.2	68.1
Top-3	22.7	69.9	14.4	22.4	56.9	14.5	66.0	25.1	83.3	16.5	24.4	67.1	15.4	66.0	26.6	92.3	17.0	25.2	70.8	15.6	66.0
Region Selector (w/o Inner)	25.2	70.6	17.7	24.1	70.6	16.8	70.2	26.8	101.5	19.6	25.6	80.5	18.0	70.2	27.4	108.0	19.7	25.8	82.2	18.2	70.2
<b>Region Selector</b>	26.2	97.0	19.2	<b>24.9</b>	<b>78.2</b>	<b>18.3</b>	<b>75.0</b>	27.6	109.2	21.1	<b>26.1</b>	<b>87.7</b>	<b>19.4</b>	<b>75.0</b>	27.9	115.3	21.0	<b>26.3</b>	<b>88.5</b>	<b>19.4</b>	<b>75.1</b>
Oracle Constraints	27.3	107.0	20.6	25.6	84.0	19.0	76.0	28.5	118.9	22.5	26.6	91.7	20.2	76.0	28.6	122.9	22.3	26.6	92.3	20.2	76.0

**Encoder.** Recent captioning literature has shown that object regions are the leading solution to encode visual inputs [4, 37, 39], followed by self-attentive layers to model region relationships [11, 13, 16, 25, 27, 32]. However, as self-attention can only encode pairwise similarities, it exhibits a significant limitation on encoding knowledge learned from data. To overcome this restraint, we enrich our encoder with memory slots [7, 11]. Specifically, we extend the set of keys and values of self-attention layers with additional learnable vectors, which are independent of the input sequence and can encode a priori information retrieved through attention.

**Decoder.** The decoder is the actual language model, conditioned on both previously generated words and image region encodings. As in the standard Transformer [33], our language model is composed of a stack of decoder layers, each performing a masked self-attention and a cross-attention followed by a position-wise feed-forward network. Specifically, for each cross-attention, keys and values are inferred from the encoder output, while for the masked self-attention, queries, keys, and values are exclusively extracted from the input sequence of the decoder. This self-attention is right-masked so that each query can only attend to keys obtained from previous words.

### 2.3 Including Lexical Constraints

To include the lexical constraints produced by the region selector when decoding from the language model, we devise a variant of the Beam Search algorithm [14, 26] which supports the adoption of single-word constraints. Given a number of word constraints  $\mathbf{W} = \{w_0, w_1, \dots, w_n\}$  and a maximum decoding length  $T$ , we frame the decoding process in a matrix  $G$  with  $n$  rows and  $T$  columns, where the horizontal axis covers the time steps in the output sequence, and the vertical axis indicates the constraints coverage. Each cell of the matrix can contain a beam of partially decoded sequences.

At iteration  $t$ , each row  $i$  of  $G[:, t]$  can be filled in two ways: either by continuing the beam contained in  $G[i, t-1]$  by sampling from the probability distribution of the language model, or by forcing the inclusion of a constraint from  $\mathbf{W}$ . In the former case, the resulting updated beam of sequences is stored in  $G[i, t]$ , while in the latter case it is stored in  $G[i+1, t]$ . At the end of the generation process, the last row of  $G$  will contain sequences that satisfy all constraints.

Algorithm 1 reports the pseudo-code of our constrained beam search procedure. There,  $k$  indicates the number of elements in each bin,  $\text{model.step}$  indicates sampling from the language model probability distribution to continue the generation of a partially-decoded

#### Algorithm 1: Grid Beam Search

```

 $G \leftarrow \text{initGrid}(n, T, k)$ 
for  $t = 1; t < T; t++$  do
  for  $c = \max(0, n + t - T); c < \min(t, n); c++$  do
     $g, s = \emptyset$  forall  $hyp$  in  $G[c, t-1]$  do
       $g \leftarrow g \cup \text{model.step}(hyp)$ 
    end
    if  $c > 0$  then
      forall  $hyp$  in  $G[c-1, t-1]$  do
         $s \leftarrow s \cup \text{model.add\_constr}(hyp, \{w_0, \dots, w_n\})$ 
      end
    end
     $G[c, t] \leftarrow k\text{-argmax}_{h \in g \cup s} (\text{model.score}(h))$ 
  end
end
 $topHyp \leftarrow \text{hasEOS}(G[n, :])$   $\triangleright$  Remove sequences w/o EOS
return  $\text{argmax}_{h \in topHyp} (\text{model.score}(h))$ 

```

sequence, while  $\text{model.add\_constr}$  indicates a function which continues a beam by adding all the available constraints, excluding those which have already been generated for a sequence. Because all the operations required to include constraints are differentiable, we call our constraint inclusion approach *Differentiable Grid Beam Search* (DGBS), and employ it to fine-tune the image captioner also when using a CIDEr-D optimization strategy.

## 3 EXPERIMENTS

### 3.1 Evaluation Protocol

**Dataset.** We conduct experiments on the *held-out* COCO dataset [12], which consists of a subset of the COCO dataset [23] for standard image captioning, where the training set excludes all image-caption pairs that mention at least one of the following eight objects: *bottle*, *bus*, *couch*, *microwave*, *pizza*, *racket*, *suitcase*, and *zebra*. We follow the splits defined in [12] and take half of COCO validation set for validation and the other half for testing.

**Metrics.** To evaluate caption quality, we use standard captioning metrics (i.e. BLEU-4 [28], METEOR [5], ROUGE [22], CIDEr [34], and SPICE [2]), while we employ F1-scores [12] to measure the model ability to incorporate new objects in generated captions.

**Implementation details.** To extract geometric features and confidence scores for our region selector, we employ Faster R-CNN [30] with ResNet-50-FPN backbone, trained on COCO [23]. For both training and inference, we discard the detections of the *person* and *background* classes. During training, we use different loss weights

Table 2: Comparison with the state of the art on the *held-out* COCO test set.

	F1 Scores									Captioning Metrics				
	F1 <sub>bottle</sub>	F1 <sub>bus</sub>	F1 <sub>couch</sub>	F1 <sub>microwave</sub>	F1 <sub>pizza</sub>	F1 <sub>racket</sub>	F1 <sub>suitcase</sub>	F1 <sub>zebra</sub>	F1 <sub>average</sub>	B-4	M	R	C	S
DCC [12]	4.6	29.8	45.9	28.1	64.6	52.2	13.2	79.9	39.8	-	21.0	-	59.1	13.4
NOC [35]	17.8	68.8	25.6	24.7	69.3	55.3	39.9	89.0	48.8	-	21.3	-	-	-
NBT [24]	14.0	74.8	42.8	63.7	74.4	19.0	44.5	92.0	53.2	-	23.9	-	84.0	16.6
CBS [3]	16.3	67.8	48.2	29.7	77.2	57.1	49.9	85.7	54.0	-	23.6	-	77.6	15.9
LSTM-C [38]	29.7	74.4	38.8	27.8	68.2	70.3	44.8	91.4	55.7	-	23.0	-	-	-
DNOC [36]	33.0	76.9	54.0	46.6	75.8	33.0	59.5	84.6	57.9	-	21.6	-	-	-
LSTM-P [21]	28.7	75.5	47.1	51.5	81.9	47.1	62.6	93.0	60.9	-	23.4	-	88.3	16.6
NBT + CBS [24]	38.3	80.0	54.0	<b>70.3</b>	81.1	74.8	67.8	<b>96.6</b>	70.3	-	24.1	-	86.0	17.4
Top-2	29.6	77.4	44.7	62.6	83.3	<b>81.2</b>	70.7	95.1	68.1	28.1	25.6	52.7	80.0	17.2
Region Selector (w/o Inner)	42.3	78.3	54.4	59.4	85.3	79.1	67.2	95.6	70.2	28.4	25.8	52.8	82.2	18.2
<b>Region Selector</b>	<b>43.9</b>	<b>83.7</b>	<b>66.8</b>	64.7	<b>88.0</b>	81.0	<b>76.9</b>	95.4	<b>75.1</b>	<b>30.3</b>	<b>26.3</b>	<b>53.8</b>	<b>88.5</b>	<b>19.4</b>

Table 3: Region selector performance evaluation using different loss weights for zero and one values.

	$\lambda_0$	$\lambda_1$	In-Domain			Out-Domain			
			M	C	S	M	C	S	F1
Region Selector (w/o Inner)	0.4	0.6	27.4	111.9	20.3	25.8	85.6	18.7	68.5
<b>Region Selector</b>	0.4	0.6	<b>28.1</b>	<b>119.2</b>	<b>21.3</b>	<b>26.0</b>	<b>89.0</b>	<b>19.4</b>	<b>70.4</b>
Region Selector (w/o Inner)	0.3	0.7	27.2	108.7	19.9	25.9	84.9	18.6	69.9
<b>Region Selector</b>	0.3	0.7	<b>28.0</b>	<b>116.4</b>	<b>21.2</b>	<b>26.2</b>	<b>88.7</b>	<b>19.4</b>	<b>74.2</b>
Region Selector (w/o Inner)	0.2	0.8	27.4	108.0	19.7	25.8	82.1	18.2	70.2
<b>Region Selector</b>	0.2	0.8	<b>27.9</b>	<b>115.3</b>	<b>21.0</b>	<b>26.3</b>	<b>88.5</b>	<b>19.4</b>	<b>75.1</b>
Region Selector (w/o Inner)	0.1	0.9	26.9	97.8	18.2	25.6	73.2	16.7	67.1
<b>Region Selector</b>	0.1	0.9	<b>27.9</b>	<b>114.3</b>	<b>20.8</b>	<b>26.2</b>	<b>87.5</b>	<b>19.2</b>	<b>75.6</b>

(i.e.,  $\lambda_0 = 0.2$  and  $\lambda_1 = 0.8$ ) to balance the importance of zero and one ground-truth values, and we limit the number of object proposals for each image to 10 according to their confidence scores. Region selector features are projected to a 128-dimensional embedding space and passed through  $N = 2$  identical layers, each composed of inner-attention, self-attention, and feed-forward.

For our image captioning model, we extract object features from Faster R-CNN [30] with ResNet-101 finetuned on Visual Genome [4, 19]. Following [11], we use three layers for both encoder and decoder and employ 40 memory vectors for each encoder layer. We represent words with GloVe word embeddings [29], using two fully-connected layers to convert between the GloVe dimensionality (i.e., 300) and the captioning model dimensionality (i.e., 512) before the first decoding layer and after the last decoding layer. Before the final softmax, we multiply with the transpose of the word embeddings. We pre-train our captioning model using cross-entropy and finetune it using RL with CIDEr-D reward. During this phase, we use the classes detected by Faster R-CNN, trained on COCO, that are mentioned in the ground-truth captions as constraints for our DGBS algorithm. We limit the number of possible constraints to 5.

All experiments are performed with a batch size equal to 50. For training the region selector and pre-training the captioning model, we use the learning rate scheduling strategy of [33] with a warmup equal to 10,000 iterations and Adam [18] as optimizer. CIDEr-D optimization is done with a learning rate equal to  $5 \times 10^{-6}$ .

### 3.2 Experimental Results

Table 1 shows the results of our model in terms of captioning metrics and F1-score averaged over the eight held-out classes, using

different strategies to train the captioning model. We compare with a variant of our region selector without inner-attention (i.e., w/o Inner) and using the top- $k$  detections, with  $k = 1, 2, 3$ , instead of our selection strategy. For reference, we also report the performance when using oracle constraints coming from ground-truth captions. As it can be seen, our solution achieves the best results in terms of both caption quality and F1-score, demonstrating the effectiveness of our region selector for choosing constraints for the captioning model and the importance of the inner-attention operator. Furthermore, by comparing the results with standard CIDEr optimization and those obtained using our DGBS algorithm during training, we can see improved results on both in-domain and out-domain captions, thus confirming the usefulness of our training strategy.

In Table 3, we show the results when using different weights to balance the importance of zero and one ground-truth values. As it can be seen, our complete region selector achieves better performance than the variant without inner-attention, thus further demonstrating the effectiveness of the proposed attention operator. Additionally, employing  $\lambda_0 = 0.2$  and  $\lambda_1 = 0.8$  provides the best balance in terms of captioning metrics and F1-score.

Finally, in Table 2, we compare our model with NOC state-of-the-art approaches. As it can be noticed, our region selector obtains the best results in terms of both F1-scores and captioning metrics, achieving a new state of the art on the *held-out* COCO dataset.

## 4 CONCLUSION

We have presented a fully-attentive approach for NOC that learns to select and describe unseen visual concepts. Our method is based on a class-independent region selector and an image captioning model trained with a differentiable grid beam search algorithm that generates sentences with given constraints, in an end-to-end fashion. Experimental results have showed that our model achieves a new state of the art on the *held-out* COCO dataset, demonstrating its effectiveness in successfully describing novel objects.

## ACKNOWLEDGMENTS

This work has been supported by “Fondazione di Modena”, by the national project “IDEHA: Innovation for Data Elaboration in Heritage Areas” (PON ARS01\_00421), cofunded by the Italian Ministry of University and Research, and by the project “Artificial Intelligence for Cultural Heritage (AI for CH)”, cofunded by the Italian Ministry of Foreign Affairs and International Cooperation.

## REFERENCES

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*.
- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Proceedings of the European Conference on Computer Vision*.
- [3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided Open Vocabulary Image Captioning with Constrained Beam Search. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [5] Satandeep Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Annual Meeting on Association for Computational Linguistics Workshops*.
- [6] Roberto Bigazzi, Federico Landi, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2020. Explore and Explain: Self-supervised Navigation and Re-counting. In *Proceedings of the International Conference on Pattern Recognition*.
- [7] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2020. SMaRT: Training Shallow Memory-aware Transformers for Robotic Explainability. In *Proceedings of the IEEE International Conference on Robotics and Automation*.
- [8] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2017. Visual saliency for image captioning in new multimedia services. In *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops*.
- [9] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 2 (2018), 1–21.
- [10] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2020. Explaining Digital Humanities by Aligning Images and Textual Descriptions. *Pattern Recognition Letters* 129 (2020), 166–172.
- [11] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [12] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [13] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image Captioning: Transforming Objects into Words. In *Advances in Neural Information Processing Systems*.
- [14] Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*.
- [15] Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2020. VIVO: Surpassing Human Performance in Novel Object Captioning with Visual Vocabulary Pre-Training. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [16] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on Attention for Image Captioning. In *Proceedings of the IEEE International Conference on Computer Vision*.
- [17] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [18] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.
- [19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.
- [20] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Proceedings of the European Conference on Computer Vision*.
- [21] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2019. Pointing Novel Objects in Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [22] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Annual Meeting on Association for Computational Linguistics Workshops*.
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*.
- [24] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural Baby Talk. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [25] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. 2021. Dual-Level Collaborative Transformer for Image Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [26] Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics* 30, 4 (2004), 417–449.
- [27] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-Linear Attention Networks for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*.
- [29] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149.
- [31] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *Proceeding of the International Conference on Image Analysis and Processing*.
- [32] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2020. A Novel Attention-based Aggregation Function to Combine Vision and Language. In *Proceedings of the International Conference on Pattern Recognition*.
- [33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.
- [34] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [35] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning Images with Diverse Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [36] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. 2018. Decoupled Novel Object Captioner. In *Proceedings of the ACM International Conference on Multimedia*.
- [37] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-Encoding Scene Graphs for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [38] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017. Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [39] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring Visual Relationship for Image Captioning. In *Proceedings of the European Conference on Computer Vision*.
- [40] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.