engineering proceedings

MDPI

Proceeding Paper

Enhancing User Profile Authenticity through Automatic Image Caption Generation Using a Bootstrapping Language–Image Pre-Training Model [†]

Smita Bharne [1,2,*] **and Pawan Bhaladhare** [1]

[1] School of Computer Sciences and Engineering, Sandip University, Nashik 422213, India; pawan_bh1@yahoo.com
[2] Ramrao Adik Institute of Technology, D. Y. Patil Deemed to Be University, Navi Mumbai 400706, India
[*] Correspondence: smita146@gmail.com

[†] Presented at the International Conference on Recent Advances in Science and Engineering, Dubai, United Arab Emirates, 4–5 October 2023.

Abstract: Generating captions automatically for images has been a challenging task, requiring the integration of image processing and natural language processing techniques. In this study, we propose a system that focuses on generating captions for online social network users' profile images using a Bootstrapping Language–Image Pre-Training Model. Our approach leverages pre-training techniques, enabling the model to learn visual and textual representations from large datasets, which are then fine-tuned on a task-specific dataset. By utilizing this methodology, our proposed system demonstrates promising performance in generating captions for online social network users' profile images. The model effectively combines visual and textual information to generate informative and contextually relevant captions. This can greatly enhance user engagement and personalization on social media platforms, as users' profile images are accompanied by meaningful captions that accurately describe the content and context of the images. The proposed system shows its performance on the task of caption generation for online social network users' profile images. Furthermore, we show that our model can be used to identify scam (fake) profiles on online social networks by generating more accurate and informative captions for real profiles than for fake ones. By leveraging the power of pre-training and bootstrapping techniques, our model showcases its potential in enhancing user experiences, improving platform security, and promoting a more trustworthy online social environment. The proposed system has the potential to improve the authenticity and trustworthiness of user profiles on online social networks.

Keywords: automatic image caption generation; online social network user; bootstrapping language–image pre-training model (BLIP); fake profiles; pre-training techniques; profile verification

check for updates

Citation: Bharne, S.; Bhaladhare, P. Enhancing User Profile Authenticity through Automatic Image Caption Generation Using a Bootstrapping Language–Image Pre-Training Model. *Eng. Proc.* **2023**, *59*, 182. https://doi.org/10.3390/engproc2023059182

Academic Editors: Nithesh Naik, Rajiv Selvam, Pavan Hiremath, Suhas Kowshik CS and Ritesh Ramakrishna Bhat

Published: 18 January 2024

1. Introduction

Humans possess the innate ability to effortlessly describe the surroundings they find themselves in. With a mere glance at an image, humans are capable of providing a wealth of information about it. This fundamental skill has been the focus of many artificial intelligence researchers, who aspire to replicate humans' capacity to comprehend and interpret the visual world. Despite notable progress in areas such as object detection, features, image, scene, and action classification [1–3], the ability for computers to generate natural-sounding sentences that describe images remains a relatively new challenge [4,5]. Picture caption generation (PCG) is the process of writing a textual description or caption that adequately conveys the meaning of a picture. The algorithm examines the image first and then extracts details from it that are pertinent to the caption. Usually, convolutional neural networks (CNNs) are used for this. Following that, a recurrent neural network (RNN) model used for natural language processing (NLP) uses the features to produce a phrase that describes the image. The aim of picture caption generation is to produce precise and educational captions for a range of applications, including social networking, image ranking, virtual supporters, and assisting those with visual impairments [4,5]. Image caption generation is a complex task that involves several sub-tasks, including:

1. Object recognition: The method for recognizing and pinpointing any items in the image.
2. Recognition of attributes: The algorithm must be able to determine the extent, color, and shape of the objects in the image.
3. Image understanding: The methodology must be able to comprehend the relationships and context of the image's visual components.
4. Feature extraction: To create the caption, the computer extracts the pertinent features from the image.
5. Natural language processing: The algorithm creates a statement that represents the image that is both grammatically and semantically sound using natural language processing techniques.
6. Caption evaluation: The algorithm must assess the effectiveness of the generated caption. This is commonly done using metrics like BLEU (bilingual evaluation understudy), which uses metrics to measure bilingualism [6].

Thus, the picture (image) caption generation is a complex process. The ability to describe the visual world using natural language is crucial to many forms of human communication. In the literature, mostly machine learning (ML) and deep learning (DL) are used for image captioning [7]. In conventional machine learning methods, features are extracted from the input data. However, utilizing machine learning for generating image captions is not very efficient because it can be challenging to extract handcrafted features, especially when working with large, diverse, and complex datasets. However, over the past six to seven years, deep learning frameworks have been proposed for image captioning [8]. Recently, based on the DL model, the Vision-Language Pre-training (VLP) models are popular types of deep learning frameworks that have shown great success in image captioning. VLP models are trained on large-scale datasets that contain both images and their corresponding captions. During training, the VLP model learns to encode both visual and textual data features, allowing it to generate captions that accurately reflect the image's content [9]. User profile PCG in an online social network (OSN) is the challenge of automatically generating a description that describes the content of a user's profile photograph on a social network platform. This task can be challenging due to the diverse and often abstract nature of profile images, as well as the need to generate captions that are both informative and engaging to other users on the platform.

One approach to this task is to use deep-learning-based frameworks, such as encoder–decoder models or attention-based models, that have been developed for image captioning tasks. These models can be trained on a dataset of user profile images and their corresponding captions and then used to generate captions for new profile images. Overall, online social network user profile image caption generation is a promising area of research with many potential applications, including improving user engagement and user experience on social media platforms. Images are crucial in online social network platforms as they can significantly impact the number of people interested in a user's profile and assist in filtering potential interactions. However, this feature can also be exploited by criminals to target vulnerable users and expand their pool of potential victims. In this paper, we generate captions for the user profile images to identify the relevance of these captions belonging to the scammer profiles or real profiles.

2. Related Work

Until recently, there has been little effort toward generating descriptions for real-life images [10]. The initial research on image captioning follows a retrieval-based approach and a template-based approach. However, these methods had limitations on pre-existing subtitles in the training set or used fixed language structures, making them inflexible. Therefore, the generated descriptions were not expressive enough, leading to limitations in the quality of the output. Several authors proposed different approaches in the literature based on retrieval-based, template-based, deep-learning-based, encoder- and decoder-based formats. We summarize the related work based on these methods.

A strategy to deal with the problem of noisy visual estimation in image-captioned algorithms based on image retrieval was proposed by the author in [11]. According to visual similarity, their method combines a collection of described photos for a query image using a Conditional Random Field (CRF) to choose the image contents to be used for the image captioning. Following a sentence template, a description is then made using the results of this inference. An approach to provide rank-based sentences for a given image was proposed by the author in [12] by integrating sentence fragments and image fragments into a single common area. The dependency tree relations of a phrase are used to produce sentence fragments, while image fragments are obtained by using the detection results using CNN.

Until recently, there has been little effort toward generating descriptions for real-life images. The initial research on image captioning follows a retrieval-based approach and a template-based approach. However, these methods had limitations on pre-existing subtitles in the training set or used fixed language structures, making them inflexible. Therefore, the generated descriptions were not expressive enough, leading to limitations in the quality of the output. Several authors proposed different approaches in the literature based on retrieval-based, template-based, deep-learning-based, encoder- and decoder-based formats [13]. We summarize the related work based on these methods.

The development of neural network-based machine translation systems has been inspired by the encoder–decoder (ED) picture captioning model. This model utilizes a DL-based framework in which the decoder generates captions using the features extracted by the encoder from the input image [14]. Encoder–decoder (ED)-based image captioning models employ a DL-based framework. The decoder builds the corresponding caption using the features that the encoder used to extract the necessary features from the input image. Long Short-Term Memory (LSTM) and Recurrent Neural Networks were used by the researchers to encode textual data, while a deep CNN was used to encode visual data. Instead of only feeding picture data into the system during the initial stage, some authors included both image features and context word features into a sequential model at every time step, similar to how neural machine translation works [15]. The authors of [16,17] employed deep CNN-based picture encoding. The extracted visual features are then decoded to create a sentence description using Long Short-Term Memory (LSTM) and RNN. To enhance the pre-training procedure, authors in [18] introduced the BLIP (Bootstrapping Language-Image Pre-training) model, which integrates contrastive learning. This model combines encoders and decoders in many modes. It uses cross-modal pre-training and fine-tuning to enhance the efficiency of image captioning. The authors [19] utilize a lightweight Querying Transformer and employ a two-stage approach to narrow down the gap between modalities. The authors follow a two-stage approach in their methodology. Firstly, a frozen image encoder is utilized to pre-train the Querying Transformer for learning vision-language representations in the initial stage. Next, in the second stage, the Querying Transformer is pre-trained for vision-to-language generative learning, utilizing a frozen Large Language Model (LLM). The authors in [20,21] have presented their work based on CNN and LSTM with integration with ML algorithms. The authors in [22] presented their work of generating the captions for the text summarization technique using an ML-based pre-trained algorithm.

3. Bootstrapping Process for Language-Image Pretraining Model

To achieve pre-training of a unified vision-language model that combines comprehension and generation abilities, the Bootstrapping Language–Image Pre-training (BLIP) model introduces a multimodal encoder–decoder architecture. This architecture serves three key functions [19,20]:

1. Unimodal Encoders: The BLIP model employs separate encoders for processing images and text. The image encoder utilizes a vision transformer, while the text encoder is built upon BERT. To summarize this in a nutshell, a [CLS] token is inserted at the start of the text input.
2. Image-Grounded Text Encoder: This feature integrates visual information by adding a cross-attention layer between the self-attention layer and the feedforward network for each transformer block within the text encoder. A task-specific [Encode] token is appended to the text, and the output embedding of [Encode] serves as the multimodal representation for the image–text pair.
3. Image-Grounded Text Decoder: In the decoder part of the BLIP model, the bi-directional self-attention layers in the text encoder are replaced with causal self-attention layers. A special [Decode] token is employed to signify the start of a sequence.

By utilizing these functions, the BLIP model can effectively pre-train by jointly processing visual and textual information. This multimodal approach empowers the model to comprehend and generate image captions by capitalizing on both visual and linguistic cues.

4. Proposed Work

Our suggested method makes use of the BLIP model, a multimodal combination of encoder and decoder created specifically for picture-captioning tasks, which is a deep learning framework. An encoder and a decoder are the two fundamental parts of most image-captioning encoder–decoder configurations. The encoder collects relevant features from the source image, and the decoder then generates a caption that matches those features. The decoder frequently uses RNN to generate the caption word per word, whereas the encoder commonly employs CNN to extract high-level visual information from the input image. The dataset was constructed using actual profile photographs from datingnmore.com [23] and fraudulent profile images from scamdigger.com [24]. The raw data are subjected to data preparation processes in order to prepare the final dataset. The total number of profile images in the dataset is 12,000+ image profiles. The process is given in the following way.

1. The process begins by taking the user profile image and passing it through a CNN encoder, resulting in a feature vector represented as 'f' (as described in Equation (1)):

$$f = CNN(I) \tag{1}$$

2. Subsequently, this feature vector is provided as input to an RNN decoder for the step-by-step generation of the image caption. Equation (2) demonstrates how the RNN decoder computes the probability of the next word (y_t) in the caption based on the previous words, the hidden state at time 't - 1' (h_{t-1}), and the feature vector 'f' from the image:

$$P(y_t \mid y_1,\ldots,y_{t-1},f) = RNN(y_{t-1}, h_{t-1}, f) \tag{2}$$

3. The RNN decoder then generates the caption by selecting the word with the highest likelihood, considering the preceding words and the image features, as depicted in Equation (3):

$$y_t = argmax\ P(y_t \mid y_1,\ldots,y_{t-1},f) \tag{3}$$

This iterative process generates the complete image caption while considering the context of the previously generated words and the underlying image features.

These models can be trained on a dataset of user profile images and their corresponding captions and then used to generate captions for new profile images. Figure 1 shows the proposed flow of the BLIP caption generation for the OSN user profiles. The steps of execution are as follows:

1. Data Collection: A dataset is created by us for scam and real user profiles, for this study from scamdigger.com and datingnmore.com. After data collection, preprocessing is done.

2. Text Pre-training: Pre-train a language model on the textual profile data classification.
3. Image Pre-training: Convolutional neural networks (CNNs) should be trained beforehand on image data using ResNet algorithms.
4. Joint Pre-training: To make the system more accurate, the pre-trained language model and CNN are combined to create a joint pre-training model. This involves fine-tuning the combined model on a dataset that includes both image and text data. The goal is to leverage the pre-trained language model's ability to understand natural language and CNN's ability to extract visual features.
5. Transfer Learning: to train the model to extract visual features, to generate captions.

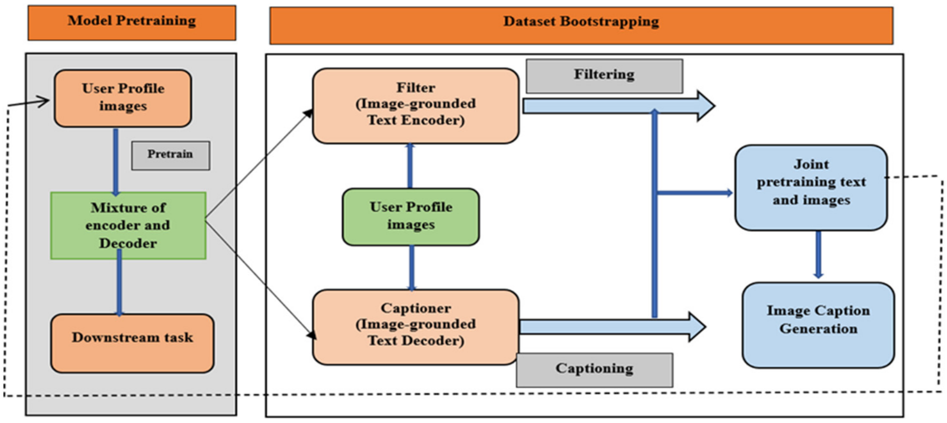


Figure 1. Framework of the proposed system.

Within the BLIP (Bootstrapping Language–Image Pre-training) process, an image is encoded into a fixed-length vector representation by leveraging its anticipated characteristics or classes. Simultaneously, a text encoder is employed to generate a fixed-length vector representation from textual descriptions. This representation is acquired through the prediction of text properties based on information from the images. The joint image–text encoder is subsequently trained using a dataset containing image–text pairs, a process aimed at enhancing its performance. This neural network takes both the image and textual description as input and produces a fixed-length vector representation as output. The principal objective of this joint encoder is to capture the semantic connection between the image and text, facilitating the generation of descriptive image captions. Notably, this approach finds utility in various applications, one of which is the generation of captions for user profile images. The algorithm for this process is outlined below.

Algorithm

Step 1: T = t1, t2,. . ., tm is a set of textual descriptions, and let I = I1, I2,. . ., In be a set of images. The pre-trained image encoder, pre-trained text encoder, and the jointly to-be-trained image–text encoder are represented by the variables f, g, and h, respectively.
Step 2: The BLIP process revolves around the core objective of training the joint image–text encoder 'h' to maximize the likelihood of the joint distribution between images and their corresponding textual descriptions. This can be mathematically expressed, as shown in Equation (4): Maximizing the sum of the logarithm of conditional probabilities for each pair of images and their related textual descriptions is the central aim:

$$\text{Maximize } \Sigma i = 1 \text{ to } n \ \Sigma j = 1 \text{ to } m \log P\ (tj \mid fi, gj) \quad (4)$$

where fi = f(Ii) is the encoded image representation, gj = g(tj) is the encoded textual representation, and P (tj | fi, gj) is the probability of the text explanation tj given the encoded image and textual representations.

Step 3: In order to learn the combined image–text encoder, h, the negative log-likelihood of the aforementioned objective is minimized, which is expressed as in Equation (5):

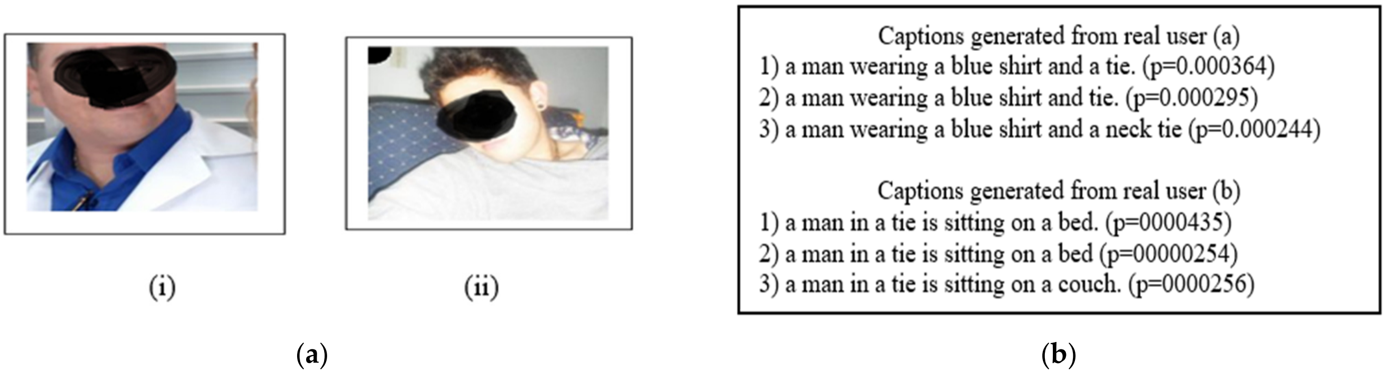
$$\text{Minimize-}\Sigma i = 1\hat{\ }n\ \Sigma j = 1\hat{\ }m \log P(tj \mid fi, gj) \quad (5)$$

In essence, the goal is to optimize the encoder to best capture the relationship between images and their associated text, ensuring a coherent and meaningful correspondence.

5. Results and Discussion

The model generates a description for each image in a profile that captures the underlying semantics of the image, based on its analysis. In Figures 2a and 3a, we can see examples of images belonging to real and scam profile images. The model then generates three possible descriptions for each image with a certain probability denoted as p. For the user's privacy purposes, we are not revealing the faces of any user belonging to either the scam or real profile category. Next, we show the complete captions generated with three probable descriptions with probability p. The captions generated from the real profile images are shown in Figure 2b, and the captions generated from the scam profile images are shown in Figure 3b, along with their probabilities. It is analyzed that the image captions generated from scam profiles illustrate their hobbies or personal interests, how scammers present themselves through their profile images. The caption classifier module is designed to analyze the textual content of image captions and determine the likelihood of a profile being a scam based on these features with machine learning algorithms. Based on generated captions, we explored Support Vector Machine (SVM), Random Forest, linear SVM, and XTree (Extra randomized trees) machine learning algorithms to classify the scam and real user profiles as one of the important features as a caption classifier. Figure 4a shows the graph visualization of captions classifier performance using an ML algorithm. Each algorithm contributes its unique approach to capturing patterns in the image captions and predicting the likelihood of fraudulent profiles. Table 1 shows the performance metrics for the caption classifier module.

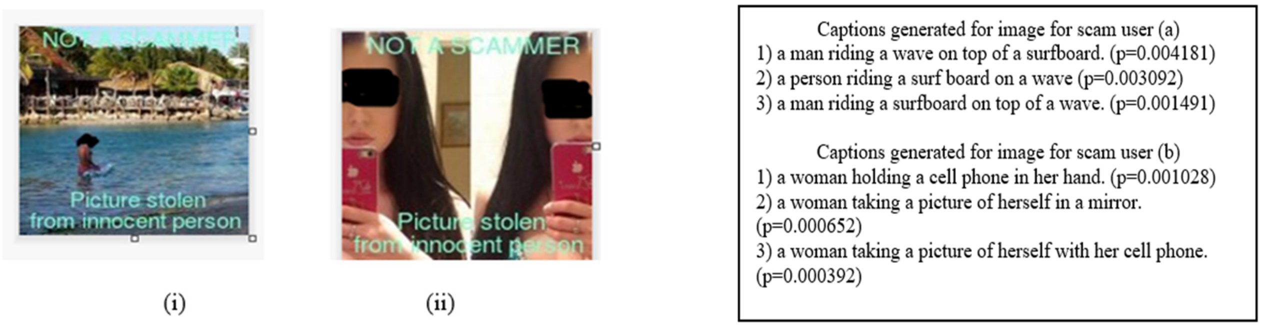
The development environment for the Microsoft Windows 10 (10.0) Professional 64-bit (Build 19045), AMD Athlon 3000G processor with Radeon Vega Graphics running at 2734.72 MHz, 2 cores, and 4 threads also utilized a GIGABYTE Technology NVIDIA GeForce RTX 3070 graphics card with 8GB DDR6 memory. We have used the programming language Python with PyCharm in an integrated development environment (IDE). Figure 4b shows the caption generation time per sample (batch size = 32 profile images) within the 5001 sample images in the train dataset. The average time per sample is 0.0.2 4 s per image, demonstrating the system's efficiency.



Captions generated from real user (a)
1) a man wearing a blue shirt and a tie. (p=0.000364)
2) a man wearing a blue shirt and tie. (p=0.000295)
3) a man wearing a blue shirt and a neck tie (p=0.000244)

Captions generated from real user (b)
1) a man in a tie is sitting on a bed. (p=0000435)
2) a man in a tie is sitting on a bed (p=00000254)
3) a man in a tie is sitting on a couch. (p=0000256)

Figure 2. (**a**) Images belonging to real user profiles (i) and (ii); (**b**) Captions generated from the real user profile images.


NOT A SCAMMER
Picture stolen from innocent person
NOT A SCAMMER
Picture stolen from innocent person

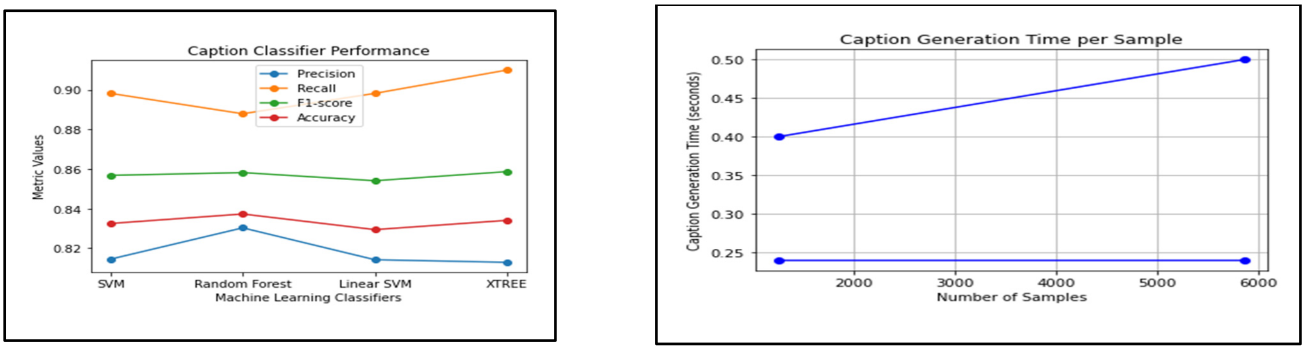
(i) (ii)

Captions generated for image for scam user (a)
1) a man riding a wave on top of a surfboard. (p=0.004181)
2) a person riding a surf board on a wave (p=0.003092)
3) a man riding a surfboard on top of a wave. (p=0.001491)

Captions generated for image for scam user (b)
1) a woman holding a cell phone in her hand. (p=0.001028)
2) a woman taking a picture of herself in a mirror. (p=0.000652)
3) a woman taking a picture of herself with her cell phone. (p=0.000392)

(**a**) (**b**)

Figure 3. (**a**) Images belonging to scam user profiles (i) and (ii); (**b**) Captions generated from the scam user profile images.



(**a**) (**b**)

Figure 4. (**a**) Caption classifier performance; (**b**) Caption generation time per sample.

Table 1. Performance Metrics for Caption Classification.

Caption Classifier	Precision	Recall	F1 score	Accuracy
SVM	0.8145	0.8982	0.8568	0.8325
Random Forest	0.8303	0.8879	0.8582	0.8373
Linear SVM	0.8142	0.8982	0.8541	0.8294
XTree	0.8129	0.91	0.8587	0.8341

Usually, the image is pre-processed before being fed into a CNN to extract characteristics. An RNN is then used to create a caption from it. The links between words and their contexts are taught to the RNN using a sizable corpus of text data. In order to determine whether the generated caption is real or fraudulent, it is compared to the actual captions of the profile image. Therefore, the researchers will undoubtedly benefit from our automatic image caption production.

6. Conclusions and Future Work

Automatic picture caption generation (APCG) is a complex topic that requires the fusion of several different methodologies. Recent development in the DL domain with the BLIP model shows prominent results in image caption generation. However, it has been observed that the solutions based on deep learning techniques have shown better performance. The goal of OSN user profile image caption generation for identifying fake profiles is to automatically generate captions that accurately describe the profile image and are consistent with the user's other profile information. This can help users and platform administrators identify fake profiles and reduce the spread of misinformation on social media.

Overall, online social network user profile image caption generation is a promising area of research with many potential applications, including improving user engagement and user experience on social media platforms. APCG using a bootstrapping language–image pre-training model for user profile images has several potential future applications and areas for further research. Some are listed below:

1. Improved User Experience: Automatic image caption generation could improve user experience by making it easier for users to search for and find relevant content on social media platforms. By generating informative and accurate captions for user profile images, users can better communicate their interests and preferences to others, leading to more meaningful interactions and connections.
2. Multimodal Understanding: an approach to data analysis that combines different types of data, such as text and image data, to gain a more comprehensive understanding of a user or a system. In addition to text and image data, future research could explore ways to integrate other types of data, such as audio or video, to create even more robust multimodal models. By incorporating a wider range of data types, these models could capture the full richness of user profiles and provide a more complete understanding of users and their behaviors.
3. Personalization: APCG could be used to personalize recommendations for users based on their interests and preferences. By analyzing the captions generated for user profile images, platforms could better understand what types of content users are interested in and tailor their recommendations accordingly.
4. Accessibility: APCG could improve accessibility for users with visual impairments or other disabilities. By generating informative and accurate captions for user profile images, platforms could make their content more accessible to a wider range of users.
5. Ethical Considerations and Potential Bias with APCG: As with any technology that uses personal data, there are ethical considerations to be addressed. APCG for user profiles raises ethical concerns regarding privacy, potential biases, accuracy, and transparency. Privacy must be protected by obtaining user consent and adhering to data protection regulations. Biases and discriminatory outputs should be addressed through diverse training data and fairness-aware algorithms. Ensuring the accuracy and reliability of generated captions is crucial to avoid misrepresentation or dissemination of false information. Future research could explore ways to ensure that user privacy and data security are maintained when generating captions for user profile images and to mitigate the risk of biases or stereotypes being introduced into the caption generation process.
6. Challenges in handling diverse user profiles and noisy image content: Noisy or low-quality images can pose a challenge for the system, as it may struggle to extract meaningful features and context from such images. This can lead to inaccurate or irrelevant captions that do not effectively describe the profile image. To overcome this problem, we can use attention mechanisms in the caption generation models to improve the model's ability to focus on relevant image regions and extract important features, even in the presence of noisy or complex visual content.

Author Contributions: Conceptualization, S.B.; methodology, S.B.; validation, S.B. and P.B.; writing—original draft preparation, S.B.; writing—review and editing, S.B. and P.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can be obtained from the corresponding author on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gan, C.; Yang, T.; Gong, B. Learning attributes equals multi-source domain generalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 87–97.
2. Maji, S.; Bourdev, L.; Malik, J. Action recognition from a distributed representation of pose and appearance. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 3177–3184.
3. Chao, Y.W.; Wang, Z.; Mihalcea, R.; Deng, J. Mining semantic affordances of visual object categories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4259–4267.
4. Huang, C.Y.; Hsu, T.Y.; Rossi, R.; Nenkova, A.; Kim, S.; Chan, G.Y.Y.; Koh, E.; Giles, L.C.; Huang, T.-H.K. Summaries as Captions: Generating Figure Captions for Scientific Documents with Automated Text Summarization. *arXiv* **2023**, arXiv:2302.12324.
5. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.* **2019**, *51*, 118.
6. Jiang, M.; Huang, Q.; Zhang, L.; Wang, X.; Zhang, P.; Gan, Z.; Gao, J. Tiger: Text-to-image grounding for image caption evaluation. *arXiv* **2019**, arXiv:1909.02050.
7. Wang, S.; Yao, Z.; Wang, R.; Wu, Z.; Chen, X. Faier: Fidelity and adequacy ensured image caption evaluation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14050–14059.
8. Liu, X.; Xu, Q.; Wang, N. A survey on deep neural network-based image captioning. In *The Visual Computer*; Springer Nature: Berlin/Heidelberger, Germany, 2018. [CrossRef]
9. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
10. Ordonez, V.; Kulkarni, G.; Berg, T.L. Im2Text: Describing images using 1 million captioned photographs. In Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11), Granada, Spain, 12–15 December 2011; Curran Associates Inc.: Red Hook, NY, USA, 2011; Volume 24, pp. 1143–1151.
11. Soh, M. *Learning CNN-LSTM Architectures for Image Caption Generation*; Stanford University: Stanford, CA, USA, 2016.
12. Hossain, M.Z. Deep Learning Techniques for Image Captioning. Ph.D. Thesis, Murdoch University, Perth, Australia, 2020.
13. Yi, J.; Wu, C.; Zhang, X.; Xiao, X.; Qiu, Y.; Zhao, W.; Hou, T.; Cao, D. MICER: A pre-trained encoder–decoder architecture for molecular image captioning. *Bioinformatics* **2022**, *38*, 4562–4572. [CrossRef] [PubMed]
14. Xiao, F.; Xue, W.; Shen, Y.; Gao, X. A New Attention-Based LSTM for Image Captioning. *Neural Process. Lett.* **2022**, *54*, 3157–3171.
15. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
16. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. *Int. Conf. Mach. Learn.* **2015**, *37*, 2048–2057.
17. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 652–663. [PubMed]
18. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
19. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv* **2023**, arXiv:2301.12597.
20. Predić, B.; Manić, D.; Saračević, M.; Karabašević, D.; Stanujkić, D. Automatic image caption generation based on some machine learning algorithms. *Math. Probl. Eng.* **2022**, *2022*, 4001460.
21. Sasibhooshan, R.; Kumaraswamy, S.; Sasidharan, S. Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction. *J. Big Data* **2023**, *10*, 18.
22. Gaurav; Mathur, P. A Survey on Various Deep Learning Models for Automatic Image Captioning. *J. Phys. Conf. Ser.* **2022**, *1950*, 012045. [CrossRef]
23. Available online: https://datingnmore.com/ (accessed on 5 March 2023).
24. Available online: http://scamdigger.com/ (accessed on 5 March 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.