Universal Captioner: Long-Tail Vision-and-Language Model Training through Content-Style Separation

Marcella Cornia\textsuperscript{1,}\textsuperscript{*}, Lorenzo Baraldi\textsuperscript{1,}\textsuperscript{*}, Giuseppe Fiameni\textsuperscript{2}, Rita Cucchiara\textsuperscript{1}

\textsuperscript{1}University of Modena and Reggio Emilia, Italy \quad \textsuperscript{2}NVIDIA AI Technology Centre, Italy

\{marcella.cornia,lorenzo.baraldi,rita.cucchiara\}@unimore.it \quad gfiameni@nvidia.com

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{sample_descriptions.png}
\caption{Sample descriptions generated by Universal Captioner. Our approach can describe long-tail concepts while maintaining the style of human-generated captions. Long-tail concepts are highlighted in color.}
\end{figure}

\section*{Abstract}

While captioning models have obtained compelling results in describing natural images, they still do not cover the entire long-tail distribution of real-world concepts. In this paper, we address the task of generating human-like descriptions with in-the-wild concepts by training on web-scale automatically collected datasets. To this end, we propose a model which can exploit noisy image-caption pairs while maintaining the descriptive style of traditional human-annotated datasets like COCO. Our model separates content from style through the usage of keywords and stylistic tokens, employing a single objective of prompt language modeling and being simpler than other recent proposals. Experimentally, our model consistently outperforms existing methods in terms of caption quality and capability of describing long-tail concepts, also in zero-shot settings. According to the CIDEr metric, we obtain a new state of the art on both COCO and nocaps when using external data.

\section{1. Introduction}

Generating textual descriptions from visual inputs is a fundamental step towards the automatic modeling of connections between the visual and textual modalities [33, 53, 68]. As such, the task of image captioning has witnessed considerable attention in recent years. This has led to the development of effective strategies for both feature extraction and cross-modal modeling, and the creation of architectures that can seamlessly describe images belonging to the training set distribution [3, 9, 33].

Despite these advances, captioning architectures still lack the ability to describe a big portion of the distribution of real-world concepts – a fundamental requirement for bringing these models into production. Ideally, in fact, a captioning model should be able to describe any object, person, or scene it encounters, and also give specific names to universally recognizable targets, such as “Harry Potter”
or the “Golden Gate bridge”.

The issue boils down to the limitations of existing human-annotated datasets like COCO [17, 34, 67], which, while being high-quality in terms of descriptive style, are limited in terms of semantic variability and size. On the other side, recent efforts have automatically collected large-scale datasets from the web, partially solving the semantic scale issue at the cost of significantly reducing the quality of the annotations [5, 38, 49, 52].

How to exploit such additional richness in semantics and concepts, without paying the price of lower description quality, is still under-researched. While direct training on such datasets produces captions with low style quality [5, 49], the recently proposed pre-training paradigm [33, 68, 69] has attempted to exploit self-supervised targets on weakly labeled data sources and adopts a fine-tuning phase in which the architecture is adapted to the semantic and style distribution of COCO. As we will show, while this increases the robustness and quality of descriptions in COCO’s semantic domain, it also prevents the architecture from learning long-tail concepts that live outside this domain.

Beyond the pre-training protocol, this can also be partially attributed to the use of object detections as feature descriptors [3, 68]. Detections are, indeed, not scalable to long-tail semantic concepts, and relying on them (sometimes also to describe objects which are not contained in the training set [1]) seems to be a limited choice in light of new multi-modal descriptors [42].

In this paper, we take a step in the direction of a “Universal” captioner that can describe in-the-wild concepts and universally recognizable targets. We propose an architecture that, by design, can separate content from descriptive style and that can be trained on large-scale web datasets while retaining the fluency and style of traditional datasets collected by hand at generation time. To this end, we exploit the extraction of textual keywords from large-scale multi-modal architectures like CLIP [42] and train our networks in a style-aware manner to foster the transfer of semantic concepts between sources. From the point of view of the architecture, we employ an encoder-decoder structure that clearly separates the visual and textual domain, in contrast to the paradigm of employing BERT-like architectures [33, 69]. Finally, our model is trained using only language modeling targets and does not require complex pre-training strategies.

In short, our proposal consists of the following components:

- **Inputs.** The framework employs scalable CNN or ViT-based [12, 42] feature extractors which can directly take raw pixels as input and avoid the need of using object detectors. Further, it extracts textual keywords from input images by running nearest neighbors searches over large-scale cross-modal models. Finally, it employs stylistic tokens as a means to separate hand-collected and web-based image-caption pairs.

- **Architecture.** We employ a fully-attentive encoder-decoder architecture that jointly encodes keywords, style, and text, and which is trained with a single objective of prompt language modeling.

- **Data.** Training is performed on a mixture of hand-collected and web-scale datasets, for a total of 36.4 million image-text pairs, publicly available. In contrast to previous works, we do not employ annotations collected for other tasks rather than image captioning (e.g. VQA [15]). To our knowledge, this is the largest publicly available dataset employed to train an image captioning model.

Fig. 1 reports some qualitative results obtained by our approach, which we name **Universal Captioner**. Experimentally, we assess the performance of the proposed approach both in a traditional, in-domain, captioning setting [1, 34] and on the description of long-tail visual concepts. Our model consistently outperforms existing proposals in terms of caption quality and for its capabilities of describing long-tail concepts, surpassing models trained on private and significantly larger datasets [59]. Overall, our work demonstrates that web-scale datasets can be properly used to increase the performance of image captioning systems without requiring complex pre-training pipelines and takes a step forward in the direction of real-world applications.

### 2. Related Work

**Visual Encoders.** Research on image captioning has jointly focused on modeling the visual encoding pipeline, the language model, and the multi-modal connections between them [53]. In terms of visual encoding, after the emergence of global [24, 46, 58] and grid-based descriptors [36, 61], the use of object detections [3, 68] has become one of the most popular approaches. Indeed, it provides clean visual elements and a partial bridge between the visual and the textual domains. While several works have encoded regions through additive attention [22, 25, 41], graph-based encodings [16, 63, 65], or self-attentive structures [9, 21, 32, 39], the emergence of self-attentive visual encoders [12] and large-scale multi-modal models like CLIP [42] is reopening the discussion on which feature model is most appropriate for image captioning, with strategies ranging from training better detectors to having end-to-end visual models trained from scratch [26, 60, 62, 68]. As shown in Shen et al. [50], features encoded by large-scale multi-modal architectures perform at least on par with detection-based approaches.

**V&I Pre-training.** While traditional captioning approaches have focused on training on curated datasets only, the recently emerged pre-training paradigm [6, 31, 35, 54, 55, 69] tries to learn from weakly labeled sources. Most of the approaches have employed BERT-like [11] architectures...
to learn cross-modal representations. The OSCAR [33] model, for instance, considers triplets of object tags, detections, and captions and trains using a combination of masked token loss and contrastive loss. VinVL [68] employs the same objectives while proposing a better object detector, and trains on 8.85 million text-image pairs. Like other works that employ encoder-decoder structures [60, 62], SimVLM [59] neglects the usage of BERT-like structures and learns visual features from scratch, training on the large, not publicly available, ALIGN [23]. Our work, in contrast, employs publicly-available data only.

**Scaling to out-of-domain concepts.** One of the most relevant goals in image captioning is that of extending the number of concepts that can be described. Traditionally, this has been addressed through the Novel Object Captioning protocol [1,2,64], which requires the captioner to name objects unseen in training captions, relying on object detection tags [20] or uni-modal pre-training [18]. In contrast to this approach, in this work, we foster the description of long-tail concepts while training on low-quality image-caption pairs.

### 3. Universal Captioner

Machine-collected caption sources like Conceptual Captions [49] or YFCC100M [56] and human-collected sources like COCO [34] or Flickr30k [67] have very different descriptive styles; the former features a correct and natural descriptive style, while the latter show higher semantic coverage, as shown in Fig. 2.

Our goal is to build a captioning model that can benefit from multi-source image-caption pairs and reproduce the descriptive style of one source at generation time. Formally, given a dataset of image-caption pairs $D = \{(v_i, t_i)\}_i$, partitioned into two subsets $D_m$ and $D_h$, respectively containing machine-collected and human-collected pairs, our model aims at learning the following probability distribution:

$$p(w_t|w_{\tau < t}, v, s),$$

where $v$ is an input image which can belong to the distribution of both $D_h$ and $D_m$, $s \in \{0, 1\}$ is a boolean variable that indicates whether $v$ should be described according to the textual style of $D_h$ or $D_m$, and $\{w_i\}_i$ is the sequence of words comprising the generated caption.

To accomplish this goal, the model should be capable of transferring descriptive styles seen during training regardless of the content of the input image. In the case where $D_m$ and $D_h$ are also different in terms of semantic content (e.g. objects and scenes), the model will also have to transfer semantic words to correctly describe the images in $D_m$ with the style of $D_h$, or vice versa.

With the ultimate goal of transferring the semantic knowledge of machine-collected datasets to the descriptive domain of human-collected datasets, we propose a new model with two strategies to promote the separation and transfer of content and style. First, we extract **textual keywords**, which can represent the content of an image in textual form regardless of the descriptive style of its associated caption; then, we employ a **stylistic token** to condition the language model with different descriptive styles.

### 3.1. Web-scale separation of content and style

**Extracting textual keywords.** The use of textual tags coming from an object detector [3,68] is a popular choice to link the visual and textual domains. In our case, extracting a condensed textual representation of the visual input aims to promote an objective transfer between visual and textual features, regardless of whether the image comes from the distribution of $D_h$ or $D_m$. At the same time, given the semantic breadth of $D_m$, it is crucial to scale beyond the limitation of object detection classes [28].

To extract keywords, we exploit a multi-modal retrieval approach trained on large-scale data, CLIP [42], which is well known for its scalability in terms of the number of concepts. Given a dictionary of possible keywords $Q$, the set of keywords for an input image $v$ is obtained by selecting the $k$ elements in $Q$ with the highest similarity to $v$, according to the matching function defined by the retrieval model itself. Although CLIP’s encoder was trained on machine-collected sentences, we found it to work well enough also in our case, in which it is fed with only one unigram.

To ensure a sufficient semantic coverage with respect to the semantic distribution of $D$, $Q$ must be large enough. In our case, we construct $Q$ with the set of all unigrams found in the OpenWebText corpus, a publicly available clone of OpenAI’s WebText dataset [43] and which contains around 38GB of text data

1. Compared to the tags extracted from oversampled regions [3,68], the keywords we extract do not refer to local regions of the input image, but rather correspond to the

---

1 Additional details about the construction of the dictionary are reported in the supplementary material.
image as a whole and have an increased richness in semantics, which can also contain universally recognizable proper nouns and concepts.

**Stylistic tokens.** As a second strategy to separate content from style, we give the language model full awareness of the descriptive style to which every training caption belongs. In addition to textual keywords, given an (image, caption) pair we also extract a stylistic token $s$ that indicates whether $t$ belongs to the descriptive style of $D_h$ or $D_m$. This is implemented with a learnable token embedding, which can be concatenated to the representation of the keywords.

### 3.2. Architecture

Universal Captioner represents each training image-caption pair as a quadruple of image, text, keywords and style $(v, t, k, s)$, where $v$ is encoded with a set of fixed-length visual descriptors (e.g., with activations from a CNN or a Visual Transformer).

The overall network (depicted in Fig. 3) follows an encoder-decoder Transformer [57] architecture, where visual features $v$ are encoded via bi-directional attention in the encoder, while the decoder performs an autoregressive factorization on the sequence of $[k; s; t]$.

Unlike a traditional decoder, however, the network is only trained to predict a left-shifted version of the caption $t$, while the sequence $[k; s]$ is treated as a prompt. In contrast to prompting in pre-trained language models [13,43], though, in our case, this prompting strategy is employed also while training the network, and without employing pre-trained weights. Finally, the network is trained by following an unidirectional language modeling loss based on cross-entropy, i.e.

$$\mathcal{L} = -\mathbb{E}_{x \sim \mathcal{D}} \log p(w_t|w_{t-1}, v, k, s),$$

where $\{w_t\}_t$ refers to the sequence of words in the ground-truth caption $t$. In the training stage, pairs of hand-collected images and captions are given the human-collected stylistic token. To allow content transfer between the two types of data sources, while maintaining a clear separation between styles, we randomly select samples in each mini-batch to have approximately $10\%$ of human-collected image-caption pairs.

### 3.3. Training corpus

We train on a mixture of publicly available datasets with image-caption pairs. In addition to COCO [34], which is employed as the sole source of manually collected captions, we use the full Conceptual Captions 3M [49] and Conceptual Captions 12M [5] datasets. Both were obtained by cleaning Image Alt-Text pairs from the web; after downloading, our snapshot contains a total of around 15.7 million pairs. In addition, we employ the Wikipedia-based Image Text (WIT) dataset [52], which provides images extracted from Wikipedia together with various metadata extracted from their corresponding pages such as section titles, alt texts, and reference descriptions. In this case, we filter all pages from the English Wikipedia and use the caption of each image as text source. After cleaning, we get about 5.3M pairs. Finally, we also add a subset of YFCC100M [56] with image descriptions and containing around 14.8M pairs.

Compared to previous works that adopted large-scale training data for Vision and Language tasks, our corpus has three key features: (i) differently from the datasets employed in concurrent works like OSCAR [33] and VinVL [68], it contains only data for the image-captioning task, thus neglecting the use of data from ancillary tasks such as VQA or GQA; (ii) it is made of publicly available data, thus allowing reproducibility, unlike proposals such as ALIGN [23,59]; (iii) with a total of 36.4 million image-text pairs, it is the largest publicly available training corpus used for image captioning.

---

2In our preliminary experiments we verified that such a percentage is enough to enable a smooth content transfer between different data sources.
3.4. Training protocol

The pre-training literature [33, 68] tends to employ three stages of training, i.e., an initial pre-training phase with self-supervised targets followed by a fine-tuning stage on COCO, and finally self-critical sequence training (SCST) [46]. Our method, in contrast, employs a two-stage training protocol in which we first train using prompt-language modeling on the entire corpus, and then fine-tune using SCST on COCO to increase the transfer in style.

Compared to the aforementioned pre-training scheme, this not only has the advantage of being simpler, but also reduces the forgetting of concepts that do not belong to COCO and provides an additional opportunity to align with the target descriptive style. During SCST fine-tuning, we employ the variant proposed in [9] that sets the baseline reward equal to the mean of rewards of generated captions inside a beam.

4. Experimental Evaluation

We conduct extensive experiments to validate the architectural choices behind Universal Captioner, and compare its performances with state-of-the-art solutions in both traditional image captioning and long-tail image description.

4.1. Implementation details

We devise three model configurations varying the number of decoding layers $L$, the model dimensionality $d$, and the number of attention heads $H$, taking loose inspiration from those used for BERT [11]: Tiny ($L = 3$, $d = 384$, $H = 6$, 52M params), Small ($L = 6$, $d = 512$, $H = 8$, 87M params), and Base ($L = 12$, $d = 768$, $H = 12$, 213M params). Regardless of the model configuration, we always employ three layers in the visual encoder and 5 textual keypoints. To represent words, we use lower-cased Byte Pair Encoding (BPE) [48] with a 49,152 vocab size and linearly project them to the input dimensionality of the model $d$. We employ classic sinusoidal positional encodings [57] to represent word positions. For computational efficiency, the max sequence length is capped at 80.

Training is performed using the LAMB optimizer [66] and following the learning rate scheduling strategy of [57] with a warmup equal to 6,000 iterations and multiplying the resulting learning rate by a factor of 5. We use a minibatch size of 1,080. ZeRo memory offloading [44] and mixed-precision [37] were used to accelerate training and save memory. During SCST fine-tuning, we use the Adam optimizer [27] and a fixed learning rate of $5 \times 10^{-6}$.

4.2. Visual features evaluation

We start by discussing the choice of the visual features which are fed to the encoder part of the model and compare traditional detection-based features with grid-based features extracted from modern uni-modal and multi-modal models. In particular, we consider object detection features extracted from a Faster R-CNN model [3] pre-trained on the Visual Genome dataset [28], and grid-based features extracted from Vision Transformers pre-trained for classification [12] or through self-supervision as in the DINO model [4]. Then, we consider visual encoders trained in a multi-modal fashion with language supervision [42].

Table 1 reports the results obtained when training our model on COCO only and without employing keywords and stylistic tokens. Given the limited amount of data, we employ the “Tiny” configuration. On nocaps, we do not use constrained beam search and we do not condition the language model on external information. Results are reported in terms of the standard set of captioning metrics [53].

As it can be seen, while the performance of a self-supervised model like DINO is still below that of detection-level Faster R-CNN features, these can be surpassed by modern ViT-like architectures, especially when using a small patch size. For instance, ViT-B16 [12] achieves 126.2 CIDEr points, 1.4 points higher than Faster R-CNN. Employing multi-modal training brings a further and significant advantage (CLIP-ViT-B16, for instance, achieves 6.5 additional CIDEr points). Among CLIP variants, while ViT-based ones achieve better performances than CLIP-RN50 and CLIP-RN101, the best performance is reached by the variants which employ an EfficientNet-style architecture scaling, i.e., CLIP-RN50×4 and CLIP-RN50×16. Overall, this brings a relative improvement of up to 9.9% with respect to traditional detection-based features.

4.3. Ablation study

We then assess the role of the key ingredients of Universal Captioner, i.e., training on web-scale data and the usage of keywords and stylistic tokens. Table 2 reports the performance of our “Small” and “Base” models when trained without web-scale data (i.e., on COCO only) and on the full
employing a NOC protocol with more data than COCO, we underline that we are not COCO-like descriptions. Since our model has been trained that are out-of-domain with respect to COCO while having provides an ideal test-bed for our setting, as it contains images part of COCO, we also test on nocaps. Indeed, nocaps performance on visual and textual concepts which are not same model is used for keywords retrieval. To evaluate the corpus without keywords or stylistic tokens. All models are trained using CLIP-RN50×16 as visual features, and the same model is used for keywords retrieval. To evaluate the performance on visual and textual concepts which are not part of COCO, we also test on nocaps. Indeed, nocaps provides an ideal test-bed for our setting, as it contains images that are out-of-domain with respect to COCO while having COCO-like descriptions. Since our model has been trained with more data than COCO, we underline that we are not employing a NOC protocol.

As it can be noticed from Table 2, using web-scale training without keywords and style on the “Small” configuration provides a non-negligible boost on COCO (2.7 CIDEr points), but significantly reduces the performances on out-of-domain images (from 87.4 to 72.6 CIDEr on nocaps). We observed that in this configuration the model is not learning to transfer descriptive style on out-of-domain images. Adding keywords, instead, increases the performances on nocaps by 10.7 CIDEr points. Employing both keywords and stylistic tokens, finally, brings a significant improvement both on COCO and nocaps images. Our model, in fact, achieves 143.0 CIDEr points on COCO and 110.5 CIDEr points on nocaps, amounting to a relative improvement of 3.7% on the COCO distribution and 26.4% on the nocaps distribution, with respect to a model trained on COCO only and without keywords and stylistic tokens. The same scenario is observable when moving to the “Base” configuration of Universal Captioner: also in this case, the combination of web-scale training, keywords, and stylistic tokens significantly boosts the performances according to all metrics. Our full model, compared to a base model employing only web-scale training, achieves 143.4 CIDEr points (1.34% relative improvement) on COCO and 114.5 CIDEr points on nocaps (37.0% relative improvement).

### 4.4. Comparison with the state of the art

We compare our full model, trained on web-scale data, with recent models trained on COCO only, i.e. Up-}

---

3In Agrawa et al. [1], the protocol we employ is referred to as nocaps XD.

---

### Table 2: Ablation study on the COCO Karpathy-test split and nocaps validation split, using the “Small” (top) and “Base” (bottom) configurations.

|                | COCO   | nocaps |
|----------------|--------|--------|
|                | B-4    | M      | R      | C      | S      | C      | S      |
| w/o web-scale training | 40.1  | 29.6  | 59.4  | 137.9 | 23.2  | 87.4  | 12.2  |
| w/o keywords and style tks | 40.7  | 30.0  | 59.7  | 140.6 | 23.7  | 72.6  | 11.5  |
| w/o stylistic tokens | 40.8  | 30.0  | 59.8  | 141.9 | 23.9  | 83.3  | 12.0  |
| UniversalCap\textsuperscript{small} | 41.2  | 30.4  | 60.2  | 143.0 | 24.2  | 110.5 | 13.8  |
| w/o keywords and style tks | 40.2  | 30.1  | 59.5  | 141.5 | 24.0  | 83.6  | 12.3  |
| w/o stylistic tokens | 40.7  | 30.2  | 59.9  | 142.7 | 23.9  | 85.3  | 12.4  |
| UniversalCap\textsuperscript{base} | 40.8  | 30.4  | 60.2  | 143.4 | 24.2  | 114.5 | 14.1  |

---

### Table 3: Comparison with the state of the art COCO Karpathy-test split in a single model setting. Our model is trained using the full Universal Captioner corpus. Top-1 results are highlighted in black, top-2 in blue.

|                | B-4    | M      | R      | C      | S      | C      | S      |
|----------------|--------|--------|--------|--------|--------|--------|--------|
| A\textsuperscript{4} Transformer\textsuperscript{[9]} | 39.1  | 29.2  | 58.6  | 131.2 | 22.1  | 7.6   | 12.6  |
| X-Transformer\textsuperscript{[39]} | 39.7  | 29.5  | 59.1  | 132.8 | 23.4  |
| AutoCaption\textsuperscript{[70]} | 40.2  | 29.9  | 59.5  | 135.8 | 23.8  |
| OSCAR\textsuperscript{base}\textsuperscript{[33]} | 40.5  | 29.7  | -     | 137.6 | 22.8  |
| VinVL\textsuperscript{base}\textsuperscript{[68]} | 40.9  | 30.9  | -     | 140.6 | 25.1  |
| SimVLM\textsuperscript{base}\textsuperscript{[59]} | 39.5  | 32.2  | -     | 134.4 | 24.0  |
| UniversalCap\textsuperscript{base}\textsuperscript{[65]} | 40.8  | 29.9  | 59.9  | 140.4 | 23.4  |
| UniversalCap\textsuperscript{small}\textsuperscript{[65]} | 41.2  | 30.4  | 60.2  | 143.0 | 24.1  |
| UniversalCap\textsuperscript{base}\textsuperscript{[65]} | 40.8  | 30.4  | 60.2  | 143.4 | 24.2  |
| SimVLM\textsuperscript{base}\textsuperscript{[59]} | 40.5  | 33.7  | -     | 143.3 | 25.4  |

---

### Table 4: Zero-shot performances on the VizWiz\textsuperscript{[17]} test split and the TextCaps\textsuperscript{[51]} validation split.

|                | B-4    | M      | R      | C      | S      | C      | S      |
|----------------|--------|--------|--------|--------|--------|--------|--------|
| Up-Down\textsuperscript{[3]} | ✔     | 11.3  | 18.9  | 5.8   | 12.4  | 24.2  | 8.7   |
| AoANet\textsuperscript{[21]} | ✔     | 13.2  | 19.4  | 6.2   | 18.1  | 32.3  | 11.2  |
| Up-Down\textsuperscript{[3]} | ✗     | 19.8  | 49.7  | 12.2  | 20.1  | 41.9  | 11.7  |
| AoANet\textsuperscript{[21]} | ✗     | 23.2  | 60.5  | 14.0  | 20.4  | 42.7  | 13.2  |
| VinVL\textsuperscript{base}\textsuperscript{[68]} | ✔     | 16.9  | 34.7  | 9.9   | 17.3  | 41.2  | 13.1  |
| VinVL\textsuperscript{base}\textsuperscript{[68]} | ✔     | 17.4  | 37.7  | 10.3  | 17.5  | 41.9  | 13.1  |
| UniversalCap\textsuperscript{large}\textsuperscript{[68]} | ✔     | 21.1  | 57.3  | 13.1  | 19.8  | 58.7  | 14.0  |
| UniversalCap\textsuperscript{small}\textsuperscript{[68]} | ✔     | 22.2  | 62.0  | 14.2  | 20.6  | 62.7  | 14.6  |
| UniversalCap\textsuperscript{base}\textsuperscript{[68]} | ✔     | 22.8  | 65.7  | 14.4  | 21.0  | 66.4  | 14.8  |

---

4.4. Comparison with the state of the art

We compare our full model, trained on web-scale data, with recent models trained on COCO only, i.e. Up-
Zero-shot performances on VizWiz and TextCaps. Beyond having good performances on in-domain captioning, the main benefit of scaling to web-scale data is the potential of zero-shot generalization. Therefore, we investigate whether the caption capabilities of our model transfer to other datasets in a zero-shot manner, by directly decoding from our pre-trained model.

We consider the VizWiz dataset [17], which contains images originating from blind people, and the TextCaps dataset [51], with images containing text. Both of them represent distinct visual and semantic distributions, clearly separated from the COCO ones. Table 4 shows a comparison when using Up-Down [3] and AoANet [21] in a zero-shot manner training on COCO only, and when employing the same models finetuned on the aforementioned datasets. We also compare with the “Base” and “Large” configurations of VinVL [68], which has been pre-trained on external data. All models are fine-tuned with SCST.

As shown, Universal Captioner consistently outperforms the performances of traditional approaches like Up-Down [3] and AoANet [21] when evaluated in a zero-shot setting. Interestingly, it also overcomes the previously mentioned approaches when they are trained or finetuned on VizWiz and TextCaps, confirming that the model is capable of properly transferring semantic concepts learned from web-scale data. Finally, Universal Captioner also beats the performances of VinVL in both configurations by a significant margin on both datasets, showing that our training and content-style separation strategies can overcome state-of-the-art pre-training paradigms.

Performances on nocaps-XD. As a second zero-shot test, we evaluate the capabilities of our model on the nocaps dataset [1]. Also in this case we employ nocaps as a pure test-bed. For fair comparison, we compare with VinVL [68] when trained on web-scale data and fine-tuned with SCST, thus using the same experimental setting.

Table 5 reports the results for both the validation and test sets. The upper section of the table refers to a pure NOC evaluation protocol, in which training is performed exclusively on COCO. The rest of the table, instead, refers to our setting in which external data is used at training time.

As can be seen, Universal Captioner can overcome the performances of VinVL when describing concepts from nocaps, especially on its “out-of-domain” portion. Further, our models achieve comparable performance using way fewer parameters than competitors. UniversalCap\textsuperscript{tiny}, for instance, achieves 105.8 CIDEr points on the validation set, compared to 103.1 CIDEr points achieved by SimVLM\textsuperscript{base}, while UniversalCap\textsuperscript{base} achieves 114.5 CIDEr points, compared to 112.2 and 105.1 achieved, respectively, by SimVLM\textsuperscript{base} and VinVL\textsuperscript{large}. This suggests that previous pre-training strategies, although employed on web-scale data, have been ineffective in learning to describe long-tail concepts that do not belong to COCO. Noticeably, the performances on the out-of-domain portion of nocaps increase as the dimensionality of the model increases, from 106.1 to 116.0 when comparing UniversalCap\textsuperscript{tiny} to UniversalCap\textsuperscript{base}, with a relative improvement of 9.3%.

The same behavior is observable also on the nocaps test set, where Universal Captioner consistently outperforms VinVL when describing the out-of-domain portion of the dataset. Finally, we notice that Universal Captioner can outperform SimVLM in this setting as well, despite being trained on significantly fewer data. For instance, in the out-of-domain portion of the dataset, Universal Captioner achieves 110.4 CIDEr points, compared to 109.5 CIDEr points achieved by SimVLM\textsuperscript{base}.

Table 5. Results on the nocaps dataset. † refers to a NOC protocol, the rest to a nocaps-XD one.

![Example captions from Universal Captioner and VinVL](image)

Figure 4. Sample descriptions generated on nocaps images.
4.5. Long-tail description

Beyond evaluating on datasets that contain COCO-like captions, we also assess the capability of our approach to name long-tail concepts in image collections that do not contain curated captions. To this aim, we conduct analyses on the validation sets of three datasets with large variety of visual concepts, i.e., Open Images V6 [30] (subset with Bounding Boxes), ImageNet-21K [10, 47], and CC3M [49].

We evaluate the compatibility between the image and the generated caption using the CLIP-Score metric [19] which is based on CLIP embeddings and does not require ground-truth captions while achieving high correlation with human judgments. Further, we employ a coverage-based metric inspired from [7, 8] between ground-truth detections or class annotations and the predicted caption. Formally, given the set of classes in the annotation and the set of words in the predicted caption, we compute the optimal assignment between them with the Hungarian algorithm [29], using GloVe embeddings [40]. Then, we compute the coverage score as the ratio between the Hungarian score and the cardinality of the set of annotations. To measure the extent and quality of the vocabulary employed by the captioner, we also count the number of unique words which do not appear in COCO at least 5 times, and the number of unique proper nouns. We do not employ standard captioning metrics when evaluating on CC3M, as these are also influenced by descriptive style.

Table 6 shows the results, comparing Universal Captioner with VinVL [68]. As it can be seen, our approach outperforms VinVL when describing images from all three datasets. For instance, on Open Images it achieves 0.739 in terms of CLIP-Score (with a 3.36% relative improvement) while having a comparable coverage score. Further, it is capable of naming significantly more words that are outside of COCO (1,071 unique out-of-COCO words, compared to 186 named by VinVL). On the same line, Universal Captioner achieves a 0.749 CLIP-Score on the validation set of CC3M (with a 5.34% relative improvement), and again employs more out-of-COCO words. Universal Captioner consistently generates more proper nouns than VinVL, which suggests an improved degree of specificity.

4.6. Qualitative results

We showcase the capabilities of Universal Captioner to name long-tail concepts and name universally recognizable objects and people through some qualitative examples. In Fig. 1, on the first page, we compare with the baseline Tiny model trained with CLIP-RN50×16 and without web-scale data and style-content separation (reported in Table 1). We observe that our approach correctly recognizes and names famous people, places, and trademarks like the Taj Mahal, Marilyn Monroe, or the Facebook logo.

The same can be observed in Fig. 4 and 5, where we compare with VinVL\textsuperscript{Large} [68] on images taken from no-caps, OpenImages, and CC3M. Again, our model can recognize and name long-tail concepts better than previous approaches, also recognizing famous people. We refer the reader to the supplementary for more qualitative samples and for a discussion on limitations and societal impact.

5. Conclusion

We proposed Universal Captioner, an approach for describing images with long-tail concepts while training on web-scale noisy data. Our approach relies on textual keywords, stylistic tokens and employs a prompt-based language modeling objective. Experimentally, it achieves state-of-the-art results in both in-domain image captioning and when tested to name in-the-wild concepts, and demonstrates to be more specific during description.

Table 6. Long-tail description performances on the validation splits of Open Images, ImageNet-21K and Conceptual Capsions 3M.
Acknowledgment

We thank CINECA, the Italian Supercomputing Center, for providing computational resources. This work has been supported by “Fondazione di Modena”, by the “Artificial Intelligence for Cultural Heritage (AI4CH)” project, co-funded by the Italian Ministry of Foreign Affairs and International Cooperation, and by the H2020 ICT-48-2020 HumanE-AI-NET and H2020 MSCA “PERSEO - European Training Network on PErsonalized Robotics as SErvice Oriented applications” projects.

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. 2, 3, 6, 7, 13
[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided Open Vocabulary Image Captioning with Constrained Beam Search. In EMNLP, 2017. 3
[3] 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, 5, 6, 7
[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. 5
[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In CVPR, 2021. 2, 4
[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Learning UNiversal Image-TExt Representations. 2019. 2
[7] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In CVPR, 2019. 8
[8] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. SMArT: Training Shallow Memory-aware Transformers for Robotic Explainability. In ICRA, 2020. 8
[9] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In CVPR, 2020. 1, 2, 5, 6, 12
[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In CVPR, 2009. 8
[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL, 2018. 2, 5
[12] 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. 2, 5
[13] Tianyu Gao, Adam Fisch, and Danqi Chen. Making Pre-trained Language Models Better Few-shot Learners. In ACL, 2021. 4
[14] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 12
[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In CVPR, 2017. 2
[16] Longteng Guo, Jing Liu, Jinjue Tang, Jiangwei Li, Wei Luo, and Hangning Lu. Aligning Linguistic Words and Visual Semantic Units for Image Captioning. In ACM Multimedia, 2019. 2
[17] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning Images Taken by People Who Are Blind. In ECCV, 2020. 2, 6, 7, 12
[18] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data. In CVPR, 2016. 3
[19] 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. 8
[20] Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. VIVO: Visual Vocabulary Pre-Training for Novel Object Captioning. AAAI, 2020. 3, 7
[21] Lun Huang, Wemnin Wang, Jie Chen, and Xiao-Yong Wei. Attention on Attention for Image Captioning. In ICCV, 2019. 2, 6, 7
[22] Lun Huang, Wemnin Wang, Yaxian Xia, and Jie Chen. Adaptively Aligned Image Captioning via Adaptive Attention Time. In NeurIPS, 2019. 2
[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunchuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In ICML, 2021. 3, 4
[24] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015. 2
[25] Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. Reflective Decoding Network for Image Captioning. In ICCV, 2019. 2
[26] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In ICML, 2021. 2
[27] Diederek P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In ICLR, 2015. 5
[28] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. IJCV, 123(1):32–73, 2017. 3, 5
[61] 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

[62] Ming Yan, Haiyang Xu, Chenliang Li, Bin Bi, Junfeng Tian, Min Gui, and Wei Wang. Grid-VLP: Revisiting Grid Features for Vision-Language Pre-training. *arXiv preprint arXiv:2108.09479*, 2021. 2, 3

[63] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-Encoding Scene Graphs for Image Captioning. In *CVPR*, 2019. 2

[64] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects. In *CVPR*, 2017. 3

[65] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring Visual Relationship for Image Captioning. In *ECCV*, 2018. 2

[66] 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. 5

[67] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 2, 3

[68] 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. 1, 2, 3, 4, 5, 6, 7, 8

[69] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*, 2020. 2

[70] Xinxin Zhu, Weining Wang, Longteng Guo, and Jing Liu. AutoCaption: Image Captioning with Neural Architecture Search. *arXiv preprint arXiv:2012.09742*, 2020. 6
A. Supplementary Material

In the following, we present additional material about Universal Captioner. In particular, we provide additional training and implementation details, further experimental results, and qualitative samples. Finally, we discuss limitations and societal impact.

A.1. Additional Implementation Details

Construction of the keyword vocabulary. As mentioned in the main paper, the keyword dictionary is composed by extracting unigrams from the OpenWebText corpus\(^4\). During pre-processing, all unigrams are lowered and first names are removed. We only keep the 10,000 most frequent unigrams. Fig. 6 reports keywords extracted from sample images of the COCO dataset. As it can be seen, although keywords can sometimes be repetitive, they provide a significant conditioning to the language model.

Additional architectural details. Following recent literature [9], we enrich all layers of our encoder with memory slots. Specifically, we extend the set of keys and values of each self-attention operation with 40 additional learnable vectors, which are independent of the input sequence and can encode a priori information retrieved through attention.

Weight initialization. We initialize all weights by drawing inspiration from GPT-2 [43]. All linear and embedding weights are initialized according to a uniform distribution and using the approach proposed by Glorot et al. [14]. Layer normalization weights are initialized to a constant value of 1. All biases are initialized to 0. We also employ the “Special Scaled Initialization”\(^5\) when initializing the output linear projection of each Transformer layer.

Proper nouns. To identify and count the proper nouns predicted by Universal Captioner and reported in Table 6, we employ the spaCy NLP toolkit\(^6\).

A.2. Additional Results

Ablation study on VizWiz and TextCaps. In Table 4, we showed how Universal Captioner can outperform existing approaches when employed with a zero-shot protocol on VizWiz [17] and TextCaps [51]. Here, we also perform an ablation experiment on the same datasets to confirm the role of keywords and stylistic tokens under this setting. Table 7 shows the results of our “Small” and “Base” models when trained without keywords and stylistic tokens, and when trained only without stylistic tokens. All models are trained using CLIP-RN50×16 as visual features. As it can be observed, employing web-scale training alone and

Table 7. Zero-shot performances on the VizWiz [17] test split and the TextCaps [51] validation split.

|                      | VizWiz | TextCaps |
|----------------------|--------|----------|
|                      | B-4   | C | S  | B-4   | C | S  |
| w/o keywords and style tks | ✓    | 14.3  | 44.2 | 11.4 | 12.7 | 47.5 | 10.4 |
| w/o stylistic tokens | ✓    | 15.7  | 47.3 | 11.6 | 15.5 | 52.8 | 11.5 |
| UniversalCap\(^\text{small}\) | ✓    | 22.2  | 62.0 | 14.2 | 20.6 | 62.7 | 14.6 |
| w/o keywords and style tks | ✓    | 16.7  | 51.6 | 12.6 | 15.2 | 56.7 | 11.6 |
| w/o stylistic tokens | ✓    | 17.3  | 52.3 | 12.5 | 15.5 | 57.9 | 11.8 |
| UniversalCap\(^\text{base}\) | ✓    | 22.8  | 65.7 | 14.4 | 21.0 | 66.4 | 14.8 |

an encoder-decoder model already provides a significant boost over VinVL\(^\text{large}\) on both datasets (37.7 vs 51.6 CIDEr points on VizWiz and 41.9 vs 56.7 on TextCaps). Introducing keywords brings a significant advantage (3.1 CIDEr points on VizWiz and 5.3 points on TextCaps, when considering UniversalCap\(^\text{small}\)), while stylistic tokens further help to align with COCO-style captions, and provide an additional increase in caption quality. We notice that the role of keywords is less evident when increasing the model size and moving to UniversalCap\(^\text{base}\), while stylistic tokens provide a fundamental addition in both cases. Overall, this experiment confirms the appropriateness of the main ingredients of Universal Captioner in zero-shot settings.

Ablation study on nocaps-XD. To further complement the ablation studies presented above and in the main paper, in Table 8 we report the full set of results obtained when conducting the ablation on nocaps. Interestingly, the role of keywords and stylistic tokens is verified on both out-of-domain and in-domain settings, and in both the validation and test sets. Again, we notice how both ingredients contribute to the final result in both UniversalCap\(^\text{small}\) and UniversalCap\(^\text{base}\), and how keywords bring a larger improvement in a small-scale setting.

Qualitative description results. Finally, we report different qualitative results obtained on images from VizWiz (Fig. 7), TextCaps (Fig. 8), nocaps (Fig. 9), Open Images and CC3M (Fig. 10). We observe how Universal Captioner can describe and mention objects, people, and scenes with a significantly increased level of detail when compared to the current state of the art and regardless of the dataset. Also, Universal Captioner qualitatively appears to be less prone to hallucination. From the reported samples it can be observed how Universal Captioner can name people, long-tail concepts, brands, and recognize logos, and it is also noticeable that it tends to read the text in images and employs the same text in the output description. We hypothesize that this is due to the capabilities of CLIP-based features, as well as to the role of keywords, which act as a conditioning element for the generative language model. Overall this provides more specific descriptions, which can be more useful in practical scenarios, ranging from assisting visually impaired people to empowering retrieval systems.

\(^4\)https://skylion007.github.io/OpenWebTextCorpus

\(^5\)A reference implementation can be found in https://huggingface.co/transformers/_modules/transformers/models/gpt2/modeling_gpt2.html

\(^6\)https://spacy.io/
A.3. Discussion

Limitations. Universal Captioner shows that better image captioning models can be built by exploiting web-scale data and proper stylistic alignment techniques. While we believe that the results showcased by Universal Captioner are of potential interest to the community, we also identify a series of limitations that might be addressed in future works. Firstly, building a large collection of noisy datasets has clearly been a crucial element to obtain long-tail descriptions. Still, the composition of this mixture of datasets has not been experimentally assessed. Investigating the contribution of each of the components will be important to better understand the role of each of them in the mixture, and potentially build better mixtures of datasets. Further, while we have made an effort to include recently collected datasets like Wikipedia-based Image Text, we expect new automatically-collected datasets to be released in the future, given the availability of web-scale data. Scaling to new datasets and further increasing the scale of data used at training time will be interesting to evaluate the robustness of our choices.

A second element that has proved to be crucial for the final performance is the usage of stylistic tokens, which aim at clearly separating the style of automatically collected captions from that of manually collected ones. This has been implemented with a single learnable token, which does not encode the distinction between different automatically-collected datasets. As future work, the literature might need to explore alternative techniques for providing the same conditioning to the language model.

Potential negative societal impact. Universal Captioner provides a captioning algorithm that exploits data collected from the web. In addition to the societal impacts that any captioning system might have, Universal Captioner heavily relies on noisy data which has not been manually assessed or cleaned. This might increase the level of social biases that the model can learn, in comparison with previous approaches. An investigation and assessment of such biases are left for future work.

Table 8. Ablation study on the nocaps dataset [1].
Figure 7. Sample descriptions generated on images from the VizWiz dataset.

Figure 8. Sample descriptions generated on images from the TextCaps dataset.
Figure 9. Sample descriptions generated on nocaps images. Long-tail concepts are highlighted in color.
Figure 10. Sample descriptions generated on images from Open Images and CC3M. Long-tail concepts are highlighted in color.