
Self-Supervised Image Captioning with CLIP

Chuanyang Jin
New York University
cj2133@nyu.edu

Abstract

Image captioning, a fundamental task in vision-language understanding, seeks to generate accurate natural language descriptions for provided images. Current image captioning approaches heavily rely on high-quality image-caption pairs, which can be hard to obtain for many domains. To address this, we introduce a self-supervised image captioning method. After learning an initial signal from a small labeled dataset, our method transitions to self-supervised learning on unlabeled data, leveraging the auxiliary task of enhancing the CLIP relevance between images and generated captions. Remarkably, despite utilizing less than 2% of the labeled COCO dataset, our method delivers a performance comparable to state-of-the-art models trained on the complete dataset. Human evaluations further reveal that our method produces captions with greater distinctiveness and informativeness, two attributes inherently challenging to achieve through supervised learning.

1 Introduction

Image captioning aims to describe images with syntactically and semantically meaningful sentences, thereby bridging the gap between vision and language. Contemporary approaches to image captioning are predominantly supervised and face several challenges.

Firstly, acquiring image-caption pairs is a challenging endeavor for many domains, thus it is desirable to have a method that doesn't need much supervision. Despite this, prevailing methods rely heavily on extensive collections of captioned images for training. In some cases, these models even necessitate multiple reference captions and additional annotations [Stefanini et al., 2022].

Second, the quality of reference captions can be less than ideal. As will be shown later, current models can produce captions that match the quality of the reference captions in well-known datasets like COCO. Nevertheless, the reliance on reference-based similarity objectives prevents these models from surpassing this quality threshold. For example, models trained on text similarity objectives tend to overlook specific details that set one image apart from others, as public datasets' reference captions typically describe only the most conspicuous and common objects [Cho et al., 2022].

In light of these challenges, our research introduces a two-stage self-supervised captioning method. Notably, our method (1) significantly reduces dependency on reference captions through a self-supervised mechanism, and (2) while leveraging insights from these captions, surpasses the quality of the original reference captions in certain evaluations.

2 Method

2.1 Image Captioning with CLIP

Large research efforts have been devoted to image captioning. Typically, an image captioning method first encodes an image into a visual representation, which is then decoded to produce the final captions. The visual encoder extracts the representation from either classification networks [Chen et al., 2017]



A large polar bear swimming in a pool of water. A man riding a wave on top of a surfboard. A city skyline and a large body of water in the sunset. A small brown squirrel standing on the grass.

Figure 1: Example captions generated by our method.

[Fang et al., 2015] [Xu et al., 2015], or object detection networks [Anderson et al., 2018] [Li et al., 2020] [Zhou et al., 2020], with the latter providing more expressive features. Some models introduce a self-attention mechanism [Herdade et al., 2019] or Vision Transformers [Dosovitskiy et al., 2020] to make better use of visual cues. When it comes to the textual decoder, different models use LSTM variants [Chen et al., 2018] [Vinyals et al., 2015] or transformer-based architectures [Herdade et al., 2019] [Luo et al., 2021]. The emergence of large language models like GPT [Brown et al., 2020] presents promising alternatives.

Contrastive Language-Image Pre-training (CLIP) [Radford et al., 2021] leverages vision-language pre-training on abundant image-caption pairs and produces cross-modal contextual representations. ClipCap [Mokady et al., 2021] leverages CLIP representations for image captioning. First, it employs CLIP’s visual encoder to generate an encoded image representation. Next, a mapping network—either a transformer or MLP—translates the CLIP embedding to the language model space, producing a prefix to the caption. Finally, the language model auto-regressively generates the predicted image captions. By leveraging multimodal understanding from existing models instead of learning new semantic entities, ClipCap produces state-of-the-art results with a simpler and faster model.

2.2 Our Method

We propose an enhancement to the ClipCap architecture to reduce its dependency on captioned images. This is achieved by introducing an auxiliary task that aims to maximize the matching, as determined by CLIP, between the image and the captions produced.

To integrate this auxiliary task as a direct replacement for the original training objective, we face a challenge during optimization: The language model generates captions by sequentially predicting each word token, which involves a probability distribution across the vocabulary for each token. The selection of the next token typically employs either greedy decoding or beam search, both relying on the Softmax operation. However, the discrete sampling operation following Softmax is non-differentiable, thus interrupting the flow of gradients. To overcome this challenge, we use the Gumbel-Softmax method [Jang et al., 2016], which provides a reparameterization trick for discrete variables. This technique allows us to transform the non-differentiable operation into a differentiable one, thereby enabling our gradient-based optimization on the auxiliary task.

Nonetheless, in the absence of initial signals, the generated text is purely random and incoherent. As a result, the image-caption relevance calculated would stay extremely low. This creates a significant challenge in further optimizing the model. Given that the search space for potential inputs is exceptionally vast and discrete, the loss will fluctuate wildly and fail to converge. To overcome this challenge, we introduce an initial supervised training stage before the self-supervised training stage. This supervised stage equips the model with an essential set of preliminary signals, thereby establishing a baseline proficiency that can be further refined during the self-supervised training stage. An ablation study highlighting the significance of this supervised stage can be found in Appendix A.

Breaking down our methodology, during the supervised training stage, an initial signal is established using a small dataset of 10,000 captioned images, constituting less than 2% of the standard training division in the COCO-captions dataset [Lin et al., 2014]. For each image paired with its corresponding caption, the captions are considered as a sequence of tokens represented as $c = c_1, \dots, c_l$. Using the CLIP model, we generate the visual embedding v . This is then transformed by our mapping network to produce a series of prefix embeddings $p = p_1, \dots, p_k$, where each shares the same dimensionality as the word embedding. Following this, we concatenate the prefix embeddings with the caption embeddings, c_1, \dots, c_{l-1} , and feed the concatenation into the language model LLaMA 2 to predict

the subsequent token c_i . We either train the mapping component or finetune the LLaMA 2 using a cross-entropy loss $\mathcal{L}_{\text{supv}} = -\sum_{i=1}^l \log p_\theta(c_i|p_1, \dots, p_k, c_1, \dots, c_{i-1})$.

In the self-supervised training stage, each step involves two main steps. First, we employ LLaMA 2 to generate a caption, $\mathbf{c}_{\text{model}}$, corresponding to a specific image. Subsequently, we train the model based on the relevance between the image and its generated caption. The loss function for this phase is crafted to be the inverse of the image-caption relevance as evaluated by the CLIP embeddings, represented as $\mathcal{L}_{\text{unsupv}} = -\cos(\mathbf{v}, \mathbf{c}_{\text{model}})$.

3 Experiments

3.1 Comprehensive Evaluation

Current methods for evaluating the quality of generated text, like in image captioning, are mainly based on comparing the text to a set of reference texts. These methods include n-gram overlap techniques such as BLEU [Papineni et al., 2002], METEOR [Denkowski and Lavie, 2014], and CIDEr [Vedantam et al., 2015]. There are also other metrics like SPICE [Anderson et al., 2016] and TIGER [Jiang et al., 2019] that go beyond simple overlap and incorporate more sophisticated models of similarity between reference and candidate texts. These metrics operate under the belief that the reference captions are of high quality and serve as a benchmark for the generated captions to reach. However, our subsequent analyses indicate that many captions created by models are actually better than reference captions, offering more precise descriptions. This suggests a need for new metrics that do not solely rely on how close the generated caption is to a reference caption.

Recent metrics suggest using relevance scores from Vision-Language Models (VLMs) like BERTScore [Zhang et al., 2019], ViLBERTScore [Lee et al., 2020], UMIC [Lee et al., 2021], and CLIPScore [Hessel et al., 2021]. These VLMs, trained on large datasets, acquire a rich semantic understanding of the relationship between images and texts. Therefore, they can evaluate captions in a zero-shot manner. By using reference captions available in the test set, we introduce the RefCompare Score. This metric offers a more consistent evaluation and delves deeper into understanding how a method measures up against the reference captions. In particular, the RefCompare Score calculates the average percentage of reference captions that have a lower CLIP relevance score for the image than the score of the generated caption.

In this context, a RefCompare Score equal to or higher than 0.5 suggests that the model’s captions are of comparable or better quality than the benchmark captions. As illustrated in Figure 2, top-performing models have the ability to create captions that are not just similar, but often better in quality than the standard references. This underscores the importance of having evaluation metrics that measure beyond mere textual similarity.

For a comprehensive assessment, we use the traditional BLEU score along with our RefCompare Score to compare our baseline methods and proposed approach. Acknowledging that existing metrics might not completely reflect the subtleties of human judgment, we also conduct a human evaluation to measure two important aspects: distinctiveness and informativeness.

3.2 Baselines and Variants of Our Method

We evaluate the performance of state-of-the-art vision-language models, notably **Oscar** [Li et al., 2020] and **VLP** [Zhou et al., 2020], adhering to the finetuning and inference strategies described in their respective publications.

For our method, we experiment with four variants: two that involve training the mapping network with a frozen language model, and two that include the finetuning of the language model. The mapping network is either architected as an MLP or a Transformer network. We refer to these as **MLP**, **Transformer**, **MLP + LLaMA finetuning**, and **Transformer + LLaMA finetuning**.

3.3 Results

We compare our method with the Oscar and VLP baselines. It is noteworthy that our model has fewer parameters and needs less time to train. In terms of performance, while our model yields

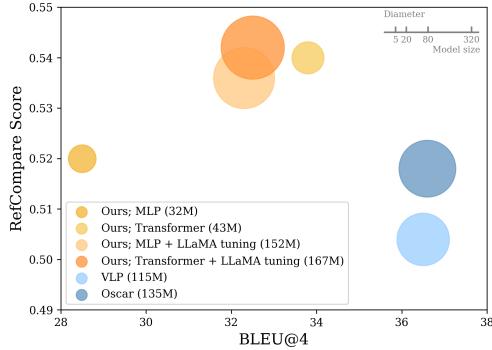


Figure 2: Comparative evaluation of two state-of-the-art models against four variants of our method.

slightly lower BLEU@4 scores, similar to the supervised ClipCap method, it stands out by securing substantially higher RefCompare Scores.

Looking at the four variants of our method, our default Transformer configuration delivers the highest BLEU@4 score and excellent RefCompare Score, while maintaining a relatively small model size. The MLP configuration, on the other hand, presents weaker performance, likely due to its limitation in expressiveness. Both the MLP + LLaMA tuning and Transformer + LLaMA tuning versions yield decent results, but they do not distinctly outperform the Transformer, even with more trainable parameters and longer training times. This observation suggests that finetuning the language model in conjunction with using a Transformer mapping network could lead to an excess of expressive power.

3.4 Human Evaluation

We conduct a human evaluation to assess preferences between captions generated by the Oscar model and those produced by our method. We looked at two key aspects: how unique the captions are and how much information they convey. For this, we sample 1000 random images from the COCO-captions dataset and create captions using both methods. We then engage five independent evaluators to compare the distinctiveness and informativeness of each set of captions without disclosing which are generated by our model. The results show that our model’s captions are perceived as more distinctive in 58.6% of the cases, and more informative in 69.2% of the cases.

We present some examples in Appendix B. As we can see, supervised methods relying on references often employ more commonplace words in their descriptions, leading to a better match with reference captions. In contrast, our method generates more distinct, semantically rich, and “human-like” interpretations. While these may not always echo the phrasing of the reference captions and might contain occasional errors, they largely resonate with human intuition and common sense. Notably, while both the human-annotated captions and those from reference-based methods tend to focus on describing the most prominent and frequently encountered objects, our method seeks to incorporate additional objects within the scene, providing a more comprehensive and detailed description.

4 Conclusion

We present a self-supervised method of image captioning that leverages the auxiliary task of CLIP relevance. Our method not only showcases remarkable performance through similarity-based metrics but also produces state-of-the-art results in Vision-Language Model (VLM) evaluations. Unlike existing methods, our method significantly diminishes the dependency on captioned images and generates captions that are more distinctive, informative, and aligned with human preference.

Acknowledgements

We thank Peiqi Liu, He He, Saining Xie, Tianmin Shu, and Hao Zhu for their valuable insights shared during our discussions.

References

- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.
- 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017.
- Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. Regularizing rnns for caption generation by reconstructing the past with the present. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 7995–8003, 2018.
- Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*, 2022.
- Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.
- Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in neural information processing systems*, 32, 2019.
- 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.
- Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
- Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. Tiger: Text-to-image grounding for image caption evaluation. *arXiv preprint arXiv:1909.02050*, 2019.
- Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 34–39, 2020.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. Umic: An un-referenced metric for image captioning via contrastive learning. *arXiv preprint arXiv:2106.14019*, 2021.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2286–2293, 2021.

Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559, 2022.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Luwei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020.

A More Results

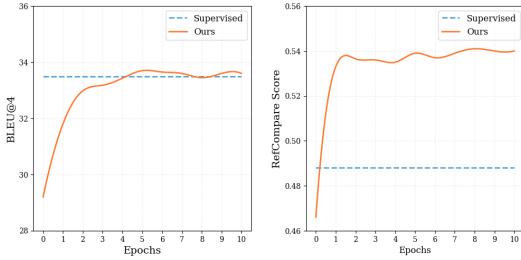


Figure 3: BLEU@4 and RefCompare Scores during the self-supervised stage.

In Figure 3, we present the model’s performance in terms of BLEU@4 and RefCompare Scores during the self-supervised training stage. Initially, the performance trails that of the supervised-only approach, largely due to the limited size of the dataset used in the supervised phase. However, as training progresses, we observe a notable increase in performance. While there are occasional fluctuations, the general trajectory is upward, stabilizing after about 10 epochs. Impressively, our approach not only narrowly exceeds the BLEU@4 score of the supervised method but also significantly outperforms it in the RefCompare Score.

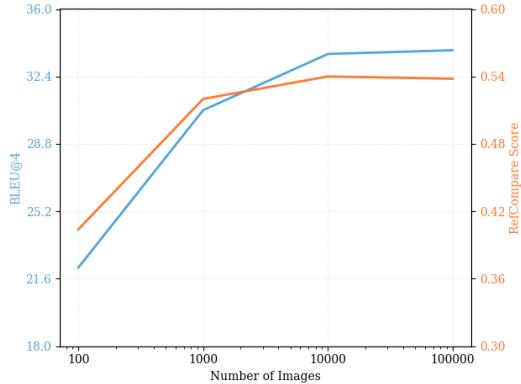


Figure 4: Final BLEU@4 and RefCompare Scores with the variation in the number of images used in the supervised stage.

In Figure 4, we depict the relationship between the size of the labeled training dataset used during the supervised stage and the model’s final performance. Notably, even a modest dataset of 100 images for supervised training is sufficient to guide the model effectively through the subsequent self-supervised stage. As we increase the size of the supervised training set, the model’s performance improves consistently. However, beyond an approximate limit of 10,000 images, adding more training data does not lead to a significant enhancement of the model’s performance.

B More Examples

In Figure 5 and Figure 6, we show more examples of captions generated through our method, the supervised method, and Oscar.

					
Reference	A puppy rests on the street next to a bicycle	A traffic light and street sign surrounded by buildings	A giraffe bending over while standing on green grass	A person up in the air, upside down while outside	A very cute cat near a bunch of birds
Supervised	A puppy sleeping on the street (0.02)	A city in the fog (0.68)	A giraffe that is standing in the grass (0.04)	A man flying through the air while riding a snowboard (1)	A white cat standing next to a bunch of pigeons (0.98)
Ours	A white dog laying on the street next to a bike (0.15)	A black and white photo of a traffic light (0.99)	A giraffe eating grass in a grassy field (0.16)	A person on a snowboard jumping over a snow-covered slope (1)	A cat approaching a bunch of pigeons (0.97)

Figure 5: Uncurated captions generated through both the supervised approach and our method for five images in the COCO-captions dataset. RefCompare scores are shown in parentheses.

					
Reference	A man with a red helmet on a small moped on a dirt road	A young girl inhales with the intent of blowing out a candle	A man on a bicycle riding next to a train	A wooden cutting board topped with sliced up food	A kitchen is shown with a variety of items on the counters
Oscar	A man riding a motorcycle down a dirt road (0.22)	A woman sitting at a table with a plate of food (0)	A woman riding a bike down a street next to a train (0.02)	A woman sitting at a table with a plate of food (0.42)	A kitchen with a sink, dishwasher and a window (0.41)
Ours	A man riding a motorbike on a dirt mountain road (0.37)	A woman is blowing out a candle on a piece of cake (0.91)	A man is riding a bike next to a train (0.15)	A lot of wooden spoons on a wooden table (1)	A kitchen with a sink, a stove and a window (0.31)

Figure 6: Uncurated captions generated through both Oscar and our method for five images in the COCO-captions dataset. RefCompare scores are shown in parentheses.