

PAPER • OPEN ACCESS

Reproducing “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”

To cite this article: Haixia Liu and Tim Brailsford 2023 *J. Phys.: Conf. Ser.* **2589** 012012

View the [article online](#) for updates and enhancements.

You may also like

- [Identifying dominant industrial sectors in market states of the S&P 500 financial data](#)

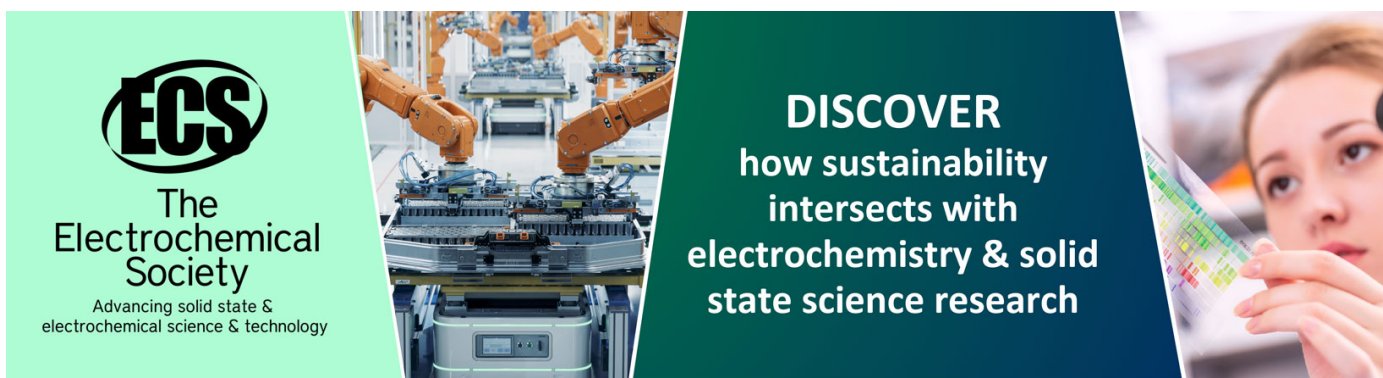
Tobias Wand, Martin Heßler and Oliver Kamps

- [Ultrasound imaging based recognition of prenatal anomalies: a systematic clinical engineering review](#)

Natarajan Sriraam, Babu Chinta, Seshadhri Suresh et al.

- [SN-SAE: a new damage diagnosis method for CFRP using Lamb wave](#)

Zhiyong Li, Zhiyong Wang, Yong Li et al.



ECS
The
Electrochemical
Society
Advancing solid state &
electrochemical science & technology

DISCOVER
how sustainability
intersects with
electrochemistry & solid
state science research

Reproducing "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention"

Haixia Liu, Tim Brailsford

Department of Computer Science and Creative Technology, University of the West of England (UWE Bristol), Frenchay Campus, Coldharbour Lane, Bristol, BS16 1QY, UK

E-mail: Haixia.Liu@uwe.ac.uk, Tim.Brailsford@uwe.ac.uk

Abstract. This paper replicates the experiment presented in the work of Xu et al. [1], and examines errors in the generated captions. The analysis of the identified errors aims to provide deeper insight into the underlying causes. This study also encompasses subsequent experiments aiming at investigating the feasibility of rectifying these errors via a post-processing stage. Image recognition and object detection models, as well as a language probability computational model were explored. The findings presented in this paper aim to contribute towards the overarching objective of Explainable Artificial Intelligence (XAI), thereby providing potential pathways to improve image captioning.

1. Introduction

Image captioning (IC) is a complex task requiring algorithms to generate concise textual descriptions of an image's content. IC is crucial for achieving comprehensive scene understanding and has applications in healthcare, education, and a wide variety of fields that involve the interpretation of images. Zhang et al. [2] demonstrated IC's potential in robot-enhanced therapy for children with autism spectrum disorder by exploring differing combinations of ResNet101 and word embedding schemes. Liu et al. [3] researched IC's application in construction activity scenes, and demonstrated the feasibility of integrating modules including: Convolutional Neural Networks (CNNs), e.g. VGG-16 and ResNet-50; Recurrent Neural Networks (RNNs), e.g. LSTM and techniques such as word embeddings with domain-specific datasets. The success of these applications depends on enhancing algorithms by acquiring a deep understanding of the underlying issues. Image captioning is a problem that requires techniques from various different areas of Computer Science including Computer Vision (CV) and Natural Language Processing (NLP). Challenges include refining the pre-processing steps, improving the CV and NLP models, multimodal integration, and evaluation. Researchers have attempted to enhance current methods, but there is still a need for error analysis of the outcomes. In this study, we perform error analysis on the output of the "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" (SAT) model, and we investigate possible reasons for errors by comparing different computer vision models in a case-study approach.



2. Related Work

2.1. Overview of Image Captioning Methods

Image captioning methods, essential in bridging the gap between visual content and natural language, have been categorized in various ways. Yao et al. [4] classified them into three primary types: template-based, search-based, and language model-based. Template-based methods [5] generate sentences by aligning sentence fragments, such as subject, verb, and object, with words detected from the image content. They then use predetermined language templates to create a coherent sentence. Although this can be quite effective, the results are heavily influenced by the sentence templates used. This limits their flexibility, and the generated captions might fail to capture the diverse and creative nature of human language. Search-based methods (such as [6]) employ models for image and sentence matching that utilize an intermediate meaning space between sentence and image spaces. By mapping both images and sentences to this shared space, the model can evaluate similarity and establish connections between them. The authors of this approach acknowledged their sentence model's oversimplification and suggested that an iterative procedure for deeper exploration of sentences and images might generate more useful results. This strategy could potentially reveal subtler connections between visual and textual information. Language model-based methods aim to learn the probability distribution in the shared space of visual content and textual sentences. This learning process enables the generation of novel sentences with adaptable syntactical structures, providing greater flexibility and creativity compared to template-based methods. Vinyals et al. [7] utilized an end-to-end neural network architecture with Long Short-Term Memory (LSTM) to generate a single sentence for a given image. This approach combines the strengths of deep learning and natural language processing to create more accurate and relevant captions. Additionally, Xu et al. [1], whose work we aim to replicate, implemented an attention mechanism, allowing the algorithm to focus on specific regions when generating corresponding words. This advancement further improves the model's ability to generate meaningful and contextually accurate captions.

2.2. Contribution

We begin by reproducing Xu et al.'s [1] language model-based work, analysing the errors in the generated captions, and identifying potential areas for improvement. Subsequently, we explore possible enhancements to the results by incorporating aspects of template-based and search-based methods into the existing framework. By combining the strengths of different approaches, we aim to develop a more robust and comprehensive image captioning model that can generate accurate, contextually relevant, and creative captions for a wide range of images.

3. Error Analysis

In error analysis, We aim to answer the following questions:

- What are the problems with the generated captions?
- Why do these problems occur?
- How can we address the issues and improve the quality of the output?

To investigate these questions, we will analyse the errors in the generated captions through manual analysis, automated analysis, and empirical experiments based on case studies

3.1. Manual Analysis

We classify the problems in the resulted captions into the following error-categories: object detection, action recognition, relations between object, place detection, facial expression detection, gender detection and Syntactic errors. We manually examined 143 cases and labeled each of them with one or multiple of the error-categories. The top three errors are: incorrect object, incorrect action, and omitted object. The less significant error is facial emotion detection.

Figure 1 shows the distribution of error categories as a bar chart, where each bar represents the count of instances of a particular error type. Note that these are absolute counts and do not consider the proportion of each error type's count among the total count of instances for each topic.

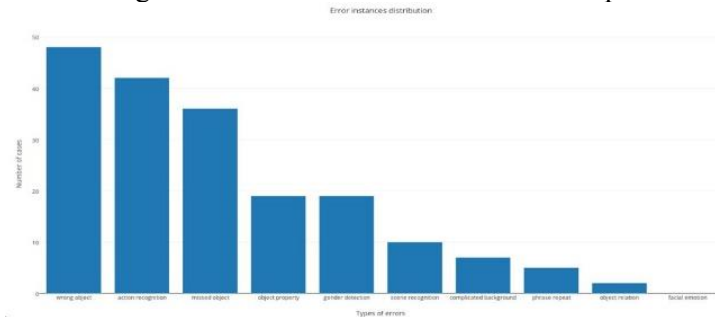


Figure 1 Distribution of error categories, covering Wrong object(incorrect object), action recognition(incorrect action), missed object(omitted object), object property, gender detection, scene recognition, complicated background, phrase repeat (Syntactic errors), object relation, facial emotion.

3.2. Automatic Analysis

After performing stemming on all image captions, NLP techniques are utilized to extract nouns and verbs from the captions. We define an incorrect noun/verb as a noun/verb that appears in the generated caption but not in the reference caption. The top incorrect nouns and verbs are shirt, man, woman, bench, field; stand, jump, run. The bar charts in Figure 2a and 2b illustrate the most commonly occurring nouns and verbs respectively in the generated captions, which are not present in the referenced captions.

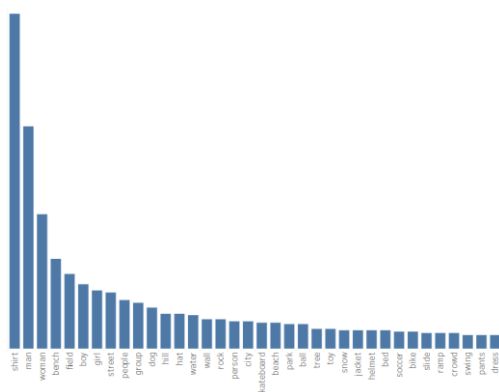


Figure 2a. Error occurrences of automatically detected nouns.

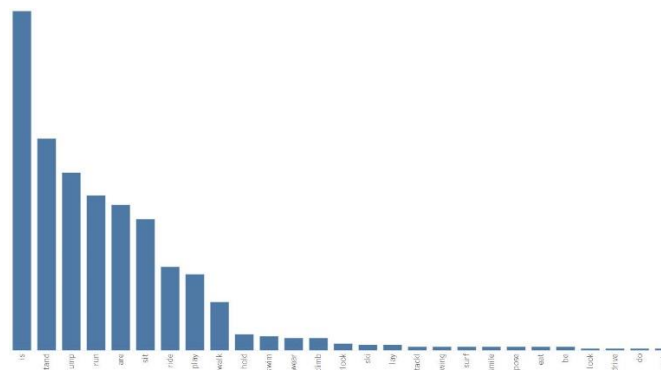


Figure 2b. Error occurrences of automatically detected verbs.

Examples with incorrect nouns and verbs are displayed in Figure 3a and Figure 3b.



Figure 3a. Examples of incorrect nouns.



Figure 3b. Examples of incorrect verbs.

3.3. Potential Causes of the Errors

The top two errors in Figure 1 suggest that there is room for improvement from an object detection and action recognition perspective. We also bear in mind that these errors may have been caused by the way the multimodal system was designed. According to Zhou et al. [8], the text input used in pre-trained vision-language models appears to significantly affect their performance on downstream datasets. For example, even adding the word “a” before the class token can result in an accuracy improvement of over 5%. The first example shown in Figure 3 indicated that the image recognition model has difficulty recognizing human faces if the rotation of the image has resulted in the human face not appearing in a straight, upright position. Additionally, the model may confuse human hair with dog hair if their colors and textures are similar. Given the possible reasons, we have developed the following questions:

- What predictions can be made by the VGG model used in the SAT paper without considering the encoded captions?
- What predictions can be made by state-of-the-art object detection models?
- What are the results of training the model using the transformed data by typed dependency parser?

4. Efforts to Reduce Error

Drawing inspiration from both template-based and search-based methods, we propose a framework (shown in Figure 4) aimed at correcting errors by post-processing generated captions. In this framework, generated captions and their corresponding image are processed in parallel, resulting in different merged sets of nouns and verbs (S_c for caption, S_i for image). Using all combinations of nouns and verbs from the generated sentence ‘template’, sentence scores are computed and compared, fulfilling the final component in the top-right of this framework. The sentence with the highest score will be chosen as the final caption. In this framework, VGG can be replaced by other computer vision models. ‘IR’ stands for Information Retrieval.

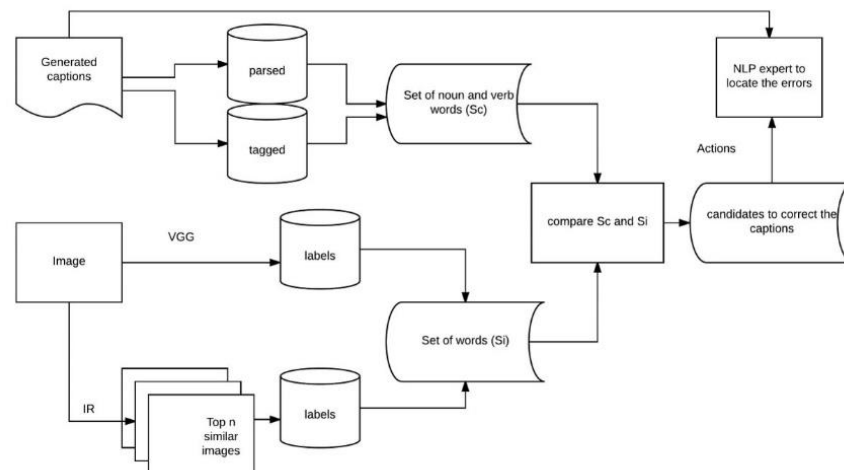


Figure 4. Framework for rectifying errors.

4.1. The Framework in Action: A Case Study

The framework was applied in a case study using the example illustrated in Figure 5.



Figure 5. An example of incorrect image caption generation in red.

The five sentences above the generated caption are referenced captions by human.

First, the generated captions are processed using part-of-speech (POS) tagging, resulting in set of nouns and verbs. We merge these two sets into one set (S_c): A/DT, baby/N N, with/IN, a/DT, baby/N N, in/IN, his/P RP, mouth/N N. The image is then sent to object detection model (e.g.: using VGG [9] or YOLO [10]), resulting in set of nouns and verbs. We merge the nouns and verbs to one set: S_i . Note that for this simple case study, we set the n in ‘Top n similar’ images to 1. The outputs from VGG19 are ‘ear’, ‘spike’, ‘capitulum’. The outputs from YOLO using coco.name are ‘person’, ‘baseball glove’. We can see that none of the models correctly predict ‘plant’, ‘grass seeds’, ‘wheat grass’ from the image. We then use the simple algorithm 1 (shown in Figure 6) attempt to find the best caption. The term ‘satCap’ refers to the captions generated by the algorithm described in the SAT (Show Attend Tell) paper that we are replicating.

Algorithm 1 Correct The Caption

```

SetPairCapSentScore  $\leftarrow \emptyset$            ▷ Set of pairs of captions and their corresponding scores
scoreC  $\leftarrow$  compute_sentence_score(satCap)
SetPairCapSentScore  $\leftarrow < \text{satCap}, \text{scoreC} >$ 
for each  $c$  in  $S_c$  do
  for each  $i$  in  $S_i$  do
    newCap  $\leftarrow$  Replace_c_with_i( $c, i$ )
    newScoreI  $\leftarrow$  compute_sentence_score(newCap)
    SetPairCapSentScore  $\leftarrow < \text{newCap}, \text{newScoreI} >$ 
return arg(max_score(SetPairCapSentScore))

```

Figure 6. An algorithm to correct the caption.

There are several ways to compute sentence score. The results using Transformer [11] are shown in Table 1 (higher scores indicate more accurate outcomes). For example: the score of the sentence ‘I love cats’ is 1547.05. The score of ‘A baby with a grass in his mouth’ is 168.2. The score of ‘A baby with a baby in his mouth’ is 59.67.

Table 1. Examples of Sentence Score Calculated by Transformer (GPT2LMHeadModel, GPT2Tokenizer): The top 3 sentences are parts of the caption-corrections based on the generated captions and the bottom 2 are the referenced ones generated by human.

Captions	Sentence Score
A ear with a baby in his mouth	122.88
A baby with a ear in his mouth	108.77
A baseball glove with a baseball glove in his mouth	61.24
A baby has some grass seeds in his mouth	158.49
A baby sticking wheat grass into his mouth	713.00

We can see from the table that the first sentence received a relatively higher score. However, the meaning of that sentence does not make sense and it contains grammatical issues.

4.2. Typed Dependency

The motivation for transforming the raw captions into dependency-words is to capture the relation between two words with a certain distance, hoping to avoid the errors caused by language model. To verify if converting the original sentences into typed-dependency [12] format, experiments were conducted by retraining the model from scratch utilizing the converted captions. An example of transformed caption and the generated caption were shown in Figure 7.



Original caption: A bald man is attempting to slam dunk a basketball in a game while people in the stands watch .

Transformed caption (dependency-words are colored):
det-man-a amod-man-bald nsubj-attempting-man nsubj-
slam-man aux-attempting-is root-root-attempting aux-slam-
to xcomp-attempting-slam dep-slam-dunk det-basketball-a
dobj-slam-basketball det-game-a prep_in-basketball-game
mark-watch-while nsubj-watch-people det-stands-the
prep_in-people-stands advcl-slam-watch .

Generated caption:

[illegible]

Figure 7. The original referenced caption and the transformed caption using typed dependency parser are shown on the top. The generated caption based on the transformed caption is demonstrated on the bottom. The generated caption based on the original captions: ‘ A man in a white shirt is jumping to a man in a white shirt.’

5. Discussion

We replicated the SAT model and analysed results both manually and automatically, investigating errors through case studies. Our caption-correction framework and algorithm, inspired by the template-based approach, leverage SAT’s smooth “templates” and incorporate computer vision and language models separately in an attempt to generate more useful outputs. The algorithm’s effectiveness depends on object, action detection, and language models’ performance, with hopes that advanced models will enhance results. Captions generated using the typed-dependency method frequently contain ‘det-shirt-a’. After converting the original captions to typed-dependency based ‘words’, the vocabulary size was increased from 28,417 to 317,119. Both ‘shirt’ and ‘det-shirt-a’ are high-frequency words. It appears that the SAT model is influenced by word frequency. For future work we propose training the model in a tree manner (root to leaf) instead of linearly (left to right in a sentence), to test whether that provides better results. This experiment may also reveal the algorithm’s difficulty in distinguishing between words like “shirt” and “outfit”, possibly caused by an imbalanced dataset with fewer images labelled “outfit” than “shirt”. It is worthwhile to explore and compare different natural language processing (NLP) techniques, such as those described in [13, 14]. We hope that insights gained from such a study will be valuable and useful to the community.

6. References

- [1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in International conference on machine learning. PMLR, 2015, pp. 2048–2057.
- [2] B. Zhang, L. Zhou, S. Song, L. Chen, Z. Jiang, and J. Zhang, “Image captioning in chinese and its application for children with autism spectrum disorder,” in Proceedings of the 2020 12th International Conference on Machine Learning and Computing, 2020, pp. 426–432.

- [3] H. Liu, G. Wang, T. Huang, P. He, M. Skitmore, and X. Luo, “Manifesting construction activity scenes via image captioning,” *Automation in Construction*, vol. 119, p. 103334, 2020.
- [4] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, “Boosting image captioning with attributes,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4894–4902.
- [5] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, and H. Daumé III, “Midge: Generating image descriptions from computer vision detections,” in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 747–756.
- [6] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, “Every picture tells a story: Generating sentences from images,” in *Computer Vision—ECCV 2010: 11th European Conference on Computer Vision*, Heraklion, Crete, Greece, September 5–11, 2010, *Proceedings, Part IV* 11. Springer, 2010, pp. 15–29.
- [7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.
- [8] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [9] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz et al., “Huggingface’s transformers: State-of-the-art natural language processing,” *arXiv preprint arXiv:1910.03771*, 2019.
- [12] M.-C. De Marneffe, B. MacCartney, C. D. Manning et al., “Generating typed dependency parses from phrase structure parses,” in *Lrec*, vol. 6, 2006, pp. 449–454.
- [13] A. S. Girsang and F. J. Amadeus, “Extractive text summarization for indonesian news article using ant system algorithm,” *Journal of Advances in Information Technology*, vol. 14, no. 2, 2023.
- [14] K. Kuppusamy and G. Aghila, “Segmentation based, personalized web page summarization model,” *Journal of Advances in Information Technology*, vol. 3, no. 3, pp. 155–161, 2012.

Acknowledgement

This work is part of AIPHES project funded by German Research Foundation, and was conducted in collaboration with the UKP Lab in the Technical University of Darmstadt.