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 relying 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].

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 relying 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].

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.

Our results show that CLIP features provide a signifi-
icant 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 [17,59], 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 pairs 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 $p(w_t|w_{<t}, V)$, where $V$ is an input image and $\{w_t\}$ 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 multimoal 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 $(V, W)$, where $V$ 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 $V = \{v_i\}_{i=1}^N$ are encoded via bi-directional attention in the encoder, while the token embeddings of the caption $W = \{w_i\}_{i=1}^L$ 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:

$$
\text{encoder} \quad \hat{v}_i = \text{MSA}(v_i, V) \\
\text{decoder} \quad O_{w_i} = \text{MSCA}(w_i, \hat{V}, \{w_i\}_{i=1}^L),
$$

where $O$ is the network output, MSA$(x, Y)$ a self-attention with $x$ mapped to query and $Y$ mapped to key-values, and MSCA$(x, Y, Z)$ a self-attention with $x$ as query and $Z$ as key-values, followed by cross-attention with $x$ as query and $Y$ 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.

$$
\mathcal{L} = -\mathbb{E}_{(V, W) \sim \mathcal{D}} \left( \sum_{i=1}^L \log p(O_{w_i} | V, W_{\tau<i}) \right),
$$

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 $\hat{w}_t$ from the output probability distribution. This is then concatenated to previously predicted tokens to form a sequence $\{\hat{w}_\tau\}_{\tau=1}^t$, 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 $V$ 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 $\{\text{CLS}\}$ 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 $5 \times 10^{-6}$.
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-SRef). 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...
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.

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
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-SRef 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 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


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


