Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 1331–1336
DOI: 10.15439/2023F7280
ISSN 2300-5963 ACSIS, Vol. 35

Temporal Image Caption Retrieval Competition – Description and Results

Jakub Pokrywka, Piotr Wierzchoń, Kornel Weryszko, Krzysztof Jassem
Adam Mickiewicz University
Faculty of Mathematics and Computer Science,
Emails: jakub.pokrywka@amu.edu.pl, wierzch@amu.edu.pl
korwer@st.amu.edu.pl, jassem@amu.edu.pl

***Abstract*—Multimodal models, which combine visual and textual information, have recently gained significant recognition. This paper addresses the multimodal challenge of Text-Image retrieval and introduces a novel task that extends the modalities to include temporal data. The Temporal Image Caption Retrieval Competition (TICRC) presented in this paper is based on the Chronicling America and Challenging America projects, which offer access to an extensive collection of digitized historic American newspapers spanning 274 years. In addition to the competition results, we provide an analysis of the delivered dataset and the process of its creation.**

I. Introduction

MULTIMODAL models are gaining great recognition, especially those combining image and text. A recent example is the image generation model, DALL·E 2 [1]. Tasks executed by such multimodal models usually consist of text-image retrieval, namely, either retrieving an image from its text description or retrieving a text caption for a given image. In this challenge, we introduce a task in the caption retrieval setup, additionally extending the model with temporal data.

Language models rarely utilize metadata, such as text domain, timestamp, or website URL. Additional temporal information may prove helpful when factual knowledge is required, and the facts rely on time (e.g., the answer to the question: "Who is the president of the U.S.A?" depends on the date). Temporal information may also be relevant in case of language semantic changes (e.g., the meaning of the word "gay" has shifted from "cheerful" to referring to homosexuality).

The presented task is based on the projects: Chronicling America [2] and Challenging America [3]. Chronicling America is an open database of over 16 million pages of digitized historic American newspapers covering 274 years. Challenging America is a set of temporal challenges based on the Chronicling America dataset.

The described competition was conducted using the Gonito platform [4], and its results are available at https://gonito.csi.wmi.amu.edu.pl/challenge/cnlps-ticrc. The competitions started on Feb 20, 2023, and ended on June 14, 2023. The training dataset was published in two batches (train and train2). Participants were allowed to use the delivered development dataset (dev) for training. The preliminary testing dataset (test-A) was available from the beginning of the competition. The final testing dataset (test-B) was released in the last two weeks of the competition. The golden truth for the testing datasets has not been made public. The Gonito platform is open to post-competition submissions.

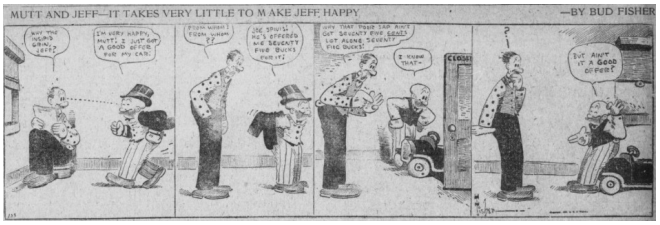


Fig. 1. Sample picture with a caption above. This picture comes from a newspaper issued dated Jan 11, 1928.

II. Motivation

From a linguistic and historical standpoint, Temporal Image Caption Retrieval (TICRC) holds significant value and brings various benefits. Firstly, TICRC facilitates the analysis of language evolution over time by associating image captions with specific temporal periods. Through this approach, researchers can investigate changes in vocabulary, grammar, and linguistic styles, thereby gaining insights into the adaptation and evolution of language across different historical contexts.

Secondly, TICRC contributes to the preservation and documentation of historical knowledge. Image captions accompanying visual content often contain valuable historical information. By leveraging TICRC, historians and researchers can effectively search and analyze these image captions, enabling a deeper understanding of specific historical periods, events, or cultural contexts. This process enhances the documentation of historical knowledge and enriches our comprehension of the past.

Furthermore, TICRC facilitates cross-referencing and integration of visual and textual sources. By associating image captions with specific temporal intervals, the competition makes it possible to establish connections between relevant textual documents, such as diaries, newspapers, or historical records. The interlinking of visual and textual data enhances contextualization and aids in interpreting and analyzing visual content from a historical perspective.

Moreover, TICRC offers valuable contextual information regarding the depicted scenes, individuals, or objects in images. By retrieving relevant captions based on temporal information,

researchers gain a more comprehensive understanding of the context in which the images were captured. This contextualization further strengthens the interpretation and analysis of visual content within its historical framework.

In summary, Temporal Image Caption Retrieval enables the analysis of language evolution, enhances historical documentation and preservation, facilitates the integration of visual and textual sources, provides contextualization of visual content, and supports the study of cultural and societal changes over time.

III. Related Work

A. Temporal language datasets and models

Several textual benchmarks concerning the date of text publication have been published in recent years. Challenging America [3] presents a set of three temporal tasks. Authors of [5] introduce a temporal question answering task and dataset, in which the query's answer depends on a year, e.g., *Who is the current president of the USA?*. Both benchmarks contain a baseline temporal language model trained on a text with a date timestamp prepended as text. In [6], the authors propose another text classification task, including temporal information. In addition to the timestamp in the textual form the model is also trained on temporal input embeddings. The authors of [7] modify the transformer architecture, proposing a temporal attention component.

B. Multimodal vision-language models

Recently, the quality of vision-language models has improved greatly thanks to introducing models such as CLIP [1], EVAL-CLIP [8], ALIGN [9], BASIC [10], LiT [11], Flamingo [12], or GPT-4 [13] and [14].

MS COCO [15] and Visual Genome [16] are two large-scale, high-quality vision datasets annotated by humans. YFCC-100M [17] is an even larger dataset that contains user data collected from Flickr, not specifically designed for model training. Authors of CC12M [18] and LAION-5B [19] apply cleaning procedures to adapt user data for the purpose of model training. The works mentioned did not prioritize the importance of temporal data.

IV. Task Definition

The task here is to retrieve a relevant caption from a caption set for the given picture from a newspaper and the newspaper's publication daily date. For each picture, only one caption is relevant.

The dataset is provided on the challenge GitHub repository https://github.com/kubapok/cnlps-ticrc.

Figure 2 presents an example source picture with a caption.

A. Sample Data

In this section, we provide sample data. A picture and the publication date (in the YYYY-MM-DD format) of a given newspaper issue are given, as well as the collection of all captions for the given dataset type (train, train2, dev-0, test-A, or test-B). In the caption collection, a newline character is represented as \n. The challenge participant is supposed to return the list of captions from the given dataset in descending probability order.

Picture: Figure 2

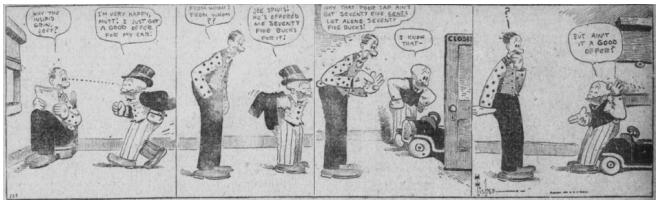


Fig. 2. Sample input picture

Date timestamp: 1928-01-11

Set of all possible captions:

- "China Dinner Sets."
- "MUTT AND JEFF — IT TAKES VERY LITTLE TO MAKE JEFF HAPPY"
- "PARIS MILLINERY\nfrom every Parisian modiste,\nof note - embracing every \nstyle tendency of the fall \nand winter season \nand \n GOWNS COATS WRAPS \nTAILORED SUITS AND \nDRESSES"
- ...

Correct Output: "MUTT AND JEFF –– IT TAKES VERY LITTLE TO MAKE JEFF HAPPY"

More examples are provided in Figure 8.

B. Metric

The metric for the competition is Mean Reciprocal Rank:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i},$$

where: |Q| — number of queries, $rank_i$ — rank position of the relevant document for the i-th query. The metric is implemented in the GEval evaluation tool [20] and available for offline use (details are provided on the competition page).

V. Data Annotation Process

The data was taken from the Challenging America project, according to the data processing rules provided there. The annotation was done manually in the Doccano [21] system, which helped effective processing of annotation pairs: image and text. The annotation platform required the annotation of the entire newspaper pages. A sample page from which a picture was selected is presented in Figure 3. The annotation of images was carried out according to given guidance rules divided into three aspects: Objects to be annotated (what to annotate), technical parameters of the image area (what technical requirements are imposed on annotated objects), and rules of text transcription (how to transcript caption texts).

These were the annotation guidance rules:

Fig. 3. Picture selected on the whole page.

a) Objects to be annotated:

- Images may be selected for annotation only if they occur along with the corresponding caption.
- The caption text should be maximum a few sentences long. In case of longer captions, the annotator should select and mark the most relevant fragment of the caption.
- The caption text should – at the discretion of the annotator – be relevant to the image in content.
- The annotator should select at most one image per page.
- If the annotator has already encountered the same image on one of the previously annotated pages, the image should not be annotated again.
- The annotator should minimize the number of portraits.

b) Technical requirements for the image area (bbox):

- The picture frame should encompass the image in its entirety (the picture should not be cut off).
- The image frame should not cover more area than the image.
- The frame must not cover the caption text.

c) Rules for text transcription:

- The transcription should preserve the character size of the original
- Punctuation and line-break characters should be preserved as in the original.
- Paragraph indentation in the text should be ignored. If the words are divided by a hyphen or line break, the original spelling (separated words) should be preserved.

The dataset was annotated mainly by one annotator, and his work took 70 hours.

TABLE I
DATA SPLIT STATISTICS

Type	Name	Instances	Ratio
Training	train	675	70.0
	train2	2054	
Development	dev-0	646	16.6
Testing	test-A	92	13.4
	test-B	435	

VI. DATA ANALYSIS

The dataset comprises 3902 instances, each consisting of a picture, a caption, and a date timestamp. The pictures and corresponding captions were extracted from scans of newspapers dating back to 1853, which appends the element of fuzziness in image recognition to the challenge and makes the temporal aspect even more relevant (as the image quality depends on the publication date).

A. Data Split

Five datasets have been prepared for the competition – two training sets (train, train2), a development set (dev-0), and two test sets (test-A, test-B). The final split ratio is illustrated in Table I. Precautions similar to those described in [3] have been taken to ensure that there is no detrimental overlap between the datasets.

B. Datasets Statistics

For the sake of statistical analysis, the two testing datasets and the development dataset have been combined into one dataset, referred to as the testing dataset in this section. Similarly, the two training datasets have been combined into one.

Figures 4 and 5 provide insight into the temporal variance in the frequency distributions of the instances. Whereas both datasets are negatively skewed (as suggested by the mean $\approx$ 1895.82 and median $=$ 1897.0 of the testing dataset and mean $\approx$ 1903.52, median $=$ 1905.0 in the case of the training dataset), the latter covers a significantly greater period containing data points between 1853 and 1922. The testing dataset spans from 1880 to 1900. Moreover, the testing dataset's standard deviation ≈ 4.18 is also less than $\frac{1}{3}$ of the training dataset's standard deviation ≈ 12.97.

The captions are measured in the number of words and characters. The captions from the testing dataset captions tend to be longer, with mean $\approx$ 11.77 and median $= 8.0$ words per caption and mean $\approx$ 66.79, median $=$ 44.0 characters per caption. The respective parameters for captions from the training dataset have the following values: mean $\approx$ 9.80, median $=$ 7.0 and mean $\approx$ 56.54, median $=$ 43.0. There is no significant difference in the corresponding frequency distributions, as can be seen in Figures 6 and 7.

VII. BASELINES

The official competition baseline is included in the competition repository and relies on the transformer model clip-ViT-B-32 [14] model without fine-tuning. The secondary baseline is the randomized caption order.

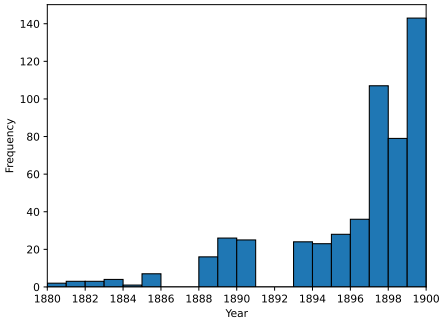


Fig. 4. Testing distribution over the years

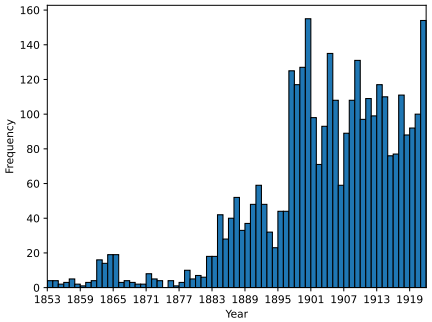


Fig. 5. Training distribution over the years

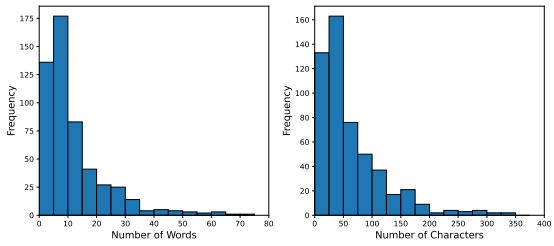


Fig. 6. Word and character per caption statistics in testing dataset

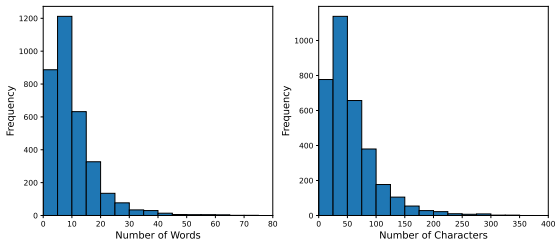


Fig. 7. Word and character per caption statistics in the training dataset

VIII. Shared Task Results

Five teams participated in the competition. Three solutions scored above the official competition baseline. The final results are provided in Table II.

TABLE II
FINAL COMPETITION RESULTS. THE TEST-B DATASET IS USED FOR WINNER DETERMINATION, WHEREAS THE TEST-A DATASET IS ONLY PRELIMINARY.

place	submitter	test-A MRR	test-B MRR	submissions
1	Kaszuba	0.6059	0.3444	6
2	s478846	0.5529	0.33850	11
3	Serba	0.3506	0.2283	1
-	**transformer baseline**	0.2697	**0.1710**	-
4	Szyszko	0.0887	0.0621	1
-	**random baseline**	0.0513	**0.0193**	-
5	s478855	0.0514	0.0137	3

The competition's winner is Patryk Kaszuba, who was invited to prepare a report for publication in the conference proceedings and presentation at FedCSIS 2023. His solution is based on EVA02_CLIP_E_psz14_plus_s9B model [8]. The model was used without fine-tuning to the competition dataset.

IX. Conclusions

In this paper, we introduced a new benchmark for temporal image caption retrieval, called TRIC (Temporal Image Caption Retrieval). TRIC includes a three-modal (vision-language-time) dataset, divided into two train sets, two test sets and a development set. The proposed task consists in selecting a caption relevant for a given image, from a given set. The temporal information is significant for the task as the data comprise scanned texts spanning the period of 274 years.

We organised the competition based on the benchmark. Five participants participated, with three of them scoring above the baseline. The benchmark is still open for further improvement of the obtained results.

We believe that TRIC will have a positive impact on the analysis of language evolution and support the study of cultural and societal changes over time.

References

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

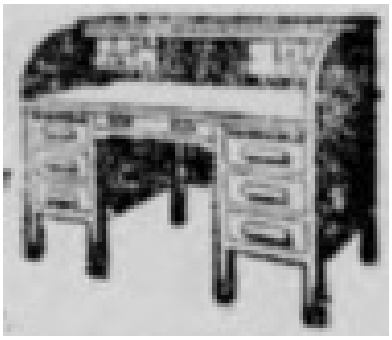
[2] B. C. G. Lee, J. Mears, E. Jakeway, M. Ferriter, C. Adams, N. Yarasavage, D. Thomas, K. Zwaard, and D. S. Weld, "The newspaper navigator dataset: Extracting headlines and visual content from 16 million historic newspaper pages in chronicling america," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, ser. CIKM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 3055–3062. [Online]. Available: https://doi.org/10.1145/3340531.3412767

[3] J. Pokrywka, F. Graliński, K. Jassem, K. Kaczmarek, K. Jurkiewicz, and P. Wierzchon, "Challenging America: Modeling language in longer time scales," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 737–749. [Online]. Available: https://aclanthology.org/2022.findings-naacl.56

(a) 1922-04-10 WINCHESTER\nGolders' Headquarters\nFor Every Golfer

(b) 1911-06-13 MICHELIN \nInner Tubes \nFor Michelin and all other Envelopes

(c) 1911-08-17 HERE'S the Shirt hit of the \nseason



(d) 1913-02-12 DESKS AND OFFICE FURNITURE.

(e) 1902-10-16 M. O'NEIL & CO. \nFurniture \nAND... \nCarpets \nWe are now showing one of the \nmost complete lines of \nParlor \nand \nLibrary \nTables

(f) 1907-06-27 Guaranteed \nPURE. \nLEAD AND ZINC PAINTS. \n"Made in BALTIMORE"

Fig. 8. Sample images from the training dataset with the corresponding date of publication caption. The images were not selectively chosen.

[4] F. Graliński, R. Jaworski, Ł. Borchmann, and P. Wierzchoń, "Gonito.net – open platform for research competition, cooperation and reproducibility," in *Proceedings of the 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, A. Branco, N. Calzolari, and K. Choukri, Eds., 2016, pp. 13–20.

[5] B. Dhingra, J. R. Cole, J. M. Eisenschlos, D. Gillick, J. Eisenstein, and W. W. Cohen, "Time-aware language models as temporal knowledge bases," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 257–273, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.15

[6] J. Pokrywka and F. Graliński, "Temporal language modeling for short text document classification with transformers," in *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2022, pp. 121–128.

[7] G. D. Rosin and K. Radinsky, "Temporal attention for language models," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1498–1508. [Online]. Available: https://aclanthology.org/2022.findings-naacl.112

[8] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023.

[9] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[10] H. Pham, Z. Dai, G. Ghiasi, K. Kawaguchi, H. Liu, A. W. Yu, J. Yu, Y.-T. Chen, M.-T. Luong, Y. Wu *et al.*, "Combined scaling for zero-shot transfer learning," *arXiv preprint arXiv:2111.10050*, 2021.

[11] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 123–18 133.

[12] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[13] OpenAI, "Gpt-4 technical report," 2023.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[15] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[16] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[17] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[18] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3558–3568.

[19] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[20] F. Graliński, A. Wróblewska, T. Stanisławek, K. Grabowski, and T. Górecki, "GEval: Tool for debugging NLP datasets and models," in

Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 254–262. [Online]. Available: https://www.aclweb.org/anthology/W19-4826

[21] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang, "doccano: Text annotation tool for human," 2018, software available from https://github.com/doccano/doccano. [Online]. Available: https://github.com/doccano/doccano