The Unreasonable Effectiveness of CLIP Features for Image Captioning: An Experimental Analysis

Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, Rita Cucchiara
University of Modena and Reggio Emilia, Italy
{name.surname}@unimore.it

Abstract

Generating textual descriptions from visual inputs is a fundamental step towards machine intelligence, as it entails modeling the connections between the visual and textual modalities. For years, image captioning models have relied on pre-trained visual encoders and object detectors, trained on relatively small sets of data. Recently, it has been observed that large-scale multi-modal approaches like CLIP (Contrastive Language-Image Pre-training), trained on a massive amount of image-caption pairs, provide a strong zero-shot capability on various vision tasks. In this paper, we study the advantage brought by CLIP in image captioning, employing it as a visual encoder. Through extensive experiments, we show how CLIP can significantly outperform widely-used visual encoders and quantify its role under different architectures, variants, and evaluation protocols, ranging from classical captioning performance to zero-shot transfer.

1. Introduction

Image captioning is a task at the intersection between vision and language, whose challenges come both from each modality and, most importantly, their interaction. In fact, to properly describe an image, not only the ability to produce meaningful and grammatical sentences is needed, but correctly understanding its content is crucial. To this end, image representation plays a key role, making this aspect of great interest to the community working on image captioning and, in general, on tasks connecting vision and language. For years, image captioning approaches have relied on visual representations based on detected visual entities [2, 27], among which relations have been modeled via graphs [49, 51] or attention mechanisms [6, 8, 28, 31].

Despite the remarkable performance of these approaches, their applicability is somewhat limited since the set of objects the detector can distinguish defines what can be described in an image. For this reason, approaches resorting to learning grid-based features from scratch are being reconsidered. In fact, this paradigm eases large-scale pre-training even on noisy image-text pairs, automatically collected from the web, resulting captioning approaches that obtained state-of-the-art performance [14, 47].

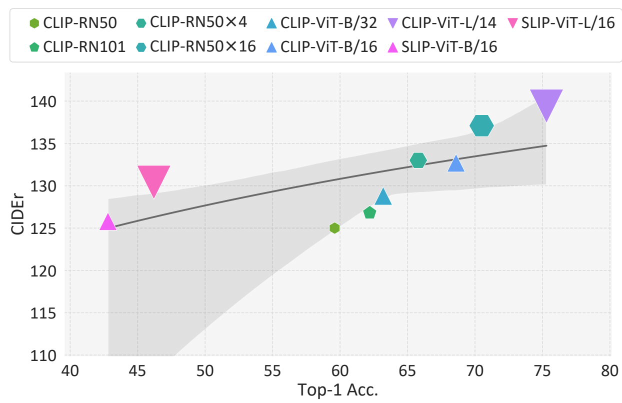

Figure 1. Relationship between zero-shot classification capability (expressed in terms of top-1 accuracy on Imagenet) and image captioning performance (expressed in terms of CIDEr on COCO) of CLIP-like features. The marker sizes are proportional to the number of parameters of the considered models.

On this line, recent developments in multi-modal contrastive training have led to effective CLIP-like models [33] able to extract rich visual features that are suitable for a variety of down-stream tasks [39]. Inspired by the role that CLIP features are exhibiting in other tasks, in this paper we investigate and quantify their role for image captioning. To this aim, we devise a general yet effective image captioning architecture, which is later employed for all experiments both in cross-entropy training and in self-critical fine-tuning [36]. We evaluate CLIP features, under different variants, in comparison with traditional detection-based features, as well as features coming from recent classification and self-supervised architectures [4, 11]. Further, we analyze both in-domain and out-of-domain performances and zero-shot transfer over a variety of datasets.

Our results show that CLIP features provide a significant improvement in terms of caption quality, out-of-domain performance, and zero-shot performance, and that are largely superior to features used in previous lines of research. Indeed, a simple Transformer-based captioner, equipped with CLIP features, can largely overcome state of the art approaches based on significantly more complex architectures. This effectiveness is quantified both as a function of the visual encoder architecture and as a function of the training data by running comparisons with models trained on fewer data. As a side contribution, we also assess the value of CLIP as a metric for image captioning.

Overall, our work is intended to be a "progress report" on visual features for image captioning, sheds new light on the role of emerging large-scale and multi-modal features, and provides effective baselines for future vision and language works.

2. Related Work

Image captioning requires both a deep understanding of the visual content within an image, its objects, attributes, and relationships and the capability of a language model to generate syntactically and semantically correct descriptions. In particular, the language model is asked to generate a sentence conditioned on the image representation, whose role is key for obtaining satisfactory results [41].

To represent the visual input, CNN-based solutions have been proposed for extracting global features [18,36] or grids of features [26, 48], and further improved through object detectors [2, 27] for obtaining a region-based features representation, and self-attention. As for the language model, in earlier works it was implemented as a recurrent neural network [15, 18, 20, 36], while more recent approaches employ Transformer-based fully-attentive models [5, 8, 28, 56]. The success of this latter strategy has also encouraged the proposal of multi-modal early-fusion strategies [14, 22, 54], which proved the effectiveness of building a semantic representation of the image by exploiting also the text at the early stages of the captioning pipeline.

Representing the image via region-based features, in combination with an attention mechanism, has been the standard design choice for years. More recently, however, fully-attentive models exploiting Transformer-like architectures [11, 43] and grid features became more popular, either combined with a CNN [56] or directly applied to image patches [25]. Thanks to the competitive performance of such models, grid features have been reconsidered [17, 39], and their suitability has been demonstrated also as starting point for large-scale vision-and-language pre-trained models such as SimVLM [47] and CLIP [33], whose features are employed in recent state-of-the-art captioning approaches [3, 7, 29, 39].

Contrastive Image-Language Pre-Training. The Contrastive Language-Image Pretraining (CLIP) paradigm adopted in [33] has allowed leveraging a large amount of weakly-labeled data for pre-training large models able to encode rich semantic information from multi-modal data. Due to the success of this idea, variants of CLIP have been developed. To gain efficiency, cross-modality parameter sharing [52] has been proposed. Moreover, some approaches go towards the direction of fully exploiting the noisy training data by modifying the training objective with self-supervision [30], within-modality loss terms [9, 23], and training data refinement [21, 44]. Finally, other approaches refine the granularity of the image-text alignment by proposing to exploit text paired with pixel [34], regions [57] or visual-textual tokens pairing [50].

The rich features from CLIP-like models can then be employed for a number of downstream tasks, both visual-only, such as image classification [33], action recognition [45], semantic segmentation [46], and visual-textual tasks such as text-guided image generation [32] and image and video captioning [7, 29, 39, 42]. In this work, we focus on the image captioning task and experimentally evaluate features from CLIP-like models to quantitatively assess their suitability for this task combining vision and language.

3. CLIP-Captioner

The goal of a captioning module is that of modeling an autoregressive distribution probability $p(\boldsymbol{w}_t|\boldsymbol{w}_{\tau<t}, \mathbf{V})$, where $\mathbf{V}$ is an input image and $\{\boldsymbol{w}_t\}_t$ is the sequence of words comprising the generated caption. This is achieved in existing models by training a language model conditioned on visual features to mimic ground-truth descriptions.

Recent works have employed both encoder-decoder [8] and encoder-only architectures [22, 54], in which multimodal connections are realized with a late-fusion or early-fusion strategy, respectively. While none of the two alternatives has demonstrated clear superiority [16], in this work we opt for an encoder-decoder model which separates visual and textual features inside the architecture and which can amplify the role of different visual descriptors. Even if a comparison with an encoder-only model is left for future work, we expect our findings to transfer seamlessly to an encoder-only model.

Architecture. We represent each training sample as a pair of image and text $(\mathbf{V}, \mathbf{W})$, where $\mathbf{V}$ is encoded with a set of fixed-length visual descriptors. The text input is tokenized with lower-cased Byte Pair Encodings [37].

For multimodal fusion, we employ an encoder-decoder Transformer [43] architecture, in which the encoder is in charge of processing visual features through multi-head self-attention (MSA) and feed-forward layers, while the decoder generates output words through multi-head self- and cross-attention (MSCA) and feed-forward layers. For enabling text generation, sequence-to-sequence attention

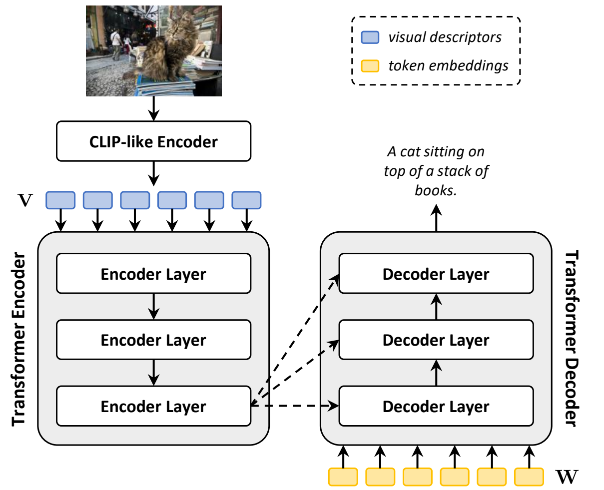

Figure 2. Overview of the considered CLIP-based captioning architecture.

masks are employed in each self-attention layer of the decoder. The visual descriptors $\mathbf{V} = \{\boldsymbol{v}_i\}_{i=1}^N$ are encoded via bi-directional attention in the encoder, while the token embeddings of the caption $\mathbf{W} = \{\boldsymbol{w}_i\}_{i=1}^L$ are inputs of the decoder, where N and L indicate the number of visual embeddings and caption tokens, respectively. The overall network operates according to the following schema:

$$\begin{aligned} &\textit{encoder} \quad && \tilde{\boldsymbol{v}}_i = \text{MSA}(\boldsymbol{v}_i, \mathbf{V}) \\ &\textit{decoder} \quad && \mathbf{O}_{\boldsymbol{w}_i} = \text{MSCA}(\boldsymbol{w}_i, \tilde{\mathbf{V}}, \{\boldsymbol{w}_t\}_{t=1}^i), \end{aligned} \tag{1}$$

where $\mathbf{O}$ is the network output, $\text{MSA}(\boldsymbol{x}, \mathbf{Y})$ a self-attention with $\boldsymbol{x}$ mapped to query and $\mathbf{Y}$ mapped to key-values, and $\text{MSCA}(\boldsymbol{x}, \mathbf{Y}, \mathbf{Z})$ a self-attention with $\boldsymbol{x}$ as query and $\mathbf{Z}$ as key-values, followed by cross-attention with $\boldsymbol{x}$ as query and $\mathbf{Y}$ as key-values. We omit feed-forward layers and the dependency between consecutive layers for ease of notation. Both encoder and decoder are, however, implemented with a sequence of Transformer layers.

Training objective. As in the case of early and late fusion, current literature has been investigating bidirectional masked losses as well as autoregressive language modeling losses. In contrast to concurrent works, which have adopted a bidirectional Masked Language Modeling objective that tends to be suboptimal for sequence generation, we train our network by following a unidirectional loss based on cross-entropy, *i.e.*

$$\mathcal{L} = -\mathbb{E}_{(\mathbf{V},\mathbf{W})\sim\mathcal{D}} \left(\sum_{t=1}^{L} \log p(\mathbf{O}_{\boldsymbol{w}_t} | \mathbf{V}, \boldsymbol{w}_{\tau<t}) \right), \tag{2}$$

where $\mathcal{D}$ indicates the training dataset.

Following a standard practice in image captioning [2, 36], after pre-training with cross-entropy we also adopt a fine-tuning stage using reinforcement learning. We employ a variant of the self-critical sequence training approach [36] on sequences sampled using beam search [2]: to decode, we sample the top-k words from the decoder probability distribution at each timestep and always maintain the top-k sequences with the highest probability. Following previous works [2], we use the CIDEr-D score as reward and baseline using the mean of the rewards in a beam [8].

Inference. Once the model is trained, at each time step t, the model samples a token $\hat{\boldsymbol{w}}_t$ from the output probability distribution. This is then concatenated to previously predicted tokens to form a sequence $\{\hat{\boldsymbol{w}}_\tau\}_{\tau=1}^t$, which is employed as the input for the next iteration. Since the representation of output tokens does not depend on subsequent tokens, the past intermediate representations are kept in memory to avoid repeated computation and increase efficiency at prediction time.

Visual features. To obtain the set of visual features $\mathbf{V}$ for an image, we employ a CLIP-like visual encoder pre-trained to match vision and language [33]. CLIP [33] and similar approaches employ either ResNet-based or ViT-based visual encoders. In the case of ViT-based architectures, we employ the grid of features coming from the last encoder layer for preserving spatial awareness and a better feature granularity. We also include the output of the `[CLS]` token, which is usually employed as a global feature vector for contrastive learning. On the other hand, in CLIP, ResNet-based backbones replace the global average pooling layer with an attention pooling mechanism. In this case, we employ the grid of features of the last residual block as visual descriptors. As in this case the global feature vector used in contrastive learning is obtained by applying an attention operator between the average pooled representation of the image and the grid of features we drop it to avoid redundancy.

4. Experimental Analysis

4.1. Implementation details

Visual features are projected into d-dimensional vectors with $d = 384$ and fed to our Transformer-based captioning model, which has three layers in the encoder and three layers with six attention heads in the decoder. For efficiency, the length of the output token sequence is limited to 80 tokens. For training with cross-entropy loss, we use the LAMB optimizer [53] and the learning rate scheduling strategy as in [43], with minibatch size equal to 1,080. For the CIDEr-based fine-tuning, we adopt the SCST strategy [36] sampling over the $k = 5$ best sequences from a beam-search scheme, with the Adam optimizer [19] and learning rate of 5×10^{-6}.

Table 1. Results on the COCO Karpathy-test split.

		Cross-Entropy Training								CIDEr Optimization							
	# Params (M)	B-4	M	R	C	S	BERT-S	CLIP-S	CLIP-S$^{\text{Ref}}$	B-4	M	R	C	S	BERT-S	CLIP-S	CLIP-S$^{\text{Ref}}$
Faster R-CNN [2, 35]	65.4	35.3	27.4	56.1	111.4	20.2	93.8	0.727	0.788	37.7	28.3	57.6	124.8	21.9	94.0	0.737	0.792
DINO-ViT-B/16 [4]	85.8	32.4	25.8	54.0	101.1	18.7	93.5	0.719	0.777	33.8	26.5	55.2	112.5	20.0	93.6	0.721	0.777
ViT-B/32 [11]	88.3	34.8	27.2	55.8	110.7	20.2	93.8	0.735	0.792	37.2	27.9	57.3	122.4	21.6	93.8	0.741	0.797
ViT-B/16 [11]	86.9	35.6	27.9	56.5	115.2	20.8	93.8	0.741	0.799	37.7	28.5	57.9	126.2	22.3	94.1	0.745	0.802
CLIP-RN50 [33]	38.3	35.6	27.5	56.4	113.1	20.5	93.8	0.739	0.796	36.8	28.0	57.4	125.0	21.5	93.5	0.739	0.794
CLIP-RN101 [33]	56.2	36.0	28.0	56.6	116.0	21.0	93.8	0.748	0.802	38.0	28.5	57.9	126.8	22.3	94.0	0.757	0.807
CLIP-RN50×4 [33]	87.1	37.3	28.3	57.4	118.9	21.3	93.9	0.743	0.801	39.2	29.2	58.8	133.0	22.7	93.8	0.753	0.808
CLIP-RN50×16 [33]	167.3	38.4	28.7	58.3	123.1	21.7	94.0	0.746	0.805	40.0	29.4	59.4	137.1	23.2	93.9	0.750	0.808
CLIP-ViT-B/32 [33]	87.8	36.0	27.8	56.5	114.9	20.8	93.9	0.750	0.803	37.9	28.5	57.7	128.0	22.3	94.0	0.757	0.807
CLIP-ViT-B/16 [33]	86.2	37.2	28.4	57.5	119.9	21.3	94.0	0.747	0.805	38.7	29.2	58.8	132.7	23.0	**94.2**	0.750	0.805
CLIP-ViT-L/14 [33]	304.0	**38.7**	**29.3**	**58.6**	**126.0**	**22.5**	**94.2**	**0.754**	**0.811**	**40.6**	**30.0**	**59.9**	**139.4**	**23.9**	94.1	**0.760**	**0.814**

4.2. Evaluation protocol

To assess the role of visual features extracted from CLIP-like models in image captioning, we consider a number of datasets employed for the task, both in its standard definition and variants, to explore the suitability of such features in standard and more challenging image captioning settings. We use the commonly adopted COCO dataset [24], by following the splits defined by Karpathy *et al.* [18]. In addition, we consider a dataset collected for studying novel object image captioning, nocaps [1]. The images in this dataset contain around 400 objects that are not in COCO and are grouped into three subsets depending on their semantic distance to COCO (*i.e.* in-domain, near-domain, and out-of-domain images). We also take into account two domain-specific datasets, *i.e.* the VizWiz [12] and TextCaps [40] datasets. The former contains images taken from visually-impaired people for everyday activities, while the latter images with text that must be included in the caption. Moreover, we consider the large-scale pre-training Conceptual Captions (CC3M) dataset [38], which contains pairs of images and a single noisy caption for each image.

In our evaluation, we express the performance in terms of the standard image captioning metrics and learning-based metrics such as BERT-S [55] and CLIP-S [13], in its standard version comparing image and generated caption directly, and its variant considering also the reference captions (CLIP-S$^{\text{Ref}}$). These learning-based metrics exploit pre-trained embeddings from the text-only BERT [10] model and the multi-modal CLIP [33], respectively.

4.3. Quantitative results

Effectiveness of CLIP features. Features based on object detections are currently the most popular choice in image captioning [14, 54] when feature learning is not performed from scratch [47]. Recent literature, however, has developed visual backbones by improving both in architectural terms, with ViT-based solutions [11] and self-supervised and multi-modal training strategies. In Table 1 we compare detection features with classification and self-supervised visual features and CLIP-based backbones.

Performance on COCO reveals that a self-supervised architecture like DINO [4] fails to provide the same performance of a Faster R-CNN trained on Visual Genome [2]. In contrast, a sufficiently large and fine-grained Vision Transformer [11] trained for classification provides a significant improvement with respect to detection features (111.4 vs. 115.2 CIDEr points, in XE). This outlines that modern grid-like features can overcome traditional detection features and that ViT is an appropriate feature extraction architecture for image captioning.

Moving to features that are trained to match vision and language, we compare different CLIP backbones based on ResNet and ViT. The smallest CLIP model in terms of number of parameters, CLIP-RN50, improves over detection-based features (111.4 vs. 113.1 CIDEr). Remarkably, this performance margin increases as model and input size increase in ResNet-based backbones. Increasing model depth from 50 to 101 layers brings CIDEr from 113.1 to 116.0, while adopting EfficientNet-style architectures further improves performance, with CLIP-RN50×16 reaching 123.1 CIDEr points. When employing ViT-like architectures, instead, we notice that reducing input patch size can provide similar results to CNN-based architectures: CLIP-ViT-B/16, for instance, reaches 119.9 CIDEr points while being comparable to CLIP-RN50×4 in terms of number of parameters. Increasing model depth and further reducing patch size clearly improves performance, with CLIP-ViT-L/14 reaching 126.0 CIDEr points after XE pre-training. Overall, this amounts to a 13.1% relative improvement over the traditional Faster R-CNN features.

The aforementioned considerations transfer seamlessly to the corresponding models trained with CIDEr optimization, highlighting that the role of visual features is maintained between the two learning stages. CLIP-ViT-L/14, in particular, attains 139.4 CIDEr points. Overall, this outlines

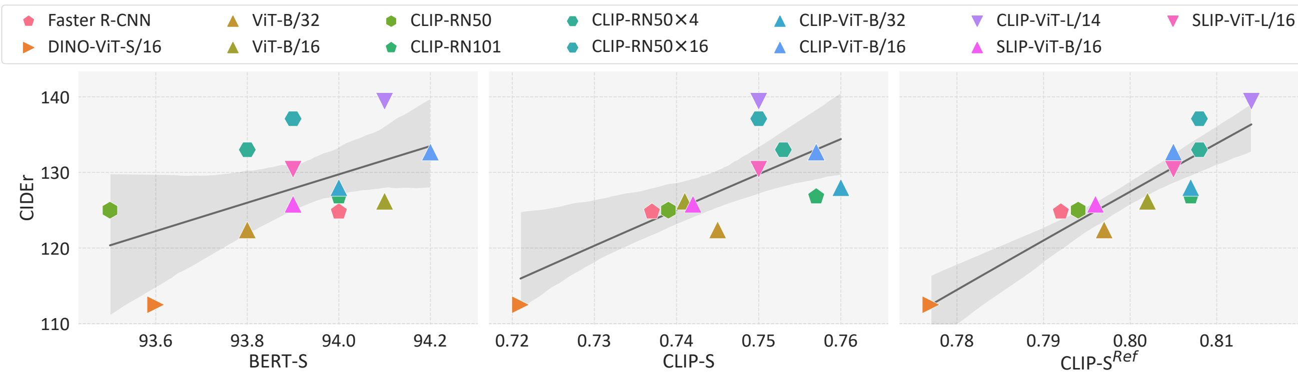

Figure 3. Relationship between CIDEr and learned image captioning metrics on the COCO Karpathy-test split.

that multi-modal features learned by matching vision and language are more effective than traditional features learned on vision only and that CLIP is one of the best visual feature extractors available at present. However, it shall be noted that there is no experimental evidence that grid-based feature extraction is superior to a detection-based strategy. Indeed, (*i*) the Faster R-CNN feature extraction employs an RN50 architecture, and thus, it could be improved in architectural terms; (*ii*) ViT and CLIP models have been trained on a significantly larger amount of data. Further research is thus needed to separate the role of data and architecture and to assess the role of detection-level pooling.

Out-of-domain performance. Beyond evaluating the performance on in-domain captioning with COCO, we also assess the role of visual features on the nocaps dataset [1] which contains both in-domain and out-of-domain images. Also in this case, CLIP shows an increased performance, especially when employing larger models or models having a higher input resolution. Interestingly, ResNet-based backbones work slightly better than ViT-based ones, given a fixed amount of parameters on out-of-domain data. For instance, CLIP-RN50×4 is superior to CLIP-ViT-B/16. The best performance, however, is again reached by CLIP-ViT-L/14. Interestingly again, in this case, ViT models trained on classification provide a significant boost on out-of-domain data compared to Faster R-CNN features, suggesting that training on large-scale data helps to obtain better features for out-of-domain captioning. This is further confirmed when comparing the performance gain provided by CLIP with respect to Faster R-CNN on nocaps (28.5 CIDEr points) and on COCO (13.6 CIDEr points).

Self-supervised and contrastive learning. In Table 3 we further compare with SLIP [30], which has been trained by using a self-supervised criterion in addition to contrastive learning, but on a subset of YFCC100M, thus on significantly less data than CLIP. SLIP features appear to be weaker than CLIP's ones, with a drop that ranges between 6.9 CIDEr points and 9.0 CIDEr points when comparing models with similar dimensionality and architecture. Considering that SLIP is superior to CLIP in zero-shot classification when trained on the same amount of data, the drop in captioning performance should be attributed to the lack of training data.

Table 2. Results on the nocaps validation set.

	In		Near		Out		Overall		
	C	S	C	S	C	S	C	S	CLIP-S
Faster R-CNN [2, 35]	81.1	12.0	64.1	11.0	34.5	8.7	60.5	10.8	0.640
DINO-ViT-B/16 [4]	70.3	10.5	54.9	10.0	31.7	8.2	52.4	9.7	0.623
ViT-B/32 [11]	82.3	11.8	68.9	11.1	46.5	9.4	66.3	10.9	0.667
ViT-B/16 [11]	86.9	12.0	72.9	11.4	51.5	9.8	70.5	11.2	0.674
CLIP-RN50 [33]	91.2	12.0	73.7	11.3	45.3	8.9	70.5	10.9	0.673
CLIP-RN101 [33]	98.0	12.9	78.5	12.0	60.0	10.2	77.6	11.8	0.699
CLIP-RN50x4 [33]	98.7	13.1	84.0	12.2	60.3	10.2	81.0	11.9	0.696
CLIP-RN50x16 [33]	100.1	12.9	86.5	12.2	64.8	10.0	83.1	11.9	0.691
CLIP-ViT-B/32 [33]	89.6	12.7	75.5	11.8	51.2	9.6	72.6	11.5	0.689
CLIP-ViT-B/16 [33]	93.2	12.3	80.4	12.2	57.1	10.2	77.5	11.8	0.683
CLIP-ViT-L/14 [33]	**104.5**	**13.5**	**92.2**	**13.0**	**67.6**	**10.7**	**89.0**	**12.6**	**0.708**

Table 3. Comparison between CLIP-based and SLIP-based encodings with the same backbones on the COCO Karpathy-test split.

	# Params (M)	B-4	M	R	C	S	CLIP-S
SLIP-ViT-B/16 [30]	85.8	36.6	28.5	57.4	125.8	22.2	0.742
CLIP-ViT-B/16 [33]	86.2	**38.7**	**29.2**	**58.8**	**132.7**	**23.0**	**0.750**
SLIP-ViT-L/16 [30]	303.3	38.5	28.9	58.5	130.4	22.6	0.750
CLIP-ViT-L/14 [33]	304.0	**40.6**	**30.0**	**59.9**	**139.4**	**23.9**	**0.760**

In Figure 1, we also compare the zero-shot classification capabilities of the aforementioned models with their CIDEr scores. As it can be seen, there is a weak but relevant correlation between the two quantities, especially when considering models with similar dimensionalities. This further highlights the dependency between CLIP's performance as a feature extractor and the data on which it has been trained and suggests that collecting large-scale datasets with sufficient quality will be crucial for further V&L research.

CLIP for image captioning evaluation. Other than as a

Table 4. Zero-shot results on CC3M, VizWiz, and TextCaps validation splits.

	CC3M						VizWiz						TextCaps					
	B-4	M	R	C	S	CLIP-S	B-4	M	R	C	S	CLIP-S	B-4	M	R	C	S	CLIP-S
DINO-ViT-B/16 [4]	1.5	5.6	15.1	18.7	6.1	0.595	11.5	12.4	36.0	18.1	5.8	0.574	12.2	13.1	34.1	24.1	8.6	0.583
ViT-B/32 [11]	1.7	6.2	16.0	24.1	7.9	0.637	14.2	14.2	38.4	26.3	7.7	0.613	13.8	14.1	35.2	29.6	10.2	0.621
ViT-B/16 [11]	1.8	6.6	16.6	27.2	8.8	0.647	14.3	14.3	38.7	28.2	7.9	0.628	14.3	14.7	36.0	31.6	10.9	0.629
CLIP-RN50 [33]	1.8	6.5	16.6	26.1	8.2	0.655	14.9	14.4	39.4	27.2	7.5	0.628	14.8	14.8	36.8	32.9	10.8	0.636
CLIP-RN101 [33]	1.9	6.8	16.8	28.2	8.8	0.674	15.0	15.1	39.7	31.1	8.3	0.653	15.3	15.2	36.7	34.8	11.4	0.657
CLIP-RN50x4 [33]	2.1	7.0	17.4	30.8	9.3	0.673	15.9	15.2	40.1	32.5	8.4	0.650	15.7	15.3	37.0	35.6	11.5	0.652
CLIP-RN50x16 [33]	2.2	7.1	17.7	31.9	9.5	0.673	17.1	15.7	41.4	36.3	9.0	0.651	16.0	15.5	37.4	36.3	11.6	0.649
CLIP-ViT-B/32 [33]	2.0	6.7	17.0	27.1	8.4	0.667	15.0	14.7	39.4	29.0	8.0	0.645	15.0	14.8	36.7	33.3	10.9	0.646
CLIP-ViT-B/16 [33]	2.0	6.9	17.0	29.0	9.1	0.662	15.4	15.3	39.6	31.4	8.8	0.640	15.0	15.0	36.4	34.5	11.4	0.641
CLIP-ViT-L/14 [33]	**2.4**	**7.5**	**18.1**	**34.9**	**10.4**	**0.688**	**16.9**	**16.1**	**41.1**	**39.9**	**9.6**	**0.657**	**16.8**	**15.7**	**37.9**	**38.7**	**11.9**	**0.659**

visual encoder, CLIP can benefit image captioning also as the building block for an evaluation score, which led to the definition of the CLIP-S [13]. The score is obtained as the adjusted cosine similarity of image and candidate caption representations, and thus, it does not need ground-truth annotations, making it applicable also in an unpaired captioning scenario. Nevertheless, if available, reference captions can be exploited in the CLIP-S$^{\text{Ref}}$ variant of the score.

Despite assessing the suitability of the CLIP-S as an image captioning metric is beyond the scope of this work, in Tables 1-4 we also report the performance of the considered models in terms of CLIP-S and deepen the analysis of its relation with the standard CIDEr metric on COCO in Fig. 3. It can be observed that CLIP-S highly correlates with the CIDEr, both in the standard and reference-based definitions, with a Pearson correlation coefficient equal to 0.72 and 0.91, respectively. This unveils that the reference-based metric could be employed as a replacement of classical metrics, while the reference-free counterpart should be used with higher caution, although providing a significant correlation with CIDEr.

In Fig. 3, we also report the relationship between the CIDEr and the single-modality learning-based BERT-S. The correlation between the two scores is weaker compared to the CLIP-based scores (with a Pearson correlation coefficient of 0.54). This suggests that the multi-modal embedding obtained from CLIP allows giving more precise insights into the performance of image captioning models even when no ground-truth captions are available, with respect to text-only embeddings comparing candidate and reference captions but disregarding the image.

Zero-shot captioning transfer. As an additional analysis, we perform experiments in a zero-shot captioning setting. In particular, we consider the web-scale CC3M dataset, the VizWiz dataset [12], which contains images originating from blind people, and TextCaps [40], with images containing text. Although all of them represent distinct visual and semantic distributions from those of COCO, images from VizWiz and TextCaps have been manually annotated, while CC3M contains automatically-collected captions obtained by cleaning alt-text pairs from the web.

Table 5. Comparison with the state of the art the COCO Karpathy-test split.

	Web-Scale Training	B-4	M	R	C	S
Up-Down [2]	-	36.3	27.7	56.9	120.1	21.4
AoANet [15]	-	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$ Transformer [8]	-	39.1	29.2	58.6	131.2	22.6
X-Transformer [31]	-	39.7	29.5	59.1	132.8	23.4
DLCT [28]	-	39.8	29.5	59.1	133.8	23.0
RSTNet [56]	-	39.3	29.4	58.8	133.3	23.0
CLIP-Captioner (RN50×16)	-	40.0	29.4	59.4	137.1	23.2
CLIP-Captioner (ViT-L/14)	-	**40.6**	**30.0**	**59.9**	**139.4**	**23.9**
OSCAR$_{\text{large}}$ [22]	✓	41.7	30.6	-	140.0	24.5
VinVL$_{\text{large}}$ [54]	✓	41.0	31.1	-	140.9	25.2
SimVLM$_{\text{huge}}$ [47]	✓	40.6	33.7	-	143.3	25.4
LEMON$_{\text{huge}}$ [14]	✓	42.6	31.4	-	145.5	25.5

Table 4 reports the results obtained by testing models trained exclusively on COCO on the validation splits of the three considered datasets. CLIP-ViT-L/14 provides the best performances on all the three considered datasets, confirming that the choice of the visual backbone does not strictly depend on that of the dataset. Further, employing CLIP-like descriptors helps also in this case and provides a large margin with respect to descriptors trained for classification or in a self-supervised manner.

Comparison to the state of the art. Finally, in Table 5 we compare with state-of-the-art approaches for image captioning, either trained exclusively on COCO (upper portion of the table) or using external data as well (bottom part of the table). The considered baseline captioner outperforms all approaches trained on COCO only and which employ detection-based features. For instance, the encoder-decoder captioner reaches 139.4 CIDEr points when used in conjunction with CLIP-ViT-L/14, which is superior to the recent RSTNet [56].

Further, we notice that the same captioner reaches similar results to OSCAR [22] and VinVL [54] in their "Large" configurations, although having a significantly lower number of parameters. Overall, a baseline captioner based on CLIP features is only 6.1 CIDEr points lower than the current state of the art, LEMON [14], which employs detection-based features and trains on large-scale data. This underlines the high-quality level reached by CLIP features and the need for revisiting captioning baselines in light of the role of visual features. Future works that will deal with out-of-domain data and that are likely to employ CLIP-based features, indeed, will need to design careful baselines to achieve fair comparisons with previous literature.

5. Conclusion

In this paper, we have extensively explored the effectiveness of CLIP features in image captioning. In particular, we have considered CLIP-like visual encoders with different backbones, both based on ResNet and ViT, and used them in conjunction with an encoder-decoder image captioning approach. The performance of these variants has been assessed on the benchmark COCO dataset and tested on other captioning datasets in a zero-shot fashion. The experimental results obtained demonstrate the superior suitability of CLIP-based encoders compared to non-CLIP-based, for a multi-modal task such as image captioning.

Acknowledgment

We thank CINECA, the Italian Supercomputing Center, for providing computational resources. This work has been partially supported by "Fondazione di Modena" and by the H2020 ICT-48-2020 HumanE-AI-NET project.

References

[1] Harsh Agrawal, Karan Desai, Xinlei Chen, Rishabh Jain, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 4, 5

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 1, 2, 3, 4, 5, 6

[3] Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. CaMEL: Mean Teacher Learning for Image Captioning. In *ICPR*, 2022. 2

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*, 2021. 1, 4, 5, 6

[5] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. SMArT: Training Shallow Memory-aware Transformers for Robotic Explainability. In *ICRA*, 2020. 2

[6] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Explaining Transformer-based Image Captioning Models: An Empirical Analysis. *AI Communications*, 2021. 1

[7] Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. Universal Captioner: Inducing Content-Style Separation in Vision-and-Language Model Training. *arXiv preprint arXiv:2111.12727*, 2022. 2

[8] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In *CVPR*, 2020. 1, 2, 3, 6

[9] Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing Contrastive Language-Image Pre-training: A CLIP Benchmark of Data, Model, and Supervision. *arXiv preprint arXiv:2203.05796*, 2022. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2018. 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 1, 2, 4, 5, 6

[12] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning Images Taken by People Who Are Blind. In *ECCV*, 2020. 4, 6

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4, 6

[14] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling Up Vision-Language Pre-training for Image Captioning. *arXiv preprint arXiv:2111.12233*, 2021. 1, 2, 4, 6, 7

[15] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on Attention for Image Captioning. In *ICCV*, 2019. 2, 6

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*, 2021. 2

[17] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *CVPR*, 2020. 2

[18] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2, 4

[19] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 3

[20] Federico Landi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. Working Memory Connections for LSTM. *Neural Networks*, 144:334–341, 2021. 2

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086*, 2022. 2

[22] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV*, 2020. 2, 6

[23] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *ICLR*, 2021. 2

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 4

[25] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. CPTR: Full Transformer Network for Image Captioning. *arXiv preprint arXiv:2101.10804*, 2021. 2

[26] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 2

[27] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural Baby Talk. In *CVPR*, 2018. 1, 2

[28] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-Level Collaborative Transformer for Image Captioning. In *AAAI*, 2021. 1, 2, 6

[29] Ron Mokady, Amir Hertz, and Amit H Bermano. ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2

[30] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision meets Language-Image Pre-training. *arXiv preprint arXiv:2112.12750*, 2021. 2, 5

[31] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-Linear Attention Networks for Image Captioning. In *CVPR*, 2020. 1, 6

[32] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *ICCV*, 2021. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 3, 4, 5, 6

[34] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting. *arXiv preprint arXiv:2112.01518*, 2021. 2

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. PAMI*, 39(6):1137–1149, 2017. 4, 5

[36] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-Critical Sequence Training for Image Captioning. In *CVPR*, 2017. 1, 2, 3

[37] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *ACL*, 2016. 2

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*, 2018. 4

[39] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 1, 2

[40] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *ECCV*, 2020. 4, 6

[41] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From Show to Tell: A Survey on Deep Learning-based Image Captioning. *IEEE Trans. PAMI*, 2022. 2

[42] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. CLIP4Caption: CLIP for Video Caption. In *ACM Multimedia*, 2021. 2

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3

[44] Jue Wang, Haofan Wang, Jincan Deng, Weijia Wu, and Debing Zhang. EfficientCLIP: Efficient Cross-Modal Pre-training by Ensemble Confident Learning and Language Modeling. *arXiv preprint arXiv:2109.04699*, 2021. 2

[45] Mengmeng Wang, Jiazheng Xing, and Yong Liu. ActionCLIP: A New Paradigm for Video Action Recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[46] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-Driven Referring Image Segmentation. *arXiv preprint arXiv:2111.15174*, 2021. 2

[47] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *ICLR*, 2022. 1, 2, 4, 6

[48] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[49] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-Encoding Scene Graphs for Image Captioning. In *CVPR*, 2019. 1

[50] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained Interactive Language-Image Pre-Training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[51] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring Visual Relationship for Image Captioning. In *ECCV*, 2018. 1

[52] Haoxuan You, Luowei Zhou, Bin Xiao, Noel C Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. MA-CLIP: Towards Modality-Agnostic Contrastive Language-Image Pre-training. 2021. 2

[53] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In *ICLR*, 2020. 3

[54] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 2, 4, 6

[55] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *ICLR*, 2020. 4

[56] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. RST-Net: Captioning With Adaptive Attention on Visual and Non-Visual Words. In *CVPR*, 2021. 2, 6

[57] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-based Language-Image Pretraining. *arXiv preprint arXiv:2112.09106*, 2021. 2