This is the peer reviewd version of the followng article:

From Show to Tell: A Survey on Deep Learning-based Image Captioning / Stefanini, Matteo; Cornia, Marcella; Baraldi, Lorenzo; Cascianelli, Silvia; Fiameni, Giuseppe; Cucchiara, Rita. - In: IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. - ISSN 0162-8828. - 45:1(2023), pp. 539-559. [10.1109/TPAMI.2022.3148210]

Terms of use:

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

19/04/2024 02:40

(Article begins on next page)

From Show to Tell: A Survey on Deep Learning-based 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, 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 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.

The early-proposed approaches to image captioning have entailed description retrieval [2], [3], [4], [5], [6], [7] or template filling and hand-crafted natural language generation techniques [8], [9], [10], [11], [12], [13], [14], [15]. While these have been treated in other surveys [16], [17], [18], image captioning is currently based on the usage of deep learning-based generative models. In its standard configuration, the task is an image-to-sequence problem whose inputs are pixels. These inputs 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 subwords decoded according to a given vocabulary.

- *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*

In these few years, the research community has improved model design considerably: from the first deep learning-based proposals adopting Recurrent Neural Networks (RNNs) fed with global image descriptors, methods have been enriched with attentive approaches and reinforcement learning up to the breakthroughs of Transformers and self-attention and 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 metrics to compare results with human-generated ground-truths. However, despite the investigation and improvements achieved in these years, image captioning is still far from being considered a solved task.

Several domain-specific proposals and variants of the task have also been investigated to accommodate for different user needs and descriptions styles. According to [19], [20], indeed, image captions can be perceptual, when focusing on low-level visual attributes; non-visual, when reporting implicit and contextual information; conceptual, when describing the actual visual content (*e.g.* visual entities and their relations). While the latter is commonly recognized as the target of the image captioning task, this definition encompasses descriptions focusing on different aspects and at various levels of detail (*e.g.* including attributes or not, mentioning named entities or high-level concepts only, describing salient parts only, or also finer details).

With the aim of providing a testament to the journey that captioning has taken so far, and with that of encouraging novel ideas, we trace a holistic overview of techniques, models, and task variants developed in the last years. Furthermore, we review datasets and evaluation metrics and perform quantitative comparisons of the main approaches. Finally, we discuss open challenges and future directions.

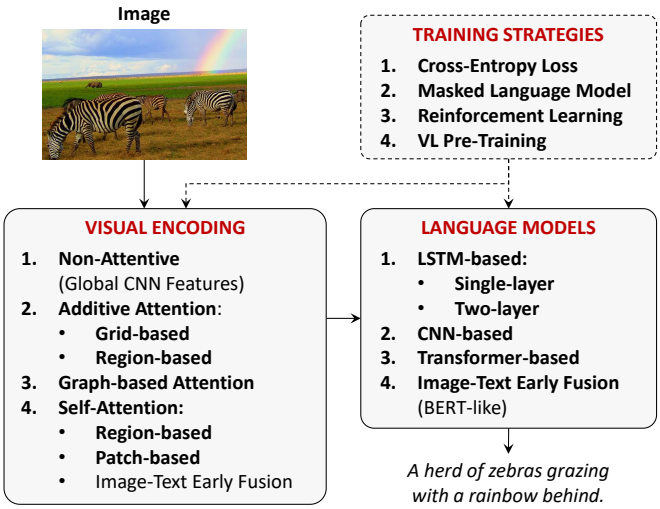


Fig. 1: Overview of the image captioning task and taxonomy of the most relevant approaches.

Contributions. To sum up, the contributions of this survey are as follows:

- Following the inherent dual nature of captioning models, we develop taxonomies for visual encoding and language modeling approaches and describe their key aspects and limitations.
- We review the training strategies adopted in the literature over the past years and the recent advancement obtained by the pre-training paradigm and masked language model losses.
- We review the main datasets used to explore image captioning, both domain-generic benchmarks and domain-specific datasets collected to investigate specific aspects.
- We analyze both standard and non-standard metrics adopted for performance evaluation and the characteristics of the caption they highlight.
- We present a quantitative comparison of the main image captioning methods considering 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.
- We give an overview of many variants of the task and discuss open challenges and future directions.

Compared to previous surveys on image captioning [16], [17], [18], [21], [22], we provide a comprehensive and updated view on deep learning-based generative captioning models. We perform a deeper analysis of proposed approaches and survey a considerably larger number of papers on the topic. Also, we cover non-standard evaluation metrics, which are disregarded by other works, discuss their characteristics, and employ them in a quantitative evaluation of state-of-the-art methods. Moreover, we tackle emerging variants of the task and a broader set of available datasets.

2 Visual Encoding

Providing an effective representation of the visual content is the first challenge of an image captioning pipeline. The current approaches for 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 representations, which are then used as a conditioning element for the language model (Fig. 2a). This is the approach employed in the seminal "Show and Tell" paper [23][1], where the output of GoogleNet [24] is fed to the initial hidden state of the language model. In the same year, Karpathy *et al.* [25] used global features extracted from AlexNet [26] as the input for a language model. Further, Mao *et al.* [27] and Donahue *et al.* [28] injected global features extracted from the VGG network [29] at each time-step of the language model.

Global CNN features were then employed in a large variety of image captioning models [30], [31], [32], [33], [34], [35], [36], [37]. Notably, Rennie *et al.* [38] introduced the FC model, in which images are encoded using a ResNet-101 [39], preserving their original dimensions. Other approaches [40], [41] integrated high-level attributes or tags, represented as a probability distribution over the most common words of the training captions.

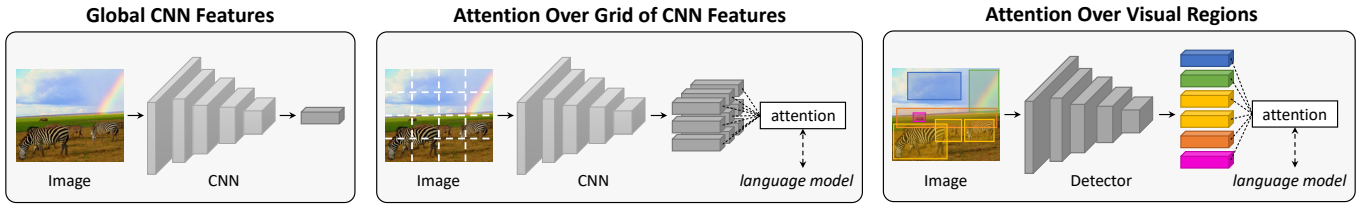
The main advantage of employing global CNN features resides in their simplicity and compactness of representation, which embraces the capacity to extract and condense information from the whole input and to consider the overall context of an image. However, this paradigm also leads to excessive compression of information and lacks granularity, 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 [38], [42], [43]. For instance, Dai *et al.* [44] have employed 2D activation maps in place of 1D global feature vectors to bring spatial structure directly in the language model. Drawing from machine translation literature, a big portion of the captioning community has instead employed the additive attention mechanism (Fig. 2b), which 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.* [45] (also known as *additive attention*), a single-layer feed-forward neural network with a hyperbolic tangent

1. The title of this survey is a tribute of this pioneering work.



(a) (b) (c)

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 the language model; **(c)** image region features coming from a detector, together with an attention mechanism.

non-linearity is used to compute attention weights. Formally, given two generic sets of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\{\mathbf{h}_1, \ldots, \mathbf{h}_m\}$, the additive attention score between $\mathbf{h}_i$ and $\mathbf{x}_j$ is computed as follows:

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

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 $p\left(\mathbf{x}_j \mid \mathbf{h}_i\right)$, 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 visual representations with the hidden states of a language model.

Attending convolutional activations. Xu *et al.* [42] 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 [29], 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.

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 [40], [43], [46], [47], [48], [49].

Review networks – For instance, Yang *et al.* [50] 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.* [51] 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.* [52] 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 with stimulus-based attention. This idea was first explored by Sugano and Bulling [53] who exploited human eye fixations for image captioning by including normalized fixation histograms over the image as an input to the soft-attention module of [42] and weighing the attended image regions based on whether these are fixated or not. Subsequent works on this line [54], [55], [56], [57] employed saliency maps as a form of additional attention source.

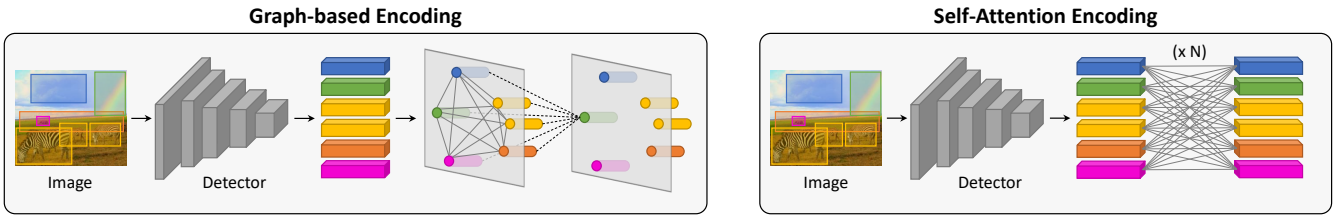
2.3 Attention Over Visual Regions

The intuition of using saliency boils down to neuroscience, which suggests that our brain integrates a top-down reasoning process with a bottom-up flow of visual signals. The top-down path consists of predicting the upcoming sensory input by leveraging our knowledge and inductive bias, while the bottom-up flow provides visual stimuli adjusting the previous predictions. Additive attention can be thought of as a top-down system. In this mechanism, the language model predicts the next word while attending a feature grid, whose geometry is irrespective of the image content.

Bottom-up and top-down attention. Differently from saliency-based approaches [57], in the solution proposed by Anderson *et al.* [58] the bottom-up path is defined by an object detector in charge of proposing image regions. This is then coupled with a top-down mechanism that learns to weigh each region for each word prediction (see Fig. 2c). In this approach, Faster R-CNN [59], [60] is adopted to detect objects, obtaining a pooled feature vector for each region 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 [61] 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 image region features 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 [62], [63], [64], [65]. 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.* [66] 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.



(a) (b)

Fig. 3: Summary of the two most recent visual encoding strategies for image captioning: **(a)** graph-based encoding of visual regions; **(b)** self-attention-based encoding over image region features.

Geometric Transforms – Pedersoli *et al.* [67] proposed to use spatial transformers for generating image-specific attention areas by regressing region proposals in a weakly-supervised fashion. 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.

Spatial and semantic graphs. The first attempt in this sense is due to Yao *et al.* [68], followed by Guo *et al.* [69], who proposed the use of a graph convolutional network (GCN) [70] to integrate both semantic and spatial relationships between objects. The semantic relationships graph is obtained by applying a classifier pre-trained on Visual Genome [61] 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.* [71] 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 relations. On the same line, Shi *et al.* [72] 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.

Hierarchical trees. As a special case of a graph-based encoding, Yao *et al.* [73] 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.

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 by Vaswani *et al.* [74] for machine translation and language understanding tasks, giving birth to the Transformer architecture and its 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 $\boldsymbol{Q}$, a set of key vectors $\boldsymbol{K}$, and a set of value vectors $\boldsymbol{V}$, both containing n_k elements. The operator takes a weighted sum of value vectors according to a similarity distribution between query and key vectors:

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

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.* [75] used a self-attentive module to encode relationships between features coming from an object detector. Later, Li *et al.* [76] 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 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 [77], [78], [79], [80], [81].

Geometry-aware encoding – Herdade *et al.* [77] 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.* [78] proposed a normalized and geometry-aware version of self-attention that makes use of the relative geometry relationships between input objects. Further, He *et al.* [82] introduced a spatial graph transformer, which considers different categories of spatial relationship between

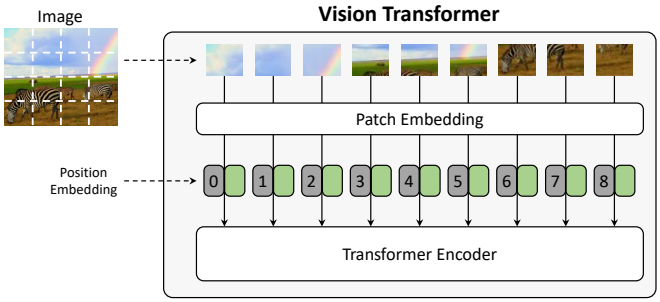


Fig. 4: Vision Transformer encoding. The image is split into fixed-size patches, linearly embedded, added to position embeddings, and fed to a standard Transformer encoder.

detections (*e.g.*, parent, neighbor, child) when performing attention.

Attention on Attention – Huang *et al.* [79] proposed an extension of the attention operator in which the final attended information is weighted by a gate guided by the context. Specifically, the output of the self-attention is concatenated with the queries, then an information and a gate vector are computed and finally multiplied together. In their encoder, they employed this mechanism to refine the visual features. This method is then adopted by later models such as [83].

X-Linear Attention – Pan *et al.* [80] 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.* [81], [84] 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.

Other self-attention-based approaches. Ji *et al.* [85] 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.* [86] 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. On a different line, the architecture proposed by Liu *et al.* [87] is based on an attention module to align grid or detection features with visual words extracted from a concept extractor and to obtain semantic-grounded encodings.

Attention on grid features and patches. Other than applying the attention operator on detections, the role of grid features has been recently re-evaluated [88]. For instance, the approach proposed by Zhang *et al.* [89] applies self-attention directly to grid features, incorporating their relative geometry relationships into self-attention computation. Transformer-like architectures can also be applied directly on image patches, thus excluding the usage of the convolutional operator [90], [91] (Fig. 4). On this line, Liu *et al.* [92] devised the first convolution-free architecture for image captioning. Specifically, a pre-trained Vision Transformer network (*i.e.* ViT [90]) is adopted as encoder, and a standard Transformer decoder is employed to generate captions. Interestingly, the same visual encoding approach has been adopted in CLIP [93] and SimVLM [94], with the difference that the visual encoder is trained from scratch on large-scale noisy data. CLIP-based features have then been used by subsequent captioning approaches [95], [96], [97].

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 [98], [99] and early-fusion strategies [100], [101]. For example, following the BERT architecture [102], Zhou *et al.* [101] 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.

On the same line, Li *et al.* [100] proposed OSCAR, a BERT-like architecture that includes object tags as anchor points to ease the semantic alignment between images and text. 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. Later, Zhang *et al.* [103] 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. On this line, Hu *et al.* [104] improved the VinVL model by scaling up its size and using larger scale noisy data to pre-train.

2.6 Discussion

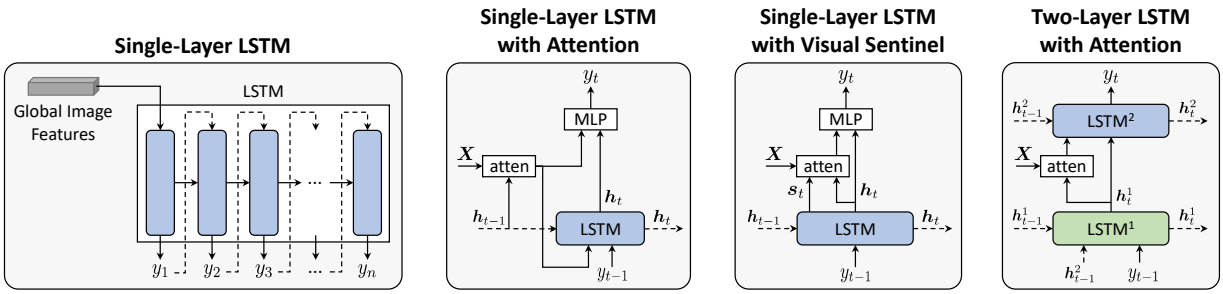
After the emergence of global features and grid features, region-based features have been the state-of-the-art choice in image captioning for years thanks to their compelling performances. Recently, however, different factors are reopening the discussion on which feature model is most appropriate for image captioning, ranging from the performance of better-trained grid features [88] to the emergence of self-attentive visual encoders [90] and large-scale multi-modal models like CLIP [93]. Recent strategies encompass training better object detectors on large-scale data [103] or employing end-to-end visual models trained from scratch [94]. 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 goal of a language model is to predict the probability of a given sequence of words to occur in a sentence. As such, it is a crucial component in image captioning, as it gives the ability to deal with natural language as a stochastic process.

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

$$P(y_1, y_2, \ldots y_n \mid \boldsymbol{X}) = \prod_{i=1}^{n} P(y_i \mid y_1, y_2, \ldots, y_{i-1}, \boldsymbol{X}), \tag{3}$$



(a) (b) (c) (d)

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 [42]; **(c)** LSTM with attention, in the variant proposed in [43]; **(d)** two-layer LSTM with attention, in the style of the bottom-up top-down approach by Anderson *et al.* [58]. In all figures, $\boldsymbol{X}$ represents either a grid of CNN features or image region features extracted by an object detector.

where $\boldsymbol{X}$ 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 usually 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 [105] 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.* [23]. 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, Xu *et al.* [42] introduced the additive attention mechanism. As depicted in Fig. 5b, in this case, the previous hidden state guides the attention mechanism over the visual features $\boldsymbol{X}$, 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 [50], [51], [67], while others have proposed significant modifications, summarized below.

Visual sentinel – Lu *et al.* [43] 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.

Hidden state reconstruction – Chen *et al.* [46] 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.* [48] enhance context modeling by 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.* [47] 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.* [49] 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.

Semantic-guided LSTM – Jia *et al.* [32] proposed an extension of LSTM that includes semantic information extracted from the image to guide the generation. Specifically, the semantic information is used as an extra input to each gate in the LSTM block.

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.* [28] 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.* [58] 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 [68], [71], [72], [73]. 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.* [106] 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.* [62] introduced two reflective modules: while the first computes the relevance between hidden states from all the past predicted words and the current one, the second improves the syntactic structure of the sentence by guiding the generation process with words common position information.

Look back and predict forward – On a similar line, Qin *et al.* [63] 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.* [64] 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.

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 models [79], [80], [83], [107]. In particular, Huang *et al.* [79] augmented the LSTM with the Attention on Attention operator, which computes another step of attention on top of visual self-attention. Pan *et al.* [80] introduced the X-Linear attention block, which enhances self-attention with second-order interactions and improves both the visual encoding and the language model. On a different line, Zhu *et al.* [107] applied the neural architecture search paradigm to select the connections between layers and the operations within gates of RNN-based image captioning language models, using a decoder enriched with self-attention [80].

3.2 Convolutional Language Models

A worth-to-mention approach is that proposed by Aneya *et al.* [108], 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.

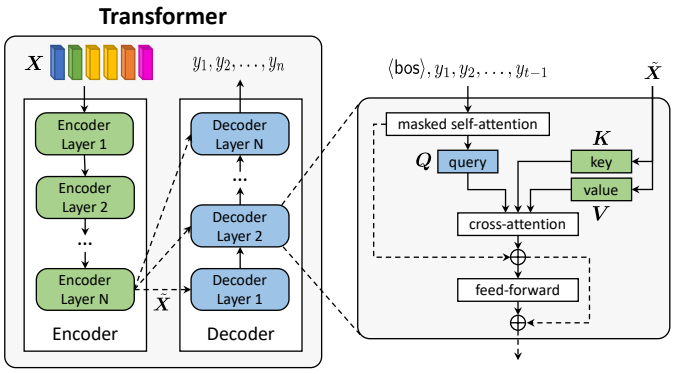


Fig. 6: Schema of the Transformer-based language model. The caption generation is performed via masked self-attention over previously generated tokens and cross-attention with encoded visual features.

3.3 Transformer-based Architectures

The fully-attentive paradigm proposed by Vaswani *et al.* [74] 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 [102] and GPT [109], and the standard de-facto architecture for many language understanding tasks. As image captioning can be cast as a sequence-to-sequence problem, 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 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 decoder has been employed in some image captioning models without significant architectural modifications [77], [78], [86], [94]. Besides, some variants have been proposed to improve language generation and visual feature encoding.

Gating mechanisms. Li *et al.* [76] 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.* [85] 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.* [81] 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

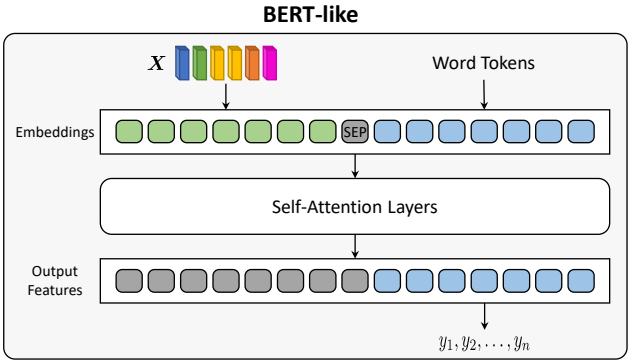


Fig. 7: Schema of a BERT-like language model. A single stream of attentive layers processes both image regions and word tokens and generates the output caption.

of all the encoding layers independently and a gate that weights these contributions guided by the text query. In [94], [97], the decoder architecture is again employed in conjunction with textual prefixes, also extracted from pre-trained visual-semantic models and employed as visual tags.

3.4 BERT-like Architectures

Despite the encoder-decoder paradigm being a common approach to image captioning, some works have revisited captioning architectures to exploit a BERT-like [102] structure in which the visual and textual modalities are fused together in the early stages (Fig. 7). 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 widely adopted in works that exploit pre-training [100], [101], [103]. The first example is due to Zhou *et al.* [101], who developed a unified model that fuses visual and textual modalities 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.* [100] 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 [110], [111], [112], [113]. 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. Subsequent methods, instead, employ reinforcement learning techniques to improve the final results. Specifically, 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 [111], [113]. These works also leverage knowledge distillation on unlabeled data and a post-processing step to remove identical consecutive tokens.

3.6 Discussion

Recurrent models have been the standard for many years, and their application brought to the development of clever and successful ideas that can be integrated also into non-recurrent solutions. However, they are slow to train and struggle to maintain long-term dependencies: these drawbacks are alleviated by autoregressive and Transformer-based solutions that recently gained popularity. 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 either encoder-decoder or BERT-like architectures, often in conjunction with textual tags. This strategy led to impressive performance, suggesting that visual and textual semantic relations can be inferred and learned also from not well-curated data [94], [100], [104]. BERT-like architectures are suitable for such a massive pre-training but are not generative architectures by design. Massive pre-training on generative-oriented architectures [94], [97] is currently a worth-exploring direction, which leads to performances that are at least on-pair with the early-fusion counterparts.

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 [114] 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*; 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:

$$L_{XE}(\theta) = -\sum_{i=1}^{n} \log \left(P\left(y_i \mid y_{1:i-1}, \boldsymbol{X} \right) \right), \tag{4}$$

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 $\boldsymbol{X}$ 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 [115] 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 [102] architecture. 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 [100], [103].

4.3 Reinforcement Learning

Given the limitations of word-level training strategies observed when using limited amounts of data, 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, and 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.* [115], which introduced the usage of the REINFORCE algorithm [116] adopting BLEU [117] and ROUGE [118] as reward signals. Ren *et al.* [119] 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.* [120] proposed to use as reward a linear combination of SPICE [121] and CIDEr [122], called SPIDEr. Finally, the most widely adopted strategy [81], [123], [124], introduced by Rennie *et al.* [38], entails using the CIDEr score, as it correlates better with human judgment [122]. The reward is normalized with respect to a baseline value to reduce variance. Formally, to compute the loss gradient, beam search and greedy decoding are leveraged as follows:

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

where $\boldsymbol{w}^i$ 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 [38], or as the average reward of the beam candidates [81].

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 Large-scale Pre-Training

In the context of vision-and-language pre-training in early-fusion architectures, 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 [102], 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 [100], [103]. 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 [125].

On the other hand, cross-entropy has also been used when pre-training on noisy captions [94], [97], sometimes also employing prefixes. PrefixLM [94] has indeed proved to be a valuable strategy that enables bidirectional attention within the prefix sequence, and thus, it is applicable for both decoder-only and encoder-decoder sequence-to-sequence language models. Noticeably, some large-scale models pre-trained on noisy data under this setting can achieve state-of-the-art performance without requiring a fine-tuning stage with Reinforcement [94].

Finally, we notice that image captioning can be used as a pre-training task to efficiently learn visual representations, which can benefit downstream tasks such as image classification, object detection, and instance segmentation [126].

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

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.

TABLE 1: Overview of the main image captioning datasets.

	Domain	Nb. Images	Nb. Caps (per Image)	Vocab Size	Nb. Words (per Cap.)
COCO [128]	Generic	132K	5	27K (10K)	10.5
Flickr30K [129]	Generic	31K	5	18K (7K)	12.4
Flickr8K [19]	Generic	8K	5	8K (3K)	10.9
CC3M [130]	Generic	3.3M	1	48K (25K)	10.3
CC12M [131]	Generic	12.4M	1	523K (163K)	20.0
SBU Captions [4]	Generic	1M	1	238K (46K)	12.1
VizWiz [132]	Assistive	70K	5	20K (8K)	13.0
CUB-200 [133]	Birds	12K	10	6K (2K)	15.2
Oxford-102 [133]	Flowers	8K	10	5K (2K)	14.1
Fashion Cap. [134]	Fashion	130K	1	17K (16K)	21.0
BreakingNews [135]	News	115K	1	85K (10K)	28.1
GoodNews [136]	News	466K	1	192K (54K)	18.2
TextCaps [137]	OCR	28K	5/6	44K (13K)	12.4
Loc. Narratives [138]	Generic	849K	1/5	16K (7K)	41.8

captions is usually long-tailed, thus, when using word-level dictionaries, the common practice is to include in the vocabulary only those terms whose frequency is above a predefined threshold. Recently, however, using subword-based tokenization approaches like BPE [127] is a popular choice that allows avoiding dataset pre-processing. 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, a procedure that guides the development of image captioning strategies by allowing to identify suitable directions. Datasets used as benchmarks should be representative of the task at hand, both in terms of the challenges and ideal expected results (*i.e.* achievable human performance). Further, they should contain a large number of generic-domain images, each associated with multiple captions.

Early image captioning architectures [25], [27], [28] were commonly trained and tested on the **Flickr30K** [129] and **Flickr8K** [19] datasets, consisting of pictures collected from the Flickr website, containing everyday activities, events, and scenes, paired with five captions each. Currently, the most commonly used dataset is **Microsoft COCO** [128], which consists of images of complex scenes with people, animals, and common everyday objects in their context. It contains more than 120,000 images, each annotated with five 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.* [25], 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 server[2].

5.1.2 Pre-training datasets

Although training on large well-curated datasets is a sound approach, some works [94], [99], [100], [104] have demonstrated the benefits of pre-training on even bigger vision-and-language datasets, which can be either image captioning datasets of lower-quality captions or datasets collected for other tasks (*e.g.* visual question answering [100], [101], text-to-image generation [139], image-caption association [93]). Among the datasets used for pre-training, that have been specifically collected for image captioning, it is worth mentioning **SBU Captions** [4], originally used for tackling image captioning as a retrieval task [19], which contains around 1 million image-text pairs, collected from the Flickr website. Similarly, **YFCC100M** [140] is composed of 100 million media objects in which 14.8 million images are available with automatically-collected textual descriptions. Later, the **Conceptual Captions** [130], [131] 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. Differently from

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

previous datasets, **Wikipedia-based Image Text** (WIT) [141] provides images coming from Wikipedia together with various metadata extracted from the original pages, with approximately 5.3 million images available with the corresponding descriptions in English. Although the large scale and variety in caption style make all these datasets particularly interesting for pre-training, the contained captions can be noisy, and the availability of images is not always guaranteed since most of them 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. Accordingly, some pre-training datasets are currently not publicly available, such as **ALIGN** [94], [142] and **ALT-200** [104], respectively containing 1.8 billion and 200 million noisy image-text pairs, or the datasets used to train DALL-E [139] and CLIP [93] consisting of 250 and 400 million pairs.

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 the visual domain is the **VizWiz Captions** [132] 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** [143] and the **Oxford-102** [144] datasets, which contain images of birds and flowers, respectively, that have been paired with ten captions each by Reed *et al.* [133]. 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 [145], visual explanation generation [146], [147], and text-to-image synthesis [148]. Another domain-specific dataset is **Fashion Captioning** [134] 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** [135] and **GoodNews** [136] 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** [137] dataset, which contains images with text, that must be "read" and included in the caption, and to **Localized Narratives** [138], 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 [121], [122], 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 [149]. However, human evaluation is costly and 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 do not rely on reference captions. Table 3.

5.2.1 Standard evaluation metrics

The first strategy adopted to evaluate image captioning performance consists of exploiting metrics designed for NLP tasks. For example, the **BLEU** score [117] and the **METEOR** [150] score were introduced for machine translation. The former is based on n-gram precision considering n-grams up to length four; the latter favors the recall of matching unigrams from the candidate and reference sentences in their exact form stemmed form and meaning. Moreover, the **ROUGE** score [118] was designed for summarization and applied also for image captioning in its variant considering the longest subsequence of tokens in the same relative order, possibly with other tokens in-between, that appears in both candidate and reference caption. Later, specific image captioning metrics have been proposed [121], [122]. The reference **CIDEr** score [122] 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. The **SPICE** score [121] considers matching tuples extracted from the candidate and the reference (or possibly directly the image) scene graphs, thus favoring the semantic content rather than the fluency.

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 [150]), both at corpus-level and caption-level [121], [151], [152]. 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.

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 [153]. 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 [154], [155], [156], [157]. Most of these metrics can potentially be calculated even when no ground-truth captions are available

TABLE 2: Performance analysis of representative image captioning approaches in terms of different evaluation metrics. The † marker indicates models trained by us with ResNet-152 features, while the ‡ marker indicates unofficial implementations. For all the metrics, the higher the value, the better (↑).

		Standard Metrics						Diversity Metrics				Embedding-based Metrics			Learning-based Metrics			
	#Params (M)	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† [23]	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)‡ [38]	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† [42]	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)‡ [38]	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‡ [58]	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 [71]	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 [72]	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 [79]	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 [80]	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 [83]	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 [107]	-	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 [77]	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 [92]	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 [81]	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 [80]	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 [101]	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 [103]	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**

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 [155]). 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 [158], [159], [160]. For example, the **WMD** score [161], originally introduced to evaluate document semantic dissimilarity, can also be applied to captioning evaluation by considering generated captions and ground-truth captions as the compared documents [162]. Moreover, the **Alignment** score [163] is based on the alignment between the sequences of nouns in the candidate and reference sentence and captures whether concepts are mentioned in a human-like order. Finally, the **Coverage** score [84], [164] expresses the completeness of a caption, which is evaluated by considering the mentioned scene visual entities. 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 [151], [152], [165], [166], [167], [168] are being investigated. To this end, it can be exploited a component of a complete captioning approach, in charge to evaluate the produced caption completeness [169] or how human-like it is [170]. Alternatively, learning-based evaluation is usually based on a pre-trained model. For example, the **BERT-S** score [171], which is used to evaluate various language generation tasks [172], exploits pre-trained BERT embeddings [102] to represent and match the tokens in the reference and candidate sentences via cosine similarity. Moreover, the **TIGEr** score [173] represents the reference and candidate captions as grounding score vectors obtained from a pre-trained model [174] that grounds their words on the image regions and scores the candidate caption based on the similarity of the grounding vectors. Further, the **CLIP-S** score [175] is a direct application of the CLIP [93] model to image captioning evaluation and consists of an adjusted cosine similarity between image and candidate caption representation. Thus, CLIP-S is designed to work without reference captions, although the CLIP-S^{Ref} variant can exploit also the reference captions.
We refer the reader to Appendix A for a deeper discussion on diversity, embedding-based, and learning-based metrics.

6 Experimental Evaluation

In Table 2, 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 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

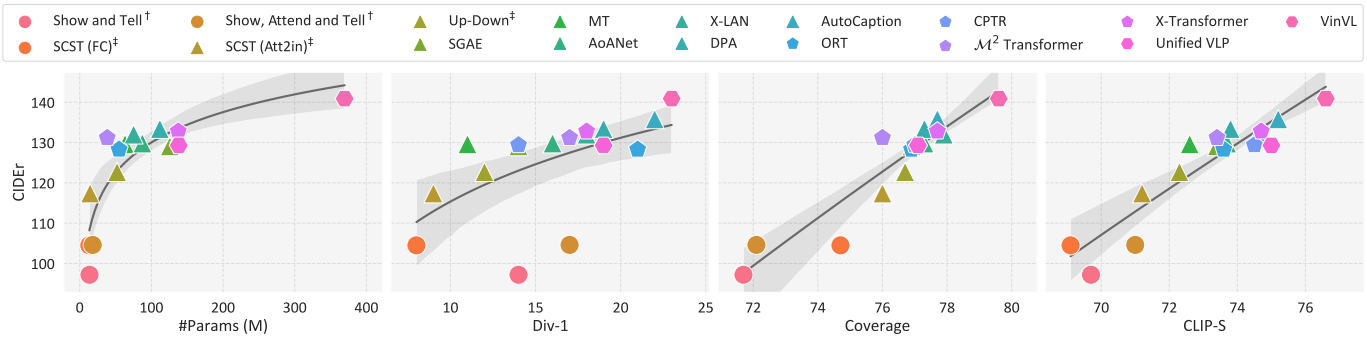


Fig. 9: Relationship between CIDEr, number of parameters and other scores. Values of Div-1 and CLIP-S are multiplied by powers of 10 for readability.

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 (consider 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 2. 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. 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. We refer the reader to Appendix B for additional performance analyses and qualitative results.

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 the lack of training data*; 2. *focusing on the visual input*; 3. *focusing on the textual output*; 4. *addressing user requirements*.

7.1 Dealing with the lack of 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 [176], [177] 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.* [176] introduced a variant of the COCO dataset [128], 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 [178]. Some approaches to this variant [179], [180] integrate copying mechanisms in the language model to select novel objects predicted from a tagger or generate a caption template with placeholders to be filled with novel objects [106], [181]. On a different line, Anderson *et al.* [182] 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.* [183] proposed a multi-layer Transformer model pre-trained by randomly masking one or more tags from image-tag pairs.

Unpaired Image Captioning. Unpaired Image Captioning approaches can be either unsupervised or semi-supervised. Unsupervised captioning aims at understanding and describing images without paired image-text training data. Following unpaired machine translation approaches, the early work [184] 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 [185], [186]. As alternative approaches, it is worth mentioning [187] that generates a caption from the image scene-graph and [188] that leverages a memory-based network. Moreover, semi-supervised approaches have been proposed, such as [189], which uses both paired and unpaired data with adversarial learning, and [190], which performs iterative self-learning.

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 [191], and the model should be able to transfer visual concepts from one to the other while enlarging its vocabulary.

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.* [192] 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. To address this task, contextual and global features [193], [194] and attribute generators [195], [196] can be exploited. Related to this variant, an important line of works [66], [197], [198], [199], [200], [201] 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.* [137] with the TextCaps dataset. Another dataset designed for pre-training for this variant is *OCR-CC* [202], which is a subset of images containing meaningful text taken from the CC3M dataset [130] 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 [203], [204], in the visual encoding [137], [205]. Another direction entails generating multiple captions describing different parts of the image, including the contained text [206].

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 [207] 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 [208] 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 [209], [210], [211] or perform multi-task learning with image retrieval as an auxiliary task [212], 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 [213] that entail dividing the beams into similar groups and encouraging diversity between groups. Other solutions have been investigated, such as contrastive learning [214], conditional GANs [154], [170], and paraphrasing [215]. However, these solutions tend to underperform in terms of caption quality, which is partially recovered by using variational auto-encoders [216], [217], [218], [219]. Another approach is exploiting multiple part-of-speech tags sequences predicted from image region classes [220] and forcing the model to produce different captions based on these sequences.

Multilingual Captioning. Since image captioning is commonly performed in English, multilingual captioning [221] 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 COCO images [222], [223], German captions for Flick30K [224]), or directly training multilingual captioning systems with unpaired captions [184], [221], [225], [226].

Application-specific Captioning. Image captioning can be applied to ease and automate activities involving text generation from images. For example, captioning systems can be applied for medical report generation, for which they need to predict disease tags and try to imitate the style of real medical reports [227], [228], [229]. Another interesting application is art description generation, which entails describing not only factual aspects of the artworks, but also their context and style, and conveyed message art description [230]. To this end, captioning systems could also rely on external knowledge, *e.g.* metadata. A similar application is automatic caption generation for news articles [135], [136], for which named entities from the article should be described [231], [232], and the rich journalistic style should be maintained [233], [234]. Another important application domain is assistive technology for the visually impaired [235], where image captioning approaches must be able to provide informative descriptions even for low-quality visual inputs [132].

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 [236], [237]. On another line, Zhang *et al.* [238] 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 [239], [240], [241], [242]. Some datasets have been collected to explore this variant, such as *InstaPIC* [236], which is composed of multiple Instagram posts from the same users, *FlickrStyle10K* [239], which contains images and textual sentences with two different styles, and *Personality-Captions* [243], 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 sparse, as selected image regions [163], [244] and user-provided visual words [220], or dense, as mouse traces [138], [245]. Eventually, the guiding signal can incorporate some form of structure, such as sequences that encode the mentioning order of concepts (part-of-speech tag as in [220]) or visual objects [163]. Guiding inputs can also encode the relation between objects that is most of interest for the user, as done for example in [246] via verbs and semantic roles (verbs represent activities in the image and semantic roles determine how objects engage in these activities) and in [247], [248] via user-generated or user-selected scene graphs. A different control signal is introduced by [249], which consist of a length-level embedding added as an additional token to each textual word, providing existing models the ability to generate length-controllable image captions.

Image Captioning Editing. Image captioning editing was proposed by Sammani *et al.* [250], following the consideration that generated captions may have repetitions and inconsistencies. This variant focuses on decoupling the decoding stage in a caption generation step and a caption polishing one to correct syntactic errors.

8 Conclusions and Future Directions

Image captioning is an intrinsically complex challenge for machine intelligence as it integrates difficulties from both Computer Vision and NLP. Further, as mentioned in the Introduction, the task itself is vaguely defined and captions can, in principle, be generated with many different styles and objectives. The presented literature review and experimental comparison show the performance improvement over the last few years on standard datasets. 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. Based on the analysis presented, we can trace three main developmental directions for the image captioning field, which are discussed in the following.

Procedural and architectural challenges. Since image captioning models are data greedy, pre-training on large-scale datasets, even if not well-curated, is becoming a solid strategy, as demonstrated in [97], [100], [101], [103]. In this regard, promoting the public release of such datasets will be fundamental to fostering reproducibility and allowing fair comparisons. The growing size of pre-training models is also a concern, and the community will probably need to investigate less computationally-intensive alternatives to promote equality in the community. In architectural terms, instead, the growing dichotomy between early-fusion strategies and the encoder-decoder paradigm is still to be solved and is currently one of the main open issues. On the other side, the supremacy of detection features is leaving space to a variety of visual encoding strategies (pre-training from scratch, using detections, using features from multi-modal models) which all appear to be on pair in terms of performance.

Generalization, diversity, long-tail concepts. While pre-training on web-scale datasets provides a promising direction to increase generalization and promoting long-tail concepts [97], specializing in particular domains and generating captions with different styles and aims is still among the main open challenges for image captioning. Although we discussed some attempts to encourage naturalness and diversity [170], [214], [216], further research is needed to design models that are suitable for real-world applications. In this sense, the emergence of models which can deal with long-tail concepts [97], [104] offers a valuable promise of modeling real-life scenarios and generalizing to different contexts. Additionally, developments in image captioning variants such as novel objects captioning or controllable captioning could help to tackle this open issue. Notably, the emergence of subword-based tokenization techniques has made it possible to handle and generate rare words.

Design of trustworthy AI solutions. Due to its potential in human-machine interaction, image captioning needs solutions that are transparent and acceptable for end-users, framed as overcome bias, and interpretable. Since most vision-and-language datasets share common patterns and regularities, datasets bias and overrepresented visual concepts are major issues for any vision-and-language task. In this sense, some effort should be devoted to the study of fairness and bias: two possible directions entail designing specific evaluation metrics and focusing on the robustness to unwanted correlations. Further, 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. In this sense, the design of appropriate and reproducible evaluation protocols [251], [252], [253] 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. Finally, since existing image captioning algorithms lack reliable and interpretable means for determining the cause of a particular output, further research is needed to shed more light on model explainability, focusing on how these deal with different modalities or novel concepts.

Acknowledgments

We thank CINECA for providing computational resources. This work has been supported by "Fondazione di Mod-

ena", by the "Artificial Intelligence for Cultural Heritage (AI4CH)" project, co-funded by the Italian Ministry of Foreign Affairs and International Cooperation, and by the H2020 ICT-48-2020 HumanE-AI-NET project. 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] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in *NeurIPS*, 2011.

[5] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: a deep visual-semantic embedding model," in *NeurIPS*, 2013.

[6] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," in *NeurIPS Workshops*, 2014.

[7] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *NeurIPS*, 2014.

[8] 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.

[9] A. Aker and R. Gaizauskas, "Generating image descriptions using dependency relational patterns," in *ACL*, 2010.

[10] Y. Yang, C. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *EMNLP*, 2011.

[11] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *CoNLL*, 2011.

[12] A. Gupta, Y. Verma, and C. Jawahar, "Choosing linguistics over vision to describe images," in *AAAI*, 2012.

[13] 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.

[14] 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.

[15] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions," *TACL*, vol. 2, pp. 351–362, 2014.

[16] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *JAIR*, vol. 55, pp. 409–442, 2016.

[17] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, 2018.

[18] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A Comprehensive Survey of Deep Learning for Image Captioning," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–36, 2019.

[19] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *JAIR*, 2013.

[20] N. Sharif, U. Nadeem, S. A. A. Shah, M. Bennamoun, and W. Liu, "Vision to Language: Methods, Metrics and Datasets," in *Machine Learning Paradigms*, 2020, pp. 9–62.

[21] X. Liu, Q. Xu, and N. Wang, "A survey on deep neural network-based image captioning," *The Visual Computer*, vol. 35, no. 3, pp. 445–470, 2019.

[22] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj, and R. K. Mishra, "Image captioning: a comprehensive survey," in *PARC*, 2020.

[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015.

[24] 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.

[25] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012.

[27] 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.

[28] 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.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[30] X. Chen and C. Lawrence Zitnick, "Mind's Eye: A Recurrent Visual Representation for Image Caption Generation," in *CVPR*, 2015.

[31] 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.

[32] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the Long-Short Term Memory model for Image Caption Generation," in *ICCV*, 2015.

[33] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *CVPR*, 2016.

[34] 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.

[35] J. Gu, G. Wang, J. Cai, and T. Chen, "An Empirical Study of Language CNN for Image Captioning," in *ICCV*, 2017.

[36] F. Chen, R. Ji, J. Su, Y. Wu, and Y. Wu, "StructCap: Structured Semantic Embedding for Image Captioning," in *ACM Multimedia*, 2017.

[37] 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.

[38] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[40] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *ICCV*, 2017.

[41] 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.

[42] 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.

[43] 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.

[44] B. Dai, D. Ye, and D. Lin, "Rethinking the form of latent states in image captioning," in *ECCV*, 2018.

[45] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2014.

[46] 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.

[47] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, "Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition," in *CVPR*, 2017.

[48] 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.

[49] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-Captioning: Coarse-to-Fine Learning for Image Captioning," in *AAAI*, 2018.

[50] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review Networks for Caption Generation," in *NeurIPS*, 2016.

[51] 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.

[52] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, "Recurrent Fusion Network for Image Captioning," in *ECCV*, 2018.

[53] Y. Sugano and A. Bulling, "Seeing with Humans: Gaze-Assisted Neural Image Captioning," *arXiv preprint arXiv:1608.05203*, 2016.

[54] H. R. Tavakoli, R. Shetty, A. Borji, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in *ICCV*, 2017.

[55] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *CVPR*, 2017.

[56] 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.

[57] S. Chen and Q. Zhao, "Boosted attention: Leveraging human attention for image captioning," in *ECCV*, 2018.

[58] 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.

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

[60] ——, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. PAMI*, vol. 39, no. 6, pp. 1137–1149, 2017.

[61] 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.

[62] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, "Reflective Decoding Network for Image Captioning," in *ICCV*, 2019.

[63] Y. Qin, J. Du, Y. Zhang, and H. Lu, "Look Back and Predict Forward in Image Captioning," in *CVPR*, 2019.

[64] L. Huang, W. Wang, Y. Xia, and J. Chen, "Adaptively Aligned Image Captioning via Adaptive Attention Time," in *NeurIPS*, 2019.

[65] L. Wang, Z. Bai, Y. Zhang, and H. Lu, "Show, Recall, and Tell: Image Captioning with Recall Mechanism," in *AAAI*, 2020.

[66] 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.

[67] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of Attention for Image Captioning," in *ICCV*, 2017.

[68] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring Visual Relationship for Image Captioning," in *ECCV*, 2018.

[69] 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.

[70] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[71] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-Encoding Scene Graphs for Image Captioning," in *CVPR*, 2019.

[72] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving Image Captioning with Better Use of Captions," in *ACL*, 2020.

[73] T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy Parsing for Image Captioning," in *ICCV*, 2019.

[74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[75] X. Yang, H. Zhang, and J. Cai, "Learning to Collocate Neural Modules for Image Captioning," in *ICCV*, 2019.

[76] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled Transformer for Image Captioning," in *ICCV*, 2019.

[77] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image Captioning: Transforming Objects into Words," in *NeurIPS*, 2019.

[78] 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.

[79] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on Attention for Image Captioning," in *ICCV*, 2019.

[80] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-Linear Attention Networks for Image Captioning," in *CVPR*, 2020.

[81] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," in *CVPR*, 2020.

[82] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," in *ACCV*, 2020.

[83] F. Liu, X. Ren, X. Wu, S. Ge, W. Fan, Y. Zou, and X. Sun, "Prophet Attention: Predicting Attention with Future Attention," in *NeurIPS*, 2020.

[84] M. Cornia, L. Baraldi, and R. Cucchiara, "SMArT: Training Shallow Memory-aware Transformers for Robotic Explainability," in *ICRA*, 2020.

[85] 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.

[86] 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.

[87] F. Liu, Y. Liu, X. Ren, X. He, and X. Sun, "Aligning visual regions and textual concepts for semantic-grounded image representations," in *NeurIPS*, 2019.

[88] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," in *CVPR*, 2020.

[89] 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.

[90] 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.

[91] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021.

[92] W. Liu, S. Chen, L. Guo, X. Zhu, and J. Liu, "CPTR: Full Transformer Network for Image Captioning," *arXiv preprint arXiv:2101.10804*, 2021.

[93] 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.

[94] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SimVLM: Simple visual language model pretraining with weak supervision," *arXiv preprint arXiv:2108.10904*, 2021.

[95] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer, "How Much Can CLIP Benefit Vision-and-Language Tasks?" *arXiv preprint arXiv:2107.06383*, 2021.

[96] R. Mokady, A. Hertz, and A. H. Bermano, "ClipCap: CLIP Prefix for Image Captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[97] M. Cornia, L. Baraldi, G. Fiameni, and R. Cucchiara, "Universal Captioner: Long-Tail Vision-and-Language Model Training through Content-Style Separation," *arXiv preprint arXiv:2111.12727*, 2021.

[98] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *EMNLP*, 2019.

[99] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019.

[100] 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.

[101] 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.

[102] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL*, 2018.

[103] 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.

[104] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, "Scaling Up Vision-Language Pre-training for Image Captioning," *arXiv preprint arXiv:2111.12233*, 2021.

[105] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[106] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural Baby Talk," in *CVPR*, 2018.

[107] X. Zhu, W. Wang, L. Guo, and J. Liu, "AutoCaption: Image Captioning with Neural Architecture Search," *arXiv preprint arXiv:2012.09742*, 2020.

[108] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *CVPR*, 2018.

[109] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[110] Z.-c. Fei, "Fast Image Caption Generation with Position Alignment," *AAAI Workshops*, 2019.

[111] 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.

[112] Z. Fei, "Iterative Back Modification for Faster Image Captioning," in *ACM Multimedia*, 2020.

[113] L. Guo, J. Liu, X. Zhu, and H. Lu, "Fast Sequence Generation with Multi-Agent Reinforcement Learning," *arXiv preprint arXiv:2101.09698*, 2021.

[114] P. Koehn, *Statistical Machine Translation*. Cambridge University Press, 2009.

[115] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *ICLR*, 2016.

[116] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, 1992.

[117] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[118] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL Workshops*, 2004.

[119] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *CVPR*, 2017.

[120] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved Image Captioning via Policy Gradient Optimization of SPIDEr," in *ICCV*, 2017.

[121] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in *ECCV*, 2016.

[122] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," in *CVPR*, 2015.

[123] 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.

[124] J. Gao, S. Wang, S. Wang, S. Ma, and W. Gao, "Self-critical n-step Training for Image Captioning," in *CVPR*, 2019.

[125] 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.

[126] K. Desai and J. Johnson, "VirTex: Learning Visual Representations From Textual Annotations," in *CVPR*, 2021.

[127] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *ACL*, 2016.

[128] 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.

[129] 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.

[130] 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.

[131] 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.

[132] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning Images Taken by People Who Are Blind," in *ECCV*, 2020.

[133] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *CVPR*, 2016.

[134] 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.

[135] 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.

[136] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, "Good news, everyone! context driven entity-aware captioning for news images," in *CVPR*, 2019.

[137] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "TextCaps: a Dataset for Image Captioning with Reading Comprehension," in *ECCV*, 2020.

[138] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," in *ECCV*, 2020.

[139] 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.

[140] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[141] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning," *arXiv preprint arXiv:2103.01913*, 2021.

[142] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[143] 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.

[144] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *ICVGIP*, 2008.

[145] 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.

[146] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *ECCV*, 2016.

[147] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations," in *ECCV*, 2018.

[148] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *ICML*, 2016.

[149] J. Kasai, K. Sakaguchi, L. Dunagan, J. Morrison, R. L. Bras, Y. Choi, and N. A. Smith, "Transparent human evaluation for image captioning," *arXiv preprint arXiv:2111.08940*, 2021.

[150] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *ACL Workshops*, 2005.

[151] N. Sharif, L. White, M. Bennamoun, and S. A. A. Shah, "NNEval: Neural network based evaluation metric for image captioning," in *ECCV*, 2018.

[152] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, "Learning to evaluate image captioning," in *CVPR*, 2018.

[153] O. Caglayan, P. Madhyastha, and L. Specia, "Curious Case of Language Generation Evaluation Metrics: A Cautionary Tale," in *COLING*, 2020.

[154] 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.

[155] E. Van Miltenburg, D. Elliott, and P. Vossen, "Measuring the diversity of automatic image descriptions," in *COLING*, 2018.

[156] Q. Wang and A. B. Chan, "Describing like humans: on diversity in image captioning," in *CVPR*, 2019.

[157] Q. Wang, J. Wan, and A. B. Chan, "On Diversity in Image Captioning: Metrics and Methods," *IEEE Trans. PAMI*, 2020.

[158] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object Hallucination in Image Captioning," in *EMNLP*, 2018.

[159] M. Jiang, J. Hu, Q. Huang, L. Zhang, J. Diesner, and J. Gao, "REO-Relevance, Extraness, Omission: A Fine-grained Evaluation for Image Captioning," in *EMNLP-IJCNLP*, 2019.

[160] Z. Wang, B. Feng, K. Narasimhan, and O. Russakovsky, "Towards unique and informative captioning of images," in *ECCV*, 2020.

[161] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *ICML*, 2015.

[162] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," in *ACL*, 2017.

[163] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions," in *CVPR*, 2019.

[164] R. Bigazzi, F. Landi, M. Cornia, S. Cascianelli, L. Baraldi, and R. Cucchiara, "Explore and Explain: Self-supervised Navigation and Recounting," in *ICPR*, 2020.

[165] 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.

[166] Y. Yi, H. Deng, and J. Hu, "Improving image captioning evaluation by considering inter references variance," in *ACL*, 2020.

[167] S. Wang, Z. Yao, R. Wang, Z. Wu, and X. Chen, "FAIEr: Fidelity and Adequacy Ensured Image Caption Evaluation," in *CVPR*, 2021.

[168] H. Lee, S. Yoon, F. Dernoncourt, T. Bui, and K. Jung, "UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning," in *ACL*, 2021.

[169] B. Dai, S. Fidler, and D. Lin, "A neural compositional paradigm for image captioning," in *NeurIPS*, 2018.

[170] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards Diverse and Natural Image Descriptions via a Conditional GAN," in *ICCV*, 2017.

[171] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *ICLR*, 2020.

[172] I. J. Unanue, J. Parnell, and M. Piccardi, "BERTTune: Fine-Tuning Neural Machine Translation with BERTScore," *ACL*, 2021.

[173] 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.

[174] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked Cross Attention for Image-Text Matching," in *ECCV*, 2018.

[175] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "CLIPScore: A Reference-free Evaluation Metric for Image Captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[176] 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.

[177] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning Images with Diverse Objects," in *CVPR*, 2017.

[178] 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.

[179] T. Yao, Y. Pan, Y. Li, and T. Mei, "Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects," in *CVPR*, 2017.

[180] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Pointing Novel Objects in Image Captioning," in *CVPR*, 2019.

[181] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, "Decoupled Novel Object Captioner," in *ACM Multimedia*, 2018.

[182] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Guided Open Vocabulary Image Captioning with Constrained Beam Search," in *EMNLP*, 2017.

[183] 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.

[184] J. Gu, S. Joty, J. Cai, and G. Wang, "Unpaired image captioning by language pivoting," in *ECCV*, 2018.

[185] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *CVPR*, 2019.

[186] I. Laina, C. Rupprecht, and N. Navab, "Towards unsupervised image captioning with shared multimodal embeddings," in *ICCV*, 2019.

[187] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," in *ICCV*, 2019.

[188] D. Guo, Y. Wang, P. Song, and M. Wang, "Recurrent relational memory network for unsupervised image captioning," *IJCAI*, 2020.

[189] 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.

[190] 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.

[191] R. Del Chiaro, B. Twardowski, A. D. Bagdanov, and J. van de Weijer, "RATT: Recurrent Attention to Transient Tasks for Continual Image Captioning," in *NeurIPS*, 2020.

[192] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional Localization Networks for Dense Captioning," in *CVPR*, 2016.

[193] L. Yang, K. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context," in *CVPR*, 2017.

[194] X. Li, S. Jiang, and J. Han, "Learning object context for dense captioning," in *AAAI*, 2019.

[195] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, "Context and attribute grounded dense captioning," in *CVPR*, 2019.

[196] 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.

[197] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *CVPR*, 2017.

[198] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition GAN for visual paragraph generation," in *ICCV*, 2017.

[199] Y. Mao, C. Zhou, X. Wang, and R. Li, "Show and Tell More: Topic-Oriented Multi-Sentence Image Captioning," in *IJCAI*, 2018.

[200] M. Chatterjee and A. G. Schwing, "Diverse and coherent paragraph generation from images," in *ECCV*, 2018.

[201] 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.

[202] 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.

[203] J. Wang, J. Tang, and J. Luo, "Multimodal Attention with Image Text Spatial Relationship for OCR-Based Image Captioning," in *ACM Multimedia*, 2020.

[204] J. Wang, J. Tang, M. Yang, X. Bai, and J. Luo, "Improving OCR-based Image Captioning by Incorporating Geometrical Relationship," in *CVPR*, 2021.

[205] Q. Zhu, C. Gao, P. Wang, and Q. Wu, "Simple is not Easy: A Simple Strong Baseline for TextVQA and TextCaps," in *AAAI*, 2021.

[206] 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.

[207] H. Jhamtani and T. Berg-Kirkpatrick, "Learning to describe differences between pairs of similar images," in *EMNLP*, 2018.

[208] D. H. Park, T. Darrell, and A. Rohrbach, "Robust Change Captioning," in *CVPR*, 2019.

[209] 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.

[210] 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.

[211] H. Kim, J. Kim, H. Lee, H. Park, and G. Kim, "Viewpoint-Agnostic Change Captioning With Cycle Consistency," in *ICCV*, 2021.

[212] M. Hosseinzadeh and Y. Wang, "Image Change Captioning by Learning from an Auxiliary Task," in *CVPR*, 2021.

[213] 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.

[214] B. Dai and D. Lin, "Contrastive learning for image captioning," in *NeurIPS*, 2017.

[215] L. Liu, J. Tang, X. Wan, and Z. Guo, "Generating diverse and descriptive image captions using visual paraphrases," in *ICCV*, 2019.

[216] L. Wang, A. G. Schwing, and S. Lazebnik, "Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space," in *NeurIPS*, 2017.

[217] J. Aneja, H. Agrawal, D. Batra, and A. Schwing, "Sequential latent spaces for modeling the intention during diverse image captioning," in *ICCV*, 2019.

[218] 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.

[219] S. Mahajan and S. Roth, "Diverse image captioning with context-object split latent spaces," in *NeurIPS*, 2020.

[220] 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.

[221] D. Elliott, S. Frank, and E. Hasler, "Multilingual image description with neural sequence models," *ICLR*, 2015.

[222] 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.

[223] T. Miyazaki and N. Shimizu, "Cross-lingual image caption generation," in *ACM Multimedia*, 2016.

[224] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30K: Multilingual English-German Image Descriptions," in *ACL Workshops*, 2016.

[225] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," in *ACM Multimedia*, 2017.

[226] Y. Song, S. Chen, Y. Zhao, and Q. Jin, "Unpaired cross-lingual image caption generation with self-supervised rewards," in *ACM Multimedia*, 2019.

[227] B. Jing, P. Xie, and E. Xing, "On the Automatic Generation of Medical Imaging Reports," in *ACL*, 2018.

[228] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation," in *CVPR*, 2021.

[229] X. Yang, M. Ye, Q. You, and F. Ma, "Writing by Memorizing: Hierarchical Retrieval-based Medical Report Generation," in *ACL-IJCNLP*, 2021.

[230] Z. Bai, Y. Nakashima, and N. Garcia, "Explain Me the Painting: Multi-Topic Knowledgeable Art Description Generation," in *ICCV*, 2021.

[231] Y. Feng and M. Lapata, "Automatic Caption Generation for News Images," *IEEE Trans. PAMI*, vol. 35, no. 4, pp. 797–812, 2012.

[232] A. Tran, A. Mathews, and L. Xie, "Transform and Tell: Entity-Aware News Image Captioning," in *CVPR*, 2020.

[233] F. Liu, Y. Wang, T. Wang, and V. Ordonez, "Visual News: Benchmark and Challenges in News Image Captioning," in *EMNLP*, 2021.

[234] X. Yang, S. Karaman, J. Tetreault, and A. Jaimes, "Journalistic Guidelines Aware News Image Captioning," *arXiv preprint arXiv:2109.02865*, 2021.

[235] S. Wu, J. Wieland, O. Farivar, and J. Schiller, "Automatic alt-text: Computer-generated image descriptions for blind users on a social network service," in *CSCW*, 2017.

[236] C. Chunseong Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in *CVPR*, 2017.

[237] C. C. Park, B. Kim, and G. Kim, "Towards personalized image captioning via multimodal memory networks," *IEEE Trans. PAMI*, 2018.

[238] 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.

[239] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "StyleNet: Generating Attractive Visual Captions with Styles," in *CVPR*, 2017.

[240] A. Mathews, L. Xie, and X. He, "Semstyle: Learning to generate stylised image captions using unaligned text," in *CVPR*, 2018.

[241] L. Guo, J. Liu, P. Yao, J. Li, and H. Lu, "Mscap: Multi-style image captioning with unpaired stylized text," in *CVPR*, 2019.

[242] W. Zhao, X. Wu, and X. Zhang, "MemCap: Memorizing style knowledge for image captioning," in *AAAI*, 2020.

[243] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging image captioning via personality," in *CVPR*, 2019.

[244] Y. Zheng, Y. Li, and S. Wang, "Intention oriented image captions with guiding objects," in *CVPR*, 2019.

[245] 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.

[246] L. Chen, Z. Jiang, J. Xiao, and W. Liu, "Human-like Controllable Image Captioning with Verb-specific Semantic Roles," in *CVPR*, 2021.

[247] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *CVPR*, 2020.

[248] Y. Zhong, L. Wang, J. Chen, D. Yu, and Y. Li, "Comprehensive image captioning via scene graph decomposition," in *ECCV*, 2020.

[249] C. Deng, N. Ding, M. Tan, and Q. Wu, "Length-controllable image captioning," in *ECCV*, 2020.

[250] F. Sammani and L. Melas-Kyriazi, "Show, edit and tell: A framework for editing image captions," in *CVPR*, 2020.

[251] M. Hodosh and J. Hockenmaier, "Focused evaluation for image description with binary forced-choice tasks," in *ACL Workshops*, 2016.

[252] 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.

[253] M. Alikhani, P. Sharma, S. Li, R. Soricut, and M. Stone, "Cross-modal Coherence Modeling for Caption Generation," in *ACL*, 2020.

[254] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: UNiversal Image-TExt Representation Learning," in *ECCV*, 2020.

[255] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*, 2017.

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, focusing 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 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. He is currently Tenure Track Assistant Professor with the 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 70 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 an Assistant Professor with the University of Modena and Reggio Emilia. She was a Visitor Researcher at the Queen Mary University of London in 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. 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.

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. She is currently Full Professor with the University of Modena and Reggio Emilia, where she heads the AImage-Lab 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. She is Member of the Advisory Board of the Computer Vision Foundation, and Director of the ELLIS Unit of Modena.

APPENDIX A
FURTHER ANALYSIS OF THE EVALUATION METRICS

In this section, we extend our analysis of the evaluation metrics for image captioning. In particular, in Table 3, we provide a taxonomy and summarize the main characteristics of the metrics presented. Moreover, in the following, we describe in more detail additional diversity metrics, embedding-based metrics, and learning-based metrics, which were only mentioned in the main paper.

Diversity metrics. Local Diversity can be quantified via the average BLEU score between each caption and the others (**mBLEU**: the lower the mBLEU, the more diverse the produced caption set is) [154]. Another approach to quantify captions diversity is to focus on the mention of all the relevant words in the produced caption, despite their rareness in the training set. In this respect, in [155], two recall-based diversity metrics have been proposed, namely **Global Recall** and **Local Recall**. The former is computed as the fraction of generated words with respect to the words appearing in both the training and validation set. The latter is computed for each test image as the fraction of generated words with respect to the words in the reference captions. Moreover, when the system can produce multiple captions for the same image, diversity can be quantified at the topics level by using the Latent Semantic Analysis-based metric **LSA** and the kernelized version of the CIDEr score **Self-CIDEr**, proposed in [156], [157].

Embedding-based metrics. Among the specific aspects of a produced caption that can e evaluated via embedding-based metrics, the hallucination rate is measured via the **CHAIR** score [158], which is expressed as the fraction of hallucinated objects among those mentioned in a caption (in the $CHAIR_i$ variant) or the fraction of captions with at least a hallucinated object among all the produced captions (in the $CHAIR_s$ variant). Other aspects that can be measured by exploiting embedding-based representation of candidate caption, reference captions, and image are **Relevance** (via cosine similarity), **Extraness**, and **Omission** (via orthogonal projections), as proposed in [159]. Moreover, to take into account the uniqueness of the generated captions, in [160] the SPICE-U score is proposed as the harmonic mean of the SPICE score and a measure of the caption uniqueness, which considers the fraction of images in the training set not containing the mentioned concepts.

Learning-based metrics. With respect to learning-based evaluation scores, **NNEval** [151] was proposed as the first learning-based image captioning evaluation approach. It considers classical metrics (BLEU, METEOR, CIDEr, SPICE, and WMD) as features describing the candidate caption when compared to reference captions and predicts its probability of being human-generated. Another early-proposed learning-based evaluation strategy is **LEIC** [152], which directly scores the probability of a caption being human-generated, conditioned on the image and eventually on a reference caption, by using a binary classifier fed with pre-trained ResNet image features and an LSTM-based encoding of the caption. As a refinement of BERT-S specific for image captioning, the **ViLBERT-S** [165] exploits the image-conditioned embedding obtained from the vision-and-language representation model ViLBERT [99]. Similar to the BERT-S, the matching between these tokens is expressed via the cosine similarity of their embeddings, and the best matching token pairs are used for computing precision, recall, and F1-score. Another variant of the BERT-S, to which we here refer to as **BERT-S^{IRV}** [166], takes into account the variability of the reference captions associated with the same image by combining them in a unique embedding vector that contains all the mentioned concepts and against which the candidate caption is compared. To evaluate the candidate caption fidelity and adequacy, the **FAIEr** [167] score exploits scene graphs matching. The references and the image scene graphs are fused in a unique scene graph, whose more relevant nodes (representing the concepts more often mentioned in the references) get more weight, and the score is obtained based on the similarity between the candidate scene graph and the unique scene graph. On a similar line, the **UMIC** score [168] exploits the pre-trained vision-and-language representation model UNITER [254], fine-tuned on carefully-designed negative samples, to score a candidate caption without the need for reference captions.

TABLE 3: Taxonomy and main characteristics of image captioning metrics.

			Inputs		
		Original Task	Pred	Refs	Image
Standard	BLEU [117]	Translation	✓	✓	
	METEOR [150]	Translation	✓	✓	
	ROUGE [118]	Summarization	✓	✓	
	CIDEr [122]	Captioning	✓	✓	
	SPICE [121]	Captioning	✓	(✓)	(✓)
Diversity	Div [154]	Captioning	✓		
	Vocab [154]	Captioning	✓		
	%Novel [154]	Captioning	✓		
	mBLEU [154]	Captioning	✓		
	LSA [156], [157]	Captioning	✓		
	Self-CIDEr [156], [157]	Captioning	✓		
	Recall [155]	Captioning	✓	(✓)	
Embedding-based	WMD [161]	Doc. Dissimilarity	✓	✓	
	Alignment [163]	Captioning	✓	✓	
	Coverage [163], [164]	Captioning	✓	(✓)	(✓)
	Relevance [159]	Captioning	✓	(✓)	✓
	Extraness [159]	Captioning	✓	(✓)	✓
	Omission [159]	Captioning	✓	(✓)	✓
	SPICE-U [160]	Captioning	✓	(✓)	✓
	CHAIR [158]	Captioning	✓	✓	✓
Learning-based	NNEval [151]	Captioning	✓	✓	
	BERT-S [171]	Text Similarity	✓	✓	
	BERT-S^{IRV} [166]	Captioning	✓	✓	
	UMIC [168]	Captioning	✓		✓
	LEIC [152]	Captioning	✓	(✓)	✓
	CLIP-S [175]	Captioning	✓	(✓)	✓
	TIGEr [173]	Captioning	✓	✓	✓
	ViLBERT-S [165]	Captioning	✓	✓	✓
	FAIEr [167]	Captioning	✓	✓	✓

APPENDIX B
FURTHER PERFORMANCE ANALYSIS

According to the taxonomies proposed in Sections 2, 3, and 4, in Table 4, we overview the most relevant surveyed methods. We report their performance in terms of BLEU-4, METEOR, and CIDEr on the COCO Karpathy 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 40.0 for those exploiting the attention and self-attention mechanisms, peaking at 42.6 in the case of vision-and-language pre-training. By looking at the performance in terms of the CIDEr score, we can notice that, 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 133.2 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. 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, it emerges that vision-and-language pre-training on large datasets allows boosting the performance and deserves further investigation (with an average CIDEr score of 140.4).

For completeness, in Fig. 10 we report the relation between the CIDEr score and all the other characteristics from Table 2. The almost-linear relation with the CIDEr score is observable also for the scores not reported in Fig. 9 in the main paper, with the only exception of the BERT-S score.

We deepen our analysis by considering also the Flickr30K dataset, and evaluate the performance of methods trained on COCO, and tested on the test set of Flickr30K, and of methods developed for COCO, but trained and tested on Flickr30K. The results of this study are reported in Table 5. Compared to what is obtained by the same models when trained and tested on COCO (in Table 2), the standard metrics and the embedding-based metrics significantly drop while diversity metrics increase. This can be imputed to the smaller size of Flickr30K compared to COCO. Learning-based metrics, especially BERT-S and CLIP-S, are more stable. As expected, training on COCO and directly testing on Flickr30K results in a performance drop under all the metrics when compared with the case in which both training and test are performed on data from Flickr30K. This confirms that generalization is still an open issue for image captioning approaches. Interestingly, VinVL surpasses the other considered approaches under all metrics in the case of the test on Flickr30K, suggesting the benefits of pre-training in terms of generalization capabilities of the resulting model.

For a deeper qualitative analysis, in Fig. 11-12, we report examples of captions generated by eight popular approaches on images from the COCO dataset. It can be observed that the produced captions have similar length and structure (*i.e.* the main subject is mentioned first, then the main action, and finally additional details concerning other visual entities in the scene). This mimics the characteristics of the majority of the ground-truth captions in COCO. Another aspect that emerges is the lack of counting capabilities (consider *e.g.* the first example in the second row and the second example in the bottom row of Fig. 12). Current approaches struggle in mentioning the right amount of instances of the same entities and generally refer to multiple instances as *a group of*. Finally, it is worth mentioning the difficulty in describing unusual concepts, both situations and visual entities (consider *e.g.* the first two examples in Fig. 11), which is a symptom of the lack of generalization capability. In fact, in such cases, unusual concepts are described as more represented concepts in the training set. For example, the *ferret* in the top-right of Fig. 11 is described as a *cat* or a *mouse*, and the historically-dressed *man* in the last example of the second row of Fig. 12 is described as a *woman*. This issue is less evident for VinVL, enforcing the role of pre-training to achieve better generalization capabilities.

Finally, in Fig. 13, we report a visualization of the attention states corresponding to the captions generated by two methods based on image regions, *i.e.* Up-Down [58], which performs attention over image regions, and $\mathcal{M}^2$ Transformer [81], which performs self-attention encoding. For visualization, we use the normalized attribution scores obtained for each image region via the Integrated Gradients approach [255] and projected between 0 and 1 by applying a contrast stretching function. In particular, we show the attended regions for each generated word and outline the region with the highest attribution score. With a focus on visual words, it can be observed that, for both approaches, the regions with the highest scores are coherent with the produced word. However, thanks to the more complex relations modeled via self-attention, $\mathcal{M}^2$ Transformer generally pays more attention to fewer, more precise regions compared to Up-Down (consider *e.g.* the region contributing the most to outputting *tracks* in the third example, or *skateboard* in the last one).

TABLE 4: Overview of deep learning-based image captioning models. Scores are taken from the respective papers. For all the metrics, the higher the value, the better (↑).

	Visual Encoding					Language Model			Training Strategies				Main Results		
Model	Global	Grid	Regions	Graph	Self-Attention	RNN/LSTM	Transformer	BERT	XE	MLM	Reinforce	VL Pre-Training	BLEU-4	METEOR	CIDEr
LEMON [104]			✓		✓			✓		✓	✓	✓	**42.6**	31.4	**145.5**
UniversalCap [97]		✓			✓		✓		✓		✓	✓	40.8	30.4	143.4
SimVLM [94]					✓		✓		✓			✓	40.6	**33.7**	143.3
VinVL [103]			✓		✓			✓		✓	✓	✓	41.0	31.1	140.9
Oscar [100]			✓		✓			✓		✓	✓	✓	41.7	30.6	140.0
Unified VLP [101]			✓		✓			✓	✓		✓	✓	39.5	29.3	129.3
AutoCaption [107]			✓		✓	✓			✓		✓		40.2	29.9	135.8
RSTNet [89]		✓			✓		✓		✓		✓		40.1	29.8	135.6
DLCT [86]		✓	✓		✓		✓		✓		✓		39.8	29.5	133.8
DPA [83]			✓		✓	✓			✓		✓		40.5	29.6	133.4
X-Transformer [80]			✓		✓		✓		✓		✓		39.7	29.5	132.8
NG-SAN [78]			✓		✓		✓		✓		✓		39.9	29.3	132.1
X-LAN [80]			✓		✓	✓			✓		✓		39.5	29.5	132.0
GET [85]			✓		✓		✓		✓		✓		39.5	29.3	131.6
$\mathcal{M}^2$ Transformer [81]			✓		✓		✓		✓		✓		39.1	29.2	131.2
AoANet [79]			✓		✓	✓			✓		✓		38.9	29.2	129.8
CPTR [92]					✓		✓		✓		✓		40.0	29.1	129.4
ORT [77]			✓		✓		✓		✓		✓		38.6	28.7	128.3
CNM [75]			✓		✓	✓			✓		✓		38.9	28.4	127.9
ETA [76]			✓		✓		✓		✓		✓		39.9	28.9	127.6
GCN-LSTM+HIP [73]			✓	✓		✓			✓		✓		39.1	28.9	130.6
MT [72]			✓	✓		✓			✓		✓		38.9	28.8	129.6
SGAE [71]			✓	✓		✓			✓		✓		39.0	28.4	129.1
GCN-LSTM [68]			✓	✓		✓			✓		✓		38.3	28.6	128.7
VSUA [69]			✓	✓		✓			✓		✓		38.4	28.5	128.6
SG-RWS [65]			✓			✓			✓		✓		38.5	28.7	129.1
LBPF [63]			✓			✓			✓		✓		38.3	28.5	127.6
AAT [64]			✓			✓			✓		✓		38.2	28.3	126.7
CAVP [66]			✓			✓			✓		✓		38.6	28.3	126.3
Up-Down [58]			✓			✓			✓		✓		36.3	27.7	120.1
RDN [62]			✓			✓			✓				36.8	27.2	115.3
Neural Baby Talk [106]			✓			✓			✓				34.7	27.1	107.2
Stack-Cap [49]		✓				✓			✓		✓		36.1	27.4	120.4
MaBi-LSTM [48]		✓				✓			✓				36.8	28.1	116.6
RFNet [52]		✓				✓			✓		✓		35.8	27.4	112.5
SCST (Att2in) [38]		✓				✓			✓		✓		33.3	26.3	111.4
Adaptive Attention [43]		✓				✓			✓				33.2	26.6	108.5
Skeleton [47]		✓				✓			✓				33.6	26.8	107.3
ARNet [46]		✓				✓			✓				33.5	26.1	103.4
SCA-CNN [51]		✓				✓			✓				31.1	25.0	95.2
Areas of Attention [67]		✓				✓			✓				30.7	24.5	93.8
Review Net [50]		✓				✓			✓				29.0	23.7	88.6
Show, Attend and Tell [42]		✓				✓			✓				24.3	23.9	-
SCST (FC) [38]	✓					✓			✓		✓		31.9	25.5	106.3
PG-SPIDEr [120]	✓					✓			✓		✓		33.2	25.7	101.3
SCN-LSTM [41]	✓					✓			✓				33.0	25.7	101.2
LSTM-A [40]	✓					✓			✓				32.6	25.4	100.2
CNN$_\mathcal{L}$+RNH [35]	✓					✓			✓				30.6	25.2	98.9
Att-CNN+LSTM [34]	✓					✓			✓				31.0	26.0	94.0
GroupCap [37]	✓					✓			✓				33.0	26.0	-
StructCap [36]	✓					✓			✓				32.9	25.4	-
Embedding Reward [119]	✓					✓			✓		✓		30.4	25.1	93.7
ATT-FCN [33]	✓					✓			✓				30.4	24.3	-
MIXER [115]	✓					✓			✓		✓		29.0	-	-
MSR [31]	✓					✓			✓				25.7	23.6	-
gLSTM [32]	✓					✓			✓				26.4	22.7	81.3
m-RNN [27]	✓					✓			✓				25.0	-	-
Show and Tell [23]	✓					✓			✓				24.6	-	-
Mind's Eye [30]	✓					✓			✓				19.0	20.4	-
DeepVS [25]	✓					✓			✓				23.0	19.5	66.0
LRCN [28]	✓					✓			✓				21.0	-	-

TABLE 5: Performance analysis of representative image captioning approaches in terms of different evaluation metrics on the Flickr30K datatset. The † marker indicates models trained by us with ResNet-152 features, while the ‡ marker indicates unofficial implementations. For all the metrics, the higher the value, the better (↑).

		Standard Metrics						Diversity Metrics				Embedding-based Metrics			Learning-based Metrics			
	Trained on	**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† [23]	COCO	51.0	11.4	13.1	34.8	22.8	7.6	0.037	0.093	331	94.8	8.6	0.019	61.9	52.9	90.6	0.604	0.656
Show, Attend and Tell† [42]	COCO	57.3	14.7	15.1	38.8	29.4	9.4	0.044	0.124	402	96.3	9.5	0.053	63.7	53.0	91.1	0.638	0.686
Up-Down‡ [58]	COCO	65.5	19.5	18.6	44.0	42.6	12.5	0.047	0.131	463	98.1	11.1	0.105	68.9	53.6	91.9	0.682	0.719
$\mathcal{M}^2$ Transformer [81]	COCO	67.9	21.0	19.4	45.3	47.4	13.0	0.048	0.150	470	98.9	11.7	0.106	67.0	53.7	91.8	0.680	0.721
VinVL [103]	COCO	**74.3**	**28.4**	**23.5**	**51.1**	**75.2**	**16.8**	**0.066**	**0.188**	**651**	**98.8**	**15.0**	**0.147**	**72.2**	**53.6**	**93.1**	**0.754**	**0.787**
Show and Tell† [23]	Flickr30K	64.1	21.5	18.3	44.4	41.7	12.2	0.037	0.075	373	84.5	11.2	0.090	64.2	53.5	92.1	0.658	0.701
Show, Attend and Tell† [42]	Flickr30K	65.6	23.6	19.2	45.4	49.1	13.3	0.045	0.096	454	90.1	11.8	0.089	64.1	53.4	92.1	0.679	0.717
Up-Down‡ [58]	Flickr30K	**72.4**	28.3	21.6	49.5	63.3	15.9	0.061	0.155	587	95.6	13.1	0.119	65.5	53.5	92.7	0.720	0.755
ORT [77]	Flickr30K	72.2	**30.1**	**22.8**	50.4	**68.8**	**16.9**	0.072	0.171	**738**	**96.1**	**13.7**	**0.129**	**67.2**	53.5	92.7	**0.728**	0.760
$\mathcal{M}^2$ Transformer [81]	Flickr30K	**72.4**	29.8	22.4	**50.6**	68.4	16.2	**0.079**	**0.196**	728	93.8	13.6	0.120	65.0	**53.6**	**92.9**	0.724	**0.763**

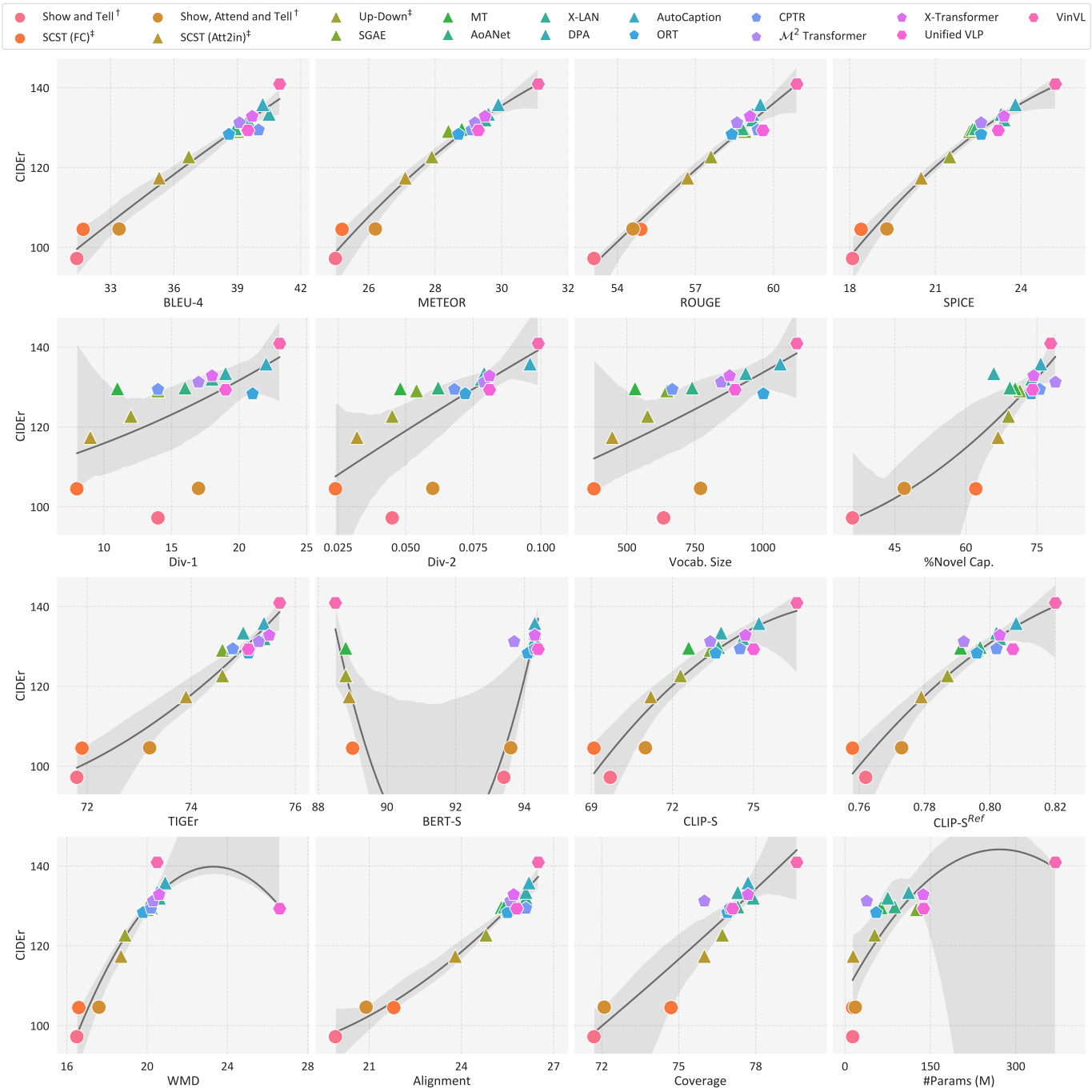


Fig. 10: Relationship between CIDEr, number of parameters and other scores. Values of Div-1, Div-2, Alignment, CLIP-S, and CLIP-S^{Ref} are multiplied by powers of 10 for readability.

Show and Tell: *A woman and a child are in a kitchen.*
Show, Attend and Tell: *A woman in a kitchen holding a dog.*
SCST (FC): *A woman sitting at a kitchen with a dog.*
Up-Down: *A woman is brushing her teeth with a dog.*
AoANet: *A woman feeding a dog in a bath tub.*
M² Transformer: *A woman is washing a dog in a sink.*
X-Transformer: *A woman feeding a dog in a bath tub.*
VinVL: *A woman bathing a dog in a bath tub.*

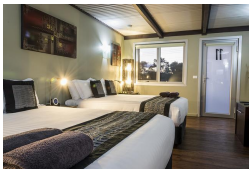
Show and Tell: *A baseball player holding a bat on a field.*
Show, Attend and Tell: *A man holding a baseball bat on a field.*
SCST (FC): *A baseball player holding a bat on a field.*
Up-Down: *A group of baseball players standing in a field.*
AoANet: *A group of men standing next to a baseball field.*
M² Transformer: *A group of men standing around a sheep.*
X-Transformer: *A group of men standing on a baseball field with a dog.*
VinVL: *A group of baseball players standing around a dog.*

Show and Tell: *A baby elephant standing next to a bigger elephant.*
Show, Attend and Tell: *An elephant standing next to a baby elephant.*
SCST (FC): *A baby elephant standing in the grass in a field.*
Up-Down: *Two elephants standing next to an adult elephant.*
AoANet: *An adult elephant standing next to a baby elephant.*
M² Transformer: *An adult elephant and a baby elephant standing in a field.*
X-Transformer: *An adult elephant standing next to a baby elephant.*
VinVL: *A baby elephant standing next to an adult elephant.*

Show and Tell: *A bed sitting in a bedroom next to a window.*
Show, Attend and Tell: *A bed that is sitting in a room.*
SCST (FC): *A person sitting on a bed with a window.*
Up-Down: *A bed with a window in a room.*
AoANet: *A bed with a bed with a table.*
M² Transformer: *A bed in a room with a window.*
X-Transformer: *A bed with a wooden frame and a table.*
VinVL: *A white bed with a wooden frame in a room.*

Show and Tell: *A cat laying on top of a remote control.*
Show, Attend and Tell: *A close up of a cat with a remote control.*
SCST (FC): *A cat laying on top of a remote control.*
Up-Down: *A white cat laying on top of a remote control.*
AoANet: *A white cat laying on top of a remote control.*
M² Transformer: *A white cat laying on top of a remote control.*
X-Transformer: *A white kitten laying on top of a remote control.*
VinVL: *A small white mouse sitting on top of a remote control.*

Show and Tell: *A couple of people that are on a horse.*
Show, Attend and Tell: *A black and white photo of a person on a horse.*
SCST (FC): *A couple of people standing next to a horse.*
Up-Down: *A woman sitting on a bench next to a horse.*
AoANet: *Two women sitting on a fence with two horses.*
M² Transformer: *A black and white photo of a woman sitting on a horse.*
X-Transformer: *A black and white photo of a man and a horse.*
VinVL: *A black and white photo of a person sitting on a horse.*



Show and Tell: *A living room with a couch and a table.*
Show, Attend and Tell: *A hotel room with two beds and a television.*
SCST (FC): *A bedroom with a bed in a room with a table.*
Up Down: *A hotel room with two beds and a window.*
AoANet: *A hotel room with two beds and a table.*
M² Transformer: *A hotel room with two beds and a door.*
X-Transformer: *A hotel room with two beds and a door.*
VinVL: *A hotel room with two beds and a window.*

Show and Tell: *A woman holding a toothbrush in her hand.*
Show, Attend and Tell: *A little girl with a hair dryer in her hair.*
SCST (FC): *A woman brushing her teeth in a bathroom mirror.*
Up-Down: *A woman brushing her hair with a baseball bat.*
AoANet: *A little girl holding a baseball bat in her hand.*
M² Transformer: *A young girl holding a baseball bat.*
X-Transformer: *A young girl holding a hair dryer in a room.*
VinVL: *A little girl brushing her hair with a brush.*

Show and Tell: *A large building with a clock on it.*
Show, Attend and Tell: *A church with a large cathedral in front of it.*
SCST (FC): *A wooden bench sitting in front of a building.*
Up-Down: *A church with a table and chairs in front of a building.*
AoANet: *A church with a bench and a table.*
M² Transformer: *A church with two benches and a fireplace.*
X-Transformer: *A church with pews and a table in a room.*
VinVL: *A church with a pews and a cross on the wall.*

Show and Tell: *A man sitting on a bench in front of a building.*
Show, Attend and Tell: *A man sitting on a bench with a tie.*
SCST (FC): *A man sitting on a bench in a building.*
Up-Down: *A man sitting on a bench in a room.*
AoANet: *A man sitting on a bench in a room.*
M² Transformer: *A man sitting on a bench in front of a.*
X-Transformer: *A man sitting on a wooden bench.*
VinVL: *A man sitting on a wooden bench.*

Show and Tell: *A man sitting on a boat with a bunch of bananas.*
Show, Attend and Tell: *A group of people that are standing in a boat.*
SCST (FC): *A group of people sitting on a boat in the water.*
Up-Down: *A group of people sitting in a boat with bananas.*
AoANet: *A group of people in a boat with food.*
M² Transformer: *Two people in a boat with baskets of food.*
X-Transformer: *A woman sitting in a boat with plates of food.*
VinVL: *A couple of people on a boat with food.*

Show and Tell: *A baby sitting on a couch with a remote.*
Show, Attend and Tell: *A little boy that is sitting on a couch.*
SCST (FC): *A young boy sitting on a couch with a remote control.*
Up-Down: *A young boy laying on a bed with a remote.*
AoANet: *A young boy laying on a couch holding a remote control.*
M² Transformer: *A boy laying on a bed holding a Nintendo Wii.*
X-Transformer: *A young boy laying on a couch holding a Wii remote.*
VinVL: *A young boy sleeping on a couch with a bottle.*

Show and Tell: *A group of people sitting on top of a bus.*
Show, Attend and Tell: *A group of people sitting on a train.*
SCST (FC): *A group of people sitting on a bus.*
Up-Down: *A man sitting on a bus with a cell phone.*
AoANet: *A a man sitting on a bus with a bus.*
M² Transformer: *A man in a hat sitting on a bus.*
X-Transformer: *A man sitting on a bus.*
VinVL: *A man sitting on a bus with yellow seats.*

Show and Tell: *A pile of oranges sitting next to each other.*
Show, Attend and Tell: *A pile of oranges sitting next to each other.*
SCST (FC): *A bunch of oranges sitting on top of a table.*
Up-Down: *A pile of oranges sitting on top of a table.*
AoANet: *A pile of oranges sitting on top of each other.*
M² Transformer: *A pile of oranges sitting in a bowl.*
X-Transformer: *A pile of oranges sitting on top of each other.*
VinVL: *A bunch of oranges sitting on a blue cloth.*

Show and Tell: *A bird sitting on top of a wooden table.*
Show, Attend and Tell: *A black bird sitting on top of a bowl.*
SCST (FC): *A bird sitting on top of a bird feeder.*
Up-Down: *A black bird sitting on a bowl of food.*
AoANet: *A black bird sitting on top of a wooden basket.*
M² Transformer: *A black bird sitting in a wooden bowl of food.*
X-Transformer: *A black bird standing in a wooden bowl.*
VinVL: *A black bird sitting in a flower pot eating food.*

Fig. 11: Additional qualitative examples from eight popular captioning models on COCO test images.

Show and Tell: *A dog sitting on top of a wooden bench.*
Show, Attend and Tell: *A dog is sitting on a wooden bench.*
SCST (FC): *A dog sitting on top of a bench.*
Up-Down: *A dog sitting on a bench in a park.*
AoANet: *A dog standing next to a wooden bench.*
M² Transformer: *A brown dog standing next to a wooden bench.*
X-Transformer: *A dog tied to a wooden bench in a park.*
VinVL: *A brown dog with a orange leash next to a bench.*

Show and Tell: *A man and a dog sitting on a bench.*
Show, Attend and Tell: *A couple of dogs are sitting on a bench.*
SCST (FC): *A man sitting on a bench with a dog.*
Up-Down: *A group of dogs are standing on a table.*
AoANet: *Two dogs sitting on a bench in front of a fruit stand.*
M² Transformer: *Three dogs on a leash sitting under a tent.*
X-Transformer: *Two dogs on leashes standing next to a fruit stand.*
VinVL: *A group of dogs on leashes sitting under a tent.*

Show and Tell: *A young boy in a white shirt and tie.*
Show, Attend and Tell: *A young boy in a white shirt and tie.*
SCST (FC): *A man wearing a shirt and a tie.*
Up-Down: *A young boy wearing a shirt and a tie.*
AoANet: *A young boy wearing a white shirt and a red tie.*
M² Transformer: *A young boy wearing a tie next to a clock.*
X-Transformer: *A young boy wearing a white shirt and red tie next to a clock.*
VinVL: *A young boy wearing a white shirt and tie standing next to*

Show and Tell: *A woman standing next to a brown horse.*
Show, Attend and Tell: *A woman standing next to a brown horse.*
SCST (FC): *A woman standing on top of a horse.*
Up-Down: *A woman is standing next to a horse.*
AoANet: *A woman standing next to a brown horse.*
M² Transformer: *A woman standing next to a horse on the beach.*
X-Transformer: *A woman standing next to a brown horse.*
VinVL: *A woman standing next to a horse on the beach.*

Show and Tell: *A dog sitting on a bench in the rain.*
Show, Attend and Tell: *A brown and white dog sitting on a wicker chair.*
SCST (FC): *A dog sitting on top of a bench.*
Up-Down: *A dog laying on a chair with a umbrella.*
AoANet: *A dog sitting in a basket with an umbrella.*
M² Transformer: *A dog laying in a basket under an umbrella.*
X-Transformer: *A brown dog laying in a basket with an umbrella.*
VinVL: *A brown dog sitting in a basket on a bike.*

Show and Tell: *A couple of giraffe standing next to each other.*
Show, Attend and Tell: *A giraffe and a giraffe in a zoo enclosure.*
SCST (FC): *A giraffe standing next to a fence.*
Up-Down: *A giraffe standing next to a tree in a field.*
AoANet: *A giraffe standing in the grass next to a tree.*
M² Transformer: *A giraffe standing next to a tree.*
X-Transformer: *A giraffe standing in a field with trees.*
VinVL: *A giraffe standing next to a tree in a field.*

Show and Tell: *A couple of kids sitting on a bench.*
Show, Attend and Tell: *A couple of girls sitting on a bench.*
SCST (FC): *A little girl sitting on a bench with a cell phone.*
Up-Down: *A group of dogs are standing on a table.*
AoANet: *Two children sitting on the ground eating a book.*
M² Transformer: *Two young children sitting on the grass eating food.*
X-Transformer: *Two children sitting on the grass eating food.*
VinVL: *A boy and a girl sitting on the ground eating food.*

Show and Tell: *A man and a woman sitting on a bed.*
Show, Attend and Tell: *A man sitting on a bed in front of a wall.*
SCST (FC): *A man sitting on a couch with a television.*
Up-Down: *A man sitting on a bed in a room.*
AoANet: *A man sitting on a bed in a room.*
M² Transformer: *A man sitting on a bed with a laptop.*
X-Transformer: *A man sitting on a bed.*
VinVL: *A man sitting on a bed in a room with a wall.*

Show and Tell: *A brown dog sitting on top of a wooden floor.*
Show, Attend and Tell: *A dog laying on the ground next to a dog.*
SCST (FC): *A couple of dogs laying on the street with a frisbee.*
Up-Down: *Two dogs laying on the ground next to a dog.*
AoANet: *A dog and a dog laying on the sidewalk.*
M² Transformer: *Two dogs laying on the sidewalk with a.*
X-Transformer: *A dog laying on the sidewalk with a toy.*
VinVL: *Two small dogs are standing next to each other.*

Show and Tell: *A woman is standing on a horse in the snow.*
Show, Attend and Tell: *A woman is standing in front of a horse.*
SCST (FC): *A woman standing on top of a horse.*
Up-Down: *A woman standing next to a white horse on the beach.*
AoANet: *A woman standing next to a horse in the snow.*
M² Transformer: *A woman standing next to a horse on the beach.*
X-Transformer: *A woman standing next to a horse in the snow.*
VinVL: *A woman standing in the snow with a white horse.*

Show and Tell: *A bike parked in front of a building.*
Show, Attend and Tell: *A bike parked in front of a building.*
SCST (FC): *A bike parked in front of a building with a bicycle.*
Up-Down: *A bike parked in front of a building.*
AoANet: *A bike parked in front of a door.*
M² Transformer: *A bike parked on the sidewalk in front of a house.*
X-Transformer: *A bike parked in front of a red building.*
VinVL: *A bike parked on the sidewalk in front of a red building.*

Show and Tell: *A close up of a pizza on a table.*
Show, Attend and Tell: *A couple of pizzas sitting on top of a table.*
SCST (FC): *A pizza sitting on top of a table.*
Up-Down: *Two pizzas sitting on top of a table.*
AoANet: *A group of pizzas sitting on a table.*
M² Transformer: *Four pizzas sitting on top of a table.*
X-Transformer: *A group of pizzas sitting on top of a table.*
VinVL: *A group of four different pizzas sitting on a pan.*

Show and Tell: *A man in a suit and tie sitting on a bench.*
Show, Attend and Tell: *A man in a suit and tie sitting on a bench.*
SCST (FC): *A couple of men standing next to each other.*
Up-Down: *Two men in suits sitting next to each other.*
AoANet: *Two men in suits standing in a field.*
M² Transformer: *Two men in suits and ties posing for a picture.*
X-Transformer: *Two men in suits posing for a picture.*
VinVL: *Two men in suits are posing for a picture.*

Show and Tell: *A black and white photo of an umbrella.*
Show, Attend and Tell: *A close up of a white umbrella on a couch.*
SCST (FC): *A umbrella sitting on top of a street.*
Up-Down: *A baseball bat is sitting in a room.*
AoANet: *A close up of a pair of scissors.*
M² Transformer: *A close up of a baseball bat.*
X-Transformer: *A close up of a vase with a tree.*
VinVL: *A close up of an umbrella on the ground.*

Show and Tell: *A bathroom with a sink and a toilet.*
Show, Attend and Tell: *A bathroom with a sink and a mirror.*
SCST (FC): *A bathroom with a sink and a mirror.*
Up-Down: *A bathroom with two sinks and a mirror.*
AoANet: *A bathroom with a sink and a mirror.*
M² Transformer: *A bathroom with a sink and a mirror.*
X-Transformer: *A bathroom with a sink and a mirror.*
VinVL: *A bathroom with a sink and a mirror.*

Fig. 12: Additional qualitative examples from eight popular captioning models on COCO test images.

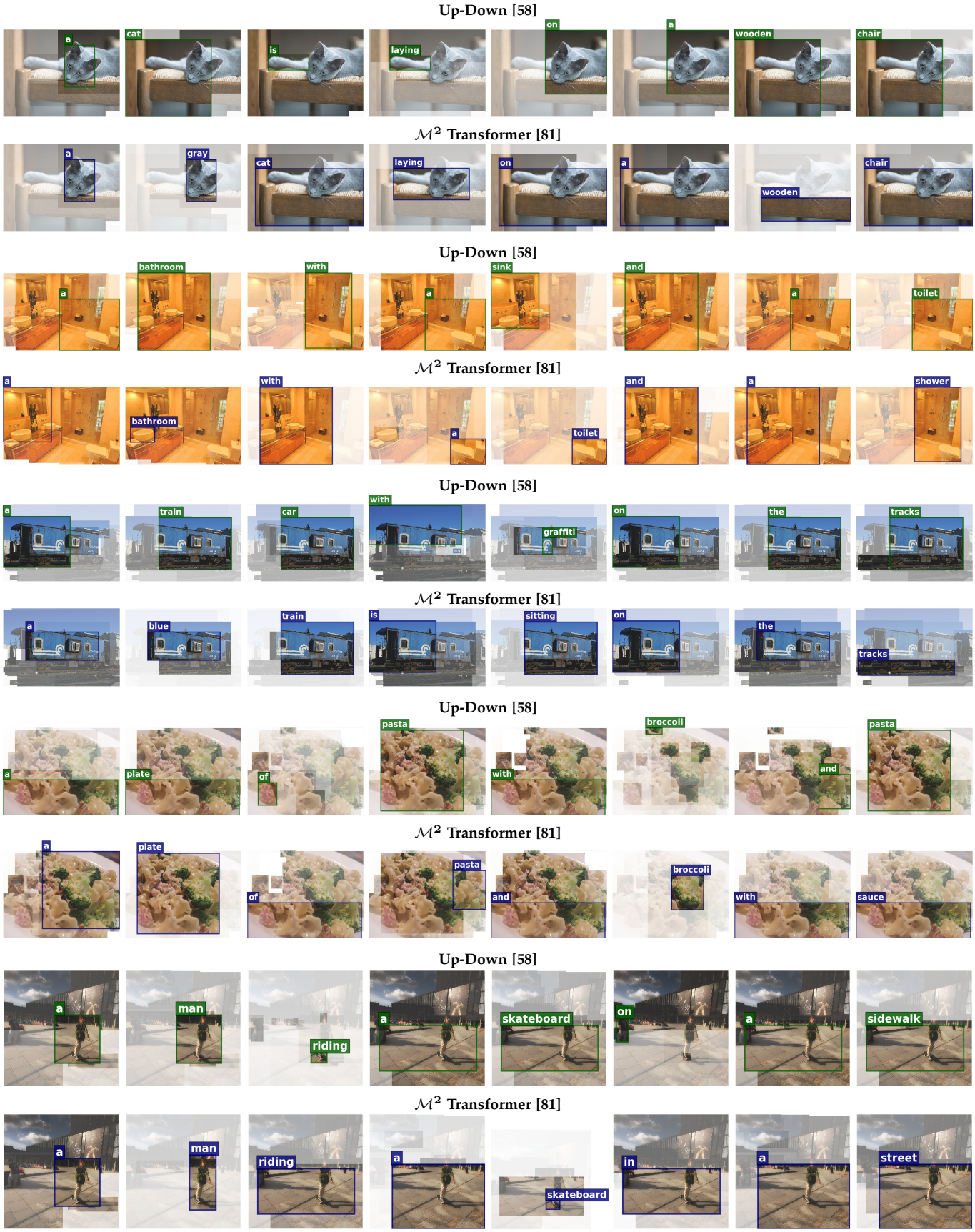


Fig. 13: Visualization of attention states for sample captions generated by Up-Down [58] and $\mathcal{M}^2$ Transformer [81]. For each generated word, we show the attended image regions, outlining the region with the maximum output attribution in green and blue, respectively.