
CAPTIONWIZARD: LEVERAGING VISION TRANSFORMERS FOR ENHANCED INSTAGRAM INTERACTIVITY

Morteza Damghani Nouri, Jeremy Jang, Deepayan Sur & Haoran Tang

Department of Computer Science

University of Southern California

Los Angeles, CA 90089, USA

{damghani, jeremyja, deepayan, haoranta }@usc.edu

ABSTRACT

Our study introduces a new approach to image captioning on Instagram, focusing on enhancing user engagement. Traditional methods prioritize descriptive accuracy but often overlook user interaction. Our methodology integrates social media metrics like likes and comments into caption generation. We train a model on Instagram posts to create captions that are both accurate and engaging. Preliminary results indicate a successful balance between accuracy and user engagement, suggesting a shift towards a more user-centric captioning model. This bridges the technical and interactive aspects of social media content.

1 INTRODUCTION

Image captioning, a critical task in the field of computer vision and natural language processing, has undergone substantial advancements in recent years. The primary objective of these developments focuses on enhancing the descriptive accuracy of captions generated for images. While these improvements have led to technically precise and contextually accurate descriptions, there remains a significant gap in terms of user engagement and appeal, particularly within social media contexts. This research aims to bridge this gap by shifting the focus from mere descriptive accuracy to the generation of captions that are more aligned with human interests and can potentially elicit greater user interaction on social media platforms, specifically Instagram.

To address this issue, this study introduces an innovative approach that integrates the analysis of social media engagement metrics with the process of image captioning. The hypothesis underpinning this research is that captions that are accurate as well as engaging and relatable to the human audience will likely achieve higher levels of user interaction, as measured by likes and comments on social media. To test this hypothesis, the research involves the collection and analysis of a unique dataset. This dataset will consist of Instagram posts, alongside corresponding engagement metrics such as the number of likes, the volume of comments, and the poster’s follower count. These metrics will serve as critical parameters in developing a threshold that defines what constitutes an engaging or ‘likable’ caption.

The primary contribution of this research is the development of a machine-learning model trained on this dataset. This model is designed to generate captions that not only accurately describe the visual content of images but also resonate with the social media audience, thereby potentially increasing user engagement. This approach introduces a unique blend of image processing, natural language processing, and social media analytics to enhance automated image captioning, with a focus on social media engagement. The code used in this research is made available ¹.

2 RELATED WORKS

The field of image captioning has seen considerable progress in terms of descriptive accuracy. Researchers have developed sophisticated algorithms capable of understanding nuanced image

¹<https://github.com/haorant14/CaptionWizard>

contexts and relationships between objects such as Li et al. (2022a), Wang et al. (2022), Luowei et al. (2019). This progress is often benchmarked against datasets such as COCO in Lin et al. (2014) and Flickr30k in Young et al. (2014), using metrics like BLEU, METEOR, and CIDEr.

Although descriptive accuracy has been a primary focus in caption generation, recent research like Kurt et al. (2019) and Park et al. (2017) explores the human-like qualities of generated captions. This includes adding elements of naturalness, sentiment, and even humor to make captions more relatable and engaging for human audiences.

In the realm of social media, particularly on platforms like Instagram, the role of image captions takes on additional significance due to their impact on user engagement. Studies in social media analytics have shown that content characteristics, including the nature of the captions, can significantly influence user interactions such as likes, comments, and shares (Zhongping et al. (2018)). Research has underscored the value of categorizing images into distinct groups for more effective analysis (Kim et al. (2020)). This approach allows for tailored analysis specific to each category, enhancing the accuracy and relevance of the insights derived.

Despite these advancements, there remains a gap in research specifically targeting the creation of image captions that are tailored to enhance user engagement in social media contexts. Most existing models prioritize technical accuracy over the ability to generate human-like, engaging text.

This project aims to address this gap by combining methodologies from both image captioning and social media analytics. By analyzing Instagram engagement metrics such as likes, comments, and follower counts, we propose a new approach to train a machine learning model that not only describes an image accurately but also resonates with social media audiences.

3 METHODOLOGY

3.1 APPROACH

We acquired a comprehensive Instagram dataset, which we processed to normalize the data based on the ratio of the combined number of likes and comments to the number of followers for each post, ensuring a standard distribution. We then focused our analysis on the top percentile of this normalized data to identify posts with the highest levels of user engagement. The next step involved using a classifier to categorize these posts based on their image content into various categories such as travel, food, and fashion. This classification enables a more detailed analysis of engagement thresholds across different content types. Finally, we trained our model on this refined dataset. The model is tailored to discern and learn the characteristics of posts correlating with high engagement levels, considering the specific nuances of each content category, thereby providing a deeper understanding of how different types of content resonate with Instagram users.

3.2 DATASET COLLECTION

Image captioning’s reliance on datasets such as COCO and Flickr has been instrumental in shaping the field, providing substantial insights through descriptive captions. However, the limitations of these datasets in encapsulating real-world communication prompted our quest for more genuine and representative sources. This study employs a comprehensive dataset from Kim et al. (2020) focused on Instagram influencers. The dataset encompasses 10,180,500 posts, accompanied by extensive metadata in JSON format (approximately 37 GB) and image files in JPEG format (around 189 GB). Each post’s metadata includes information such as captions, user tags, hashtags, timestamps, sponsorship details, likes, and comments, providing a rich basis for engagement analysis. These metadata parameters proved instrumental in quantitatively ranking and evaluating the posts within the dataset, enhancing our ability to discern and analyze the posts based on their engagement metrics.

To address posts with multiple images, the dataset includes a JSON-Image mapping file, linking each post’s metadata with its corresponding image files. This approach ensures accurate analysis of posts with varying numbers of images.

3.3 PREPROCESSING

Building upon S. Kim’s work (Kim et al. (2023)), our research sought to establish a unique metric, `likability_score`, for evaluating engagement in the dataset’s posts. This metric, shaped by log transformations and adjusted for post interactions relative to follower count, aims to balance high engagement frequency with its overall impact, considering both the quantity and quality of interactions.

$$\text{likability_score} = \log_{10} \left(\frac{\text{No of Likes} + \text{No of Comments} - \text{No of Followers} \times 0.1}{\text{No of Followers}} \right) \quad (1)$$

This scoring mechanism facilitated a standardized assessment of posts within the dataset, ensuring a more balanced representation for subsequent analyses generating a Gaussian distribution as shown in the graph 1. We focused on posts that exhibited heightened engagement, and we concentrated our attention on the top 25% of the dataset, highlighting content that showcased substantial appeal and interaction across the Instagram platform.

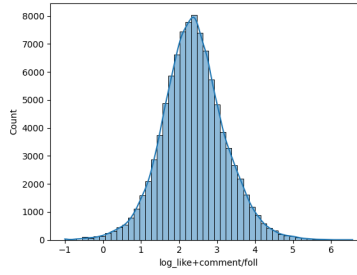


Figure 1: Distribution of the Likability Score in the dataset

Upon finalizing the dataset, a notable impediment emerged, characterized by the prevalence of non-English and multilingual code-mixed text. Given our explicit focus on utilizing English text and its associated context, we endeavored to devise an innovative solution drawing inspiration from Dutta’s work as referenced in Dutta (2021). Our approach involved harnessing a BiLSTM (Bidirectional Long Short-Term Memory) model, augmented by the utilization of a sentence piece model for subword generation and subsequent embedding creation as shown in Figure 2. To bolster the effectiveness of this model, training was conducted on the LiNCE dataset, renowned for its extensive multilingual data, which holds prominence in the language detection domain. The integration of this model proved instrumental, enabling us to discern and segregate English-only captions with a high degree of accuracy and efficiency, effectively addressing the challenge posed by the presence of multilingual content within the dataset.

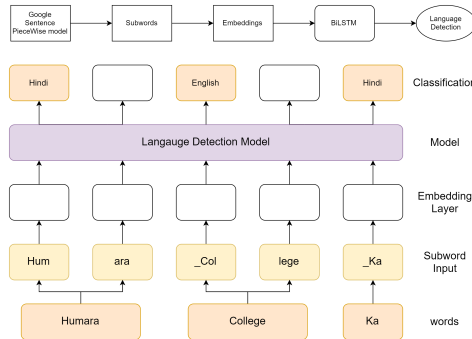


Figure 2: Model Architecture for Language Detection

3.4 IMAGE CLASSIFICATION

To enhance our dataset’s utility, we integrate an image classifier categorizing images into thirteen themes inspired by Karma (2019). These categories reflect common Instagram content, aiding in analyzing the link between image content and user engagement.

Our classification method involves a two-stage process, using the CoCa model Yu et al. (2022) for caption generation and BART-large-mnli Lewis et al. (2019) for zero-shot text classification. This approach taps into the contextual capabilities of language models, offering a unique take on image classification.

By combining CoCa’s effective visual content interpretation with BART-large-mnli’s text categorization, we achieve a comprehensive understanding of the images. Accurate captions are crucial before classification to ensure consistency and relevance.

While direct image classification might be more straightforward, our focus on captioning, given the limited pre-trained models for Instagram images, drives our project’s direction. We thus prioritize advancing captioning techniques, recognizing their potential despite not developing a new classification model.

3.5 MODEL

Our research endeavors to advance the state-of-the-art in this domain by amalgamating the prowess of vision encoder-decoder models. We used transfer learning and more specifically fine-tuning for training our model. ViT-GPT2-Image-Captioning was used as the base model. The significance of employing pre-trained checkpoints in initializing image-to-text-sequence models has been established in contemporary literature, with studies like Li et al. (2022b) highlighting its efficacy. Leveraging this knowledge, we implemented a unique architecture using a Vision Transformer (ViT) as the image encoder and the GPT-2 language model as the auto-regressive decoder. Furthermore, our utilization of a sequence-to-sequence trainer, tailored for sequence-to-sequence tasks, accentuates its suitability in our pursuit of a sequential decoder for the generation of likable captions. Our proposed architecture capitalizes on the potential of a vision encoder-decoder model for image captioning. The image encoder, based on the Dosovitskiy et al. (2021), processes visual information effectively, extracting salient features from images. These encoded features serve as input to the GPT-2 decoder, a renowned auto-regressive language model adept at generating coherent and contextually relevant captions. The seamless integration of these components forms the crux of our image-to-text generation framework. Notably, the choice of a sequence-to-sequence trainer, tailored for sequence-to-sequence tasks, aligns with our objective of employing a sequential decoder, ensuring the generation of Instagram-likable captions that encapsulate visual content effectively.

4 EXPERIMENT

For our experiment, we trained our model, CaptionWizard, using 15200 image-caption pairs and tested it with 1237 pairs. We evaluated performance using BLEU and METEOR metrics, comparing captions generated by our model and the base model (ViT-GPT2-Image-Captioning) with the dataset’s likable ground truth captions. Each of the 1237 test cases involved this comparison to assess the effectiveness of CaptionWizard relative to the base model. The results of this comparison, highlighting the performance of both models in relation to the ground truth captions, are detailed in a table.

In our initial model, we employed a CNN as an image feature extraction encoder and an LSTM for decoding. However, this setup proved computationally expensive, making it challenging to achieve satisfactory results within a reasonable timeframe. Consequently, we pivoted to using the ViT-GPT2-Image-Captioning model. This shift allowed for more efficient computation while maintaining the quality of the caption generation, aligning better with our project’s resources and objectives.

4.1 ViT-GPT2-IMAGE-CAPTIONING

In our research, we initially consider the Vision Transformer (ViT) as a foundational baseline model, renowned for its effectiveness in processing and interpreting visual information. The standard ViT, as established in the work of Dosovitskiy et al. (2021), serves as a robust starting point due to its

proficiency in extracting salient features from images. However, our model diverges from this baseline by incorporating a specialized training approach. We have fine-tuned the Vision Transformer on a unique dataset comprised of real Instagram posts, selected for their high engagement and 'likability'. This tailored training enables our model to not only harness the inherent strengths of the ViT but also to specialize in generating captions that resonate more effectively with social media audiences. Thus, while the ViT provides the structural foundation, our model represents an evolution tailored to the specific nuances and demands of social media captioning.

4.2 CoCa

Our baseline model employs the "laion/mscoco_finetuned_CoCa-ViT-L-14-laion2B-s13B-b90k" from Hugging Face to generate captions for our Instagram dataset. This model was chosen due to its advanced capabilities and fine-tuning on relevant datasets. We utilize these generated captions to assess the model's performance, applying metrics such as BLEU and METEOR, alongside BERT embeddings for a more nuanced evaluation. This detailed evaluation will provide insights into the model's effectiveness in caption generation and its relevance to our specific dataset, setting a foundational benchmark for future improvements and comparisons.

5 RESULTS

In our study, we evaluated the performance of different models, including the baseline model, ViT-GPT2, CoCa, and our proprietary model, CaptionWizard, using the HuggingFace BLEU and METEOR scoring systems. The scores for each model are presented in Table 5 for comparison. Additionally, we implemented an embedding feature using BERT to compute the embeddings of both the labels and predicted captions. We then compared these embeddings to measure similarity, providing a deeper insight into the performance of each model. The results of this comparison are also illustrated in Table 5.

Model	BLEU	METEOR	Embedding Similarity Score (using BERT)
ViT-GPT2 (baseline)	0.000114	0.031327	0.872778
CoCa (baseline)	0.000249	0.034501	0.875791
CaptionWizard (our model)	0.017658	0.162218	0.910315

Table 1: Evaluating model compared to baselines

The improved scores of CaptionWizard in BLEU, METEOR, and Embedding Similarity indicate its superior performance. The BLEU score, measuring the closeness of the generated captions to the ground truth, shows CaptionWizard's enhanced accuracy in likable caption generation. The higher METEOR score reflects better alignment with the syntactic and semantic qualities of the ground truth captions. Finally, the Embedding Similarity Score using BERT, which assesses the semantic similarity of the captions, suggests that CaptionWizard is more effective in capturing the nuanced meaning and context of the images, resulting in captions that are semantically richer and more closely aligned with the human-generated ground truth. Some of the generated captions by CaptionWizard are shown in Table 2.

6 CONCLUSION

In this work, we proposed an image captioning approach that specializes in creating realistically engaging captions for social media posts, moving beyond the descriptive, neutral captions generated from traditional image captioning methods. We leveraged an existing image captioning model, vit-image-captioning, and trained it on a real-world Instagram dataset that we deemed to be "likable" based on our likeability_score metric.



Image	Description
	<p>Base model caption: a woman in a black jacket and black boots standing on a snow covered slope</p> <p>CaptionWizard caption: i am not sure what i am going to do next but i do know that i will be spending a lot of time on my skis it is definitely going to be a busy day</p>
	<p>Base model caption: a plate of french fries and a hamburger</p> <p>CaptionWizard caption: i am loving this grilled cheese burger and it is ready to be eaten it is served in a waffle iron skillet it has a side of fries and mayo on it</p>

Table 2: Generated Outputs from CaptionWizard

7 FUTURE WORK

For future development, zero-shot learning offers the potential for classifying images without explicit labels, using knowledge from labeled data to identify unseen categories. Additionally, semi-supervised learning could utilize a few labeled examples to guide the classification of largely unlabeled datasets. These methods, combined with unsupervised techniques like clustering, could reveal natural data groupings for more efficient and adaptable image classification.

An additional area for future work involves developing techniques for creating captions that encapsulate a collection of images. This is particularly relevant for social media platforms where posts often include multiple images that collectively convey a narrative or a 'story'. Future research could focus on developing algorithms capable of understanding the thematic and narrative connections between a series of images and generating a unified, coherent caption that effectively represents the entire set. This approach would require a sophisticated understanding of visual storytelling and could significantly enhance the way we interact with and interpret multi-image posts on social media.

Incorporating user-provided context into image captioning can significantly enhance the process, allowing for personalized, relevant, and engaging captions. This user input, including event details, humor, or cultural references, bridges the gap between automated captioning and the subtleties of human storytelling.

AUTHORS CONTRIBUTIONS

This section lists the author contributions of each author.

- Haoran Tang: Developed classifier, generated CoCa captions, Developed the ViT-GPT2 model, implemented evaluation metric
- Morteza Damghani Nouri: Preprocessed the raw captions in the dataset, Implemented and trained an image caption generator using CNN and LSTM, Implemented and trained the fine-tuned model, Fine-tuned model evaluation
- Jeremy Jang: Initial processing and refinement of the raw dataset, implemented and trained CNN-LSTM image captioning model, and assessed CaptionWizard’s performance with BLEU and METEOR scores.
- Deepayan Sur: Data Collection, Data Cleaning, Data Preprocessing, Developed the Code Mix detection BiLSTM model using SentencePiece Embeddings, Implemented a version of CNN-LSTM model, Finetuned the ViT-GPT2 model and Developed the BERT Embedding Evaluation Metric

REFERENCES

- 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. 2021.
- Aparna Dutta. Word-level language identification using subword embeddings for code-mixed bangla-english social media data. 2021.
- Fanpage Karma. Top 10 most relevant topics on instagram. <https://blog.fanpagekarma.com/2019/04/18/top-10-most-relevant-topics-on-instagram/>, 2019.
- Seungbae Kim, Jyun-Yu Jiang, Masaki Nakada, Jinyoung Han, and Wei Wang. Multimodal post attentive profiling for influencer marketing. In *Proceedings of The Web Conference 2020*, pp. 2878–2884, 2020.
- Seungbae Kim, Jyun-Yu Jiang, Jinyoung Han, and Wei Wang. Influncerrank: Discovering effective influencers via graph convolutional attentive recurrent neural networks. 2023.
- Shuster Kurt, Humeau Samuel, Hu Hexiang, Bordes Antoine, and Weston Jason. Engaging image captioning via personality. In *Facebook AI Research*, 2019.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019. URL <http://arxiv.org/abs/1910.13461>.
- Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022a.
- Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. 2022b.
- T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.
- Zhou Luowei, Palangi Hamid, Zhang Lei, Hu Houdong, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019.

-
- Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. *arXiv*, 2017.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022.
- P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics*, volume 2, pp. 67–78, 2014.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. 2022.
- Zhang Zhongping, Chen Tianlang, Zhou Zheng, Li Jiaxin, and Luo Jiebo. How to become instagram famous: Post popularity prediction with dual-attention. *University of Rochester*, 2018.