Knowledge-Augmented Figure Caption Generation

Zhishen Yang[1], Raj Dabre[2], Hideki Tanaka[2], Naoaki Okazaki[1]
School of Computing, Tokyo Institute of Technology[1]
National Institute of Information and Communications Technology [2]
zhishen.yang@nlp.c.titech.ac.jp {raj.dabre, hideki.tanaka}@nict.go.jp
okazaki@c.titech.ac.jp

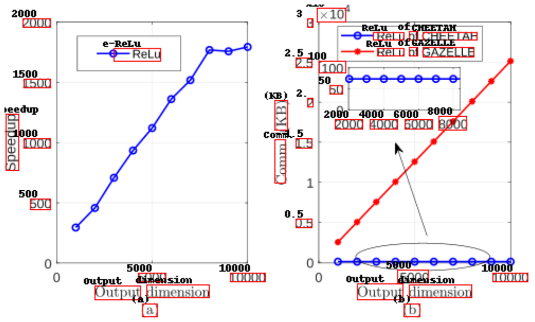
Abstract

In scholarly documents, figures are mediums for communicating scientific findings to readers in a straightforward way. Automating the generation of figure captions helps authors write informative captions that make communicating scientific findings easier to understand. In this study, figure captioning is treated as a knowledge-enhanced captioning task. To this end, we create SciCap+ by extending the large-scale SciCap dataset [1] with mention-paragraphs (paragraphs mentioning figures) and OCR tokens. Experimental results show that the mentions-paragraphs, as additional contextual knowledge, significantly improve the caption generation quality compared to the figure-only baseline. Datasets and models will be publicly released.

1 Introduction

Figures in scientific papers provide visual representations of complex information that help to share scientific findings with readers efficiently and straightforwardly. The standard practice for scientific writing is to write a caption for each figure, accompanied by paragraphs with detailed explanations. Helping authors write appropriate and informative captions for figures will improve the quality of scientific documents, thereby facilitating scientific communication. In this study, we focus on automating the generation of captions for figures in scientific papers.

Scientific figure captioning is a variant of the image captioning task with two unique challenges: 1. Figures are not natural images: In contrast to natural images, visual objects are texts and data points in scientific figures. 2. The captions of the figures should explain: Instead of simply identifying objects and texts in the figures, the caption should contain an analysis that the authors intend to present and highlight findings.



Caption:
Fig. 7. (a) Speedup of CHEETAH over GAZELLE for computing ReLu. (b) Comparison of communication cost for ReLu.

Mention-paragraph:
Fig. 7 plots the speedup and communication cost as a function of the output dimension. Similarly, CHEETAH achieves an outstanding speedup with much smaller communication cost, independent of the output dimension, compared with GAZELLE.
……

Figure 1 An example figure [2] with its captions and mention-paragraph and the text tokens recognized via OCR. Without referring to the mention-paragraph and the OCR tokens to tie the figure and the mention, we cannot have a proper interpretation of the data presented in the figure, which is communication cost comparison and speed up of CHEETAH over GAZELLE.

The previous study [1] defines the scientific figure captioning task as the figure-to-caption task. Their experiments report relatively lower automatic evaluation scores, indicating considerable room for improvement. Intuitively, humans need to have sufficient background knowledge to interpret figures. As figure 1 shows, only by looking at the figure, we do not know what “comm.(KB)” stands for; therefore, lacking the knowledge to write informative captions is challenging. However, the mention-paragraph contains “communication cost” and this is also present in the caption, indicating that such background knowledge should help in writing accurate captions.

Based on above observations, we hypothesized that gen-

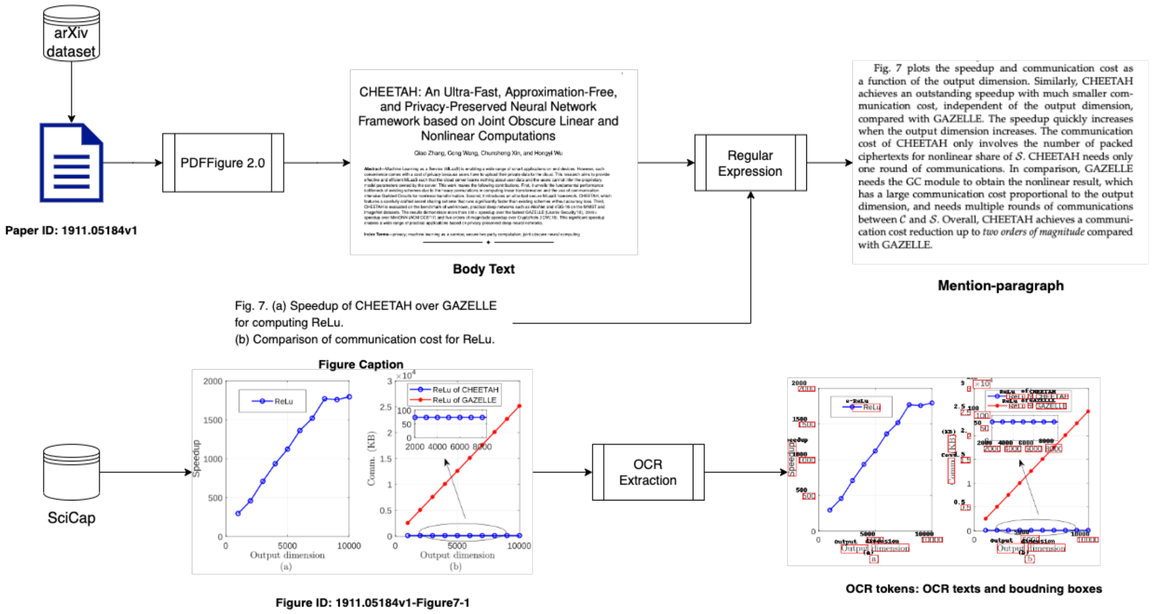

arXiv dataset
Paper ID: 1911.05184v1
PDFFigure 2.0
CHEETAH: An Ultra-Fast, Approximation-Free, and Privacy-Preserved Neural Network Framework based on Joint Obscure Linear and Nonlinear Computations
Qiao Zhang, Cong Wang, Chunsheng Xin, and Hongyi Wu
Body Text
Regular Expression
Mention-paragraph
Fig. 7. (a) Speedup of CHEETAH over GAZELLE for computing ReLu.
(b) Comparison of communication cost for ReLu.
Figure Caption
SciCap
Speedup
ReLu
Output dimension
(a)
Comm. (KB)
ReLu of CHEETAH
ReLu of GAZELLE
Output dimension
(b)
Figure ID: 1911.05184v1-Figure7-1
OCR Extraction
OCR tokens: OCR texts and boudning boxes

Fig. 7 plots the speedup and communication cost as a function of the output dimension. Similarly, CHEETAH achieves an outstanding speedup with much smaller communication cost, independent of the output dimension, compared with GAZELLE. The speedup quickly increases when the output dimension increases. The communication cost of CHEETAH only involves the number of packed ciphertexts for nonlinear share of $\mathcal{S}$. CHEETAH needs only one round of communications. In comparison, GAZELLE needs the GC module to obtain the nonlinear result, which has a large communication cost proportional to the output dimension, and needs multiple rounds of communications between $\mathcal{C}$ and $\mathcal{S}$. Overall, CHEETAH achieves a communication cost reduction up to *two orders of magnitude* compared with GAZELLE.

Figure 2 The overall workflow of the data augmentation for creating the SciCap+ dataset. For each figure in the SciCap+, we extracted its mention-paragraphs and OCR tokens (OCR texts and bounding boxes).

erating appropriate captions is infeasible without adding context knowledge to the caption generation model. This context takes two forms: background knowledge of running text and OCR tokens in a figure, which should help provide additional context to the model. We then presented scientific figure captioning as a multimodal summarization task and used the M4C captioning model [3] (a model that uses multimodal knowledge to generate captions) as a starting point to investigate the scientific figure captioning task. The experimental result of automatic evaluation demonstrates that using knowledge embedded across modalities, especially in the form of mention-paragraphs and OCR tokens, significantly boosts performance.

2 Problem Formulation

[1] pose scientific figure captioning as an image captioning task as: Given a figure I, the model generates a caption $C = [c_0, c_1, ..., c_N]$. However, we reframe this task as a knowledge-augmented image captioning task that requires knowledge extracted from text and vision modalities. Given a scientific figure I and knowledge extracted from text and vision modality: $K_{text} = (M_t, O_t)$, and $K_{vision} = (I_v, O_v, O_p)$, where M_t and O_t are text features extracted from paragraphs that mention figures (mention-paragraphs) and OCR texts. I_v and O_v are visual features obtained from figures and OCR texts. O_p represents locations of OCR texts in figures. We define the figure caption generation task as modelling the conditional probability: $P(C|I, K_{text}, K_{vision})$.

Split	Figures	Words
Training	394,005	12,336,511
Test	10,336	323,382
Validation	10,468	329,072

Table 1 Statistics of the SciCap+ dataset.

3 SciCap+ Dataset

SciCap is a large-scale figure-caption dataset comprising graph plots extracted from 10 years of collections of arXiv computer science papers. We used around 414k figures from SciCap and augment each figure with its mention-paragraphs and OCR tokens with metadata. This section details the data set creation and data augmentation processes. Figure 2 shows the overall workflow behind the creation of the SciCap+.

Data Statistics We split figures at the document level and kept all original captions and figures (graph plots, with/without sub-figures). For a figure, we kept only the first paragraph that mentions it in the body text. Table 1 shows statistics of the SciCap+ dataset.

Mention-paragraph Extraction We first obtained papers in PDF format from Kaggle arXiv dastaset [1]. The reason for using PDFs is that not all papers have source

[1] https://www.kaggle.com/datasets/Cornell-University/arxiv

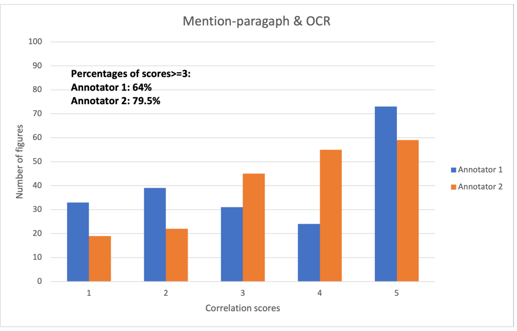


Figure 3 Score distribution on correlations between mention–paragraph, OCR tokens and figure captions. Both evaluators judged most of the figures with at least moderate correlations with captions.

files and some are complicated to parse. After obtaining PDFs, we used PDFFigures 2.0 [4] [2] to extract the body text of each paper. PDFFigure 2.0 is a tool that extracts figures, captions, tables, and text from scholarly PDFs in computer science. In scholarly documents, authors label figures with numbers (e.g. Figure 1. Fig. 1). For a figure, we used its figure number in a regular expression to locate a paragraph that mentions it.

OCR Extraction The SciCap dataset also provides texts extracted from figures as metadata, but does not provide location information for each text. To include location information for each text in a figure, we used Google Vision OCR API to extract text tokens from each figure with its coordinates of bounding boxes.

3.1 Dataset Quality Evaluation

We conducted human evaluations of the SciCap+ where we checked the mention-paragraphs and OCR tokens extraction quality. The aim was to establish whether the mention-paragraphs and OCR tokens were extracted correctly and relevant to the figure and its caption. To this end, we randomly selected 200 figures from the training set and for each figure, we asked two human evaluators to give scores of 1-5 (1 represents no relevance and 5 is highly relevant) for relevance between a caption of a figure and its mention-paragraphs and OCR tokens.

Figure 3 shows the distributions of the relevance scores. We can observe that two evaluators gave most of the figures (evaluator 1: 64% and evaluator 2: 79.5%) with relevance scores greater than 3 and a cohen kappa score of 0.28. This evaluation result indicates that the mention-paragraphs and OCR tokens have a satisfactory extraction quality and that the annotators considered most of them as relevant to the figure and its caption. However, the two annotators seem to have a relatively lower agreement (0.28) regarding which figures and captions are relevant to their mention-paragraphs and OCR tokens. We attribute this to the fact that evaluations of figure captions are highly subjective.

4 Experimental Results

This section reports experimental results using M4C-Captioner [3] as the baseline model to study the challenge of scientific figure captioning using the SciCap+ dataset. Please refer to the appendix for implementation and training details.

4.1 Main Result

The experimental results in table 2 demonstrate that using the mention-paragraph and OCR tokens significantly improves scores on all five metrics compared to the figure-only baseline. The experimental results align with our hypothesis that scientific figure captioning is a knowledge-augmented image captioning task, OCR tokens and knowledge embedded in mention-paragraphs help in composing informative captions.

We established a baseline M4C-Captioner (Figure only) with figures as the only input modality to the M4C-Captioner model in row #1. This baseline is in the non-knowledge setting. Therefore, low scores in all metrics show that the model needs knowledge of other modalities. Using the mention only in row #2 shows that the mention certainly contains a lot of useful information, as evidenced by the increase in performance. When OCR features are added to the figure input in row #3, scores for all metrics have significant gains compared to the figure-only baseline, but are still weaker than when only mentions are used. This motivates the combination of mentions and OCR features and in row #4, compared to the figure-only baseline and figure-OCR-only baseline, the performance further improves. Perhaps the most interesting result is in row #5 where we only use the mentions and OCR features but not the figure and get the best performance, particularly for SPICE and CIDEr, albeit comparable to when the figure is included in row #4. All these results indicate that explicitly extracted multimodal knowledge helps to compose informative captions.

[2] https://github.com/allenai/pdffigures2

Model	BLEU-4	METEOR	ROUGE-L	SPICE	CIDEr
1. M4C-Captioner (Figure Only)	1.5	5.6	15.4	4.3	4.6
2. M4C-Captioner (Mention Only)	5.3	11.0	27.4	14.3	49.0
3. M4C-Captioner (Figure and OCR features)	2.6	7.6	20.5	10.1	22.2
4. M4C-Captioner (Mention, Figure and OCR features)	6.3	**12.0**	29.2	15.8	55.8
Ablation Study on Figures					
5. M4C-Captioner (Mention and OCR features)	6.3	**12.0**	**29.3**	**16.1**	**56.4**
Ablation Study on OCR features					
6. M4C-Captioner (Mention, Figure and w/o OCR features)	**6.4**	11.5	27.9	14.6	50.5
7. M4C-Captioner (Mention, Figure and OCR spatial features)	5.8	11.1	27.3	14.1	48.0
8. M4C-Captioner (Mention, Figure and OCR (w/o spatial features) features)	**6.4**	**12.0**	29.1	15.7	54.6
9. M4C-Captioner (Mention, Figure and OCR (w/o visual features) features)	6.2	11.9	28.9	15.6	54.1

Table 2 Automatic evaluation scores of M4C-captioning on SciCap dataset. Aggregate knowledge from text and vision modalities significantly boosts the model performance compared to the figure-only baseline.

4.2 Ablation Studies

We first performed an ablation study on figures by removing visual feature vectors, the CIDEr score increases slightly, indicating that the visual feature is more like noise for the model. This is likely because the Resnet-152 visual encoder we used was not trained on figures.

We enriched the representations of the OCR features by adding text, visual, and spatial features. Ablation studies aim to reveal impacts of each OCR token feature. All comparisons are with row #4 even though row #5 gives slightly better scores. With OCR features completely removed in row #6, the CIDEr scores decrease by 5.3. Using only OCR spatial features in row #7, the CIDEr score dropped by 7.8. Removing OCR spatial features in row #8, the CIDEr scores dropped by 1.2. Upon removal of OCR visual features in row #9, the CIDEr score is close to removing spatial features.

The above ablation study indicates that the enriched OCR contributes to the informativeness of generated captions. Unlike OCR features, where appearance features are helpful to the model, removing visual features of figures increases CIDEr scores, further indicating that we need a specific vision encoder for figures to provide meaningful features.

5 Related Work

Unlike natural image captioning, figure captioning has been less explored. SciCap [1] is the most recent work on scientific figure captioning, comprising a large-scale scientific figure captioning dataset that includes figures from academic papers in arXiv dataset. Before SciCap, FigCAP [5] [6] and FigureQA [7] are two figure captioning datasets, but their figures are synthesized. We decided to extend and study on SciCap dataset, since its figures are from real-world scientific papers.

The closest multimodal task to figure captioning is image captioning. Recent works on integrating texts in natural images for visual question answering and image captioning tasks are based on transformer architecture augmented with a pointer network [8, 9]. The transformer enriches representations by integrating knowledge from both text and visual modality. The pointer network dynamically selects words from the fixed dictionary or OCR tokens during generation.

6 Conclusion

In this paper, we focus on scientific figure captioning as a knowledge augmented image captioning task. We transform the SciCap dataset [1]into SciCap+, by augmenting figures with their mention-paragraphs and OCR tokens. We then benchmark SciCap+ using the M4C-Captioner as the baseline model to utilize knowledge across three modalities: mention-paragraphs, figures, and OCR tokens. The experimental results reveal that using knowledge significantly improves evaluation metric scores. The release of the SciCap+ dataset promotes the further development of scientific figure captioning. For future work, we are interested in how to use multimodal pretraining strategies in this task.

Acknowledgment

These research results were partly obtained from the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), Japan, and partly obtained from the first author's internship research under NICT.

References

[1] Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. SciCap: Generating captions for scientific figures. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 3258–3264, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Qiao Zhang, Cong Wang, Chunsheng Xin, and Hongyi Wu. Cheetah: An ultra-fast, approximation-free, and privacy-preserved neural network framework based on joint obscure linear and nonlinear computations. **arXiv preprint arXiv:1911.05184**, 2019.

[3] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In **European conference on computer vision**, pp. 742–758. Springer, 2020.

[4] Christopher Clark and Santosh Divvala. Pdffigures 2.0: Mining figures from research papers. In **2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)**, pp. 143–152. IEEE, 2016.

[5] Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. Figure captioning with reasoning and sequence-level training. **arXiv preprint arXiv:1906.02850**, 2019.

[6] Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, and Ryan Rossi. Figure captioning with relation maps for reasoning. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**, March 2020.

[7] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. **arXiv preprint arXiv:1710.07300**, 2017.

[8] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, 2020.

[9] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioningwith reading comprehension. 2020.

[10] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf, 2020.

[11] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[13] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. **TACL**, Vol. 5, pp. 135–146, 2017.

[14] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. **IEEE transactions on pattern analysis and machine intelligence**, Vol. 36, No. 12, pp. 2552–2566, 2014.

[15] 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**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[16] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In **Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization**, pp. 65–72, 2005.

[17] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In **Text summarization branches out**, pp. 74–81, 2004.

[18] 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**, pp. 4566–4575, 2015.

[19] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In **European conference on computer vision**, pp. 382–398. Springer, 2016.

A Appendix

A.1 Implementation and Training

Our implementation of M4C-Captioner is based on the MMF framework [10] and Pytorch. The implementation allows users to specify diverse pre-trained encoders for each modality, which can be fine-tuned or frozen during training. The M4C-captioner itself has $D = 768$ hidden dimension size, $K = 4$ transformer layers and 12 attention heads. We used sentencepiece [11] to obtain a dictionary of 32000 subwords built from both mention-paragraphs and OCR tokens. This is used as the M4C-captioner's vocabulary. We followed the BERT-BASE hyperparameter setting and trained from scratch.

Regarding the encoders that feed features to M4C-captioner, we used pre-trained Resnet-152 as the figure's vision encoder. For each figure, we applied a 2D adaptive average pooling over outputs from layer 5 to obtain a global visual feature vector with a dimension of 2048. Layers 2, 3 and 4 layers were fine-tuned during training. For mention-paragraph features, SciBERT [12] was used to encode[3] it into 758-dimensional feature vectors. The number of vectors equals the number of sub-word tokens in the mention-paragraph, which we limit to 192. The mention-paragraph encoder is also fine-tuned during training. Finally, for OCR tokens, we use both text and visual features. We selected FastText [13] as the word encoder and Pyramidal Histogram of Characters (PHOC) [14] as the character encoder. Regarding the visual feature encoder of OCR tokens, we first extracted Faster R-CNN fc6 features and then applied fc7 weights to it to obtain 2048-dimensional appearance features for bounding boxes of OCR tokens. The fc7 weights were fine-tuned during training. We kept a maximum of 95 OCR tokens per figure.

We trained a model on a GPU server with 8 Nvidia Tesla V100 GPUs. Training a model with a complete set of features took 13 hours. During training, we used a batch size of 128. We selected CIDEr as the evaluation metric. The evaluation interval is every 2000 iterations, we stop training if CIDEr score does not improve for 4 evaluation intervals. The optimizer is Adam with a learning rate of 0.001 and $\epsilon = 1.0\text{E}{-}08$. We also used a multistep learning rate schedule with warmup iterations of 1000 and a warmup factor of 0.2. We kept the maximum number of decoding steps at the decoding time as 67.

For evaluation, we used five standard metrics for evaluating image captions: BLEU-4 [15], METEOR [16], ROUGE-L [17], CIDEr [18] and SPICE [19]. Since figure captions contain scientific terms which can be seen as uncommon words, among all five metrics, we are particularly interested in CIDEr since it emphasizes them.

[3] We only used the first 3 layers of SciBERT for lightweightness.