SynthCap: Augmenting Transformers with Synthetic Data for Image Captioning

Davide Caffagni[1 [0009−0002−3279−6480]], Manuele Barraco[1 [0000−0001−6075−2309]], Marcella Cornia[1 [0000−0001−9640−9385]], Lorenzo Baraldi[1 [0000−0001−5125−4957]], and Rita Cucchiara[1,2 [0000−0002−2239−283X]]

[1] University of Modena and Reggio Emilia, Modena, Italy
{name.surname}@unimore.it
[2] IIT-CNR, Pisa, Italy

Abstract. Image captioning is a challenging task that combines Computer Vision and Natural Language Processing to generate descriptive and accurate textual descriptions for input images. Research efforts in this field mainly focus on developing novel architectural components to extend image captioning models and using large-scale image-text datasets crawled from the web to boost final performance. In this work, we explore an alternative to web-crawled data and augment the training dataset with synthetic images generated by a latent diffusion model. In particular, we propose a simple yet effective synthetic data augmentation framework that is capable of significantly improving the quality of captions generated by a standard Transformer-based model, leading to competitive results on the COCO dataset.

Keywords: Image Captioning · Synthetic Data · Vision-and-Language.

1 Introduction

Image captioning is a complex task that involves the description of an image in natural language, posing challenges at the intersection of Computer Vision and Natural Language Processing fields. The most promising solutions to tackle the task are represented by deep learning-based captioning architectures which have become the de facto standard for the task [46]. Despite achieving state-of-the-art results, it is becoming difficult to further improve their performance, primarily because of the struggles in finding datasets containing a satisfactory amount of image-caption pairs. To overcome this issue, the predominant approach in the field is to train captioning networks [13,20,51,58] on large-scale datasets collected from the web [42, 44], usually downloading an image along with the description provided in its "alt" tag. As a matter of fact, there is no surprise in witnessing more and more advanced deep learning-based models being trained on web-collected data, especially after the spread of large-scale language models [10, 59] and cross-modal architectures [36]. The knowledge found on the web, indeed, excels for size and variety, stimulating the robustness and sensibility of deep learning models to long-tail concepts. However, its quality and ethics might be

questionable, especially for image captioning which requires proper alignment between visual and textual contents. Although there are successful attempts to refine or distinguish web-based information [13,24], it is unfeasible to completely filter out wrong and noisy data when its extent grows too much.

Synthetic data seems an appealing alternative to match the scaling requirements of modern neural networks while attenuating the drawbacks of web-crawled data. In fact, synthetic data can be produced on-demand, are virtually infinite, and their annotations are in most cases at no cost. Moreover, from an ethical perspective, they usually offer better control over biases than their web counterparts. While the usage of synthetically generated data has led to promising results in various Computer Vision tasks [1,5,9,11,16], limited research efforts have been done in the context of image captioning.

Motivated by the recent advancements in Generative AI, in this work we explore the usage of synthetic images to boost the performance of captioning architectures. In particular, we leverage the well-known Stable Diffusion model [39] to generate synthetic images associated with human-annotated textual sentences and employ these newly generated data to augment the most widely used dataset in the image captioning field (*i.e.* COCO [28]). From a technical point of view, we introduce a simple yet effective framework to employ synthetic data that probabilistically replace real pictures with fake ones and apply it to a standard Transformer-based architecture [48]. To validate our proposal, we conduct extensive experiments to evaluate whether synthetic images can be leveraged to improve the quality of generated captions. Experimental results on the popular COCO dataset [28] demonstrate the effectiveness of our solution, which achieves better results than a baseline model without synthetic data augmentation and competitive performance compared to previously proposed approaches. We believe that our analysis can serve as a starting point for employing synthetically generated images as an effective data augmentation strategy in the field of image captioning and other vision-and-language tasks.

2 Related Work

Image Captioning. Early deep learning-based image captioning models were based on a basic encoder-decoder scheme, with the use of RNNs and LSTMs as popular choices for the text generation part along with CNNs to encode the visual content [22, 38, 50]. Following these initial attempts, subsequent techniques have steadily advanced both the image encoding and language generation stages. Regarding the image encoding, remarkable progress has been achieved through the introduction of additive attention mechanisms to incorporate spatial knowledge, first from a grid of CNN features [55] and later utilizing image regions extracted from pre-trained object detectors [4], eventually considering their semantic and spatial relationships encoded by graph neural networks [56, 57]. Nowadays, Transformer-based architectures [48], initially designed for machine translation and language comprehension purposes and then employed in a variety of tasks [15,35,47], have been adopted in the domain of image captioning as well.

These models are commonly used both in the visual encoding stage [12,21,29,53] and as language models [14,17,30,60], also leading to the design of effective variants of the self-attention operator [14, 21, 33].

Recent advancements have been obtained by large-scale vision-and-language pre-training which usually employs noisy image-text pairs to increase the number of training samples, thus further enhancing the performance of fully-attentive image captioning models [13, 20, 51, 58]. Effective alternatives also involve the use of visual features from large-scale cross-modal architectures [7, 8, 45] like CLIP [36]. These multimodal architectures also allow for the enrichment of predicted textual sentences employing retrieval components, that can be added to the captioning model, and external knowledge from which to extract additional information to improve the final performance [26, 32, 41].

Synthetic Data. To the best of our knowledge, there is a limited amount of works that explore the usage of synthetic data in image captioning. In particular, Hossain *et al.* [19] introduced artificial images into a captioning system, by creating new pictures thanks to generative adversarial networks. More recently, Xiao *et al.* [54] leveraged a latent diffusion model [39] to augment the training dataset, also employing paraphrasing sentences to pair with the generated pictures. However, they only achieved promising results when using limited training instances or when switching to an unpaired image captioning setting. Concurrently, Li *et al.* [25] proposed to employ fake images as a replacement for difficult samples to finetune a large-scale vision-and-language model for captioning. In this work, we stick with the same latent diffusion model to generate fake images (*i.e.* Stable Diffusion [39]), but we do not require any additional textual data outside of captions from the COCO dataset, demonstrating the effectiveness of synthetic data augmentation for the standard image captioning task.

3 Proposed Method

In this section, we introduce SynthCap, a novel image captioning architecture trained with the proposed synthetic augmentation strategy. Fig. 1 shows an overview of our complete model.

3.1 Model Architecture

Visual encoder. Our architecture is based on a fully-attentive Transformer network that takes as input visual features extracted from a pre-trained visual encoder. For the latter, we leverage the image encoder of a pre-trained CLIP-based model [36] and we freeze its weights throughout all the experiments. Specifically, we opt for the CLIP ViT-L/14 version which is based on the Vision Transformer (ViT) backbone [15].

Transformer model. Our language model is a standard encoder-decoder Transformer network [48]. Each encoder layer is made of a self-attention block followed by a feed-forward layer. The former refines the supplied visual tokens via bi-directional self-attention. The latter operates on single tokens with two dense

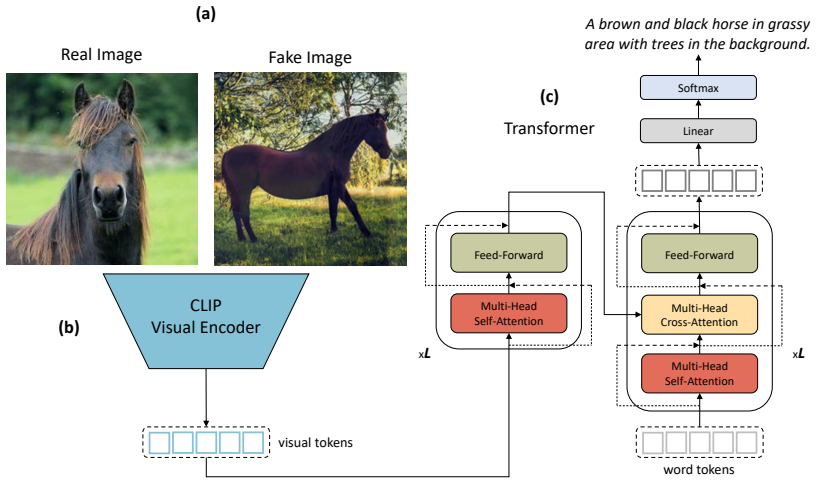

Fig. 1. Overview of the proposed method: (a) we select either a real or a synthetic image, according to a λ_s weight; (b) the CLIP-based visual encoder converts the input image into a sequence of visual tokens; (c) the encoder-decoder Transformer network generates the caption grounded on the visual token.

layers, featuring a GELU non-linearity in between. The output of each block is summed along with its input through a residual connection and then normalized. The decoder network shows a similar architecture to the encoder, but it comprises a cross-attention block interposed between the self-attention and feed-forward block. This additional component is critical, as here occurs the cross-modal integration between visual and textual modalities. In detail, the tokens representing the partial caption generated by the decoder up to time t act as queries, that attend the visual tokens from the encoder, *i.e.* keys and values. Unlike the encoder self-attention block, the decoder self-attention requires a causal mask to prevent tokens from attending to the future. Specifically, masking is implemented by artificially zeroing the entries of the self-attention matrix with row-column indexes $(i, j)_{\forall j>i}$. The output of the decoder is a token sequence $\widetilde{\mathbf{x}} = \{\widetilde{x}_t\}_{t=1,\dots,N}$ whose length is equal to the input. To select the next word $\widetilde{x}_{t+1}$, we sample from a probability distribution over all the possible words in the reference vocabulary, obtained by feeding $\widetilde{x}_t$ to a linear and a softmax layer. At inference time, the decoder works in an auto-regressive manner, meaning that the token produced at time t will be included in the input for time $t + 1$.

3.2 Synthetic Data Augmentation

Our goal is to probe whether synthetic images can be a valuable source of information to train captioning algorithms. We leverage Stable Diffusion [39] to generate fake images to extend the training set of the COCO dataset [22], which is originally composed of more than half a million image-caption pairs (I^r, c_k), with $k = 1, 2, 3, 4, 5$, *i.e.* there are five different reference descriptions available for each image. By conditioning the Stable Diffusion model on c_k, we build an extra

dataset of synthetic (or fake) images paired with the original captions (I_k^s, c_k). As we show in the experimental section, the synthetically generated images prove to have a good correspondence with the captions they have been generated from and therefore can be a valuable data augmentation strategy to train an image captioning model. Conversely, training a model exclusively on synthetic images and corresponding captions leads to unsatisfactory results. Therefore, we argue that both real and artificial pictures are useful for the task of image captioning, and they may be complementary to each other.

In our training framework, we propose to probabilistically replace a real image with its fake counterpart during each training iteration. When we feed the model with a real image I^r, one reference caption is sampled among the five ground-truth sentences available in the dataset. When instead a synthetic image I_k^s is given as input, the network should only focus on the words specifically mentioned in c_k, as c_k alone has been considered by the Stable Diffusion model when generating I_k^s. Formally, given a caption c_k, we build an image-text pair (I, c_k), in which the visual component is chosen as follows:

$$I = \begin{cases} I_k^s & \text{if } \epsilon < \lambda_s \\ I^r & \text{otherwise,} \end{cases} \tag{1}$$

where λ_s is a hyperparameter controlling the probability of using synthetic data at each training iteration and $\epsilon \sim U(0, 1)$. When we set $\lambda_s = 0$, the training set is the original one without any synthetic data augmentation, while when $\lambda_s = 1$ the training set is composed only of fake images and corresponding textual sentences. Note that, regardless of λ_s, the amount of processed samples per epoch remains the same as in the original training process.

Training procedure. We adhere to the two-phase training typically used in image captioning [46] which consists of a pre-training step with cross-entropy loss followed by a finetuning phase based on the self-critical sequence training (SCST) proposed in [38], which optimizes the captioning model with reinforcement learning using the CIDEr metric [49] as a reward.

During SCST optimization, the baseline reward is chosen as the average score over all the sequences sampled using beam search within the same beam, following [14]. According to this setup, whenever we require a synthetic image to replace its associated real one, we opt to randomly draw from the five available fake images. Formally, $I_k^s \sim \{I_1^s, I_2^s, I_3^s, I_4^s, I_5^s\}$. Note that, although for each k, the synthetic image I_k^s has been created from a single description c_k, the CIDEr metric still measures the consensus of the captions generated by our model among all five reference captions $c_{k=1,\dots,5}$.

4 Experimental Evaluation

4.1 Implementation Details

Dataset and evaluation metrics. We evaluate our proposal on the Microsoft COCO dataset [28], using the standard Karpathy splits [22]. We report the

results according to evaluation metrics typically used for image captioning: BLEU [34], METEOR [6], ROUGE [27], CIDEr [49], and SPICE [3].

Architecture. Before being fed to the CLIP visual encoder, each input image undergoes a pre-processing pipeline. The first step involves a resize to reduce the longer side length to a maximum of 224 pixels, keeping the original aspect ratio. It follows a center crop plus a channel-wise normalization. The resulting input is a tensor with shape $3 \times 224 \times 224$, from which the ViT-based CLIP encoder extracts a grid of 256×1024 features, *i.e.* the visual tokens. Our Transformer-based image captioning network comprises $L = 3$ layers in both the encoder and decoder, operating on a hidden size $d = 512$. We therefore apply a linear projection over the CLIP visual features to match this dimensionality. We employ multi-head attention with 8 different heads in each attentive layer, plus dropout with probability 0.1. To convert words into tokens, we leverage the same byte-pair encoding (BPE) tokenizer [43] used by the CLIP text module.

Training details. During cross-entropy optimization, we stick with the setup suggested in [26] using a batch size of 32 and the learning rate scheduling strategy of [48] with warmup equal to $20,000$ iterations. In the SCST phase, we use a batch size of 16, a constant learning rate of 10^{-6}, and apply beam search decoding with a beam size equal to 5. For both training phases, we employ Adam [23] as optimizer. All experiments have been carried out with mixed precision [31] and ZeRO memory offloading [37], using the Huggingface Transformers library [52].

Synthetic data generation. All synthetic images are generated following [2], by feeding Stable Diffusion with the reference captions from the COCO Karpathy training split using the standard prompt "*An image of*". As Stable Diffusion model, we employ the implementation provided by the Huggingface library[3].

4.2 Ablation Studies and Analysis

In this section, we conduct ablation studies to discuss the main design choice of our proposal and validate the proposed synthetic data augmentation strategy.

Overall validation of synthetic images. We first validate the correspondence of generated synthetic images with associated textual sentences by computing the image-text similarity between cross-modal embeddings extracted from CLIP-based visual and textual backbones. As demonstrated in recent literature [18,40], this image-text similarity is effective for evaluating image captioning models. As shown in Table 1, on average, synthetic images seem to have a slightly higher affinity with their descriptions compared to the real ones. This suggests that they could be a valuable source of information to feed an image captioning model during training.

Percentage of synthetic data. In our framework, we control the probability to replace a real image with a synthetic one thanks to λ_s. Table 2 presents the results when varying this parameter in comparison with a baseline model trained without synthetic data. When $\lambda_s = 1.0$, we entirely rely on synthetic images and

[3] `https://huggingface.co/CompVis/stable-diffusion-v1-4`

Table 1. CLIP-based image-text similarity scores for real and synthetic images and corresponding textual sentences.

	Mean	Median	Min	Max
Real images	0.256	0.257	0.004	**0.463**
Synthetic images	**0.263**	**0.262**	**0.098**	0.437

Table 2. Analysis using different percentages of synthetic data. Results are reported after cross-entropy pre-training.

Synth. Data	λ_s	B-1	B-4	M	R	C	S
✗	-	77.5	37.2	30.0	58.6	126.5	23.3
✓	0.1	77.3	37.1	30.3	58.8	127.2	23.5
✓	0.2	77.5	37.9	30.3	59.0	128.1	23.4
✓	0.3	77.7	37.7	30.3	59.1	127.7	23.5
✓	0.4	77.8	37.8	30.4	59.0	128.3	23.5
✓	0.5	77.7	37.6	30.3	58.9	**128.6**	23.4
✓	0.6	77.9	37.6	30.1	58.8	127.5	23.3
✓	0.7	77.4	37.0	30.0	58.7	126.5	23.4
✓	1.0	72.7	29.2	25.5	53.1	100.2	19.0

experience a consistent drop with respect to the baseline. This behavior can be due to the reality gap between real and synthetic images which prevents the model to generalize on real data when it is trained on synthetically generated samples only. This means that synthetic images, despite the advancements in Generative AI, are still far from exactly mimicking pictures from the natural distribution. On the other hand, all other models benefit from augmented training with synthetic images. In detail, we reach the highest CIDEr score when feeding the model with fake images half of the time (*i.e.* $\lambda_s = 0.5$), but we still observe improvements with up to 60% of synthetic images. The positive effects of synthetic data appear to worsen with $\lambda_s = 0.7$, even though the performance is still competitive against the baseline without synthetic data augmentation.

Effectiveness of synthetic data. To prove that the observed improvements truly come from using synthetic images to augment our training set, we repeat the setup explained in Sec. 3.2 but change the source of visual input for augmentation. Since a synthetic image is naturally similar to the original image, a reasonable comparison should rely on visually similar but real images. Thus, in this case, given an image I from the COCO dataset, we replace it with probability λ_s with I_k^r, that corresponds to a real image randomly selected among the top-k similar images with respect to I. In particular, following [41], we extract a feature vector for each image from a pre-trained CLIP model. Then, given an encoded query image, the k most similar ones are retrieved with $k = 1, 3, 5$, using the cosine similarity between pairs of feature vectors as a similarity measure. For this experiment, we employ $\lambda_s = 0.5$ that corresponds to the configuration leading to the highest CIDEr score in the previous analysis. According to the results reported in Table 3, we can notice that our synthetic data augmentation

Table 3. Analysis using our best configuration (*i.e.* $\lambda_s = 0.5$), replacing synthetic images with real ones selected among the top-k similar images. Results are reported after cross-entropy pre-training.

	Synth. Data	B-1	B-4	M	R	C	S
Transformer	✗	77.5	37.2	30.0	58.6	126.5	23.3
Transformer (w/ similar images, $k = 1$)	✗	77.6	37.0	29.8	58.3	125.1	22.8
Transformer (w/ similar images, $k = 2$)	✗	76.7	37.0	30.0	58.5	125.5	23.1
Transformer (w/ similar images, $k = 3$)	✗	76.8	37.0	29.8	58.2	124.6	22.9
SynthCap	✓	**77.7**	**37.6**	**30.3**	**58.9**	**128.6**	**23.4**

Table 4. Comparison with the state of the art on the COCO Karpathy test.

	Cross-Entropy Loss						**CIDEr Optimization**					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
Up-Down [4]	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
GCN-LSTM [57]	77.3	36.8	27.9	57.0	116.3	20.9	80.9	38.3	28.6	58.5	128.7	22.1
SGAE [56]	77.6	36.9	27.7	57.2	116.7	20.9	81.0	39.0	28.4	58.9	129.1	22.2
AoANet [21]	77.4	37.2	28.4	57.5	119.8	21.3	80.2	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$ Transformer [14]	-	-	-	-	-	-	80.8	39.1	29.2	58.6	131.2	22.6
X-Transformer [33]	77.3	37.0	28.7	57.5	120.0	21.8	80.9	39.7	29.5	59.1	132.8	23.4
DLCT [30]	-	-	-	-	-	-	81.4	39.8	29.5	59.1	133.8	23.0
RSTNet [60]	-	-	-	-	-	-	81.8	40.1	29.8	59.5	135.6	23.3
DIFNet [53]	-	-	-	-	-	-	81.7	40.0	29.7	59.4	136.2	23.2
CaMEL [8]	78.3	39.1	29.4	58.5	125.7	22.2	82.8	41.3	30.2	60.1	140.6	23.9
COS-Net [26]	**79.2**	**39.2**	29.7	**58.9**	127.4	22.7	82.7	42.0	30.6	60.6	141.1	24.6
Transformer	77.5	37.2	30.0	58.6	126.5	23.3	82.9	42.2	30.7	60.9	141.9	24.6
SynthCap	77.7	37.6	**30.3**	**58.9**	**128.6**	**23.4**	**83.0**	**42.4**	**30.8**	**61.1**	**143.1**	**24.7**

strategy achieves the best performance compared to both the baseline and the employed retrieval-based augmentation solution.

4.3 Comparison to the State of the Art

We now test SynthCap against other state-of-the-art captioning models. In our analysis, we include earlier approaches featuring LSTM as language models and attention over image regions, like Up-Down [4], eventually boosted with graph-based encoding (GCN-LSTM [57] and SGAE [56]) or self-attention, such as AoANet [21]. Further, we include more recent proposals that rely on the Transformer network, namely $\mathcal{M}^2$ Transformer [14], X-Transformer [33], DLCT [30], RSTNet [60], DIFNet [53], CaMEL [8], and COS-Net [26]. We report the results in Table 4. As it can be seen, SynthCap beats the baseline across all the metrics, in both the cross-entropy pre-training and CIDEr-based optimization stages. Compared to the other better-performing approaches, our framework achieves competitive results, while being based on a simple encoder-decoder Transformer model without any other specific architectural component.

To further confirm the effectiveness of our data augmentation strategy, we report the results on the COCO online test server in Table 5. Following previous literature, we leverage an ensemble of four models trained using different random

Table 5. Leaderboard of various methods on the online COCO test server.

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Up-Down [4]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
SGAE [56]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoANet [21]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
$\mathcal{M}^2$ Transformer [14]	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
X-Transformer [33]	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
RSTNet [60]	82.1	96.4	67.0	91.3	52.2	83.0	40.0	73.1	29.6	39.1	59.5	74.6	131.9	134.0
DLCT [30]	82.4	96.6	67.4	91.7	52.8	83.8	40.6	74.0	29.8	39.6	59.8	75.3	133.3	135.4
COS-Net [26]	83.3	96.8	68.6	92.3	54.2	84.5	42.0	74.7	30.4	40.1	60.6	76.4	136.7	138.3
CaMEL [8]	83.2	97.3	68.3	92.7	53.6	84.8	41.2	74.9	30.2	39.7	60.2	75.6	137.5	140.0
SynthCap	**83.7**	**97.6**	**69.2**	**93.5**	**54.9**	**86.3**	**42.8**	**77.1**	**30.9**	**41.3**	**61.4**	**77.7**	**140.1**	**142.6**

Fig. 2. Qualitative comparison between SynthCap and the baseline on sample images from the COCO dataset.

seeds. Also in this setting, SynthCap achieves the best results according to all evaluation metrics. Finally, in Fig. 2, we show some qualitative results on sample images from the COCO dataset, comparing captions generated by our model with those generated by the baseline without synthetic data augmentation.

5 Conclusion

In this work, we propose a novel image captioning framework enhanced with a synthetic data augmentation strategy. In particular, we leverage the well-known Stable Diffusion model to generate additional images that can be effectively employed as additional training samples. The proposed strategy is widely usable, given the easy accessibility of advanced text-to-image generative models and their increasingly impressive results. Experimentally, the proposed solution is capable of boosting the performance of a standard Transformer-based model, working only at the data level and maintaining the exact same network.

Acknowledgements This work has partially been supported by the European Commission under the PNRR-M4C2 (PE00000013) project "FAIR - Future Artificial Intelligence Research", by the Horizon Europe project "European Lighthouse on Safe and Secure AI (ELSA)" (HORIZON-CL4-2021-HUMAN-01-03), co-funded by the European Union, and by the PRIN project "CREATIVE: CRoss-modal understanding and gEnerATIon of Visual and tExtual content" (CUP B87G22000460001), co-funded by the Italian Ministry of University.

References

1. Allegretti, S., Bolelli, F., Cancilla, M., Pollastri, F., Canalini, L., Grana, C.: How does connected components labeling with decision trees perform on GPUs? In: CAIP (2019)
2. Amoroso, R., Morelli, D., Cornia, M., Baraldi, L., Del Bimbo, A., Cucchiara, R.: Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images. arXiv preprint arXiv:2304.00500 (2023)
3. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: Semantic Propositional Image Caption Evaluation. In: ECCV (2016)
4. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
5. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic Data from Diffusion Models Improves ImageNet Classification. arXiv preprint arXiv:2304.08466 (2023)
6. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshops (2005)
7. Barraco, M., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: The Unreasonable Effectiveness of CLIP Features for Image Captioning: An Experimental Analysis. In: CVPR Workshops (2022)
8. Barraco, M., Stefanini, M., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: CaMEL: Mean Teacher Learning for Image Captioning. In: ICPR (2022)
9. Bolelli, F., Allegretti, S., Grana, C.: One DAG to rule them all. IEEE Trans. PAMI **44**(7), 3647–3658 (2021)
10. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)
11. Chen, Y., Li, W., Chen, X., Gool, L.V.: Learning Semantic Segmentation From Synthetic Data: A Geometrically Guided Input-Output Adaptation Approach. In: CVPR (2019)
12. Cornia, M., Baraldi, L., Cucchiara, R.: Explaining Transformer-based Image Captioning Models: An Empirical Analysis. AI Communications **35**(2), 111–129 (2022)
13. Cornia, M., Baraldi, L., Fiameni, G., Cucchiara, R.: Universal Captioner: Inducing Content-Style Separation in Vision-and-Language Model Training. arXiv preprint arXiv:2111.12727 (2022)
14. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-Memory Transformer for Image Captioning. In: CVPR (2020)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: ICLR (2021)
16. Fabbri, M., Brasó, G., Maugeri, G., Cetintas, O., Gasparini, R., Ošep, A., Calderara, S., Leal-Taixé, L., Cucchiara, R.: MOTSynth: How Can Synthetic Data Help Pedestrian Detection and Tracking? In: ICCV (2021)
17. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image Captioning: Transforming Objects into Words. In: NeurIPS (2019)
18. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: EMNLP (2021)

19. Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H., Bennamoun, M.: Text to image synthesis for improved image captioning. IEEE Access **9**, 64918–64928 (2021)
20. Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L.: Scaling Up Vision-Language Pre-training for Image Captioning. In: CVPR (2022)
21. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on Attention for Image Captioning. In: ICCV (2019)
22. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
23. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: ICLR (2015)
24. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: ICML (2022)
25. Li, W., Lotz, F.J., Qiu, C., Elliott, D.: Data curation for image captioning with text-to-image generative models. arXiv preprint arXiv:2305.03610 (2023)
26. Li, Y., Pan, Y., Yao, T., Mei, T.: Comprehending and ordering semantics for image captioning. In: CVPR (2022)
27. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: ACL Workshops (2004)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV (2014)
29. Liu, W., Chen, S., Guo, L., Zhu, X., Liu, J.: CPTR: Full Transformer Network for Image Captioning. arXiv preprint arXiv:2101.10804 (2021)
30. Luo, Y., Ji, J., Sun, X., Cao, L., Wu, Y., Huang, F., Lin, C.W., Ji, R.: Dual-Level Collaborative Transformer for Image Captioning. In: AAAI (2021)
31. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H.: Mixed Precision Training. In: ICLR (2018)
32. Moratelli, N., Barraco, M., Morelli, D., Cornia, M., Baraldi, L., Cucchiara, R.: Fashion-Oriented Image Captioning with External Knowledge Retrieval and Fully Attentive Gates. Sensors **23**(3), 1286 (2023)
33. Pan, Y., Yao, T., Li, Y., Mei, T.: X-Linear Attention Networks for Image Captioning. In: CVPR (2020)
34. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL (2002)
35. Pipoli, V., Cappelli, M., Palladini, A., Peluso, C., Lovino, M., Ficarra, E.: Predicting gene expression levels from DNA sequences and post-transcriptional information with Transformers. Computer Methods and Programs in Biomedicine **225**, 107035 (2022)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (2021)
37. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: ZeRO: Memory optimizations Toward Training Trillion Parameter Models. In: SC (2020)
38. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-Critical Sequence Training for Image Captioning. In: CVPR (2017)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
40. Sarto, S., Barraco, M., Cornia, M., Baraldi, L., Cucchiara, R.: Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In: CVPR (2023)

41. Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.: Retrieval-Augmented Transformer for Image Captioning. In: CBMI (2022)
42. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
43. Sennrich, R., Haddow, B., Birch, A.: Neural Machine Translation of Rare Words with Subword Units. In: ACL (2016)
44. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In: ACL (2018)
45. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K.: How Much Can CLIP Benefit Vision-and-Language Tasks? In: ICLR (2022)
46. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From Show to Tell: A Survey on Deep Learning-based Image Captioning. IEEE Trans. PAMI **45**(1), 539–559 (2022)
47. Stefanini, M., Lovino, M., Cucchiara, R., Ficarra, E.: Predicting gene and protein expression levels from DNA and protein sequences with Perceiver. Computer Methods and Programs in Biomedicine **234**, 107504 (2023)
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
49. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: Consensus-based Image Description Evaluation. In: CVPR (2015)
50. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR (2015)
51. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In: ICLR (2022)
52. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-Art Natural Language Processing. In: EMNLP (2020)
53. Wu, M., Zhang, X., Sun, X., Zhou, Y., Chen, C., Gu, J., Sun, X., Ji, R.: DIFNet: Boosting Visual Information Flow for Image Captioning. In: CVPR (2022)
54. Xiao, C., Xu, S.X., Zhang, K.: Multimodal Data Augmentation for Image Captioning using Diffusion Models. arXiv preprint arXiv:2305.01855 (2023)
55. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
56. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-Encoding Scene Graphs for Image Captioning. In: CVPR (2019)
57. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring Visual Relationship for Image Captioning. In: ECCV (2018)
58. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: VinVL: Revisiting visual representations in vision-language models. In: CVPR (2021)
59. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: OPT: Open Pre-trained Transformer Language Models. arXiv preprint arXiv:2205.01068 (2022)
60. Zhang, X., Sun, X., Luo, Y., Ji, J., Zhou, Y., Wu, Y., Huang, F., Ji, R.: RSTNet: Captioning With Adaptive Attention on Visual and Non-Visual Words. In: CVPR (2021)