Deep Image Captioning: Models, Data and Evaluation

Ingrid Hrga
Juraj Dobrila University of Pula, Department of Information and Communication Technologies
Zagrebačka 30, 52100 Pula, Croatia
ingrid.hrga@unipu.hr

***Abstract* - As a problem that resides at the intersection of Computer Vision and Natural Language Processing, image captioning has witnessed a rapid progress in a very short time, from initial template-based models to the current ones, based on deep neural networks. This paper gives an overview of current issues and recent research on image captioning, with a special emphasis on models employing deep encoder-decoder architectures. We discuss the advantages and disadvantages of different approaches, along with reviewing some of the most commonly used datasets and evaluation metrics. We point out to some open questions and conclude with directions for future research.**

Keywords – image captioning, attention mechanism, deep neural networks, encoder-decoder framework

I. INTRODUCTION

Recent success of deep learning methods in perceptual tasks, such as image classification [36, 65, 62] and object detection [61, 21, 59] have encouraged researchers to tackle some of the more demanding problems for which recognition is just a step towards a more complex reasoning about our visual world [35]. Image captioning[1], as a task of automatically describing an image with one or more natural language sentences, although relatively novel [26], has already gained a lot of attention in the research community, proving to have many useful applications. Massive amounts of unstructured and semi-structured data, large portions of which come in the form of images and videos, are available, practically everywhere, today. In order to leverage their value, an efficient access is required. But retrieval of such data becomes even more challenging if it's not accompanied by any additional information explaining its content, something that has often been the case because of the speed at which it is being generated. It would be quite impractical, if not impossible, to manually annotate all those large collections, even if the cost of such endeavor wouldn't be an issue. Therefore, other solutions should be found. But automatically summarizing content within an image or other media does not limit its usefulness only to content-based retrieval. Connecting media and natural language help sell products or can serve as an aid for the visually impaired in performing daily tasks, can be used for question answering, or even in robots navigation. Benefits of systems capable of generating media descriptions are certainly not questionable. But something that seemed so easy when being done by a human, turned out to be extremely difficult for computers.

As a problem that integrates vision and language understanding, its main challenges arise from the need of translating between two different, but usually paired, modalities [33]. It was shown [19] that just a fraction of a second is sufficient for a human to capture the meaning of the scene in order to be able to describe it accurately. This includes not only to discern most salient objects and their attributes but also reasoning about intricate relationships and interactions between them [35]. Even more so, people describing an image usually rely on common sense knowledge for adding context, or are capable of using imagination for making descriptions vivid and interesting. Something that still poses a problem for computers.

Some of the earliest attempts at connecting vision and language date back to the 1970s [73]. More recently, some researchers have used images for word sense disambiguation [6], while others were focused on annotating images with individual words [5, 71]. But it wasn't until just recently that researchers have started to address the problem of generating full sentence image descriptions. Sentences are richer than lists of words [18] so several presumptions should have been met to make more substantial progress in image captioning. Advances in computer hardware and the ability to leverage GPUs for acceleration of parallel calculations, combined with the availability of new datasets with millions of labeled examples [60], enabled the training of advanced models based on deep neural networks [36, 79], which particularly favored the development of Computer Vision, as well as Natural Language Processing. Through the interaction of these two fields, novel vision-to-language (V2L) tasks have emerged, such as video description [76] and visual question answering [47], along with image captioning [33, 34, 17, 49, 12, 13, 30, 69, 74, 46, 78, 2].

The rest of the paper is organized as follows: next section begins with a systematization of different approaches to image captioning, followed by some background information on neural networks currently employed for this task. Section ends with an overview of related work grouped on the basis of the reviewed systems architectures. Third section presents some of the currently most important datasets, along with discussing different ways of collecting them. Section four points out to problems arising when evaluating generative approaches, while section five discusses some other open problems. We conclude with directions for future research.

II. MODELS

In general, image captioning models can be divided into two broad categories: (1) generative models that

1 **Captions or descriptions.** In [8] authors make a clear distinction between captions and descriptions. According to them, captions provide context while descriptions verbalize the literal content of an image. Since other authors mentioned in this paper don't differentiate between these two terms, we follow their approach and use them interchangeably.

generate novel captions and (2) retrieval-based models that rank a set of existing captions.

Early generative approaches relied on the use of predefined templates, which followed some specific rule of grammar and were filled in based on the results of the detection of scene elements [37, 52, 75]. However, the advantage of such bottom-up approaches [78] in terms of the ability in capturing some subtle details, was not enough to keep them in the focus of research interest. Generated sentences were too simple, lacking the fluency of the human-written ones. Moreover, such systems were heavily hand-designed, which constrained their flexibility [69].

In order to overcome aforementioned problems, along with additional difficulties in evaluating such systems (see section IV.), authors in [26] suggested to frame image description as a **ranking task**. In this approach, images and sentences are embedded into the same vector space so that their similarity can be directly compared. For a given test image, a set of semantically similar images, along with their descriptions, is retrieved from a pool of existing ones. Of these existing ones, either the most compatible is then directly transferred to the test image [26, 63, 18] or a new one, obtained by re-composing sentence fragments [44, 38]. More recent approaches use neural networks to co-embed images and sentences [63] or image regions and sentence segments [31] into the same multimodal space.

An advantage of ranking-based approaches lies in their ability to retrieve images based on a description query, or to retrieve descriptions based on an image query. Although avoiding some of the problems that plagued generative approaches, as well as always returning well-formed sentences, they lack the ability to generate novel sentences or to describe compositionally novel images [48], i.e. those containing objects that were observed during training but appear in different combinations on the test image. Moreover, they require large amounts of human-written training data, making them hard to scale [69].

Today's state-of-the-art models are **generative and neural networks based**, more specifically, they employ an encoder-decoder architecture by combining a Convolutional Neural Network with a Recurrent Neural Network. As top-down approaches [78], they directly map from images to sentences, making such systems end-to-end trainable.

Some background information regarding deep neural networks will be given in the upcoming sections, and will be subsequently followed by an overview of recent research. A more detailed survey of some of the earlier approaches can be found in [8].

A. Background

Convolutional Neural Network (CNN). As a type of feedforward neural network, introduced in 1990s [40, 41], CNN is adapted to work with input data with a strong local structure, such as 2D images. Typical CNN architecture consists of multiple convolutional and pooling layers, alternating several times, followed by a few fully-connected layers and a soft-max layer (Figure 1). Unlike the fully-connected layers, in which each neuron is connected to all neurons from a previous layer, the convolutional layer neurons receive input only from

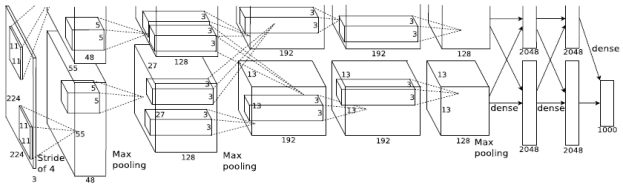

Figure 1. Architecture of AlexNet [36], a deep convolutional neural network that won the ImageNet ILSVRC challenge [60] in 2012. It consists of eight layers, five convolutional layers followed by three fully connected layers.

those in their receptive field. This forces the extraction of local features and reduces the number of learned parameters. Going from low-level features extracted at the first layer and combining them into higher-level features at subsequent layers, a fixed vector representation of the most salient aspects of the image for a given task, regardless of the exact location, is obtained. This resembles the functioning of the mammals visual cortex [27].

Neurons in a convolutional layer are organized in feature maps. All neurons in a feataure map perform the same operation at different parts of the image and share same weights. Instead by matrix multiplication, weights are calculated by convolving a convolutional layer with a set of local filters, hence the name.

Convolutional layers are followed by a pooling layer that performs local averaging and subsampling. This reduces the dimensionality of the output and helps obtaining invariance to translation or distortion.

Since 2012. CNNs achieve superior results on large-scale object recognition tasks [36, 79, 62, 65, 24].

Recurrent Neural Network (RNN). Recurrent Neural Network [15] is a neural network with feedback, designed to model sequences of data, such as words (sequences of characters) or sentences (sequences of words). RNN maintains an internal hidden state that stores context information, i.e. information computed from past inputs. In its simplest formulation, given a sequence of inputs $(x_1, ..., x_T)$, RNN computes a sequence of outputs $(y_1, ..., y_T)$ by applying the recurrence formula at every time step [13]:

$$h_t = \phi(h_{t-1}, x_t) \tag{1}$$

where h_t is internal hidden state at time step *t*, ϕ is a nonlinear activation function, h_{t-1} is hidden state at previous time step. Activation function may be a logistic sigmoid function σ or hyperbolic tangent $\tanh$ applied elementwise such that

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \tag{2}$$

where W_{xh} and W_{xh} are learned weight matrices, b_h is bias. Same parameters are shared across all time steps. Output y_t at time step *t* is calculated as:

$$y_t = W_{hy}h_t + b_y. \tag{3}$$

By maintaining an internal hidden state *h,* as a function of all the inputs from previous time steps, RNN can easily learn short-term dependencies. But for longer sequences RNNs become difficult to train [7, 55] due to the exploding and the vanishing gradient problems caused by RNNs iterative nature [29]. The exploding gradient

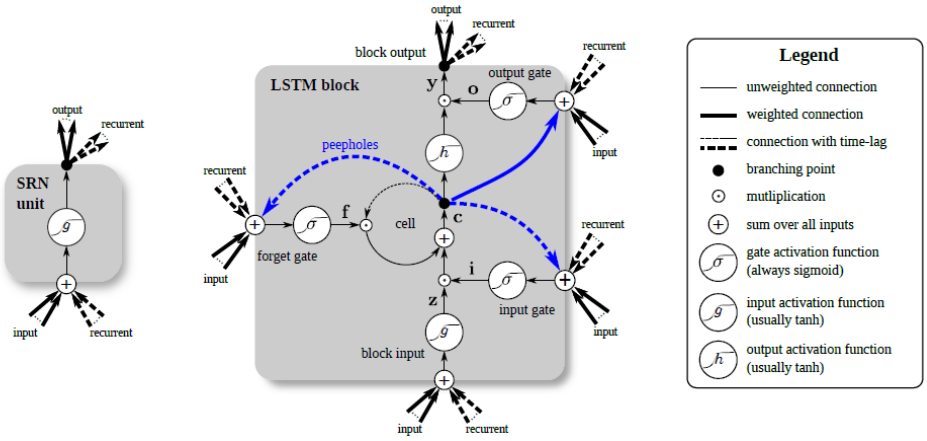

Figure 2. Comparison of the Simple Recurrent Network (SRN) unit (left) and a Long Short-Term Memory block (right) [23]. (Note: hidden state *h* here is denoted as *y*.)

problem can be addressed by a technique known as gradient clipping [55]. However, the vanishing gradient problem is more challenging [29].

Long Short Term-Memory (LSTM). Long Short Term-Memory architecture was developed by [25] and, in a somewhat modified version [22], became the standard way of dealing with the vanishing gradient problem [29]. Mathematically, the LSTM architecture is defined as [23]:

$$z_t = g(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \quad (4)$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + p_i \odot c_{t-1} + b_i) \quad (5)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + p_f \odot c_{t-1} + b_f) \quad (6)$$

$$c_t = i_t \odot z_t + f_t \odot c_{t-1} \quad (7)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + p_o \odot c_t + b_o) \quad (8)$$

$$h_t = o_t \odot \tanh(c_t) \quad (9)$$

where x_t is the input vector at time step *t*, *W* are weight matrices, *z* is input modulation gate, *i, f, o* are input, forget and output gates, *c* is memory cell, *p* are peephole weight vectors and *b* are biases. Functions σ, *g* and $\tanh$ are non-linear activation functions applied element-wise, $\odot$ denotes element-wise multiplication.

Hidden state in the LSTM consists of two states: a "fast" state *h* and a "slow" state *c* that helps alleviate the vanishing gradients problem [29]. Flow of information in a LSTM memory cell *c* is controlled by gates which can open or close, depending on their weights, so that information can be stored in, written to or read from a cell (Figure 2). An addition instead of multiplication at the memory cell is a key to preserving constant error flow when it must be propagated at depth.

Encoder-decoder architecture. Inspired by its success in Neural Machine Translation (NMT) [64] most of the current state-of-the-art models for image captioning employ the encoder-decoder architecture. In this architecture an encoder is used to map the input, i.e. a sentence in a source language, into its real-valued fixed-dimensional vector representation. A decoder then generates output, i.e. a sentence in the target language, conditioned on the representation produced by the encoder. Main advantage of such a system is that it can be trained end-to-end, meaning that the parameters of the whole network are learned together, thereby avoiding the problem of aligning several independent components.

Perceived as a task of translating one modality, i.e. a picture, to another modality, i.e. its description, the encode-decoder architecture was also successfully adopted in vision-to-language problems (Figure 3), such as image captioning [69, 13, 74], video description [76] or visual question answering [47].

For the task of image captioning, a CNN is employed on the encoder side, which acts as a feature extractor. CNN is usually pre-trained on a large dataset for a classification task [60]. A feature map from a convolutional layer or the vector representation from a fully-connected layer is then used for image representation.

On the decoder side, a RNN, or one of its variants such as LSTM or GRU [10], is employed for language modeling. A RNN is trained to predict the next word y_t conditioned on all the previously predicted words $(y_1, y_2, ..., y_{t-1})$ and the context vector *c* produced by the encoder [3]:

$$p(y_t|y_1, y_2, ..., y_{t-1}, c) = g(y_{t-1}, h_t, c) \quad (10)$$

where *g* is a nonlinear function that outputs probability of y_t, h_t is the hidden state of the RNN.

Main advantage of using neural language models, instead of e.g. n-gram based models, is in reducing the curse of dimensionality problem through the use of distributed word representations [33]. Words are represented as real-valued fixed-dimensional vectors and projected into low dimensional space so that similar words are clustered together. Instead of random initializing their weights, pre-trained word vectors [50, 56] can be used.

Attention mechanism. Some limitations of the general encoder-decoder framework have motivated the development of different extensions among which, the addition of the attention mechanism has emerged as the most important.

It was demonstrated in [3] that the fixed-length vector

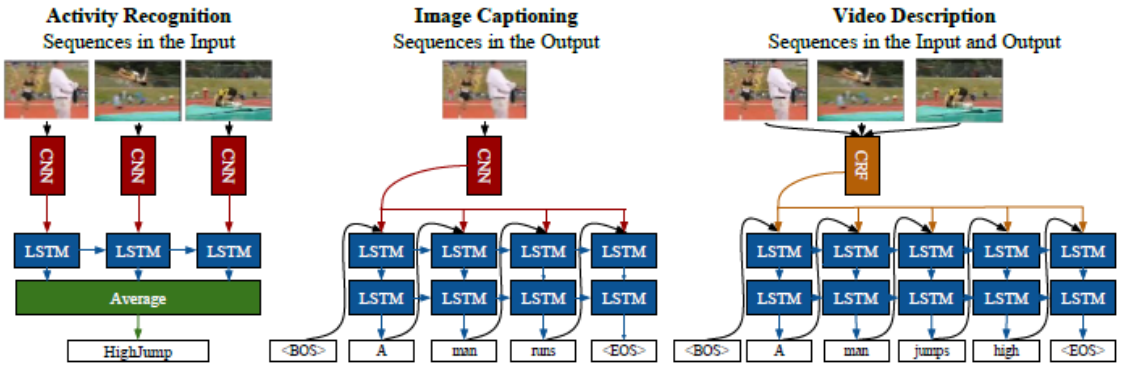

Figure 3. Encoder-decoder architectures for three vision problems involving recognition and description (activity recognition, image caption generation and video description) [13]. Proposed LRCNs combine a CNN with a stack of LSTMs to process (possibly) variable-length inputs into variable-length predictions.

representation produced by the encoder is responsible for the degradation of the performance occurring as the length of input increases. Regardless of the size of the input, in the general encoder-decoder framework all the information is compressed into a context vector c of a predefined size. Instead, authors proposed to encode the input into a set of context vectors $c = \{c_1, c_2, ..., c_M\}$ from which a subset is chosen so to adaptively attend to the most important parts of the input while generating next word of the output.

Attention mechanism is widely adopted for the task of image captioning and different variants are being developed, such as spatial attention, semantic attention or adaptive attention, to name just a few.

B. *Related work*

Encoder-decoder framework. Building on promising results that deep learning methods already demonstrated on the task of learning representations from multiple modalities, in [33] authors proposed a multimodal neural language model that can be conditioned on high-level image features learned from a deep convolutional neural network (CNN). In this model, image features extracted from the top fully connected layer of a CNN, words represented as real-valued feature vectors and all model parameters are learned together by jointly training a language model with a convolutional network. Authors introduced two multimodal log-bilinear model (MLBL) based methods that differ by the way the outputs of the CNN are used. In modality-biased log-bilinear model (MLBL-B) images are input as additive bias. The more powerful factored 3-way log-bilinear model (MLBL-F) uses gating. Although proposed MLBL models can generate descriptions by directly sampling from a language model, as opposed to earlier approaches that relied on the use of templates [37, 52, 75] or other constraints, it is mostly focused on retrieval.

To overcome the inability of retrieval-based approaches to produce novel captions or to describe images with previously unseen combinations of objects, in [49] and in [48], a subsequent version of their work, authors propose a multimodal Recurrent Neural Network (m-RNN) framework that directly models the probability distribution of a word given previous words and an image. The model consists of two sub-networks, a CNN and a RNN, modeled through five layers: two word embedding layers for learning dense word representations, a recurrent layer in which semantic temporal context is stored, a multimodal layer for connecting the language model part with a deep CNN which in turn generates image representations, and ending with a soft max layer to output probabilities. Storing context in the recurrent layer allows the use of arbitrary context length, as opposed in [33] where the model used fixed-length context. The decision not to store image information in a recurrent layer but to directly input it through a multimodal layer, combined with the use of two-layer word embedding, led to substantial improvement in performance.

In [34] authors go further and build on [33] by replacing log-bilinear model, a type of a feedforward network, with the recurrent neural network. The proposed encoder-decoder system unifies joint image-text embedding models (encoder part) with the new structure-content multimodal neural language model (SC-NLM) (decoder part). SC-NLM extracts content from sentence structure conditioned on the representation produced by the LSTM encoder. This way, the structure variables that correspond to part-of-speech for words found in the description, and serve as a soft template that steers the process in order to generate grammatically correct sentences, are obtained. An advantage of the SC-NLM is that it can be trained only on text, without the need of images, allowing additional text corpora to be provided for the purpose of improving the quality of the language model.

Previous approaches use image features extracted at the global level (full-frame image features) but [31] showed an alternative that takes advantage of object detections results from a Region CNN (RCNN) detector. Similarly, in an approach that won the Microsoft 2015 COCO Caption Challenge[2], [17] uses weakly-supervised learning to create detectors for visual words, i.e. nouns, verbs and adjectives, found in image regions, which are then mapped to words that are likely to appear in captions. With the goal of covering each of the detected words exactly once, a maximum entropy language model (ME LM), which generates candidate captions from a set of training descriptions conditioned on visually detected words, is employed. The final step in the process uses Minimum Error Rate Training (MERT) to re-rank candidate sentences by utilizing an additional Deep Multimodal Similarity Model (DMSM) based feature.

2 http://cocodataset.org/#captions-challenge2015

For each modality, DMSM learns separate neural networks, which then project image and text fragments to a common vector space and score their similarity.

In [17], image region features are utilized for the purpose of generating whole image captions. In contrast, the goal in [30] is to generate image region descriptions (Figure 4) by first aligning contiguous segments of image descriptions with image regions that they describe and then using this inferred alignments to generate novel descriptions. Following [31], objects in images are detected with a RCNN. Word representations are enriched with a variably sized context around each word and computed with a Bidirectional RNN (BRNN) [22]. Instead of random initializing weights of word embedding matrices, as in [48], BRNN exploits the benefits of pre-computed word vectors. Words and images are mapped into a common multimodal embedding so that semantically similar concepts occupy nearby regions of the space. After grounding fragments of most compatible sentence in image regions, a second model, a Multimodal Recurrent Neural Network, uses these learned correspondences as the training data in the generative process that is additionally conditioned on the image input via bias interaction on the first time step.

Experiments showed that the model is able to detect visual-semantic correspondence for small or rare objects that would be missed by full frame models. Proposed model outperformed full-frame and ranking baselines.

A limitation of the architecture proposed in [30] is that it uses two separate models. In [69] authors introduce an end-to-end trainable Neural Image Caption (NIC) system, similar in approach to [33] or [48] but in contrast, it uses a LSTM variant of a RNN. Additionally, images are input only once, at the first time step, to give LSTM an overview of the image content, and directly, not as additive bias, which allows RNN not to lose sight of the objects already mentioned in captions. Experiments showed that inputting image at every time step leads to a system more prone to overfitting.

A similar end-to-end system, combining a CNN encoder and a LSTM decoder, is introduced in [13] but with a difference that in the proposed Long-Term Recurrent Convolutional Network (LRCN) (Figure 3) image features are input at every time step. Additionally, factored representation is explored by concatenating

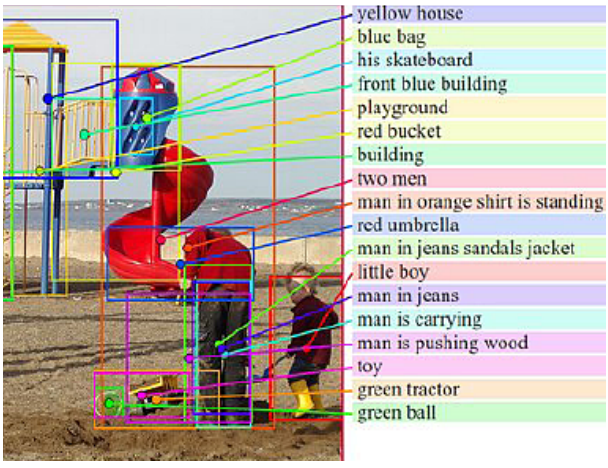

Figure 4. Image region descriptions as generated by a region-level multimodal RNN [30].

image features with the hidden state output of previous LSTM of the stack. Authors investigated the effect of different architectures and found that using LSTM instead of simple RNN combined with the more powerful CNN were the most important factors contributing to better performance. Stacking additional LSTM layers didn't bring expected improvements.

State-of-the-art results in image captioning that achieved image-conditioned language models, motivated authors in [12] to directly compare two dominant approaches: one that uses maximum-entropy language model (ME LM) to generate sentences based on a set of discrete detections, as proposed in [17], and a second one that uses a RNN LM conditioned on the continuous valued CNN activations, referred to as Multimodal Recurrent Neural Network (MRNN) [30, 48, 13]. Study showed that MRNN achieve better results when measured by an automatic metric (BLUE [54]) but tend to reproduce previously seen captions from the training set, while ME LM generate most novel captions. Authors additionally performed human evaluations which showed disparity between human and automatic scores, a known problem when evaluating automatic generated captions [26] (more on this problem in section IV.). The best results in terms of human judgments were obtained by ME LM that leverages scores from a DMSM [17].

Spatial attention. The first one to employ attention mechanism on the task of image captioning was the work of [74]. Inspired by similar approaches that already demonstrated a substantial contribution to a better performance on tasks such as image classification [53] and multiple object recognition [4], or in the context of machine translation [3] where a soft-attention mechanism was proposed, authors in [74] introduce two variants of extensions to the simple encoder-decoder model for image captioning: a soft-attention mechanism, trainable by back-propagation [39], and a hard-attention mechanism, trainable by maximizing a variational lower bound by REINFORCE learning rule [72] (Figure 5).

As authors point out, one drawback of using image representations from the top fully connected layer of a CNN is in losing some subtle details which in turn could help in the generation of more expressive and human-like captions. Instead, they extract image features from a lower convolutional layer as a set of annotation vectors that summarize a pre-defined spatial location of the image. Each annotation vector is assigned a positive weight that is computed by an attention model (a multilayer perceptron). These weights can be interpreted as probabilities of being attended by the decoder when generating next word (hard, stochastic attention mechanism) or as the relative importance to give to a location (soft, deterministic attention mechanism) [74]. After obtaining attention weights, attention mechanism computes the context vector as a dynamic representation of the relevant part of the image input at individual time step. Experiments showed that both models achieve comparable state-of-the-art results by an approach that is more flexible than [30] in that the model can attend even to “non-object” salient regions. Additional benefit is the ability to visualize the attention weights associated with the word emitted at the same time step which enables us to gain an intuition how the attention is shifting when generating individual words (Figure 5) and why some errors were made.

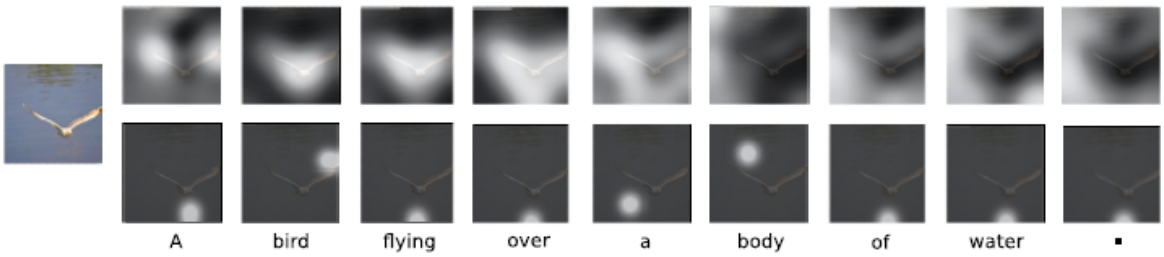

Figure 5. An attention-based model is able to attend to the most salient parts of the input while generating next word of the output. Showed are examples produced by a soft attention (top row) and a hard attention (bottom row) mechanism [74].

Semantic attention. Different from [74], where attention is modeled spatially so that pre-trained features and attention weights correspond to particular parts of the image, in [78] authors propose a novel semantic attention model on the notion that only the most semantically important parts are mentioned by people describing an image. Authors define semantic attention as the "ability to provide a detailed, coherent description of semantically important objects that are needed exactly when they are needed". By combining different sources of visual information through a feedback process, semantic attention model is able to attend to fine details all while having an end-to-end trainable system. All visual features are fed to RNN. Top-down features, extracted from the last convolutional layer of a CNN, and input only once to inform the RNN of the image content, serve as a guide where and when to attend. A set of bottom-up attributes are detected as candidates for attention. Those with highest attention scores are then used by the attention mechanism which learns to attend to semantically important concepts. Since irrelevant attributes may redirect attention to wrong concepts, attribute prediction plays a crucial role.

Similar in approach is [2] where authors combine top-down and bottom-up attention processing to calculate attention on the object-level. Instead of treating detected objects as bag-of-words that don't retain spatial information, they propose a different, feature-based approach. Bottom-up attention mechanism, based on Faster R-CNN [59], proposes a set of salient image regions represented by feature vectors indicating that some concepts belong to the same object. Combined with the more traditional top-down approach, this allows the structure of the scene to be better uncovered.

Adaptive attention. Spatial attention models have a limitation in that they cannot selectively decide whether they need to attend to the image. In [46] authors argue that attending to the image at every time step becomes unnecessary for words that don't have a corresponding visual signal such as "a ", "for" etc. They introduce an adaptive attention encode-decoder framework that, while generating next word in the caption, automatically decides whether to attend to the image or to rely solely on the language model. An LSTM extension, called sentinel gate, produces an additional visual sentinel vector extracted from linguistic information stored in decoder's memory which is then used when the model decides not to attend to the image. To be able to determine how much information should be drawn from the image and how much from what's already stored in the memory, the new adaptive context vector is modeled as a combination of the context vector of the spatial attention model and the visual sentinel vector.

III. DATA

The development of a research field greatly benefits from the availability of large datasets, which in addition to its size, should also be appropriate in terms of quality and suitability for a particular task. Large-scale datasets, such as ImageNet [60] with more than 14 millions of annotated images, organized into 22k categories according to WordNet [51] hierarchies, have already helped move the boundaries of some Computer Vision subfields and similar trends are being observed in dealing with other vision & language problems as well.

Since most models are supervised, datasets for image captioning consists of image-caption pairs. Unlike some earlier approaches [37, 75], in [17] authors showed the benefits of directly using captions in training.

A. Collecting datasets

Images are collected primarily from photo-sharing services, mostly Flickr[3]. For this purpose, some authors retrieve them by issuing specific queries [43, 58, 77], while other augment existing datasets [58, 57, 35]. However, obtaining appropriate image descriptions turned out to be much more challenging.

As [26] point out, captions provided by users of photo-sharing websites are not suitable for the training of automatic image captioning systems. Such captions usually provide context, i.e. additional information that cannot be obtained by the image alone. Describing an image with a sentence like "a woman standing in front of a tall building" would not be perceived by a human as providing valuable information. When people describe something to other people, they usually avoid mentioning the obvious [26]. But for automatic systems it is much more appropriate than just saying "my sister yesterday in Paris". Even if the algorithm could learn to recognize, just by looking at enough photos of the Eiffel Tower, that the depicted woman is actually in Paris, certainly it would not be able to recognize, just by looking at the same picture, that she is someone's sister or that the photo was taken "yesterday". And this is something that people would not be able to do either. If it is not obvious for humans, how could it be for a machine? [58]

Instead of using non-visual descriptions, [26] suggest to focus on general conceptual descriptions, i.e. those that refer to objects, attributes, events and other literal content of the image. Such descriptions are collected on a large-scale through crowdsourcing services, such as Amazon Mechanical Turk (AMT) [58, 26, 9] which involves defining a task that is performed by untrained workers

3 https://www.flickr.com

[8]. Due to the low cost and high speed, this became the preferred way of collecting image descriptions at scale. However, given the lack of control over who can participate, which in turn can negatively affect the quality of the collected descriptions, [58] suggest the use of qualification tests.

B. Datasets

UIUC PASCAL Sentences [58] was one of the first image-caption datasets, consisting of 1,000 images randomly selected from the PASCAL 2008-VOC dataset [16] and associated with five different descriptions collected via crowdsourcing. It was used by early image captioning systems [18, 37, 52, 75], but due to its limited domain, small size, and relatively simple captions it is not used anymore.

Flickr 8k [26, 58] is a larger and more diverse dataset consisting of 8,092 images collected from Flickr and focusing on people or animals performing some action. Five different captions per image, describing depicted entities and events, were collected via crowdsourcing.

Flickr 30k [77] includes and extends previous Flickr 8k dataset. 31,783 images of everyday activities, events and scenes are described by 158,915 captions obtained via crowdsourcing.

Microsoft COCO Captions [9] datasets extend the Microsoft COCO dataset [43] consisting of images of complex everyday scenes and common objects in their natural context. By the addition of human generated captions, two datasets were created. MS COCO c5 contains five captions for every of the more than 300k images in the MS COCO dataset and, since it was observed [68] that some evaluation metrics benefit from more reference captions, an additional, MS COCO c40 dataset was created by randomly choosing 5,000 images and annotating them with 40 different captions.

Flickr30k Entities [57] augments the 158k captions from the Flickr 30k dataset with 244k coreference chains and 275k bounding boxes, linking mentions of the same entities across captions and grounding those entities in image regions.

Visual Genome (VG) [35] is a novel, region captions dataset consisting of 94k images taken from the intersection of MS COCO and YFCC100M [66] datasets, along with 43.5 crowdsourced annotations per image. VG tries to overcome limitations of the previously described datasets in emphasizing only the most prominent aspects of the image. Since real-world scenes are complex, in VG great importance is assigned to attributes and relationships among objects in the scene. A useful consequence is that each image can be displayed as a scene graph [28].

Flickr 30k and MS COCO Captions are widely adopted as benchmark datasets for image captioning by most models employing deep neural networks. A survey of some earlier datasets is provided in [20], along with some quality criteria for evaluating and analyzing them.

IV. Evaluation

Correctly describing an image requires: (1) summarizing its salient content in terms of objects, attributes, relations along with deducing what is novel or interesting [17], (2) expressing this semantic content with properly formed sentences [69] that are also appropriate for the image they describe [26].

When such descriptions need to be evaluated, certain problems arise. It is clear that by placing the emphasis on one or the other aspect, the resulting sentences may vary considerably while at the same time being perfectly correct. Two captions can quite differently express the same content or in contrary, they can share most of the words and convey completely different meaning, making the evaluation of novel sentences generated by image captioning systems challenging, something about which many authors agree [68, 1].

Evaluation of novel captions [12, 13, 17, 30, 33, 34, 46, 48, 70, 74, 78] can be performed by human subjects, either by experts [26] or by untrained workers through crowdsourcing platforms [17, 13, 45]. However, human-based evaluations are associated with additional costs, they are slow and difficult to reproduce [26, 32]. A better alternative would be the use of automatic metrics, which, in turn, are fast, accurate and inexpensive [1]. Additionally, they should satisfy two criteria [45]: (1) captions that are considered good by humans should achieve high scores, (2) captions that achieve high scores should be considered good by humans. A goal that has proved to be difficult to achieve.

Image captioning sometimes is compared [14] to translating an image to its description [37, 69] or as summarizing the content of the image [75], which motivated the adoption of automatic metrics originally developed for the evaluation of models build for other tasks [54, 11, 42]. All these metrics output a score indicating a similarity between the candidate sentence and reference sentences.

BLEU (BiLingual Evaluation Understudy) [54] is a popular metric for machine translation evaluation and one of the first metrics used to evaluate image descriptions. It computes the geometric mean of n-gram (1 to 4-gram) precision scores multiplied by a brevity penalty in order to avoid overly short sentences.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) [11] is another machine translation metric. It relies on the use of stemmers, WordNet [51] synonyms and paraphrase tables to identify matches between candidate sentence and reference sentences. For the aligned sentence pair, an F-measure is calculated along with a fragmentation penalty which accounts for gaps and differences in word order.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [42] is a package of measures originally developed for the evaluation of text summaries. For the purpose of image captioning, a variant ROUGEL is usually used, which computes F-measure based on the Longest Common Subsequence (LCS) i.e. a set of words shared by two sentences which occur in the same order, without requiring consecutive matches.

CIDEr (Consensus-based Image Description Evaluation) [68] is a metric specifically designed for the evaluation of automatic generated image captions. It measures similarity between the candidate sentence and a set of human-written sentences by performing a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram. A preferred variant is CIDEr-D, since it is more robust to gaming, a situation that

occurs when a caption scored high by an automatic metric receives low scores when judged by humans.

SPICE (Semantic Propositional Image Caption Evaluation) [1] is another metric designed for image caption evaluation. It measures the quality of generated captions by computing an F-measure based on the semantic propositional content of candidate and reference sentences represented as scene graphs [28]. An advantage of this metric is that it can detect models that understand colors or those that can count.

The aforementioned evaluation metrics represent a standard set of metrics usually reported in papers, although SPICE to a lesser extent since it is relatively novel. Their popularity is partly the result of the fact that they are available through the Microsoft COCO caption evaluation server [9], which was built for the first Microsoft COCO 2015 Captioning Challenge and is still available, enabling a consistent comparison of different models using an uniform implementation of the specified metrics.

However, it was shown in [14] and [26] that automatic metrics don't always correlate with human judgments, something that was particularly evident during Microsoft COCO 2015 Captioning Challenge. Some models outperformed human upper bound according to automatic metrics, but human judges demonstrated preference for human-written captions [70]. As stated in [68] "humans don't always like what is human-like".

The authors in [32] report some additional problems with automatic metrics. They found that replacing some words with their synonyms causes the scores of all metrics to decrease. Similarly, word ordering matters for BLEU, ROUGE and CIDEr, but SPICE scores remain unaffected by changes in word order.

Since there is no best metric, some authors [1, 32] advise the use of an ensemble of metrics capturing various dimensions, such as grammaticality, saliency, correctness/truthfulness. Others propose novel metrics. One potentially interesting is **SPIDEr** [45] as a linear combination of CIDEr and SPICE, capturing the best of both worlds. The SPICE score ensures that captions capture semantic content of the image, while CIDEr score ensures their syntactic correctness.

Evaluation of ranking based systems [26, 31, 63] is performed directly on existing, human-written captions by evaluating how well the system ranks the caption of a test image over the captions of all other test images [26]. Recall@k and median rank are usual metrics employed in this setting. Some generative models can also be optimized for ranking [13, 30, 33, 34, 48].

V. OPEN PROBLEMS

A modestly set goal in [45] to generate image descriptions that would be judged by humans as "not bad", suggests that even the state-of-the-art models are still far from being perfect. Apart from previously mentioned issues with the evaluation process, there are other open problems that still need to be addressed in the future research. Some of them are inherent to the analyzed models, others are caused by external factors, such as datasets used. Authors in [33] report problems detecting colors or clothing. Similarly, models proposed in [74] were not able to recognize texture or fine-grained categories. Moreover, authors reported problems with counting. Those and other similar errors motivated authors in [67] to conduct a detailed error analysis. They found that 80% of the analyzed captions contained some errors (on average 1.56) and that 26% of generated captions were unrelated to the image.

Figure 5. Examples of generated captions, without errors (left) and with some minor errors (right) [69].

The authors in [69] investigated problems from a different perspective. They studied the effect of transferring a model to a new dataset, which resulted in the decrease of BLEU scores. In [12] authors found their models reproducing captions from the training set, suggesting lack of diversity in the training data. A similar problem was also observed in [69] where the system reproduced training captions 80% of the time.

VI. CONCLUSION

This paper presents an overview of recent advances in image captioning research, with a special focus on models employing deep encoder-decoder architectures. Main advantage of such architectures is in that they are trainable end-to-end, mapping directly from images to sentences. An important extension of the basic encoder-decoder framework is the attention mechanism, which enables to focus on the most salient parts of the input while generating the next word of the output. In this paper spatial, semantic and adaptive attention mechanisms are described. Large vision & language datasets have also contributed significantly to the development of the field. Additional features provided by novel datasets, such as coreference chains or image region captions, will certainly stimulate even faster advances in the periods to come. One important area that still remains an open problem is the evaluation of generated captions. While new evaluation metrics are being proposed, their adoption will depend on their availability through evaluation servers.

Most of the literature deals with models that generate image descriptions in the English language, emphasized by the fact that the descriptions used in training and for benchmarking are also in the English language. Our future research will be focused on developing models adapted to the generation of captions in other languages, primarily in Croatian, while simultaneously addressing some of the aforementioned issues. One direction that will be explored is the use of reinforcement learning techniques as an extension to the general encoder-decoder model. Such systems should rely less on paired training data.

REFERENCES

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould,

"Spice: Semantic propositional image caption evaluation," in *Computer Vision - ECCV 2016*, 2016, pp. 382–398.

[2] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and vqa," *arXiv preprint arXiv:1707.07998*, 2017.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[4] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[5] K. Barnard, P. Duygulu, D. Forsyth, N. D. Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[6] K. Barnard, M. Johnson, and D. Forsyth, "Word sense disambiguation with pictures," in *Proceedings of the HLT-NAACL 2003 workshop on Learning Word Meaning from Non-Linguistic Data*, 2003, vol. 6, pp. 1–5.

[7] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[8] R. Bernardi *et al.*, "Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures," *J. Artif. Intell. Res.(JAIR)*, vol. 55, pp. 409–442, 2016.

[9] X. Chen *et al.*, "Microsoft COCO captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[10] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[11] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth Workshop on Statistical Machine Translation*, 2014, pp. 376–380.

[12] J. Devlin *et al.*, "Language models for image captioning: The quirks and what works," *arXiv preprint arXiv:1505.01809*, 2015.

[13] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[14] D. Elliot and F. Keller, "Comparing automatic evaluation measures for image description," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers*, 2014, pp. 452–457.

[15] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

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

[17] H. Fang *et al.*, "From captions to visual concepts and back," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1473–1482.

[18] A. Farhadi *et al.*, "Every picture tells a story: Generating sentences from images," in *European Conference on Computer Vision*, 2010, pp. 15–29.

[19] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?," *Journal of Vision*, vol. 7, no. 1, pp. 1–29, 2007.

[20] F. Ferraro, N. Mostafazadeh, L. Vanderwende, J. Devlin, M. Galley, and M. Mitchell, "A survey of current datasets for vision and language research," *arXiv preprint arXiv:1506.06833*, 2015.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[22] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[23] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, 2017.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*, 2014, pp. 346–361.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

[27] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[28] J. Johnson *et al.*, "Image retrieval using scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3668–3678.

[29] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 2342–2350.

[30] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[31] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in Neural Information Processing Systems*, 2014, pp. 1889–1897.

[32] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," *arXiv preprint arXiv:1612.07600*, 2016.

[33] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 595–603.

[34] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[35] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012, vol. 1, pp. 1097–1105.

[37] G. Kulkarni *et al.*, "Babytalk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[38] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "TREETALK: Composition and Compression of Trees for Image Descriptions," *TACL*, vol. 2, no. 10, pp. 351–362, 2014.

[39] Y. A. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*, Springer Berlin Heidelberg, 2012, pp. 9–48.

[40] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The Handbook of Brain Theory and Neural Networks*, vol. 3361, no. 10, 1995.

[41] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[42] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop*, 2004, vol. 8.

[43] T. Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*,

2014, pp. 740–755.

[44] S. Li, G. Kulkarni, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *In Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 2011, pp. 220–228.

[45] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved Image Captioning via Policy Gradient Optimization of SPIDEr," *arXiv preprint arXiv:1612.00370*, 2016.

[46] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via A visual sentinel for image captioning," *arXiv preprint arXiv:1612.01887*, 2016.

[47] M. Malinovski, M. Rohrbach, and M. Fritz, "Ask your neurons: A deep learning approach to visual question answering," *arXiv preprint arXiv:1605.02697*, 2016.

[48] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[49] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille, "Explain images with multimodal recurrent neural networks," *arXiv preprint arXiv:1410.1090*, 2014.

[50] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[51] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: An on-line lexical database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[52] M. Mitchell *et al.*, "Midge: Generating image descriptions from computer vision detections," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 747–756.

[53] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014, vol. 2, pp. 2204–2212.

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

[55] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, 2013, pp. 1310–1318.

[56] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[57] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2641–2649.

[58] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's Mechanical Turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 139–147.

[59] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[60] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[61] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[63] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.

[64] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014, vol. 2, pp. 3104–3112.

[65] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–19.

[66] B. Thomee *et al.*, "YFCC100M: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[67] E. van Miltenburg and D. Elliot, "Room for improvement in automatic image description: an error analysis," *arXiv preprint arXiv:1704.04198*, 2017.

[68] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

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

[70] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 652–663, 2017.

[71] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: learning to rank with joint word-image embeddings," *Machine Learning*, vol. 81, no. 1, pp. 21–35, 2010.

[72] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3–4, pp. 229–256, 1992.

[73] T. Winograd, "Understanding natural language," *Cognitive Psycology*, vol. 3, no. 1, pp. 1–191, 1972.

[74] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[75] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 444–454.

[76] L. Yao *et al.*, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4507–4515.

[77] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

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

[79] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014, pp. 818–833.