

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/353284955>

From Show to Tell: A Survey on Image Captioning

Preprint · July 2021

CITATIONS

0

READS

509

6 authors, including:



Matteo Stefanini

Università degli Studi di Modena e Reggio Emilia

13 PUBLICATIONS 309 CITATIONS

[SEE PROFILE](#)



Marcella Cornia

Università degli Studi di Modena e Reggio Emilia

62 PUBLICATIONS 1,462 CITATIONS

[SEE PROFILE](#)



Lorenzo Baraldi

Università degli Studi di Modena e Reggio Emilia

102 PUBLICATIONS 2,214 CITATIONS

[SEE PROFILE](#)



Silvia Cascianelli

Università degli Studi di Modena e Reggio Emilia

38 PUBLICATIONS 242 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Visual Saliency Prediction [View project](#)



Scene Understanding for HRI [View project](#)

From Show to Tell: A Survey on Image Captioning

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli,
Giuseppe Fiameni, and Rita Cucchiara

Abstract—Connecting Vision and Language plays an essential role in Generative Intelligence. For this reason, large research efforts have been devoted to image captioning, *i.e.* describing images with syntactically and semantically meaningful sentences. Starting from 2015 the task has generally been addressed with pipelines composed of a visual encoder and a language model for text generation. During these years, both components have evolved considerably through the exploitation of object regions, attributes, and relationships and the introduction of multi-modal connections, fully-attentive approaches, and BERT-like early-fusion strategies. However, regardless of the impressive results, research in image captioning has not reached a conclusive answer yet. This work aims at providing a comprehensive overview of image captioning approaches, from visual encoding and text generation to training strategies, datasets, and evaluation metrics. In this respect, we quantitatively compare many relevant state-of-the-art approaches to identify the most impactful technical innovations in image captioning architectures and training strategies. Moreover, many variants of the problem and its open challenges are discussed. The final goal of this work is to serve as a tool for understanding the existing literature and highlighting the future directions for a research area where Computer Vision and Natural Language Processing can find an optimal synergy.

Index Terms—Image Captioning, Vision-and-Language, Deep Learning, Survey.

1 INTRODUCTION

IMAGE captioning is the task of describing the visual content of an image in natural language, employing a visual understanding system and a language model capable of generating meaningful and syntactically correct sentences. Neuroscience research has clarified the link between human vision and language generation only in the last few years [1]. Similarly, in Artificial Intelligence, the design of architectures capable of processing images and generating language is a very recent matter. The goal of these research efforts is to find the most effective pipeline to process an input image, represent its content, and transform that into a sequence of words by generating connections between visual and textual elements while maintaining the fluency of language. In its standard configuration, image captioning is an image-to-sequence problem whose inputs are pixels. These are encoded as one or multiple feature vectors in the visual encoding step, which prepares the input for a second generative step, called the language model. This produces a sequence of words or sub-words decoded according to a given vocabulary. In these few years, the research community improved the models considerably: from the first deep learning-based proposals adopting Recurrent Neural Networks (RNNs) fed with global image descriptor extracted by a Convolutional Neural Network (CNN), methods have been enriched with attentive approaches and reinforcement learning, up to the breakthroughs of Transformers and self-attention to single-stream BERT-like approaches. At the

same time, the Computer Vision and Natural Language Processing (NLP) communities have addressed the challenge of building proper evaluation protocols and evaluation metrics to compare results with human-generated ground-truths. Moreover, several domain-specific scenarios and variants of the task have been investigated. However, the achieved results are still far from setting an ultimate solution. With the aim of providing a testament to the journey that captioning has taken so far, and with that of encouraging novel ideas, in this paper, we trace a holistic overview of the models developed in the last years.

Following the inherent dual nature of captioning models, we develop a taxonomy of both the visual encoding and the language modeling approaches, focusing on their key aspects and limitations. We also focus on the training strategies followed in the literature over the past years, from cross-entropy loss to reinforcement learning and the recent advancement obtained by the pre-training paradigm. Furthermore, we review the main datasets used to explore image captioning, from domain-generic benchmarks to domain-specific datasets collected to investigate specific aspects of the problem and analyze standard and non-standard metrics adopted for performance evaluation, which capture different aspects of the produced caption quality. An additional contribution of this work is a quantitative comparison of main image captioning methods, which considers both standard and non-standard metrics, and a discussion on their relationships, which sheds light on performance, differences, and characteristics of the most important models. Finally, we give an overview of many variants of the problem and discuss some open challenges and future directions.

- M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, and R. Cucchiara are with the Department of Engineering “Enzo Ferrari”, University of Modena and Reggio Emilia, Modena, Italy.
E-mail: {matteo.stefanini, marcella.cornia, lorenzo.baraldi, silvia.cascianelli, rita.cucchiara}@unimore.it.
- G. Fiameni is with NVIDIA AI Technology Centre, Italy.
E-mail: gfiameni@nvidia.com

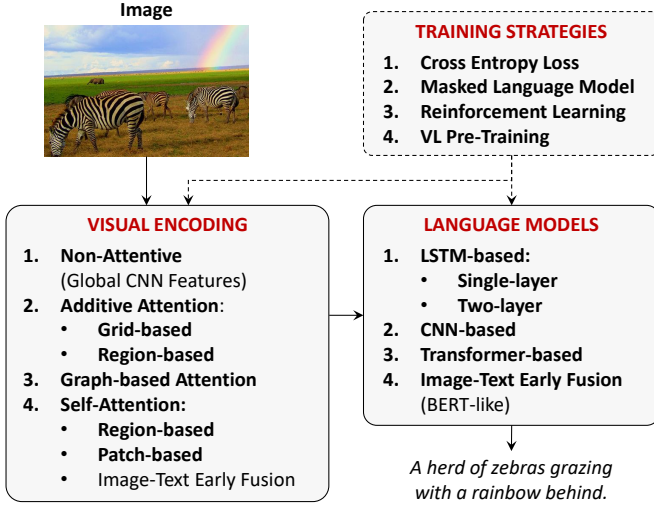


Fig. 1: Overview of the image captioning task and a simple taxonomy of the most relevant approaches.

2 VISUAL ENCODING

Providing an effective representation of the visual content is the first challenge of an image captioning pipeline. Excluding the earliest image captioning works [2], [3], [4], [5], [6], [7], [8], [9], [10], we focus on deep learning-based solutions.

The current approaches of visual encoding can be classified as belonging to four main categories: 1. *non-attentive methods* based on global CNN features; 2. *additive attentive methods* that embed the visual content using either grids or regions; 3. *graph-based methods* adding visual relationships between visual regions; and 4. *self-attentive methods* that employ Transformer-based paradigms, either by using region-based, patch-based, or image-text early fusion solutions. This taxonomy is visually summarized in Fig. 1.

2.1 Global CNN Features

With the advent of CNNs, all models consuming visual inputs have been improved in terms of performance. The visual encoding step of image captioning is no exception. In the most simple recipe, the activation of one of the last layers of a CNN is employed to extract high-level and fixed-sized representations, which are then used as a conditioning element for the language model (Fig. 2a). This is the approach employed in the seminal paper “Show and Tell” [11]¹, where the output of a GoogleNet [12] pre-trained on ImageNet [13] is fed to the initial hidden state of the language model. In the same year, Karpathy *et al.* [14] used global features extracted from AlexNet [15] as the input for a language model. Further, Mao *et al.* [16] and Donahue *et al.* [17] injected global features extracted from the VGG network [18] at each time-step of the language model.

Global CNN features were then employed in a large variety of image captioning models [19], [20], [21], [22], [23], [24], [25], [26]. Notably, Rennie *et al.* [27] introduced the FC model, in which images are encoded using a ResNet-101 [28], preserving their original dimensions. Other approaches [29], [30] integrated high-level attributes or tags,

represented as a probability distribution over the most common words of the training captions.

The main advantage of employing global CNN features resides in their simplicity and compactness of representation, which embraces the capacity to extract and condense information from the whole input and considering the overall context of an image. However, this paradigm also leads to excessive compression of information and lacks granularity: all salient objects and regions are fused in a single vector, making it hard for a captioning model to produce specific and fine-grained descriptions.

2.2 Attention Over Grid of CNN Features

Motivated by the drawbacks of global representations, most of the following approaches have increased the granularity level of visual encoding [27], [31], [32] (Fig. 2b). Drawing from machine translation, the additive attention mechanism has demonstrated remarkable performance in a wide range of tasks and has endowed image captioning architectures with time-varying visual features encoding, enabling greater flexibility and finer granularity.

Definition of additive attention. The intuition behind attention boils down to weighted averaging. In the first formulation proposed for sequence alignment by Bahdanau *et al.* [33] (also known as *additive attention*), a single-layer feed-forward neural network with a hyperbolic tangent non-linearity is used to compute attention weights. Formally, given two generic sets of vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ and $\{\mathbf{h}_1, \dots, \mathbf{h}_m\}$, the additive attention score between the \mathbf{h}_i ’s and the \mathbf{x}_j ’s is computed as follows:

$$f_{\text{att}}(\mathbf{h}_i, \mathbf{x}_j) = \mathbf{W}_3^\top \tanh(\mathbf{W}_1 \mathbf{h}_i + \mathbf{W}_2 \mathbf{x}_j), \quad (1)$$

where \mathbf{W}_1 and \mathbf{W}_2 are weight matrices, and \mathbf{W}_3 is a weight vector that performs a linear combination. A softmax function is then applied to obtain a probability distribution $p(\mathbf{x}_j | \mathbf{h}_i)$, representing how much the element encoded by \mathbf{x}_j is relevant for \mathbf{h}_i .

Although the attention mechanism was initially devised for modeling the relationships between two sequences of elements (*i.e.* hidden states from a recurrent encoder and a decoder), it can be adapted to connect a set of fine-grained visual representations with the hidden states of a language model. From the point of view of visual features extraction, a global representation may lead to sub-optimal results due to noisy contexts or irrelevant regions while generating a specific word. Employing an attention mechanism over a set of features, instead, allows retaining richer information useful for a more comprehensive sentence generation.

Attending convolutional activations. Xu *et al.* [31] introduced the first method leveraging the additive attention over the spatial output grid of a convolutional layer. This allows the model to selectively focus on certain elements of the grid by selecting a subset of features for each generated word. Specifically, the model first extracts the activation of the last convolutional layer of a VGG network [18], then uses additive attention to compute a weight for each grid element, interpreted as the relative importance of that element for generating the next word.

1. Actually, the title of this survey is a tribute of this pioneering work.

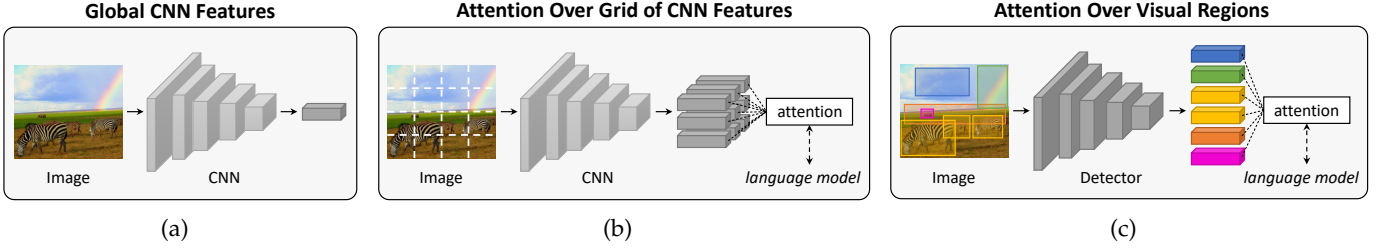


Fig. 2: Three of the most relevant visual encoding strategies for image captioning: **(a)** global CNN features; **(b)** fine-grained features extracted from the activation of a convolutional layer, together with an attention mechanism guided by a query vector produced by the language model; **(c)** pooled convolutional features from image regions, together with an attention mechanism.

Other approaches. The solution based on additive attention over a grid of features has been widely adopted by several following works with minor improvements in terms of visual encoding [29], [32], [34], [35], [36], [37].

Review networks – For instance, Yang *et al.* [38] supplemented the encoder-decoder framework with a recurrent review network. This performs a given number of review steps with attention on the encoder hidden states and outputs a “thought vector” after each step, which is then used by the attention mechanism in the decoder.

Multi-level features – Chen *et al.* [39] proposed to employ channel-wise attention over convolutional activations, followed by a more classical spatial attention. They also experimented with using more than one convolutional layer to exploit multi-level features. On the same line, Jiang *et al.* [40] proposed to use multiple CNNs in order to exploit their complementary information, then fused their representations with a recurrent procedure.

Exploiting human attention – Some works also integrated saliency information (*i.e.* what do humans pay more attention to in a scene) to guide caption generation. This idea was first proposed by Sugano and Bulling [41] who exploited human eye-fixation information for image captioning by including normalized fixation histograms over the image as an input to the soft-attention module of [31] and weighing the attended image regions based on whether these are fixated or not. Subsequent works on this line [42], [43], [44] used predicted saliency information in place of eye-fixation information.

2.3 Attention Over Visual Regions

Although the intuition of using saliency boils down to neuroscience, the same discipline suggests that our brain constantly integrates a top-down reasoning process with a bottom-up flow of visual input signals. The top-down path consists of predicting the upcoming sensory input by leveraging our knowledge and inductive bias. On the other side, the bottom-up flow constantly provides visual stimuli adjusting the previous predictions, passing from input signals to their interpretation. The captioning models mentioned so far exploit an attention mechanism that can be thought of as a top-down system. In this mechanism, the language model predicts the next word based on its learned assumptions while attending a feature grid, whose geometry is irrespective of the image content.

Bottom-up and top-down attention. The solution proposed by Anderson *et al.* [45] entails integrating an additional bottom-up path, defined by an object detector in charge of proposing image regions, coupled with the top-down mechanism that learns to weigh each region for each word prediction (see Fig. 2c). In this approach, Faster R-CNN [46], [47] is adopted to detect objects in two stages: the first, called Region Proposal Network, produces object proposals rolling over intermediate features of a CNN; the second operates a pooling of the region of interest to extract a feature vector for each proposal. One of the key elements of this approach resides in its pre-training strategy, where an auxiliary training loss is added for learning to predict attribute classes alongside object classes on the Visual Genome [48] dataset. This allows the model to predict a dense and rich set of detections, including both salient object and contextual regions and favors the learning of better feature representations.

Other approaches. Employing pooled vectors from image regions has demonstrated its advantages when dealing with the raw visual input and has been the standard de-facto in image captioning for years. As a result, many of the following works have based the visual encoding phase on this strategy [49], [50], [51], [52]. Among them, we point out two remarkable variants.

Visual Policy – While typical visual attention points to a single image region at every step, the approach proposed by Zha *et al.* [53] introduces a sub-policy network that interprets also the visual part sequentially by encoding historical visual actions (*e.g.* previously attended regions) via an LSTM to serve as context for the next visual action.

Geometric Transforms – Pedersoli *et al.* [54] proposed to use spatial transformers for generating image-specific attention areas by regressing region proposals in a weakly-supervised fashion (relying on the captioning training loss only). Specifically, a localization network learns an affine transformation or each location of the feature map, and then a bilinear interpolation is used to regress a feature vector for each region with respect to anchor boxes.

2.4 Graph-based Encoding

To further improve the encoding of image regions and their relationships, some studies consider using graphs built over image regions (see Fig. 3a) to enrich the representation by including semantic and spatial connections.

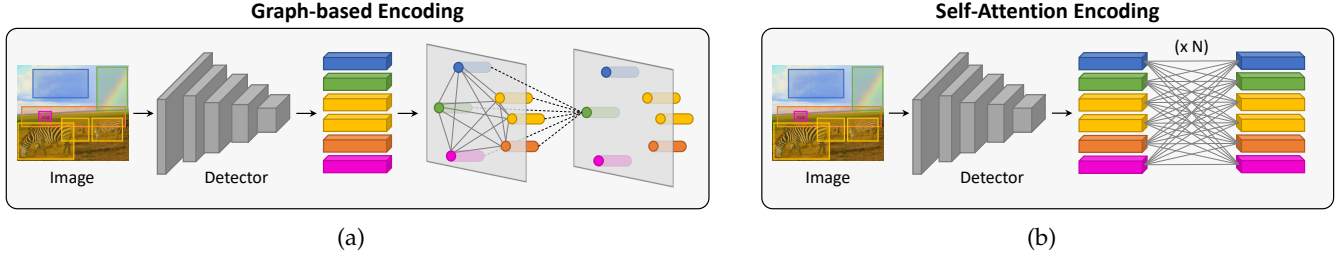


Fig. 3: Summary of the two most recent visual encoding strategies for image captioning: (a) graph-based encoding of visual regions, where each region is encoded through convolutions over hand-designed graphs; (b) self-attention which defines a fully-connected graph between regions.

Spatial and semantic graphs. The first attempt in this sense is due to Yao *et al.* [55], followed by Guo *et al.* [56], who proposed the use of a graph convolutional network (GCN) [57] to integrate both semantic and spatial relationships between objects. The semantic relationships graph is obtained by applying a classifier pre-trained on Visual Genome [48] that predicts an action or an interaction between object pairs. The spatial relationships graph is instead inferred through geometry measures (*i.e.* intersection over union, relative distance, and angle) between bounding boxes of object pairs.

Scene graphs. With a focus on modeling semantic relations, Yang *et al.* [58] proposed to integrate semantic priors learned from text in the image encoding by exploiting a graph-based representation of both images and sentences. The representation used is the scene graph, *i.e.* a directed graph connecting the objects, their attributes, and their mutual relations. Along the same line, Shi *et al.* [59] represented the image as a semantic relationship graph but proposed to train the module in charge of predicting the predicate nodes directly on the ground-truth captions rather than on external datasets [48]. The obtained graph is then fed to a GCN for encoding and is also exploited at the decoding stage.

Hierarchical trees. As a special case of a graph-based encoding, Yao *et al.* [60] employed a tree to represent the image as a hierarchical structure. The root represents the image as a whole, intermediate nodes represent image regions and their contained sub-regions, and the leaves represent segmented objects in the regions. The image encoding is then obtained by feeding the image tree to a TreeLSTM [61].

Graph encodings brought a mechanism to leverage relationships between detected objects, which allows the exchange of information in adjacent nodes and thus in a local manner. Further, it seamlessly allows the integration of external semantic information. On the other hand, manually building the graph structure can limit the interactions between visual features. This is where self-attention proved to be more successful by connecting all the elements with each other in a complete graph representation.

2.5 Self-Attention Encoding

Self-attention is an attentive mechanism where each element of a set is connected with all the others, and that can be adopted to compute a refined representation of the same set of elements through residual connections (Fig. 3b). It was first introduced in 2017 by Vaswani *et al.* [62] for machine translation and language understanding tasks, giving birth

to the Transformer architecture and its subsequent variants, which have dominated the NLP field and later also Computer Vision.

Definition of self-attention. Formally, self-attention makes use of the scaled dot-product mechanism, *i.e.* a multiplicative attention operator that handles three sets of vectors: a set of n_q query vectors Q , a set of key vectors K , and a set of value vectors V , both containing n_k elements. The operator takes a weighted sum of value vectors according to a similarity distribution between query and key vectors, *i.e.*

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (2)$$

where d_k is a scaling factor. In the case of self-attention, the three sets of vectors are obtained as linear projections of the same input set of elements. The success of the Transformer demonstrates that leveraging self-attention allows achieving superior performances compared to attentive RNNs.

Early self-attention approaches. Among the first image captioning models leveraging this approach, Yang *et al.* [63] employed a self-attentive module to encode relationships between features resulting from an object detector. Later, Li *et al.* [64] proposed a Transformer model with a visual encoder for the region features coupled with a semantic encoder that exploits knowledge from an external tagger. Both encoders are based on self-attention and feed-forward layers. Their output is then fused in the decoder through a gating mechanism governing the propagation of visual and semantic information.

Variants of the self-attention operator. Other works proposed variants or modifications of the self-attention operator tailored for image captioning [65], [66], [67], [68], [69].

Geometry-aware encoding – Herdade *et al.* [65] introduced a modified version of self-attention that takes into account the spatial relationships between regions. In particular, an additional geometric weight is computed between object pairs and is used to scale the attention weights. On a similar line, Guo *et al.* [66] proposed a normalized and geometry-aware version of self-attention that makes use of the relative geometry relationships between input objects.

Attention on Attention – Huang *et al.* [67] proposed an extension of the attention operator, named “Attention on Attention”, in which the final attended information is weighted by a gate guided by the context. Specifically, they concatenate the output of the self-attention with the queries,

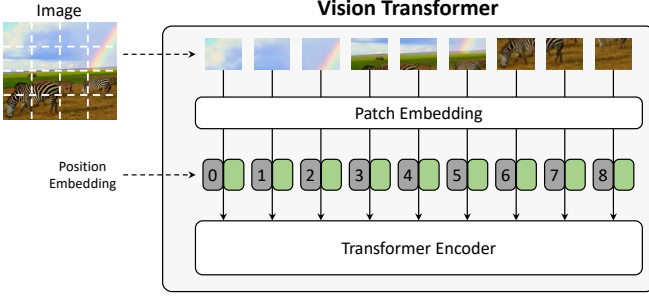


Fig. 4: Vision Transformer encoding. The image is split into fixed-size patches, linearly embedded, added to position embeddings, and fed to a standard Transformer encoder.

then compute an information vector and a gate vector that are finally multiplied together. In their visual encoder, they employ this mechanism in order to refine the visual features. This method is then adopted by later models such as [70].

X-Linear Attention – Pan *et al.* [68] proposed to use bilinear pooling techniques to strengthen the representative capacity of the output attended feature. Notably, this mechanism encodes the region-level features with higher-order interaction, leading to a set of enhanced region-level and image-level features.

Memory-augmented Attention – Cornia *et al.* [69], [71] proposed a Transformer-based architecture where the self-attention operator of each encoder layer is augmented with a set of memory vectors. Specifically, the set of keys and values is extended with additional “slots” learned during training, which can encode multi-level visual relationships with a priori knowledge.

Other self-attention-based approaches. Ji *et al.* [72] proposed to improve self-attention by adding to the sequence of feature vectors a global vector computed as their average. A global vector is computed for each layer, and the resulting global vectors are combined via an LSTM, thus obtaining an inter-layer representation. Luo *et al.* [73] proposed a hybrid approach that combines region and grid features to exploit their complementary advantages. Two self-attention modules are applied independently to each kind of features, and a cross-attention module locally fuses their interactions. Finally, the approach proposed by Zhang *et al.* [74] completely disregards region features and applies self-attention directly to grid features, incorporating their relative geometry relationships into self-attention computation.

Vision Transformer. Transformer-like architectures can also be applied directly on image patches, thus excluding or limiting the usage of the convolutional operator [75], [76] (Fig. 4). On this line, Liu *et al.* [77] devised the first convolution-free architecture for image captioning. Specifically, a pre-trained Vision Transformer network (*i.e.* ViT [75]) is adopted as encoder, and a standard Transformer decoder is employed to generate captions.

Early fusion and vision-and-language pre-training. Other works using self-attention to encode visual features achieved remarkable performance also thanks to vision-and-language pre-training [78], [79] and early-fusion strategies [80], [81]. For example, following the BERT architecture [82], Zhou *et al.* [81] combined encoder and decoder

into a single stream of Transformer layers, where region and word tokens are early fused together into a unique flow. This unified model is first pre-trained on large amounts of image-caption pairs to perform both bidirectional and sequence-to-sequence prediction tasks and then fine-tuned for image captioning. On the same line, Li *et al.* [80] proposed OSCAR, a BERT-like architecture that also includes objects tags as anchor points in order to ease the semantic alignment between images and text. These tags are extracted from an object detector and concatenated with the image regions and word embeddings fed to the model. They also performed a large-scale pre-train with 6.5 million image-text pairs, with a masked token loss similar to the BERT mask language loss and a contrastive loss for distinguishing aligned words-tags-regions triples from polluted ones. Moreover, Zhang *et al.* [83] proposed VinVL, built on top of OSCAR, introducing a new object detector capable of extracting better visual features and a modified version of the vision-and-language pre-training objectives. Specifically, the object detector presents minor changes with respect to Faster R-CNN and is pre-trained on a large corpus consisting of four public datasets. The vision-and-language pre-training objectives are the same masked token loss and a 3-way contrastive loss that takes into account two types of triples: words-tags-regions from captioning datasets and question-answer-regions from visual question answering datasets.

2.6 Discussion

Global CNN features are a simple and compact way to encode the visual information but have been proven to be insufficient. Indeed, for image captioning, the information about the visual entities in the scene, detected in image regions, is essential. Almost all the surveyed approaches adopt the same model (*i.e.* Faster R-CNN trained on Visual Genome) as the object detector backbone for its remarkable performance. Nonetheless, since the set of objects the backbone can distinguish defines what can be described in an image, applying more and more general detectors will extend the domain application of image captioning approaches. Not only distinguishing each visual entity is important, but also encoding their spatial and contextual relation has been proven to boost the captioning performance. In this sense, explicit graph-based representations, in conjunction with attentive mechanisms and, later, implicit self-attentive solutions, had more success than global representation. This fact clearly suggests that a viable direction is designing visual encoders that model mutual relations between objects. Moreover, the success of BERT-like solutions performing image and text early-fusion indicates the suitability of visual representations that also integrate textual information.

3 LANGUAGE MODELS

The primary goal of a language model is to predict the probability of a given sequence of words to occur in a sentence. As such, it represents a crucial component of many NLP tasks, as it gives a machine the ability to understand and deal with natural language as a stochastic process.

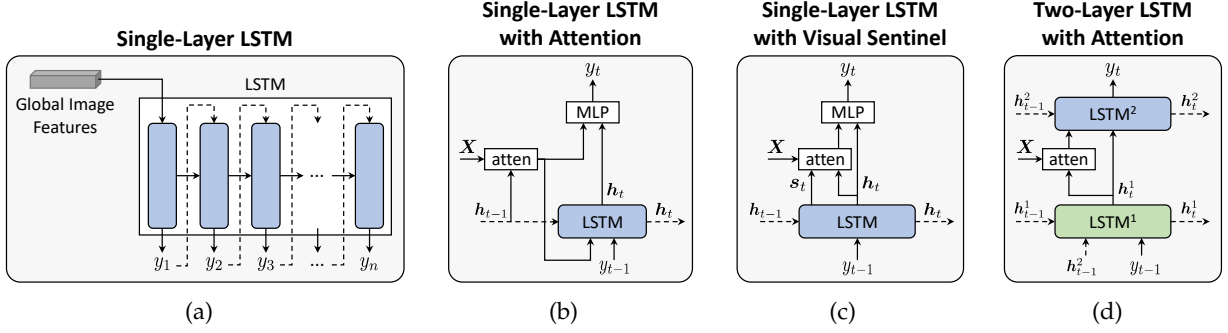


Fig. 5: LSTM-based language modeling strategies: **(a)** Single-Layer LSTM model conditioned on the visual feature; **(b)** LSTM with attention, as proposed in the Show, Attend and Tell model [31]; **(c)** LSTM with attention, in the variant proposed in [32]; **(d)** two-layer LSTM with attention, in the style of the bottom-up top-down approach by Anderson *et al.* [45]. In all figures, \mathbf{X} represents either a grid of CNN features or image region features extracted by an object detector.

Formally, given a sequence of n words, the language model component of an image captioning algorithm assigns a probability $P(y_1, y_2, \dots, y_n | \mathbf{X})$ to the sequence as:

$$P(y_1, y_2, \dots, y_n | \mathbf{X}) = \prod_{i=1}^n P(y_i | y_1, y_2, \dots, y_{i-1}, \mathbf{X}), \quad (3)$$

where \mathbf{X} represents the visual encoding on which the language model is specifically conditioned. Notably, when predicting the next word given the previous ones, the language model is auto-regressive, which means that each predicted word is conditioned on the previous ones. The language model also decides when to stop generating caption words by outputting a special end-of-sequence token.

The main language modeling strategies applied to image captioning can be categorized as: 1. *LSTM-based* approaches, which can be either single-layer or two-layer; 2. *CNN-based* methods that constitute a first attempt in surpassing the fully recurrent paradigm; 3. *Transformer-based* fully-attentive approaches; 4. *image-text early-fusion* (BERT-like) strategies that directly connect the visual and textual inputs. This taxonomy is visually summarized in Fig. 1.

3.1 LSTM-based Models

As language has a sequential structure, RNNs are naturally suited to deal with the generation of sentences. Among RNN variants, LSTM [84] has been the predominant option for language modeling.

3.1.1 Single-layer LSTM

The most simple LSTM-based captioning architecture is based on a single-layer LSTM and was proposed by Vinyals *et al.* [11]. As shown in Fig. 5a, the visual encoding is used as the initial hidden state of the LSTM, which then generates the output caption. At each time step, a word is predicted by applying a softmax activation function over the projection of the hidden state into a vector of the same size as the vocabulary. During training, input words are taken from the ground-truth sentence, while during inference, input words are those generated at the previous step.

Shortly after, with the seminal work “Show, Attend and Tell”, Xu *et al.* [31] introduced the additive attention mechanism, a dynamic and time-varying representation of the image that replaced the static global vector and improved the

alignment between words and visual content. As depicted in Fig. 5b, in this case, the previous hidden state guides the attention mechanism over the visual features \mathbf{X} , computing a context vector which is then fed to the MLP in charge of predicting the output word.

Other approaches. Many subsequent works have adopted a decoder based on a single-layer LSTM, mostly without any architectural changes [38], [39], [54], while others have proposed significant modifications, summarized below.

Visual sentinel – Lu *et al.* [32] augmented the spatial image features with an additional learnable vector, called visual sentinel, which can be attended by the decoder in place of visual features while generating “non-visual” words (e.g. “the”, “of”, and “on”), for which visual features are not needed (Fig. 5c). At each time step, the visual sentinel is computed from the previous hidden state and generated word. Then, the model generates a context vector as a combination of attended image features and visual sentinel, whose importance is weighted by a learnable gate. Many subsequent works [85], [86] confirmed the utility of this additional vector.

Hidden state reconstruction – Chen *et al.* [34] proposed to regularize the transition dynamics of the language model by using a second LSTM for reconstructing the previous hidden state based on the current one. Ge *et al.* [36] proposed to better capture context information by using a bidirectional LSTM with an auxiliary module. The auxiliary module in a direction approximates the hidden state of the LSTM in the other direction. Finally, a cross-modal attention mechanism combines grid visual features with the two sentences from the bidirectional LSTM to obtain the final caption.

Multi-stage generation – Wang *et al.* [35] proposed to generate a caption from coarse central aspects to finer attributes by decomposing the caption generation process into two phases: skeleton sentence generation and attributes enriching, both implemented with single-layer LSTMs. On the same line, Gu *et al.* [37] devised a coarse-to-fine multi-stage framework using a sequence of LSTM decoders, each operating on the output of the previous one to produce increasingly refined captions.

3.1.2 Two-layer LSTM

LSTMs can be expanded to multi-layer structures to augment their capability of capturing higher-order relations. Donahue *et al.* [17] firstly proposed a two-layer LSTM as a language model for captioning, stacking two layers, where the hidden states of the first are the input to the second.

Two-layers and additive attention. Anderson *et al.* [45] went further and proposed to specialize the two layers to perform visual attention and the actual language modeling. As shown in Fig. 5d, the first LSTM layer acts as a top-down visual attention model which takes the previously generated word, the previous hidden state, and the mean-pooled image features. Then, the current hidden state is used to compute a probability distribution over image regions with an additive attention mechanism. The so-obtained attended image feature vector is fed to the second LSTM layer, which combines it with the hidden state of the first layer to generate a probability distribution over the vocabulary.

Variants of two-layers LSTM. Because of their representation power, LSTMs with two-layers and internal attention mechanisms represent the most employed language model approach before the advent of Transformer-based architectures [55], [58], [59], [60]. As such, many other variants have been proposed to improve the performance of this approach.

Neural Baby Talk – To ground words into image regions, Lu *et al.* [85] incorporated a pointing network that modulates the content-based attention mechanism. In particular, during the generation process, the network predicts slots in the caption, which are then filled with the image region classes. For non-visual words, a visual sentinel is used as dummy grounding. This approach leverages the object detector both as a feature region extractor and as a visual word prompter for the language model.

Reflective attention – Ke *et al.* [49] introduced two reflective modules: the first computes the relevance between hidden states from all the past predicted words and the current one, thus modeling longer dependencies and fostering historical coherence. The second improves the syntactic structure of the sentence by guiding the generation process with words common position information (e.g. subjects are usually at the beginning, while predicates in the middle).

Look back and predict forward – On a similar line, Qin *et al.* [50] used two modules: the look back module that takes into account the previous attended vector to compute the next one, and the predict forward module that predicts the new two words at once, thus alleviating the accumulated errors problem that may occur at inference time.

Adaptive attention time – Huang *et al.* [51] proposed an adaptive attention time mechanism, in which the decoder can take an arbitrary number of attention steps for each generated word, determined by a confidence network on top of the second-layer LSTM.

Recall mechanisms – Wang *et al.* [52] introduced a recall mechanism modeled with a text retrieval system, which provides the model with useful words for each image. An auxiliary word distribution is obtained from the recalled words and used as a semantic guide.

3.1.3 Boosting LSTM with Self-Attention

Some works adopted the self-attention operator in place of the additive attention one in LSTM-based language mod-

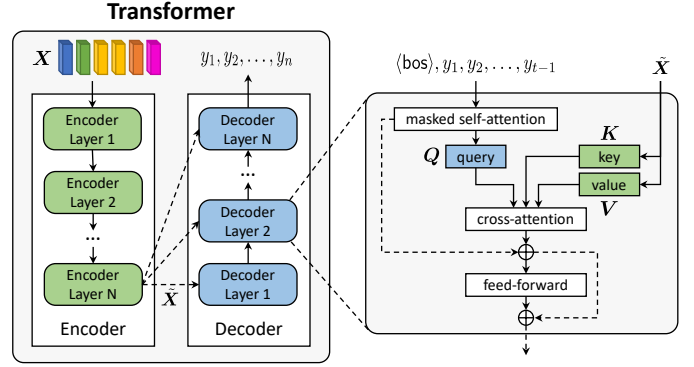


Fig. 6: Schema of the Transformer-based language model. The caption generation is performed via masked self-attention over previously generated tokens and cross-attention with encoded visual features.

els [67], [68], [70], [87]. In particular, Huang *et al.* [67] augmented the LSTM with the Attention on Attention operator, which computes another step of attention on top of visual self-attention. Pan *et al.* [68] introduced the X-Linear attention block, which enhances self-attention with second-order interactions and improves both the visual encoding and the language model.

3.1.4 Neural Architecture Search for RNN

Zhu *et al.* [87] applied the neural architecture search paradigm to select the connections between layers and the operations within gates of RNN-based image captioning language models. To evaluate their method, they considered the decoder of the X-LAN architecture [68], which includes a variant of the self-attention operator.

3.2 Convolutional Language Models

A worth-to-mention approach is that proposed by Aneya *et al.* [88], which uses convolutions as a language model. In particular, a global image feature vector is combined with word embeddings and fed to a CNN, operating on all words in parallel during training and sequentially in inference. Convolutions are right-masked to prevent the model from using the information of future word tokens. Despite the clear advantage of parallel training, the usage of the convolutional operator in language models has not gained popularity due to the poor performance and the advent of Transformer architectures.

3.3 Transformer-based Architectures

The fully-attentive paradigm proposed by Vaswani *et al.* [62] in the seminal paper “Attention is all you need” has completely changed the perspective of language generation. Shortly after, the Transformer model became the building block of other breakthroughs in NLP, such as BERT [82] and GPT [89], and the standard de-facto architecture for many language understanding tasks. As image captioning can be cast as a set-to-sequence problem when using image regions, the Transformer architecture has been employed also for this task. The standard Transformer decoder performs a masked self-attention operation, which is applied to words, followed

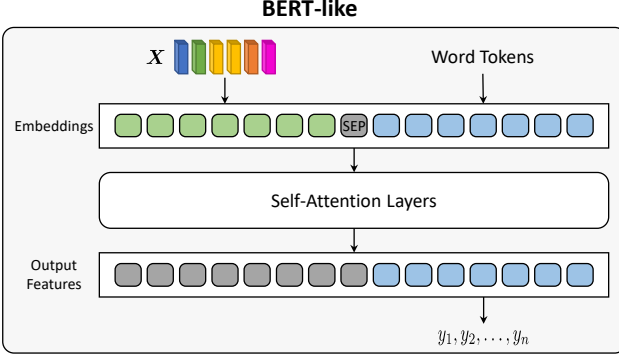


Fig. 7: Schema of a BERT-like language model. A single stream of attentive layers processes both image regions and word tokens and generates the output caption.

by a cross-attention operation, where words act as queries and the outputs of the last encoder layer act as keys and values, plus a final feed-forward network (Fig. 6). During training, a masking mechanism is applied to the previous words to constrain a unidirectional generation process. The original Transformer architecture has been employed in some image captioning models without significant architectural modifications [65], [66], [73]. Besides, some variants have been proposed to improve language generation and visual feature encoding.

Gating mechanisms. Li *et al.* [64] proposed a gating mechanism for the cross-attention operator, which controls the flow of visual and semantic information by combining and modulating image regions representations with semantic attributes coming from an external tagger. On the same line, Ji *et al.* [72] integrated a context gating mechanism to modulate the influence of the global image representation on each generated word, modeled via multi-head attention. Cornia *et al.* [69] proposed to take into account all encoding layers in place of performing cross-attention only on the last one. To this end, they devised the meshed decoder, which contains a mesh operator that modulates the contribution of all the encoding layers independently and a gate that weights these contributions guided by the text query.

3.4 BERT-like Architectures

Despite the encoder-decoder paradigm is a common approach to image captioning, some works have revisited captioning architectures to exploit a BERT-like [82] structure in which the visual and textual modalities are fused together in the early stages (Fig. 7). When employed as a language model for captioning, the main advantage of this architecture is that layers dealing with text can be initialized with pre-trained parameters learned from massive textual corpora. Therefore, the BERT paradigm has been adopted mainly in works that exploit pre-training [80], [81], [83].

In this section, we refer to a BERT-like approach when the overall architecture does not have a clear distinction between the encoding and decoding phases and when inputs coming from both modalities are processed together in a single stream made of Transformer layers.

The first example is due to Zhou *et al.* [81], who developed a unified model that fuses visual and textual modal-

ties into a BERT-like architecture for image captioning. The model consists of a shared multi-layer Transformer encoder network for both encoding and decoding, pre-trained on a large corpus of image-caption pairs and then fine-tuned for image captioning by right-masking the tokens sequence to simulate the unidirectional generation process. Further, Li *et al.* [80] introduced the usage of object tags detected in the image as anchors points for learning a better alignment in vision-and-language joint representations. To this end, their model represents an input image-text pair as a word tokens-object tags-region features triple, where the object tags are the textual classes proposed by the object detector.

3.5 Non-autoregressive Language Models

Thanks to the parallelism offered by Transformers, non-autoregressive language models have been proposed in machine translation to reduce the inference time by generating all words in parallel. Some efforts have been made to apply this paradigm to image captioning [90], [91], [92], [93], [94]. The first approaches towards a non-autoregressive generation were composed of a number of different generation stages, where all words were predicted in parallel and refined at each stage. Gao *et al.* [90] adopted a multi-stage masking procedure where mask tokens are given as inputs to the decoder, with different masking ratios for each stage. Similarly, Fei *et al.* [93] proposed to iteratively refine the captions with an additional length predictor module which predicts the total number of words and adjusts the final length of the generated sequence.

Another line of work involves using reinforcement learning, with significant performance improvements. These approaches treat the generation process as a cooperative multi-agent reinforcement system, where the positions in of the words in the target sequence are viewed as agents that learn to cooperatively maximize a sentence-level reward [92], [94]. These works also leverage knowledge distillation on unlabeled data and a post-processing step to remove identical consecutive tokens.

3.6 Discussion

Recurrent models based on LSTM have been the standard for many years. Their application for language modeling brought to the development of clever and successful ideas that can be integrated also into non-recurrent solutions. For example, the generation to increasingly detailed captions by successive refinements, the grounding of generated words, the reconstruction of the internal state as a regularization strategy, the neural architecture search strategy applied to language models. The main disadvantage of recurrent models is that they are slow to train and struggle to maintain long-term dependencies between the generated words. These drawbacks are alleviated by autoregressive Transformer-based solutions that gained popularity on many natural language generation tasks, including image captioning. Inspired by the success of pre-training on large, unsupervised corpora for NLP tasks, massive pre-training has been applied also for image captioning by employing BERT-like architectures. This strategy led to impressive performance, suggesting that visual and textual semantic relations can be inferred and learned also from not

well-curated data. BERT-like architectures are suitable for such a massive pre-training but are not generative architectures by design. This fact makes their success in image captioning more imputable to the pre-training than to the architecture design and suggests that massive pre-training on generative-oriented architectures different from BERT-like ones would be a worth-exploring direction.

4 TRAINING STRATEGIES

An image captioning model is commonly expected to generate a caption word by word by taking into account the previous words and the image. At each step, the output word is sampled from a learned distribution over the vocabulary words. In the most simple scenario, *i.e.* the greedy decoding mechanism, the word with the highest probability is output. The main drawback of this setting is that possible prediction errors quickly accumulate along the way. To alleviate this drawback, one effective strategy is to use the beam search algorithm [95] that, instead of outputting the word with maximum probability at each time step, maintains k sequence candidates (those with the highest probability at each step) and finally outputs the most probable one.

During training, the captioning model must learn to properly predict the probabilities of the words to appear in the caption. To this end, the most common training strategies are based on 1. *cross-entropy loss*; 2. *masked language model strategy*; 3. *reinforcement learning* that allows directly optimizing for captioning-specific non-differentiable metrics; 4. *vision-and-language pre-training objectives* (see Fig. 1).

4.1 Cross-Entropy Loss

The cross-entropy loss is the first proposed and most used objective for image captioning models. With this loss, the goal of the training, at each timestep, is to minimize the negative log-likelihood of the current word given the previous ground-truth words. Given a sequence of target words $y_{1:T}$, the loss is formally defined as:

$$L_{XE}(\theta) = - \sum_{i=1}^n \log (P(y_i | y_{1:i-1}, \mathbf{X})), \quad (4)$$

where P is the probability distribution induced by the language model, y_i the ground-truth word at time i , $y_{1:i-1}$ indicate the previous ground-truth words, and \mathbf{X} the visual encoding. The cross-entropy loss is designed to operate at word level and optimize the probability of each word in the ground-truth sequence without considering longer range dependencies between generated words. The traditional training setting with cross-entropy also suffers from the exposure bias problem [96] caused by the discrepancy between the training data distribution as opposed to the distribution of its own predicted words.

4.2 Masked Language Model (MLM)

The first masked language model has been proposed for training the BERT [82] architecture, with the aim of learning a bidirectional representation for language. The main idea behind this optimization function consists in randomly masking out a small subset of the input tokens sequence and

training the model to predict masked tokens while relying on the rest of the sequence, *i.e.* both previous and subsequent tokens. As a consequence, the model learns to employ contextual information to infer missing tokens, which allows building a robust sentence representation where the context plays an essential role. Since this strategy considers only the prediction of the masked tokens and ignores the prediction of the non-masked ones, training with it is much slower than training for complete left-to-right or right-to-left generation. Notably, some works have employed this strategy as a pre-training objective, sometimes completely avoiding the combination with the cross-entropy [80], [83].

4.3 Reinforcement Learning

Given the limitations of word-level training strategies, a significant improvement was achieved by applying the reinforcement learning paradigm for training image captioning models. Within this framework, the image captioning model is considered as an agent whose parameters determine a policy. At each time step, the agent executes the policy to choose an action, *i.e.* the prediction of the next word in the generated sentence. Once the end-of-sequence is reached, the agent receives a reward depending on the generated sentence. The aim of the training is to optimize the agent parameters to maximize the expected reward. Many works harnessed this paradigm and explored different sequence-level metrics as rewards. The first proposal is due to Ranzato *et al.* [96], which introduced the usage of the REINFORCE algorithm [97], [98] adopting BLEU [99] and ROUGE [100] as reward signals. Ren *et al.* [101] experimented using visual-semantic embeddings obtained from a network that encodes the image and the so far generated caption in order to compute a similarity score to be used as reward. Liu *et al.* [102] proposed to use as reward a linear combination of the SPICE [103] and CIDEr [104] metrics, called SPIDEr. Finally, the most widely adopted reinforcement learning-based strategy [69], [105], [106], introduced by Rennie *et al.* [27], entails using the CIDEr score as reward, as it correlates well with human judgment [104]. The reward is normalized with respect to a baseline value to reduce the reward variance. Formally, to compute the loss gradient, beam search and greedy decoding are leveraged as follows:

$$\nabla_{\theta} L(\theta) = -\frac{1}{k} \sum_{i=1}^k ((r(\mathbf{w}^i) - b) \nabla_{\theta} \log P(\mathbf{w}^i)), \quad (5)$$

where \mathbf{w}^i is the i -th sentence in the beam or a sampled collection, $r(\cdot)$ is the reward function, *i.e.* the CIDEr computation, and b is the baseline, computed as the reward of the sentence obtained via greedy decoding [27], or as the average reward of the beam candidates [69].

Note that, since it would be difficult for a random policy to improve in an acceptable amount of time, the usual procedure entails pre-training with cross-entropy or masked language model first, and then fine-tuning stage with reinforcement learning by employing a sequence level metric as reward. This ensures the initial reinforcement learning policy to be more suitable than the random one.

4.4 Vision-and-Language Pre-Training

In the context of vision-and-language pre-training, one of the most common pre-training objectives is the masked contextual token loss, where tokens of each modality (visual and textual) are randomly masked following the BERT strategy [82], and the model has to predict the masked input based on the context of both modalities, thus connecting their joint representation. Another largely adopted strategy entails using a contrastive loss, where the inputs are organized as image regions-captions words-object tags triples, and the model is asked to discriminate correct triples from polluted ones, in which tags are randomly replaced [80], [83]. Other objectives take into account the text-image alignment at a word-region level and entail predicting the original word sequence given a corrupted one [107].

5 EVALUATION PROTOCOL

As for any data-driven task, the development of image captioning has been enabled by the collection of large datasets and the definition of quantitative scores to evaluate the performance and monitor the advancement of the field.

5.1 Datasets

Image captioning datasets contain images and one or multiple captions associated with them. Having multiple ground-truth captions for each image helps to capture the variability of human descriptions. Other than the number of available captions, also their characteristics (*e.g.* average caption length and vocabulary size) highly influence the design and the performance of image captioning algorithms. Note that the distribution of the terms in the datasets captions is usually long-tailed, thus, the common practice is to include in the vocabulary only those terms whose frequency is above a pre-defined threshold. The threshold must be chosen as a trade-off between numerical tractability and capability to mimic the lexical richness and diversity of human descriptions. The available datasets differ both on the images contained (for their domain and visual quality) and on the captions associated with the images (for their length, number, relevance, and style). A summary of the most used public datasets is reported in Table 1, and some sample image-caption pairs are reported in Fig. 8, along with some word clouds obtained from the 50 most used visual words in the captions.

5.1.1 Standard captioning datasets

Standard benchmark datasets are used by the community to compare their approaches on a common test-bed. As a matter of fact, these comparisons guide the development of image captioning strategies by allowing to identify suitable directions. Therefore, datasets used as benchmarks should be representative of the task at hand, both in terms of the challenges it poses and of the ideal expected results (*i.e.* achievable human performance). In this sense, benchmark datasets should contain a large number of generic-domain images, each associated with multiple captions.

Early image captioning architectures [14], [16], [17] were commonly trained and tested on the **Flickr30K** [109] and **Flickr8K** [110] datasets, consisting of pictures collected from

TABLE 1: Overview of the main image captioning datasets.

	Domain	Nb. Images	Nb. Caps (per Image)	Vocab Size	Nb. Words (per Cap.)
MS COCO [108]	Generic	132K	5	27K (10K)	10.5
Flickr30K [109]	Generic	31K	5	18K (7K)	12.4
Flickr8K [110]	Generic	8K	5	8K (3K)	10.9
CC3M [111]	Generic	3.3M	1	48K (25K)	10.3
CC12M [112]	Generic	12.4M	1	523K (163K)	20.0
SBU Captions [113]	Generic	1M	1	238K (46K)	12.1
VizWiz [114]	Assistive	70K	5	20K (8K)	13.0
CUB-200 [115]	Birds	12K	10	6K (2K)	15.2
Oxford-102 [115]	Flowers	8K	10	5K (2K)	14.1
Fashion Cap. [116]	Fashion	130K	1	17K (16K)	21.0
BreakingNews [117]	News	115K	1	85K (10K)	28.1
GoodNews [118]	News	466K	1	192K (54K)	18.2
TextCaps [119]	OCR	28K	5/6	44K (13K)	12.4
Loc. Narratives [120]	Generic	849K	1/5	16K (7K)	41.8

the Flickr website, containing everyday activities, events, and scenes, paired with five captions each. Currently, the most commonly used dataset for image captioning is **Microsoft COCO** [108], which consists of images of complex scenes with people, animals, and common everyday objects in their natural context. It contains more than 120,000 images, each of them annotated with five different captions, divided into 82,783 images for training and 40,504 for validation. For ease of evaluation, most of the literature follows the splits defined by Karpathy *et al.* [14], where 5,000 images of the original validation set are used for validation, 5,000 for test, and the rest for training. The dataset has also an official test set, composed of 40,775 images paired with 40 private captions each, and a public evaluation server² to measure the performance.

5.1.2 Pre-training datasets

Although training on large well-curated datasets is a sound approach, some works [79], [80] have demonstrated the benefits of pre-training on even bigger vision-and-language datasets, which can be either image captioning datasets or less diverse and lower-quality captions or datasets collected for other tasks (*e.g.* visual question answering [80], [81], text-to-image generation [121], image-caption association [122]). Among the datasets used for pre-training, that have been specifically collected for image captioning, it is worth mentioning **SBU Captions** [113], originally used for tackling image captioning as a retrieval task [110], which contains around 1 million image-text pairs, collected from the Flickr website. Later, the **Conceptual Captions** [111], [112] datasets have been proposed, which are collections of around 3.3 million (CC3M) and 12 million (CC12M) images paired with one weakly-associated description automatically collected from the web with a relaxed filtering procedure. Although the large scale and variety in caption style make Conceptual Captions particularly interesting for pre-training, the contained captions are simple and availability of images is not always guaranteed since they are provided as URLs.

Pre-training on such datasets requires significant computational resources and effort to collect the data needed. Nevertheless, this strategy represents an asset to obtain state-of-the-art performances. For this reason, some vision-

2. <https://competitions.codalab.org/competitions/3221>

Fig. 8: Qualitative examples from some of the most common image captioning datasets: **(a)** image-caption pairs; **(b)** word clouds of the captions most common visual words.

and-language pre-training datasets are not always publicly available [121], [122].

5.1.3 Domain-specific datasets

While domain-generic benchmark datasets are important to capture the main aspects of the image captioning task, domain-specific datasets are also important to highlight and target specific challenges. These may relate to the visual domain (*e.g.* type and style of the images) and the semantic domain. In particular, the distribution of the terms used to describe domain-specific images can be significantly different from that of the terms used for domain-generic images.

An example of dataset specific in terms of visual domain is the **VizWiz Captions** [114] dataset, collected to favor the image captioning research towards assistive technologies. The images in this dataset have been taken by visually-impaired people with their phones, thus, they can be of low quality and concern a wide variety of everyday activities, most of which entail reading some text.

Some examples of specific semantic domain are the **CUB-200** [123] and the **Oxford-102** [124] datasets, which contain images of birds and flowers, respectively, that have been paired with ten captions each by Reed *et al.* [115]. Given the specificity of these datasets, rather than for standard image captioning, they are usually adopted for different related tasks such as cross-domain captioning [125], visual explanation generation [126], [127], and text-to-image synthesis [128]. Another domain-specific dataset is **Fashion Captioning** [116] that contains images of clothing items in different poses and colors that may share the same caption. The vocabulary for describing these images is somewhat smaller and more specific than for generic datasets. Differently, datasets as **BreakingNews** [117] and **GoodNews** [118] enforce using a richer vocabulary since their images, taken from news articles, have long associated captions written by expert journalists. The same applies to the **TextCaps** [119] dataset, which contains images with text, that must be “read” and included in the caption, and to **Localized Narratives** [120], whose captions have been collected by recording people freely narrating what they see in the images.

Collecting domain-specific datasets and developing solutions to tackle the challenges they pose is crucial to extend the applicability of image captioning algorithms.

5.2 Evaluation Metrics

Evaluating the quality of a generated caption is a tricky and subjective task [103], [104], complicated by the fact that captions cannot only be grammatical and fluent but need to properly refer to the input image. Arguably, the best way to measure the quality of the caption for an image is still carefully designing a human evaluation campaign in which multiple users score the produced sentences. However, human evaluation is lengthy and costly and, most importantly, is not reproducible – which prevents a fair comparison between different approaches. Automatic scoring methods exist that are used to assess the quality of system-produced captions, usually by comparing them with human-produced reference sentences, although some metrics can also be applied without relying on reference captions. A taxonomy and main characteristics of the most commonly used metrics are summarized in Table 2.

5.2.1 Standard evaluation metrics

The first strategy adopted to evaluate image captioning performance consists of exploiting metrics designed for NLP tasks such as machine translation and summarization [99], [100], [129]. Later, specific image captioning metrics have been proposed [103], [104]. The most commonly used ones capture different aspects of the caption quality based on n -gram precision and recall. For fair comparison among different approaches, the common practice is to use the implementation provided in the Microsoft COCO caption evaluation repository³.

As expected, metrics designed for image captioning usually correlate better with human judgment than those borrowed from other NLP tasks (with the exception of METEOR [129]), both at corpus-level and caption-level [103].

3. <https://github.com/tylin/coco-caption>

TABLE 2: Taxonomy and main characteristics of image captioning metrics.

		Original Task	Inputs		
			Pred	Refs	Image
Standard	BLEU [99]	Translation	✓	✓	
	METEOR [129]	Translation	✓	✓	
	ROUGE [100]	Summarization	✓	✓	
	CIDEr [104]	Captioning	✓	✓	
	SPICE [103]	Captioning	✓	(✓)	(✓)
Diversity	Div [132]	Captioning	✓		
	Vocab [132]	Captioning	✓		
	%Novel [132]	Captioning	✓		
Embedding-based	WMD [133]	Doc. Dissimilarity	✓	✓	
	Alignment [86]	Captioning	✓	✓	
	Coverage [86], [134]	Captioning	✓	(✓)	(✓)
Learning-based	TIGER [135]	Captioning	✓	✓	✓
	BERT-S [136]	Text Similarity	✓	✓	
	CLIP-S [137]	Captioning	✓	(✓)	✓

[130], [131]. Correlation with human judgment is measured via statistical correlation coefficients (such as Pearson’s, Kendall’s, and Spearman’s correlation coefficients) and via the agreement with humans’ preferred caption in a pair of candidates, all evaluated on sample captioned images.

BLEU [99]. BLEU is a precision-oriented metric designed for machine translation. To obtain the score, n -gram precision is calculated for each n -gram up to length four, with a minor modification to prevent an n -gram from appearing in the candidate more often than in the reference. Finally, the n -gram precision values are combined via a weighted sum.

METEOR [129]. METEOR is a precision and recall-based machine translation evaluation metric. Unigram precision and unigram recall are calculated by matching unigrams in the candidate and reference sentences based on their exact form, stemmed form, and meaning. Then, an F-mean is obtained, weighing the recall more than the precision. In addition, a multiplicative factor is used to reward identically ordered contiguous matched unigrams.

ROUGE [100]. ROUGE is a recall-oriented metric designed for summarization. It is based on the idea that ideal candidate summaries should overlap the reference summary. For image captioning evaluation, precision and recall are obtained based on the longest subsequence of tokens in the same relative order, possibly with other tokens in-between, that appears in both candidate and reference caption, and an F-mean is computed favoring the recall.

CIDEr [104]. CIDEr was designed to correlate well with human judgment on image captions quality. It is based on the cosine similarity between the Term Frequency-Inverse Document Frequency weighted n -grams in the candidate caption and in the set of reference captions associated with the image, thus taking into account both precision and recall. In addition, a Gaussian penalty factor rewards length similarity between candidate and reference sentences.

SPICE [103]. SPICE is specifically designed for captioning evaluation and considers the candidate caption semantic content rather than its grammaticality and fluency. The captions are represented as sets of tuples extracted from their scene graphs, and precision and recall are calculated over matching tuples in these sets. Note that two tuples match if their elements match or are synonyms. SPICE is obtained as the F1-mean. This score can be quantified for certain objects, attributes, and relations separately, and, by

definition, it could also work by directly comparing the scene graphs of the image and the candidate caption.

5.2.2 Diversity metrics

To better assess the performance of a captioning system, it is common practice to consider a set of the above-mentioned standard metrics. Nevertheless, these are somehow gameable because they favor word similarity rather than meaning correctness [138]. Another drawback of the standard metrics is that they do not capture (but rather disfavor) the desirable capability of the system to produce novel and diverse captions, which is more in line with the variability with which humans describe complex images. This consideration brought to the development of diversity metrics [132], [139], [140], [141]. Most of these metrics can potentially be calculated even when no ground-truth captions are available at test time. However, since they overlook the syntactic correctness of the captions and their relatedness with the image, it is advisable to combine them with other metrics.

The overall performance of a captioning system can be evaluated in terms of corpus-level diversity or, when the system can output multiple captions for the same image, single image-level diversity (termed as *global diversity* and *local diversity*, respectively, in [139]). To quantify the former, it can be considered the number of unique words used in all the generated captions (**Vocab**) and the percentage of generated captions that were not present in the training set (**%Novel**). For the latter, it can be used the ratio of unique captions unigrams or bigrams to the total number of captions unigrams (**Div-1** and **Div-2**).

5.2.3 Embedding-based metrics

An alternative approach to captioning evaluation consists in relying on captions semantic similarity or other specific aspects of caption quality, which are estimated via embedding-based metrics [71], [86], [133], [134].

WMD [133]. WMD was introduced to evaluate document semantic dissimilarity but can also be applied to captioning evaluation (converted into a similarity score via a negative exponential) by considering generated captions and ground-truth captions as the compared documents [142]. The captions are represented by normalized bag-of-words, not including stopwords. Their WDM is defined as the minimum cumulative sum of the pairwise euclidean distance of their word embeddings [143], weighted by a term representing the contribution of the generated caption word to the probability mass of the ground-truth caption word.

Alignment [86]. Alignment was introduced to evaluate controllable captioning but can also be applied to standard captioning evaluation. The produced caption and the ground-truth caption are represented as the sequence of their contained nouns, and an alignment score is calculated via the Needleman-Wunsch algorithm, where the noun matching score is given by the cosine similarity of the nouns word vectors represented with GloVe embeddings [144]. The noun alignment score is given by the alignment score normalized by the maximum length of the compared sequences.

Coverage [71], [134]. Coverage expresses the completeness of a caption, which is evaluated by considering the mentioned visual entities. To this end, scene object categories

TABLE 3: Overview of deep learning-based image captioning models. Scores are taken from the respective papers.

Model	Visual Encoding					Language Model			Training Strategies				Main Results		
	Global	Grid	Regions	Graph	Self-Attention	RNN/LSTM	Transformer	BERT	XE	MLM	Reinforce	VL Pre-Training	BLEU-4	METEOR	CIDEr
VinVL [83]			✓		✓			✓		✓	✓	✓	41.0	31.1	140.9
Oscar [80]			✓		✓			✓		✓	✓	✓	41.7	30.6	140.0
Unified VLP [81]			✓		✓			✓	✓		✓	✓	39.5	29.3	129.3
AutoCaption [87]			✓		✓	✓			✓		✓		40.2	29.9	135.8
RSTNet [74]	✓				✓		✓		✓		✓		40.1	29.8	135.6
DLCT [73]	✓		✓		✓		✓		✓		✓		39.8	29.5	133.8
DPA [70]			✓		✓	✓			✓		✓		40.5	29.6	133.4
X-Transformer [68]			✓		✓		✓		✓		✓		39.7	29.5	132.8
NG-SAN [66]			✓		✓		✓		✓		✓		39.9	29.3	132.1
X-LAN [68]			✓		✓	✓			✓		✓		39.5	29.5	132.0
GET [72]			✓		✓		✓		✓		✓		39.5	29.3	131.6
\mathcal{M}^2 Transformer [69]			✓		✓		✓		✓		✓		39.1	29.2	131.2
AoANet [67]			✓		✓	✓			✓		✓		38.9	29.2	129.8
CPTR [77]					✓		✓		✓		✓		40.0	29.1	129.4
ORT [65]			✓		✓		✓		✓		✓		38.6	28.7	128.3
CNM [63]			✓		✓	✓			✓		✓		38.9	28.4	127.9
ETA [64]			✓		✓		✓		✓		✓		39.9	28.9	127.6
GCN-LSTM+HIP [60]			✓	✓		✓			✓		✓		39.1	28.9	130.6
MT [59]			✓	✓		✓			✓		✓		38.9	28.8	129.6
SGAE [58]			✓	✓		✓			✓		✓		39.0	28.4	129.1
GCN-LSTM [55]			✓	✓		✓			✓		✓		38.3	28.6	128.7
VSUA [56]			✓	✓		✓			✓		✓		38.4	28.5	128.6
SG-RWS [52]			✓			✓			✓		✓		38.5	28.7	129.1
LBPF [50]			✓			✓			✓		✓		38.3	28.5	127.6
AAT [51]			✓			✓			✓		✓		38.2	28.3	126.7
CAVP [53]			✓			✓			✓		✓		38.6	28.3	126.3
Up-Down [45]			✓			✓			✓		✓		36.3	27.7	120.1
RDN [49]			✓			✓			✓				36.8	27.2	115.3
Neural Baby Talk [85]			✓			✓			✓				34.7	27.1	107.2
Stack-Cap [37]	✓					✓			✓		✓		36.1	27.4	120.4
MaBi-LSTM [36]	✓					✓			✓				36.8	28.1	116.6
RFNet [40]	✓					✓			✓		✓		35.8	27.4	112.5
SCST (Att2in) [27]	✓					✓			✓		✓		33.3	26.3	111.4
Adaptive Attention [32]	✓					✓			✓				33.2	26.6	108.5
Skeleton [35]	✓					✓			✓				33.6	26.8	107.3
ARNet [34]	✓					✓			✓				33.5	26.1	103.4
SCA-CNN [39]	✓					✓			✓				31.1	25.0	95.2
Areas of Attention [54]	✓					✓			✓				30.7	24.5	93.8
Review Net [38]	✓					✓			✓				29.0	23.7	88.6
Show, Attend and Tell [31]	✓					✓			✓				24.3	23.9	-
SCST (FC) [27]	✓					✓			✓		✓		31.9	25.5	106.3
PG-SPIDER [102]	✓					✓			✓		✓		33.2	25.7	101.3
SCN-LSTM [30]	✓					✓			✓				33.0	25.7	101.2
LSTM-A [29]	✓					✓			✓				32.6	25.4	100.2
CNN _L +RNH [24]	✓					✓			✓				30.6	25.2	98.9
Att-CNN+LSTM [23]	✓					✓			✓				31.0	26.0	94.0
GroupCap [26]	✓					✓			✓				33.0	26.0	-
StructCap [25]	✓					✓			✓				32.9	25.4	-
Embedding Reward [101]	✓					✓			✓		✓		30.4	25.1	93.7
ATT-FCN [22]	✓					✓			✓				30.4	24.3	-
MIXER [96]	✓					✓			✓		✓		29.0	-	-
MSR [20]	✓					✓			✓				25.7	23.6	-
gLSTM [21]	✓					✓			✓				26.4	22.7	81.3
m-RNN [16]	✓					✓			✓				25.0	-	-
Show and Tell [11]	✓					✓			✓				24.6	-	-
Mind's Eye [19]	✓					✓			✓				19.0	20.4	-
DeepVS [14]	✓					✓			✓				23.0	19.5	66.0
LRCN [17]	✓					✓			✓				21.0	-	-

and caption nouns are represented as word vectors [144] and matched via the Hungarian algorithm. The noun coverage score is defined as the sum of the assignment scores normalized by the number of object categories in the scene. Since this score considers visual objects directly, it can be applied even when no ground-truth caption is available.

5.2.4 Learning-based evaluation

As a further development towards captions quality assessment, learning-based evaluation strategies [130], [131], [135], [136], [137], [145], [146], [147] are being investigated that evaluate how human-like a caption is [148].

TIGER [135]. TIGER represents the reference and candidate captions as grounding score vectors obtained from a pre-trained model [149] that grounds their words on the image

regions and scores the candidate caption based on the similarity of the grounding vectors. This is computed by taking into account region rank similarity (*i.e.* how similarly the image regions are ranked based on the grounding scores in the vectors) and weight distribution similarity (*i.e.* how similar the grounding score distributions in the vectors are).

BERT-S [136]. BERT Score is a metric to evaluate various language generation tasks [150], including image captioning. It exploits pre-trained BERT embeddings [82] to represent and match the tokens in the reference and candidate sentences via cosine similarity. The best matching token pairs are used for computing precision, recall, and F1-score.

CLIP-S [137]. CLIP Score is a direct application of the CLIP [122] cross-modal retrieval model, which leverages large-scale vision-and-language pre-training, to image cap-

TABLE 4: Performance analysis of representative image captioning approaches in terms of different evaluation metrics. The \dagger marker indicates models trained by us with ResNet-152 features, while the \ddagger marker indicates unofficial implementations.

	#Params (M)	Standard Metrics						Diversity Metrics				Embedding-based Metrics			Learning-based Metrics			
		B-1	B-4	M	R	C	S	Div-1	Div-2	Vocab	%Novel	WMD	Alignment	Coverage	TIGer	BERT-S	CLIP-S	CLIP-S ^{Ref}
Show and Tell [†] [11]	13.6	72.4	31.4	25.0	53.1	97.2	18.1	0.014	0.045	635	36.1	16.5	0.199	71.7	71.8	93.4	0.697	0.762
SCST (FC) [‡] [27]	13.4	74.7	31.7	25.2	54.0	104.5	18.4	0.008	0.023	376	60.7	16.8	0.218	74.7	71.9	89.0	0.691	0.758
Show, Attend and Tell [†] [31]	18.1	74.1	33.4	26.2	54.6	104.6	19.3	0.017	0.060	771	47.0	17.6	0.209	72.1	73.2	93.6	0.710	0.773
SCST (Att2in) [‡] [27]	14.5	78.0	35.3	27.1	56.7	117.4	20.5	0.010	0.031	445	64.9	18.5	0.238	76.0	73.9	88.9	0.712	0.779
Up-Down [‡] [45]	52.1	79.4	36.7	27.9	57.6	122.7	21.5	0.012	0.044	577	67.6	19.1	0.248	76.7	74.6	88.8	0.723	0.787
SGAE [58]	125.7	81.0	39.0	28.4	58.9	129.1	22.2	0.014	0.054	647	71.4	20.0	0.255	76.9	74.6	94.1	0.734	0.796
MT [59]	63.2	80.8	38.9	28.8	58.7	129.6	22.3	0.011	0.048	530	70.4	20.2	0.253	77.0	74.8	88.8	0.726	0.791
AoANet [67]	87.4	80.2	38.9	29.2	58.8	129.8	22.4	0.016	0.062	740	69.3	20.0	0.254	77.3	75.1	94.3	0.737	0.797
X-LAN [68]	75.2	80.8	39.5	29.5	59.2	132.0	23.4	0.018	0.078	858	73.9	20.6	0.261	77.9	75.4	94.3	0.746	0.803
DPA [70]	111.8	80.3	40.5	29.6	59.2	133.4	23.3	0.019	0.079	937	65.9	20.5	0.261	77.3	75.0	94.3	0.738	0.802
AutoCaption [87]	-	81.5	40.2	29.9	59.5	135.8	23.8	0.022	0.096	1064	75.8	20.9	0.262	77.7	75.4	94.3	0.752	0.808
ORT [65]	54.9	80.5	38.6	28.7	58.4	128.3	22.6	0.021	0.072	1002	73.8	19.8	0.255	76.9	75.1	94.1	0.736	0.796
CPTR [77]	138.5	81.7	40.0	29.1	59.4	129.4	-	0.014	0.068	667	75.6	20.2	0.261	77.0	74.8	94.3	0.745	0.802
\mathcal{M}^2 Transformer [69]	38.4	80.8	39.1	29.2	58.6	131.2	22.6	0.017	0.079	847	78.9	20.3	0.256	76.0	75.3	93.7	0.734	0.792
X-Transformer [68]	137.5	80.9	39.7	29.5	59.1	132.8	23.4	0.018	0.081	878	74.3	20.6	0.257	77.7	75.5	94.3	0.747	0.803
Unified VLP [81]	138.2	80.9	39.5	29.3	59.6	129.3	23.2	0.019	0.081	898	74.1	26.6	0.258	77.1	75.1	94.4	0.750	0.807
VinVL _L [83]	369.6	82.0	41.0	31.1	60.9	140.9	25.2	0.023	0.099	1125	77.9	20.5	0.265	79.6	75.7	88.5	0.766	0.820

tioning evaluation. The score consists of an adjusted cosine similarity of image and candidate caption representation. Thus, CLIP-S is designed to work without reference captions. Nonetheless, the CLIP-S^{Ref} variant can exploit also the reference captions by considering the harmonic mean between the image-candidate CLIP-S score and the maximum cosine similarity between the candidate caption and the reference ones.

6 EXPERIMENTAL EVALUATION

According to the taxonomies proposed in Sections 2, 3, and 4, in Table 3, we overview the most relevant surveyed methods. We report their performance in terms of BLEU-4, METEOR, and CIDEr on the MS COCO Karpathy split test set and their main features in terms of visual encoding, language modeling, and training strategies. In the table, methods are clustered based primarily on their visual encoding strategy and ordered based on the obtained scores. Methods exploiting vision-and-language pre-training are further separated from the others. Image captioning models have reached impressive performance in just a few years: from an average BLEU-4 of 25.1 for the methods using global CNN features to an average BLEU-4 of 35.3 and 39.8 for those exploiting the attention and self-attention mechanisms, peaking at 41.7 in case of vision-and-language pre-training. By looking at the performance in terms of the more representative CIDEr score, we can notice the same positive trend and make the following considerations on the design choices adopted in the surveyed works. As for the visual encoding, the more complete and structured information about semantic visual concepts and their mutual relation is included, the better is the performance (consider that methods applying attention over a grid of features reach an average CIDEr score of 105.8, while those performing attention over visual regions 121.8, further increased for graph-based approaches and methods using self-attention, which reach 130.4 on average). As for the language model, LSTM-based approaches combined with strong visual encoders are still competitive with subsequent fully-attentive methods in terms of performance. These methods are slower to train but are generally smaller than Transformer-based ones (apart from the optimized \mathcal{M}^2 Transformer model [69]).

As for the training strategy, sentence-level fine-tuning with reinforcement learning leads to significant performance improvement (consider that methods relying only on the cross-entropy loss obtain an average CIDEr score of 92.3, while those combining it with reinforcement learning fine-tuning reach 125.1 on average). Moreover, the collected results show that the Masked Language loss can be a valid alternative to the cross-entropy loss. Finally, it emerges that vision-and-language pre-training on large datasets allows boosting the performance and deserves further investigation.

Furthermore, in Table 4, we analyze the performance of some of the main approaches in terms of all the evaluation scores presented in Section 5.2 to take into account the different aspects of caption quality these express and report their number of parameters to give an idea of the computational complexity and memory occupancy of the models. The data in the table have been obtained either from the model weights and captions files provided by the original authors or from our best implementation. Given its large use as a benchmark in the field, we consider the domain-generic MS COCO dataset also for this analysis. In the table, methods are clustered based on the information included in the visual encoding and ordered by CIDEr score. It can be observed that standard and embedding-based metrics all had a substantial improvement with the introduction of region-based visual encodings. Further improvement was due to the integration of information on inter-objects relations, either expressed via graphs or self-attention. Notably, CIDEr, SPICE, and Coverage most reflect the benefit of vision-and-language pre-training. Moreover, as expected, it emerges that the diversity-based scores are correlated, especially Div-1 and Div-2 and the Vocab Size. The correlation of this family of scores and the others is almost linear, except for early approaches, which perform averagely well in terms of Diversity despite lower values for standard metrics. From the trend of learning-based scores, it emerges that exploiting models trained on textual data only (BERT-S, reported in the table as its F1-score variant) does not help discriminating among image captioning approaches. On the other hand, considering as reference only the visual information and disregarding the ground-truth captions is possible with the appropriate vision-and-language pre-trained model (con-

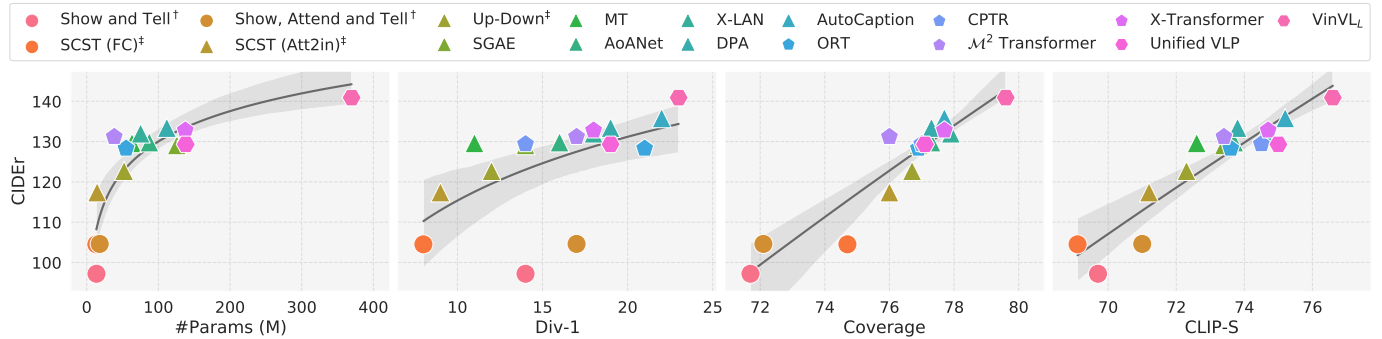


Fig. 9: Relationship between CIDEr, number of parameters and other scores. Values of Div-1 and CLIP-S are multiplied by powers of 10 for readability.

sider that CLIP-S and CLIP-S^{Ref} are linearly correlated). This is a desirable property for an image captioning evaluation score since it allows estimating the performance of a model without relying on reference captions that can be limited in number and somehow subjective.

For readability, in Fig. 9 we highlight the relation between the CIDEr score and other characteristics from Table 4 (*i.e.* number of parameters and representative non-standard metrics). We chose CIDEr as this score is commonly regarded as one of the most relevant indicators of image captioning systems performance. The first plot, depicting the relation between model complexity and performance, shows that more complex models do not necessarily bring to better performance. Consider for example \mathcal{M}^2 Transformer [69] and X-LAN [68] in comparison with X-Transformer [68] and Unified VLP [81]. All methods achieve CIDEr higher than 130, but the first two are much more compact architectures (38.8M and 75.2M parameters compared to over 137M). The other plots describe an almost-linear relation between CIDEr and the other scores, with some flattening for high CIDEr values. These trends confirm the suitability of the CIDEr score as an indicator of the overall performance of an image captioning algorithm, whose specific characteristics in terms of the produced captions would still be expressed more precisely in terms of non-standard metrics.

7 IMAGE CAPTIONING VARIANTS

Beyond general-purpose image captioning, several specific sub-tasks have been explored in the literature. These can be classified into four categories according to their scope: 1. *dealing with lacking training data*; 2. *focusing on the visual input*; 3. *focusing on the textual output*; 4. *addressing user requirements*.

7.1 Dealing with lacking training data

Paired image-caption datasets are very expensive to obtain. Thus, some image captioning variants are being explored that limit the need for full supervision information.

Novel Object Captioning. Novel object captioning focuses on describing objects not appearing in the training set, thus enabling a zero-shot learning setting that can increase the applicability of the models in the real world. Early approaches to this task [151], [152] tried to transfer knowledge from out-domain images by conditioning the model on external unpaired visual and textual data at training time.

To explore this strategy, Hendricks *et al.* [151] introduced a variant of the MS COCO dataset [108], called *held-out COCO*, in which image-caption pairs containing one of eight pre-selected object classes were removed from the training set but not from the test set. To further encourage research on this task, the more challenging *nocaps* dataset, with nearly 400 novel objects, has been introduced [153], where images are grouped into three subsets depending on their semantic distance to MS COCO (*i.e.* in-domain, near-domain, and out-of-domain images). Some approaches to this variant [154], [155] integrate copying mechanisms in the language model to select novel objects predicted from a tagger. Other methods generate a caption template with placeholders to be filled with novel objects [85], [156] or replace ambiguous words with novel objects in a second stage [157]. On a different line, Anderson *et al.* [158] devised the Constrained Beam Search algorithm to force the inclusion of selected tag words in the output caption, following the predictions of a tagger. Moreover, following the pre-training trend with BERT-like architectures, Hu *et al.* [159] proposed a multi-layer Transformer model pre-trained by randomly masking one or more tags from image-tag pairs.

Unpaired Captioning. Unpaired captioning aims at understanding and describing images without paired image-text training data. Following unpaired machine translation approaches, the early work [160] proposes to generate captions in a pivot language and then translate predicted captions to the target language. After this work, the most common approach focuses on adversarial learning by training an LSTM-based discriminator to distinguish whether a caption is real or generated [161], [162]. Some alternatives exist, such as [163] that generates a caption from the image scene-graph, [164] that leverages a memory-based network, or [165], [166] that propose semi-supervised strategies.

Continual Captioning. Continual captioning aims to deal with partially unavailable data by following the continual learning paradigm to incrementally learn new tasks without forgetting what has been learned before. In this respect, new tasks can be represented as sequences of captioning tasks with different vocabularies, as proposed in [167], and the model should be able to transfer visual concepts from one to the other while enlarging its vocabulary. To this end, continual captioning approaches focus on techniques to address the catastrophic forgetting, such as freezing part of the model during training, employing pseudo-labels, or

knowledge distillation [168].

7.2 Focusing on the visual input

Some sub-tasks focus on making the textual description more correlated with visual data.

Dense Captioning. Dense captioning was proposed by Johnson *et al.* [169] and consists of concurrently localizing and describing salient image regions with short natural language sentences. In this respect, the task can be conceived as a generalization of object detection, where caption replaces object tags, or image captioning, where single regions replace the full image. The main challenge of this task is the aperture problem, *i.e.* the lack of contextual information for each region with the surrounding ones. To face this issue, contextual and global features [170], [171] and attribute generators [172], [173] can be exploited. The performance on this task is commonly evaluated on the Visual Genome [48] dataset since it contains a large number of images with region-level annotated captions. Related to this variant, an important line of works [53], [174], [175], [176], [177], [178], [179], [180] focuses on the generation of textual paragraphs that densely describe the visual content as a coherent story.

Text-based Image Captioning. Text-based image captioning, also known as OCR-based image captioning or image captioning with reading comprehension, aims at reading and including the text appearing in images in the generated descriptions. The task was introduced by Sidorov *et al.* [119] with the TextCaps dataset. Another dataset designed for pre-training for this variant is OCR-CC [181], which is a subset of images containing meaningful text taken from the CC3M dataset [111] and automatically annotated through a commercial OCR system. The common approach to this variant entails combining image regions and text tokens, *i.e.* groups of characters from an OCR, possibly enriched with mutual spatial information [182], [183], in the visual encoding [119], [184]. Another direction entails generating multiple captions describing different parts of the image, including the contained text [185].

Change Captioning. Change captioning targets changes that occurred in a scene, thus requiring both accurate change detection and effective natural language description. The task was first presented in [186] with the *Spot-the-Diff* dataset, composed of pairs of frames extracted from video surveillance footages and the corresponding textual descriptions of visual changes. To further explore this variant, the *CLEVR-Change* dataset [187] has been introduced, which contains five scene change types on almost 80K image pairs. The proposed approaches for this variant apply attention mechanisms to focus on semantically relevant aspects without being deceived by distractors such as viewpoint changes [188], [189] or perform multi-task learning with image retrieval as an auxiliary task [190], where an image must be retrieved from its paired image and the description of the occurred changes.

7.3 Focusing on the textual output

Since every image captures a wide variety of entities with complex interactions, human descriptions tend to be diverse and grounded to different objects and details. Some image captioning variants explicitly focus on these aspects.

Diverse Captioning. Diverse image captioning tries to replicate the quality and variability of the sentences produced by humans. The most common technique to achieve diversity is based on variants of the beam search algorithm [191] that entail dividing the beams into similar groups and encouraging diversity between groups. Other solutions have been investigated, such as contrastive learning [192], conditional GANs [132], [148], and paraphrasing [193]. However, these solutions tend to underperform in terms of caption quality, which is partially recovered by using variational auto-encoders [194], [195], [196], [197]. Another approach is exploiting multiple part-of-speech tags sequences predicted from image region classes [198] and forcing the model to produce different captions based on these sequences.

Multilingual Captioning. Since image captioning is commonly performed in English, multilingual captioning [199] aims to extend the applicability of captioning systems to other languages. The two main strategies entail collecting captions in different languages for commonly used datasets (*e.g.* Chinese and Japanese captions for MS COCO images [200], [201], German captions for Flickr30K [202]), or directly training multilingual captioning systems with unpaired captions [160], [199], [203], [204], [205], which requires specific evaluation protocols [206].

7.4 Addressing user requirements

Regular image captioning models generate factual captions with a neutral tone and no interaction with end-users. Instead, some image captioning sub-tasks are devoted to coping with user requests.

Personalized Captioning. Humans consider more effective the captions that avoid stating the obvious and that are written in a style that catches their interest. Personalized image captioning aims at fulfilling this requirement by generating descriptions that take into account the user’s prior knowledge, active vocabulary, and writing style. To this end, early approaches exploit a memory block as a repository for this contextual information [207], [208]. On another line, Zhang *et al.* [209] proposed a multi-modal Transformer network that personalizes captions conditioned on the user’s recent captions and a learned user representation. Other works have instead focused on the style of captions as an additional controllable input and proposed to solve this task by exploiting unpaired stylized textual corpus [210], [211], [212], [213] and adversarial learning [212]. Some datasets have been collected to explore this variant, such as *InstaPIC* [207], which is composed of multiple Instagram posts from the same users, *FlickrStyle10K* [210], which contains images and textual sentences with two different styles, and *Personality-Captions* [214], which contains triples of images, captions, and one among 215 personality traits to be used to condition the caption generation.

Controllable Captioning. Controllable captioning puts the users in the loop by asking them to select and give priorities to what should be described in an image. This information is exploited as a guiding signal for the generation process. The signal can be in the form of part-of-speech tag sequences, as in [198], of sets or sequences of image regions corresponding to a sentence noun chunk (*i.e.* a noun with its modifiers) as in [86], of mouse traces, producing a dense visual grounding



Fig. 10: Qualitative examples from four popular captioning models on COCO test images. Errors are highlighted in red.

between words and visual elements, as in [120], or of verbs and semantic roles, where verbs represent activities in the image and semantic roles determine how objects engage in these activities, as in [215]. A step further is proposed by Meng *et al.* [216], which also incorporates controlled trace generation and joint caption-trace generation tasks.

8 CONCLUSIONS AND FUTURE DIRECTIONS

Image captioning is an intrinsically complex challenge for machine intelligence as it integrates difficulties from both Computer Vision and Natural Language Generation. While most approaches keep the visual encoding and language modeling steps distinguished, the single-stream trend of BERT-like architectures entails performing early-fusion of visual and textual data. This strategy allows achieving remarkable performance but is usually combined with massive pre-training. Thus, it is worth investigating whether standard encoder-decoder methods enriched with pre-training could achieve similar results. Nonetheless, methods based on the classical two-stream paradigm are more explainable, both for model designers and end-users. The presented literature review and experimental comparison show the performance improvement over the last few years. However, many open challenges remain since accuracy, robustness, and generalization results are far from satisfactory. Similarly, requirements of fidelity, naturalness, and diversity are not yet met. In this respect, since image captioning has been conceived for improving human-machine interaction, the possibility to include the user in the loop is promising. Based on the analysis presented, we can trace three main developmental directions for the image captioning field, which are discussed in the following.

8.1 Procedural and architectural changes

As emerged from the analysis in Section 6, a paradigm shift is needed in order to boost the achievable performance.

Large-scale vision-and-language pre-training. Since image captioning models are data greedy, training on standard datasets can be limiting. Thus, pre-training on large-scale vision-and-language datasets, even if not well-curated, is a solid strategy for improving the captioning capabilities, as demonstrated in [80], [81], [83]. Moreover, new pre-training strategies could be devised to leverage the data available in

a self-supervised fashion, *e.g.* by reconstructing the inputs or predicting correlations, finally boosting the performance on downstream tasks such as image captioning.

Novel architectures and training strategies. The best performing paradigm for image captioning is currently the bottom-up one, which leverages object detectors for image regions encoding. Nonetheless, the surveyed work [77] explores a fully-Transformer paradigm, where image patches are directly applied to Transformer encoders, as in the Vision Transformer proposed in [75]. Although this first attempt underperforms the majority of previous works, it suggests that hybrid solutions, possibly integrating a Transformer-based object detector, might be a worthwhile future direction. Other promising directions entail exploring Neural Architecture Search [87], and applying the distillation mechanism to autoregressive models, as suggested by the results achieved for non-autoregressive models in [92], [94]. Finally, a promising exploration line is the design of new objectives functions for training. In particular, when a reinforcement learning phase is performed, rewards based on human feedback or interaction can be considered [217].

8.2 Focus on the open challenges

Generalization to different domains and increased diversity and naturalness of the generated captions are among the main open challenges for image captioning.

Generalizing to different domains. Image captioning models are usually trained on datasets that do not cover all possible real-life scenarios and, therefore, cannot generalize well to different contexts. As an example, in Fig. 10, we report some qualitative results with clear errors, indicating the difficulties in dealing with rare visual concepts. In this sense, further research efforts are needed towards a robust representation of visual concepts. Moreover, developments in image captioning variants such as novel objects captioning or controllable captioning could help to tackle this open issue. This would be strategic for adopting image captioning in specific applications, such as medicine, industrial products description, or cultural heritage.

Diversity and natural generation. As argued in [148], image captioning models should produce descriptions with three properties: semantic fidelity, *i.e.* reflecting the actual visual content, naturalness, *i.e.* reading as if they were written

by a human, and diversity, *i.e.* expressing notably different concepts as different humans would describe. However, most of the existing approaches emphasize only semantic fidelity. Although we discussed some attempts to encourage naturalness and diversity with conditional GANs [148], contrastive learning [192], variational auto-encoders [194], part-of-speech tagging [198], or word latent spaces [195], further research is needed to design models that are suitable for real-world applications.

8.3 Design of trustworthy AI solutions

Due to its potential in human-machine interactions, image captioning needs solutions that are transparent and acceptable for end-users, framed as interpretable results, overcome bias, and adequate evaluation.

The need for interpretability. People can naturally give explanations, highlight proofs, and express confidence in what they predict, also recognizing the need for more information before reaching a conclusion. Conversely, existing image captioning algorithms lack reliable and interpretable means for determining the cause of a particular output. In this respect, a possible strategy can be based on attention visualization, which loosely couples word predictions and image regions, indicating correlations and grounding [69]. However, further research is needed to shed more light on models explainability, focusing on how these deal with data from different modalities or novel concepts.

Tackling datasets bias. Since most vision-and-language datasets share common patterns and regularities, memorizing those patterns gives algorithms a shortcut to exploit unwanted correspondences. Therefore, datasets bias in human textual annotations or overrepresented visual concepts are major issues for any vision-and-language task. This topic has been investigated in the context of language generation [218] but is even more challenging in image captioning [219], where the joined ambiguity of visual and textual data must be taken into account. In this sense, some effort should be devoted to the study of fairness and bias in the image-description pairs. In this regard, two possible directions entail designing specific evaluation metrics and focusing on the robustness to unwanted correlations.

The role of evaluation. Despite the promising performance on the benchmark datasets, state-of-the-art approaches are not yet satisfactory when applied in the wild. A possible reason for this is the evaluation procedures used and their impact on the training approaches currently adopted. Captioning algorithms are trained to mimic ground truth sentences, which is somewhat a different task from understanding the visual content and expressing it in text. For this reason, the design of appropriate and reproducible evaluation protocols [220], [221], [222] and insightful metrics remains an open challenge in image captioning. Moreover, since the task is currently defined as a supervised one and thus is strongly influenced by the training data, the development of scores that do not need reference captions for assessing the performance would be key for a shift towards unsupervised image captioning.

ACKNOWLEDGMENTS

We thank CINECA, the Italian Supercomputing Center, for providing computational resources. This work has been supported by “Fondazione di Modena”, by the “Artificial Intelligence for Cultural Heritage (AI4CH)” project, cofunded by the Italian Ministry of Foreign Affairs and International Cooperation, and by the H2020 ICT-48-2020 HumanE-AI-NET and ELISE projects. We also want to thank the authors who provided us with the captions and model weights for some of the surveyed approaches.

REFERENCES

- [1] A. Ardila, B. Bernal, and M. Rosselli, “Language and visual perception associations: meta-analytic connectivity modeling of Brodmann Area 37,” *Behavioural Neurology*, 2015.
- [2] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, “Automatic image captioning,” in *ICME*, 2004.
- [3] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, “Every picture tells a story: Generating sentences from images,” in *ECCV*, 2010.
- [4] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu, “GCap: Graph-based Automatic Image Captioning,” in *CVPR Workshops*, 2004.
- [5] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, “I2t: Image parsing to text description,” *Proceedings of the IEEE*, 2010.
- [6] A. Gupta, Y. Verma, and C. Jawahar, “Choosing linguistics over vision to describe images,” in *AAAI*, 2012.
- [7] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, “Composing simple image descriptions using web-scale n-grams,” in *CoNLL*, 2011.
- [8] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, “Babytalk: Understanding and generating simple image descriptions,” *IEEE Trans. PAMI*, 2013.
- [9] A. Aker and R. Gaizauskas, “Generating image descriptions using dependency relational patterns,” in *ACL*, 2010.
- [10] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, and H. Daumé III, “Midge: Generating image descriptions from computer vision detections,” in *ACL*, 2012.
- [11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *CVPR*, 2015.
- [12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *CVPR*, 2015.
- [13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.
- [14] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *CVPR*, 2015.
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *NeurIPS*, 2012.
- [16] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, “Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN),” in *ICLR*, 2015.
- [17] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *CVPR*, 2015.
- [18] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *ICLR*, 2015.
- [19] X. Chen and C. Lawrence Zitnick, “Mind’s Eye: A Recurrent Visual Representation for Image Caption Generation,” in *CVPR*, 2015.
- [20] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, “From captions to visual concepts and back,” in *CVPR*, 2015.
- [21] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, “Guiding the Long-Short Term Memory model for Image Caption Generation,” in *ICCV*, 2015.
- [22] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in *CVPR*, 2016.

- [23] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What Value Do Explicit High Level Concepts Have in Vision to Language Problems?" in *CVPR*, 2016.
- [24] J. Gu, G. Wang, J. Cai, and T. Chen, "An Empirical Study of Language CNN for Image Captioning," in *ICCV*, 2017.
- [25] F. Chen, R. Ji, J. Su, Y. Wu, and Y. Wu, "StructCap: Structured Semantic Embedding for Image Captioning," in *ACM Multimedia*, 2017.
- [26] F. Chen, R. Ji, X. Sun, Y. Wu, and J. Su, "GroupCap: Group-based Image Captioning with Structured Relevance and Diversity Constraints," in *CVPR*, 2018.
- [27] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017.
- [28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [29] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *ICCV*, 2017.
- [30] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic Compositional Networks for Visual Captioning," in *CVPR*, 2017.
- [31] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.
- [32] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017.
- [33] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2014.
- [34] X. Chen, L. Ma, W. Jiang, J. Yao, and W. Liu, "Regularizing RNNs for Caption Generation by Reconstructing The Past with The Present," in *CVPR*, 2018.
- [35] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, "Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition," in *CVPR*, 2017.
- [36] H. Ge, Z. Yan, K. Zhang, M. Zhao, and L. Sun, "Exploring Overall Contextual Information for Image Captioning in Human-Like Cognitive Style," in *ICCV*, 2019.
- [37] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-Captioning: Coarse-to-Fine Learning for Image Captioning," in *AAAI*, 2018.
- [38] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review Networks for Caption Generation," in *NeurIPS*, 2016.
- [39] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning," in *CVPR*, 2017.
- [40] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, "Recurrent Fusion Network for Image Captioning," in *ECCV*, 2018.
- [41] Y. Sugano and A. Bulling, "Seeing with Humans: Gaze-Assisted Neural Image Captioning," *arXiv preprint arXiv:1608.05203*, 2016.
- [42] H. R. Tavakoli, R. Shetty, A. Borji, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in *ICCV*, 2017.
- [43] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *CVPR*, 2017.
- [44] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention," *ACM TOMM*, vol. 14, no. 2, pp. 1–21, 2018.
- [45] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.
- [46] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015.
- [47] —, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. PAMI*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [48] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2017.
- [49] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, "Reflective Decoding Network for Image Captioning," in *ICCV*, 2019.
- [50] Y. Qin, J. Du, Y. Zhang, and H. Lu, "Look Back and Predict Forward in Image Captioning," in *CVPR*, 2019.
- [51] L. Huang, W. Wang, Y. Xia, and J. Chen, "Adaptively Aligned Image Captioning via Adaptive Attention Time," in *NeurIPS*, 2019.
- [52] L. Wang, Z. Bai, Y. Zhang, and H. Lu, "Show, Recall, and Tell: Image Captioning with Recall Mechanism," in *AAAI*, 2020.
- [53] Z.-J. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning," *IEEE Trans. PAMI*, 2019.
- [54] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of Attention for Image Captioning," in *ICCV*, 2017.
- [55] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring Visual Relationship for Image Captioning," in *ECCV*, 2018.
- [56] L. Guo, J. Liu, J. Tang, J. Li, W. Luo, and H. Lu, "Aligning linguistic words and visual semantic units for image captioning," in *ACM Multimedia*, 2019.
- [57] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ICLR*, 2017.
- [58] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-Encoding Scene Graphs for Image Captioning," in *CVPR*, 2019.
- [59] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving Image Captioning with Better Use of Captions," in *ACL*, 2020.
- [60] T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy Parsing for Image Captioning," in *ICCV*, 2019.
- [61] K. S. Tai, R. Socher, and C. D. Manning, "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks," in *ACM Multimedia*, 2015.
- [62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.
- [63] X. Yang, H. Zhang, and J. Cai, "Learning to Collocate Neural Modules for Image Captioning," in *ICCV*, 2019.
- [64] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled Transformer for Image Captioning," in *ICCV*, 2019.
- [65] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image Captioning: Transforming Objects into Words," in *NeurIPS*, 2019.
- [66] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and Geometry-Aware Self-Attention Network for Image Captioning," in *CVPR*, 2020.
- [67] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on Attention for Image Captioning," in *ICCV*, 2019.
- [68] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-Linear Attention Networks for Image Captioning," in *CVPR*, 2020.
- [69] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," in *CVPR*, 2020.
- [70] F. Liu, X. Ren, X. Wu, S. Ge, W. Fan, Y. Zou, and X. Sun, "Prophet Attention: Predicting Attention with Future Attention," *NeurIPS*, 2020.
- [71] M. Cornia, L. Baraldi, and R. Cucchiara, "SMaRT: Training Shallow Memory-aware Transformers for Robotic Explainability," in *ICRA*, 2020.
- [72] J. Ji, Y. Luo, X. Sun, F. Chen, G. Luo, Y. Wu, Y. Gao, and R. Ji, "Improving Image Captioning by Leveraging Intra- and Inter-layer Global Representation in Transformer Network," in *AAAI*, 2021.
- [73] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-Level Collaborative Transformer for Image Captioning," in *AAAI*, 2021.
- [74] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, and R. Ji, "RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words," in *CVPR*, 2021.
- [75] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.
- [76] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021.
- [77] W. Liu, S. Chen, L. Guo, X. Zhu, and J. Liu, "CPTR: Full Transformer Network for Image Captioning," *arXiv preprint arXiv:2101.10804*, 2021.
- [78] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *EMNLP*, 2019.
- [79] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019.
- [80] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *ECCV*, 2020.

- [81] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified Vision-Language Pre-Training for Image Captioning and VQA," in *AAAI*, 2020.
- [82] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL*, 2018.
- [83] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "VinVL: Revisiting visual representations in vision-language models," in *CVPR*, 2021.
- [84] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [85] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural Baby Talk," in *CVPR*, 2018.
- [86] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions," in *CVPR*, 2019.
- [87] X. Zhu, W. Wang, L. Guo, and J. Liu, "AutoCaption: Image Captioning with Neural Architecture Search," *arXiv preprint arXiv:2012.09742*, 2020.
- [88] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *CVPR*, 2018.
- [89] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.
- [90] J. Gao, X. Meng, S. Wang, X. Li, S. Wang, S. Ma, and W. Gao, "Masked Non-Autoregressive Image Captioning," *arXiv preprint arXiv:1906.00717*, 2019.
- [91] Z.-c. Fei, "Fast Image Caption Generation with Position Alignment," *AAAI Workshops*, 2019.
- [92] L. Guo, J. Liu, X. Zhu, X. He, J. Jiang, and H. Lu, "Non-autoregressive image captioning with counterfactuals-critical multi-agent learning," *IJCAI*, 2020.
- [93] Z. Fei, "Iterative Back Modification for Faster Image Captioning," in *ACM Multimedia*, 2020.
- [94] L. Guo, J. Liu, X. Zhu, and H. Lu, "Fast Sequence Generation with Multi-Agent Reinforcement Learning," *arXiv preprint arXiv:2101.09698*, 2021.
- [95] P. Koehn, *Statistical Machine Translation*. Cambridge University Press, 2009.
- [96] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *ICLR*, 2016.
- [97] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, 1992.
- [98] W. Zaremba and I. Sutskever, "Reinforcement learning neural turing machines-revised," *arXiv preprint arXiv:1505.00521*, 2015.
- [99] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *ACL*, 2002.
- [100] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL Workshops*, 2004.
- [101] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *CVPR*, 2017.
- [102] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved Image Captioning via Policy Gradient Optimization of SPIDER," in *ICCV*, 2017.
- [103] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in *ECCV*, 2016.
- [104] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," in *CVPR*, 2015.
- [105] L. Zhang, F. Sung, F. Liu, T. Xiang, S. Gong, Y. Yang, and T. M. Hospedales, "Actor-Critic Sequence Training for Image Captioning," in *NeurIPS*, 2017.
- [106] J. Gao, S. Wang, S. Wang, S. Ma, and W. Gao, "Self-critical n-step Training for Image Captioning," in *CVPR*, 2019.
- [107] Q. Xia, H. Huang, N. Duan, D. Zhang, L. Ji, Z. Sui, E. Cui, T. Bharti, and M. Zhou, "XGPT: Cross-modal Generative Pre-Training for Image Captioning," *arXiv preprint arXiv:2003.01473*, 2020.
- [108] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014.
- [109] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *TACL*, 2014.
- [110] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *JAIR*, 2013.
- [111] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018.
- [112] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts," in *CVPR*, 2021.
- [113] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *NeurIPS*, 2011.
- [114] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning Images Taken by People Who Are Blind," in *ECCV*, 2020.
- [115] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *CVPR*, 2016.
- [116] X. Yang, H. Zhang, D. Jin, Y. Liu, C.-H. Wu, J. Tan, D. Xie, J. Wang, and X. Wang, "Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards," in *ECCV*, 2020.
- [117] A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk, "BreakingNews: Article Annotation by Image and Text Processing," *IEEE Trans. PAMI*, vol. 40, no. 5, pp. 1072–1085, 2017.
- [118] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, "Good news, everyone! context driven entity-aware captioning for news images," in *CVPR*, 2019.
- [119] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "TextCaps: a Dataset for Image Captioning with Reading Comprehension," in *ECCV*, 2020.
- [120] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," in *ECCV*, 2020.
- [121] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-Shot Text-to-Image Generation," *arXiv preprint arXiv:2102.12092*, 2021.
- [122] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020*, 2021.
- [123] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," California Institute of Technology, Tech. Rep., 2010.
- [124] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *ICVGIP*, 2008.
- [125] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," in *ICCV*, 2017.
- [126] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *ECCV*, 2016.
- [127] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations," in *ECCV*, 2018.
- [128] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *ICML*, 2016.
- [129] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *ACL Workshops*, 2005.
- [130] N. Sharif, L. White, M. Bennamoun, and S. A. A. Shah, "Nneval: Neural network based evaluation metric for image captioning," in *ECCV*, 2018.
- [131] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, "Learning to evaluate image captioning," in *CVPR*, 2018.
- [132] R. Shetty, M. Rohrbach, L. Anne Hendricks, M. Fritz, and B. Schiele, "Speaking the same language: Matching machine to human captions by adversarial training," in *ICCV*, 2017.
- [133] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *ICML*, 2015.
- [134] R. Bigazzi, F. Landi, M. Cornia, S. Cascianelli, L. Baraldi, and R. Cucchiara, "Explore and Explain: Self-supervised Navigation and Recounting," in *ICPR*, 2020.
- [135] M. Jiang, Q. Huang, L. Zhang, X. Wang, P. Zhang, Z. Gan, J. Diesner, and J. Gao, "TIGer: text-to-image grounding for image caption evaluation," in *ACL*, 2019.
- [136] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *ICLR*, 2020.
- [137] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "CLIP-Score: A Reference-free Evaluation Metric for Image Captioning," *arXiv preprint arXiv:2104.08718*, 2021.

- [138] O. Caglayan, P. Madhyastha, and L. Specia, "Curious Case of Language Generation Evaluation Metrics: A Cautionary Tale," *COLING*, 2020.
- [139] E. Van Miltenburg, D. Elliott, and P. Vossen, "Measuring the diversity of automatic image descriptions," in *COLING*, 2018.
- [140] Q. Wang and A. B. Chan, "Describing like humans: on diversity in image captioning," in *CVPR*, 2019.
- [141] Q. Wang, J. Wan, and A. B. Chan, "On Diversity in Image Captioning: Metrics and Methods," *IEEE Trans. PAMI*, 2020.
- [142] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," in *ACL*, 2017.
- [143] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NeurIPS*, 2013.
- [144] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *EMNLP*, 2014.
- [145] H. Lee, S. Yoon, F. Dernoncourt, D. S. Kim, T. Bui, and K. Jung, "ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT," in *EMNLP Workshops*, 2020.
- [146] S. Wang, Z. Yao, R. Wang, Z. Wu, and X. Chen, "FAIEr: Fidelity and Adequacy Ensured Image Caption Evaluation," in *CVPR*, 2021.
- [147] H. Lee, S. Yoon, F. Dernoncourt, T. Bui, and K. Jung, "UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning," in *ACL*, 2021.
- [148] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards Diverse and Natural Image Descriptions via a Conditional GAN," in *ICCV*, 2017.
- [149] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked Cross Attention for Image-Text Matching," in *ECCV*, 2018.
- [150] I. J. Unanue, J. Parnell, and M. Piccardi, "BERTTune: Fine-Tuning Neural Machine Translation with BERTScore," *ACL*, 2021.
- [151] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data," in *CVPR*, 2016.
- [152] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning Images with Diverse Objects," in *CVPR*, 2017.
- [153] H. Agrawal, K. Desai, X. Chen, R. Jain, D. Batra, D. Parikh, S. Lee, and P. Anderson, "nocaps: novel object captioning at scale," in *ICCV*, 2019.
- [154] T. Yao, Y. Pan, Y. Li, and T. Mei, "Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects," in *CVPR*, 2017.
- [155] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Pointing Novel Objects in Image Captioning," in *CVPR*, 2019.
- [156] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, "Decoupled Novel Object Captioner," in *ACM Multimedia*, 2018.
- [157] Q. Feng, Y. Wu, H. Fan, C. Yan, M. Xu, and Y. Yang, "Cascaded revision network for novel object captioning," *IEEE TCSVT*, 2020.
- [158] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Guided Open Vocabulary Image Captioning with Constrained Beam Search," in *EMNLP*, 2017.
- [159] X. Hu, X. Yin, K. Lin, L. Wang, L. Zhang, J. Gao, and Z. Liu, "VIVO: Visual Vocabulary Pre-Training for Novel Object Captioning," *AAAI*, 2020.
- [160] J. Gu, S. Joty, J. Cai, and G. Wang, "Unpaired image captioning by language pivoting," in *ECCV*, 2018.
- [161] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *CVPR*, 2019.
- [162] I. Laina, C. Rupprecht, and N. Navab, "Towards unsupervised image captioning with shared multimodal embeddings," in *ICCV*, 2019.
- [163] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," in *ICCV*, 2019.
- [164] D. Guo, Y. Wang, P. Song, and M. Wang, "Recurrent relational memory network for unsupervised image captioning," *IJCAI*, 2020.
- [165] H. Ben, Y. Pan, Y. Li, T. Yao, R. Hong, M. Wang, and T. Mei, "Unpaired Image Captioning with Semantic-Constrained Self-Learning," *IEEE Trans. Multimedia*, 2021.
- [166] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Image Captioning with Very Scarce Supervised Data: Adversarial Semi-Supervised Learning Approach," in *EMNLP*, 2019.
- [167] R. Del Chiaro, B. Twardowski, A. D. Bagdanov, and J. van de Weijer, "RATT: Recurrent Attention to Transient Tasks for Continual Image Captioning," *NeurIPS*, 2020.
- [168] G. Nguyen, T. J. Jun, T. Tran, T. Yalaw, and D. Kim, "ContCap: A scalable framework for continual image captioning," *arXiv preprint arXiv:1909.08745*, 2019.
- [169] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional Localization Networks for Dense Captioning," in *CVPR*, 2016.
- [170] L. Yang, K. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context," in *CVPR*, 2017.
- [171] X. Li, S. Jiang, and J. Han, "Learning object context for dense captioning," in *AAAI*, 2019.
- [172] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, "Context and attribute grounded dense captioning," in *CVPR*, 2019.
- [173] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Dense Relational Captioning: Triple-Stream Networks for Relationship-Based Captioning," in *CVPR*, 2019.
- [174] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *CVPR*, 2017.
- [175] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition GAN for visual paragraph generation," in *ICCV*, 2017.
- [176] Y. Mao, C. Zhou, X. Wang, and R. Li, "Show and Tell More: Topic-Oriented Multi-Sentence Image Captioning," in *IJCAI*, 2018.
- [177] L. Melas-Kyriazi, A. M. Rush, and G. Han, "Training for diversity in image paragraph captioning," in *EMNLP*, 2018.
- [178] M. Chatterjee and A. G. Schwing, "Diverse and coherent paragraph generation from images," in *ECCV*, 2018.
- [179] Y. Luo, Z. Huang, Z. Zhang, Z. Wang, J. Li, and Y. Yang, "Curiosity-driven reinforcement learning for diverse visual paragraph generation," in *ACM Multimedia*, 2019.
- [180] W. Che, X. Fan, R. Xiong, and D. Zhao, "Visual relationship embedding network for image paragraph generation," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2307–2320, 2019.
- [181] Z. Yang, Y. Lu, J. Wang, X. Yin, D. Florencio, L. Wang, C. Zhang, L. Zhang, and J. Luo, "TAP: Text-Aware Pre-training for Text-VQA and Text-Caption," in *CVPR*, 2021.
- [182] J. Wang, J. Tang, and J. Luo, "Multimodal Attention with Image Text Spatial Relationship for OCR-Based Image Captioning," in *ACM Multimedia*, 2020.
- [183] J. Wang, J. Tang, M. Yang, X. Bai, and J. Luo, "Improving OCR-based Image Captioning by Incorporating Geometrical Relationship," in *CVPR*, 2021.
- [184] Q. Zhu, C. Gao, P. Wang, and Q. Wu, "Simple is not Easy: A Simple Strong Baseline for TextVQA and TextCaps," in *AAAI*, 2021.
- [185] G. Xu, S. Niu, M. Tan, Y. Luo, Q. Du, and Q. Wu, "Towards Accurate Text-based Image Captioning with Content Diversity Exploration," in *CVPR*, 2021.
- [186] H. Jhamtani and T. Berg-Kirkpatrick, "Learning to describe differences between pairs of similar images," in *EMNLP*, 2018.
- [187] D. H. Park, T. Darrell, and A. Rohrbach, "Robust Change Captioning," in *CVPR*, 2019.
- [188] X. Shi, X. Yang, J. Gu, S. Joty, and J. Cai, "Finding It at Another Side: A Viewpoint-Adapted Matching Encoder for Change Captioning," in *ECCV*, 2020.
- [189] Q. Huang, Y. Liang, J. Wei, C. Yi, H. Liang, H.-f. Leung, and Q. Li, "Image Difference Captioning with Instance-Level Fine-Grained Feature Representation," *IEEE Trans. Multimedia*, 2021.
- [190] M. Hosseinzadeh and Y. Wang, "Image Change Captioning by Learning from an Auxiliary Task," in *CVPR*, 2021.
- [191] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. J. Crandall, and D. Batra, "Diverse Beam Search for Improved Description of Complex Scenes," in *AAAI*, 2018.
- [192] B. Dai and D. Lin, "Contrastive learning for image captioning," *NeurIPS*, 2017.
- [193] L. Liu, J. Tang, X. Wan, and Z. Guo, "Generating diverse and descriptive image captions using visual paraphrases," in *ICCV*, 2019.
- [194] L. Wang, A. G. Schwing, and S. Lazebnik, "Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space," *NeurIPS*, 2017.
- [195] J. Aneja, H. Agrawal, D. Batra, and A. Schwing, "Sequential latent spaces for modeling the intention during diverse image captioning," in *ICCV*, 2019.

- [196] F. Chen, R. Ji, J. Ji, X. Sun, B. Zhang, X. Ge, Y. Wu, F. Huang, and Y. Wang, "Variational structured semantic inference for diverse image captioning," in *NeurIPS*, 2019.
- [197] S. Mahajan and S. Roth, "Diverse image captioning with context-object split latent spaces," *NeurIPS*, 2020.
- [198] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. Forsyth, "Fast, diverse and accurate image captioning guided by part-of-speech," in *CVPR*, 2019.
- [199] D. Elliott, S. Frank, and E. Hasler, "Multilingual image description with neural sequence models," *ICLR*, 2015.
- [200] X. Li, C. Xu, X. Wang, W. Lan, Z. Jia, G. Yang, and J. Xu, "COCO-CN for cross-lingual image tagging, captioning, and retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2347–2360, 2019.
- [201] T. Miyazaki and N. Shimizu, "Cross-lingual image caption generation," in *ACM Multimedia*, 2016.
- [202] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30K: Multilingual English-German Image Descriptions," in *ACL Workshops*, 2016.
- [203] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," in *ACM Multimedia*, 2017.
- [204] Y. Song, S. Chen, Y. Zhao, and Q. Jin, "Unpaired cross-lingual image caption generation with self-supervised rewards," in *ACM Multimedia*, 2019.
- [205] Y. Wu, S. Zhao, J. Chen, Y. Zhang, X. Yuan, and Z. Su, "Improving captioning for low-resource languages by cycle consistency," in *ICME*, 2019.
- [206] A. Chen, X. Huang, H. Lin, and X. Li, "Towards annotation-free evaluation of cross-lingual image captioning," in *ACM Multimedia Asia*, 2021.
- [207] C. Chunseong Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in *CVPR*, 2017.
- [208] C. C. Park, B. Kim, and G. Kim, "Towards personalized image captioning via multimodal memory networks," *IEEE Trans. PAMI*, 2018.
- [209] W. Zhang, Y. Ying, P. Lu, and H. Zha, "Learning Long-and Short-Term User Literal-Preference with Multimodal Hierarchical Transformer Network for Personalized Image Caption," in *AAAI*, 2020.
- [210] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "StyleNet: Generating Attractive Visual Captions with Styles," in *CVPR*, 2017.
- [211] A. Mathews, L. Xie, and X. He, "Semstyle: Learning to generate stylised image captions using unaligned text," in *CVPR*, 2018.
- [212] L. Guo, J. Liu, P. Yao, J. Li, and H. Lu, "Mscap: Multi-style image captioning with unpaired stylized text," in *CVPR*, 2019.
- [213] W. Zhao, X. Wu, and X. Zhang, "MemCap: Memorizing style knowledge for image captioning," in *AAAI*, 2020.
- [214] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging image captioning via personality," in *CVPR*, 2019.
- [215] L. Chen, Z. Jiang, J. Xiao, and W. Liu, "Human-like Controllable Image Captioning with Verb-specific Semantic Roles," in *CVPR*, 2021.
- [216] Z. Meng, L. Yu, N. Zhang, T. Berg, B. Damavandi, V. Singh, and A. Bearman, "Connecting What to Say With Where to Look by Modeling Human Attention Traces," in *CVPR*, 2021.
- [217] T. Shen, A. Kar, and S. Fidler, "Learning to caption images through a lifetime by asking questions," in *ICCV*, 2019.
- [218] O. U. Florez, "On the Unintended Social Bias of Training Language Generation Models with Data from Local Media," in *ICLR*, 2019.
- [219] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *ECCV*, 2018.
- [220] M. Hodosh and J. Hockenmaier, "Focused evaluation for image description with binary forced-choice tasks," in *ACL Workshops*, 2016.
- [221] H. Xie, T. Sherborne, A. Kuhnle, and A. Copestake, "Going beneath the surface: Evaluating image captioning for grammaticality, truthfulness and diversity," *arXiv preprint arXiv:1912.08960*, 2019.
- [222] M. Alikhani, P. Sharma, S. Li, R. Soricut, and M. Stone, "Cross-modal coherence modeling for caption generation," in *ACL*, 2020.



Matteo Stefanini received the M.Sc. degree in Computer Engineering cum laude from the University of Modena and Reggio Emilia, in 2018. He is currently pursuing a PhD degree in Information and Communication Technologies at the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia. His research activities involve the integration of vision and language modalities, with a primary focus on image captioning and Transformer-based architectures.



Marcella Cornia received the M.Sc. degree in Computer Engineering and the Ph.D. degree cum laude in Information and Communication Technologies from the University of Modena and Reggio Emilia, in 2016 and 2020, respectively. She is currently a Postdoctoral Researcher with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia. She has authored or coauthored more than 30 publications in scientific journals and international conference proceedings.



Lorenzo Baraldi received the M.Sc. degree in Computer Engineering and the Ph.D. degree cum laude in Information and Communication Technologies from the University of Modena and Reggio Emilia, in 2014 and 2018, respectively. He is currently an Assistant Professor with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia. He was a Research Intern at Facebook AI Research (FAIR) in 2017. He has authored or coauthored more than 60 publications in scientific journals

and international conference proceedings.



Silvia Cascianelli received the M.Sc. degree in Information and Automation Engineering and the Ph.D. degree cum laude in Information and Industrial Engineering from the University of Perugia, in 2015 and 2019, respectively. She is currently an Assistant Professor with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia. She was a Visitor Researcher at the Queen Mary University of London between 2017 and 2018. She has authored or coauthored more than 30 publications

in scientific journals and international conference proceedings.



Giuseppe Fiameni is a Data Scientist at NVIDIA where he oversees the NVIDIA AI Technology Centre in Italy, a collaboration among NVIDIA, CINI and CINECA to accelerate academic research in the field of AI through collaboration projects. He has been working as HPC specialist at CINECA, the largest HPC facility in Italy, for more than 14 years providing support for large-scale data analytics workloads. Research interests include large scale deep learning models, system architectures, massive data engineering,

video action detection and anticipation.



Rita Cucchiara received the M.Sc. degree in Electronics Engineering and the Ph.D. degree in Computer Engineering from the University of Bologna, in 1989 and 1992, respectively. She is currently a Full Professor of Computer Engineering with the University of Modena and Reggio Emilia, where she heads the ImageLab Laboratory. She has authored or coauthored more than 400 papers in journals and international proceedings, and has been a coordinator of several projects in computer vision and pattern recognition. Prof. Cucchiara is the Director of the Italian Laboratory in Artificial Intelligence and Intelligent Systems, a Member of the Advisory Board of the Computer Vision Foundation, and the Director of the ELLIS Unit of Modena (Unimore).