DOI: 10.15439/2023F4192
ISSN 2300-5963 ACSIS, Vol. 35

Multimodal Neural Networks in the Problem of Captioning Images in Newspapers

Patryk Kaszuba
Faculty of Mathematics and Computer Science
Adam Mickiewicz University
Poznań, Poland

***Abstract*—This paper presents the effectiveness of different multimodal neural networks in captioning newspaper scan images. These methods were evaluated on a dataset created for the Temporal Image Caption Retrieval Competition, which is a part of the FedCSIS 2023 conference. The task was to predict a relevant caption for a picture taken from a newspaper, chosen from a given list of captions. The results we obtained show the promising potential of image captioning using CLIP architectures and emphasize the importance of developing new multimodal methods for problems that combine multiple disciplines, such as computer vision with natural language processing.**

I. Introduction

IMAGE captioning is the task of transforming the visual information of an image into a natural language description of the image. This process combines the fields of natural language processing and computer vision. Artificial intelligence models, similarly to humans, can describe images with varying levels of detail. The variation in image descriptions generated by different models is due to differences between model architectures and training data sets. These factors affect the models' ability to extract different image features and focus attention on different aspects, resulting in diverse interpretations and semantics in the generated descriptions. Early methods were based on feature extraction techniques in which low-level visual features such as Histogram of Oriented Gradients (HOG) descriptor [1], attribute representation [2] or Support Vector Machine (SVM) [3] were combined with language models to generate captions. These methods had difficulties capturing higher-level semantic terms and processing images with varying content. The development of neural networks in the past decade led to the development of more successful methods in image captioning. Using deep neural networks eliminated the need for manual feature extraction, which resulted in the automatic creation of better representations and improved results. The first models used a combination of Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) containing layers such as Long Short-Term Memory (LSTM) [4] or Gated Recurrent Units (GRU) [5]. Later models used attention mechanisms [6] or reinforcement learning [7] [8].

In this paper, we will focus on the use of multimodal neural networks in the problem of image captioning. In the following sections, we will discuss in detail the competition in which we participated, describe the methods that utilized three popular pre-trained neural network models CLIP, and in the last sections, present the results and describe the conclusions.

II. Related work

Understanding and interpreting the meaning of the content in an image based on the image itself is one of the more challenging problems in the field of artificial intelligence. However, in recent years, the development of deep neural networks has brought remarkable advancements in this field, and as a result, multimodal neural networks have emerged. Combining text and image representations in a joint embedding space results in significant improvements in image captioning, as demonstrated by methods such as those described in [9] or [10]. Nevertheless, the most significant results have been achieved using contrastive learning in papers presenting methods such as VILLA [11], ERNIE-ViL [12], Oscar [13], ALIGN [14] and CLIP [15].

III. FedCSIS 2023 Competition

A. Problem description

In the Temporal Image Caption Retrieval Competition, organized during FedCSIS 2023, the goal is to select the correct caption for the image. The dataset contains temporal information along with images, which can be used to accurately assign the most relevant captions to each image based on historical data.

The evaluation metric for this competition is Mean Reciprocal Rank (MRR).

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

where $|Q|$ is the total number of images in the dataset and $rank_i$ is the position of the correct caption in the ranked list for each image.

B. Dataset description

The competition dataset is based on the project "Challenging America" [16], which was initially created for three tasks. The first task, known as "RetroTemp", focused on temporal classification. The objective was to predict the publication date based on given newspaper titles and text excerpts. In the second task, "RetroGeo", the goal of the task was to predict the latitude and longitude coordinates of the place of issue using normalized newspaper titles, text excerpts, and fractional publishing dates. The last task, "RetroGap", involved predicting the missing word within a provided normalized newspaper title, text excerpt, and year of publication in fractional format.

For the competition, the organizers expanded the original dataset with test sets that had never been published before. The purpose of this action was to prevent participants from accessing the data during the competition.

All the collected data for the dataset comes from the "Chronicling America" [17] database, which contains digitized newspapers from 1690 until now, encompassing approximately 150,000 bibliographic title entries, as well as 600,000 library holdings records.

C. Dataset structure

The organizers of the competition split the dataset into 5 sets as follows: two training sets *train* and *train2*, a development set *dev-0*, and two test sets *test-A* and *test-B*. The total number of samples in the entire dataset was 3902 samples.

Each of the splits contained the following amounts of data:

- train - 675 samples
- train2 - 2054 samples
- dev-0 - 646 samples
- test-A - 92 samples
- test-B - 435 samples

Every single record consisted of three features: a picture, a caption text, and a publication date. The images were in grayscale, with a minimum and maximum width of 2 and 1162 pixels, respectively. For the height, the minimum value was 5 pixels, and the maximum was 1592 pixels. The header text contained both lowercase and uppercase letters, as well as symbols and special characters. The number of words in the headers varied, with the shortest containing 1 word, and the longest containing 83 words. For the publication date, the ISO 8601 (YYYY-MM-DD) format was used. The oldest publication date was 1853, and the latest one was 1922.

IV. Methods

Our solution is based on three different multimodal neural networks: **CLIP-ViT**, **OpenCLIP** [18] and **EVA-CLIP** [19]. We used the above pre-trained models for a zero-shot classification task, experimenting with their various parameter variants. As part of data preprocessing, we converted newline characters to spaces. The solution is described in the form of pseudocode in Algorithm 1.

In the initial step, we preprocess all the captions and extract their embedded vector representations obtained from the neural network's output. Then, for each image, we execute the same process to obtain its embedded vector representation. Finally, we calculate cosine similarity between an individual embedded image vector and all embedded caption vectors to determine the most similar images with captions, which we then sort based on their similarity values in descending order.

Algorithm 1 Pseudocode of our solution for image captioning

```
Require: Image vector $I = (I_1, I_2, ..., I_n)$, caption vector $T = (T_1, T_2, ..., T_m)$
  for each $t \subset T$ do
    $t \leftarrow preprocess(t)$
    $Emb_t \leftarrow CLIP(t)$
  end for
  for each $i \subset I$ do
    $Emb_i \leftarrow CLIP(i)$
    for each $t \subset Emb_t$ do
      $sim \leftarrow cosinesimilarity(Emb_i, t))$
      $Y_i.insert(sim)$
    end for
  end for
  for each $c \subset Y$ do
    $Y_c \leftarrow sort(c)\ descending$
  end for
```

A. CLIP-ViT models

We utilize the CLIP-ViT pre-trained models, based on the Vision Transformer architecture. These models were pre-trained by OpenAI on a set derived from a subset of the YFCC100M [23] dataset, with four different model parameters:

- **ViT-B-16** - 12 vision layers, 12 text layers, 512 embedding dimensions, image patch size 16x16, image resolution 224^2
- **ViT-B-32** - 12 vision layers, 12 text layers, 512 embedding dimensions, image patch size 32x32, image resolution 224^2
- **ViT-L-14** - 24 vision layers, 12 text layers, 768 embedding dimensions, image patch size 14x14, image resolution 224^2
- **ViT-L-14-336** - 24 vision layers, 12 text layers, 768 embedding dimensions, image patch size 14x14, image resolution 336^2

B. OpenCLIP models

The main difference between CLIP-ViT by OpenAI is that these models were pre-trained on the LAION-2B [24] dataset. Three new models have been created with the following parameters:

- **ViT-H-14** - 32 vision layers, 24 text layers, 1024 embedding dimensions, image patch size 14x14, image resolution 224^2
- **ViT-g-14** - 40 vision layers, 24 text layers, 1024 embedding dimensions, image patch size 14x14, image resolution 224^2
- **ViT-G-14** - 48 vision layers, 32 text layers, 1280 embedding dimension, image patch size 14x14, image resolution 224^2

C. EVA-CLIP models

The models differ from the previous ones by the implied techniques, such as the LAMB [20] optimizer, random input token dropping [21], and flash attention [22]. The **EVA02_CLIP_E_psz14_plus_s9B** model, just like the previous OpenCLIP models, was pre-trained on the LAION-2B dataset, but in the case of models **EVA02_CLIP_B_psz16_s8B** and **EVA02_CLIP_L_psz14_s4B**, they were pre-trained on the Merged-2B dataset, which combines 1.6 billion samples from the LAION-2B dataset with 0.4 billion samples from the COYO-700M dataset. The models have the following parameters:

- **EVA02_CLIP_B_psz16_s8B** - 12 vision layers, 12 text layers, 512 embedding dimension, image patch size 16x16, image resolution 224^2
- **EVA02_CLIP_L_psz14_s4B** - 24 vision layers, 12 text layers, 768 embedding dimension, image patch size 14x14, image resolution 224^2
- **EVA02_CLIP_E_psz14_plus_s9B** - 64 vision layers, 32 text layers, 1024 embedding dimension, image patch size 14x14, image resolution 224^2

V. Results

The results from the evaluated models on three subsets are presented in Table I. The metric provided in the results is the same as the one used in the competition ranking. All models evaluated by us achieved a higher score than the baseline. The best result on the test-B set, which was 0.344423 MRR, was achieved by the **EVA02_CLIP_E_psz14_plus_s9B** model, due to having the highest number of parameters among all the other models.

We also conducted an error analysis for images on which our top model struggled the most. The four images that achieved the worst MRR score are shown in Fig1. The model had difficulty choosing the correct caption for the images in cases where the caption was the author's subjective interpretation of the image and did not directly relate to the description of the elements in the photo. This can be observed in Fig. 1a and Fig. 1b. Another problem related to the model was low-resolution images, which could result in difficulties in object detection and, consequently, making inferior decisions regarding the accurate labeling of the image, as seen in Fig. 1c and Fig. 1d.

VI. Conslusion

In this paper, we presented our solution for the Temporal Image Caption Retrieval Competition. We evaluated various multimodal pre-trained models with different parameter sizes. The model with the highest Mean Reciprocal Rank metric on the dev-0 set was submitted to the competition system and ranked first place. Our approach indicates that multimodal neural networks are effective for image captioning in newspapers. For future work, we suggest improving results by fine-tuning the pre-trained models using the training data provided by the organizers. Additionally, better results may be achieved by using temporal data as an extended input for the neural network and making predictions based on historical information.

TABLE I: Experiment results

Model	**MRR**		
	dev-0	**test-A**	**test-B**
Baseline	0.156270	0.269739	0.171050
ViT-B-32 *openai*	0.162395	0.328729	0.171469
ViT-B-16 *openai*	0.193840	0.389401	0.201968
ViT-B-32 *laion2b_s34b_b79k*	0.208152	0.436798	0.221351
EVA02_CLIP_B_psz16_s8B	0.221110	0.395650	0.229678
ViT-L-14 *openai*	0.243495	0.466656	0.242640
ViT-B-16 *laion2b_s34b_b88k*	0.239205	0.430418	0.255294
ViT-L-14-336 *openai*	0.259092	0.459236	0.255777
ViT-L-14 *laion2b_s32b_b82k*	0.273631	0.505207	0.291728
ViT-g-14 *laion2b_s34b_b88k*	0.296378	0.485058	0.300874
ViT-H-14 *laion2b_s32b_b79k*	0.275778	0.490147	0.313473
ViT-bigG-14 *laion2b_s39b_b160k*	0.308845	0.572111	0.319987
EVA02_CLIP_L_psz14_s4B	0.321763	0.503575	0.332623
EVA02_CLIP_E_psz14_plus_s9B	**0.339309**	**0.605919**	**0.344423**

References

[1] A. Farhadi, S. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every Picture Tells a Story: Generating Sentences from Images. ECCV (4) , volume 6314 of Lecture Notes in Computer Science, page 15-29. Springer, (2010)

[2] Vicente Ordonez, Girish Kulkarni, Tamara L. Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. Neural Information Processing Systems(NIPS), 2011.

[3] A. Farhadi, S. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. ECCV (4) , volume 6314 of Lecture Notes in Computer Science, page 15-29. Springer, (2010)

[4] A. Karpathy, and F. Li. Deep visual-semantic alignments for generating image descriptions. CVPR , page 3128-3137. IEEE Computer Society, (2015)

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. CVPR, "Show and tell: A neural image caption generator.", page 3156-3164. IEEE Computer Society, (2015)

[6] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , page 6077-6086. IEEE Computer Society, (2018)

[7] N. Xu, H. Zhang, A. Liu, W. Nie, Y. Su, J. Nie, Y. Zhang, "Multi-Level Policy and Reward-Based Deep Reinforcement Learning Framework for Image Captioning," in IEEE Transactions on Multimedia, vol. 22, no. 5, pp. 1372-1383, May 2020

[8] Zhang, L; Sung, F; Liu, F; Xiang, T; Gong, S; Yang, Y; Hospedales, TM, Actor-Critic Sequence Training for Image Captioning. ; Volume. abs/1706.09601 ; Journal. CoRR

[9] Madhyastha et al., "End-to-end Image Captioning Exploits Distributional Similarity in Multimodal Space", EMNLP, pages 381–383, 2018

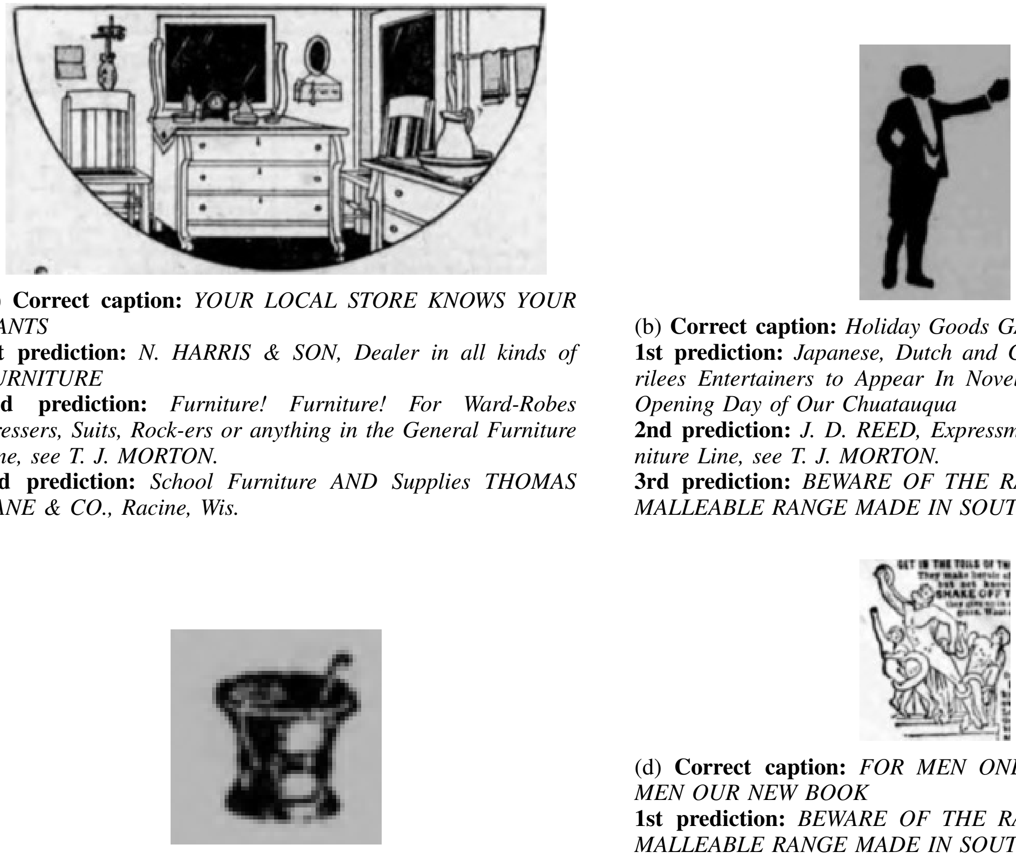
[10] YC Chen, L Li, L Yu, A El Kholy, F Ahmed, Z Gan, Y Cheng, J Liu, UNITER: universal image-text representation learning. In ECCV, vol. 12375, pages 104–120. 2020

[11] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, "Large-scale adversarial training for vision-and-language representation ". In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, NY, USA, Article 555, 6616–6628, 2020.

[12] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, & Haifeng Wang. (2021). ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph

[13] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, & Jianfeng Gao. (2020). Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks.

[14] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. ICML, 2021



(a) **Correct caption:** *YOUR LOCAL STORE KNOWS YOUR WANTS*
1st prediction: *N. HARRIS & SON, Dealer in all kinds of FURNITURE*
2nd prediction: *Furniture! Furniture! For Ward-Robes Dressers, Suits, Rock-ers or anything in the General Furniture Line, see T. J. MORTON.*
3rd prediction: *School Furniture AND Supplies THOMAS KANE & CO., Racine, Wis.*

(b) **Correct caption:** *Holiday Goods GALORE*
1st prediction: *Japanese, Dutch and Colonial Sketches Merrilees Entertainers to Appear In Novel Musical Program on Opening Day of Our Chuatauqua*
2nd prediction: *J. D. REED, Expressman and Drayman Furniture Line, see T. J. MORTON.*
3rd prediction: *BEWARE OF THE RANGE PEDLER! THE MALLEABLE RANGE MADE IN SOUTH BEND*

(c) **Correct caption:** *Down go the Prices AT THE Drug Store!*
1st prediction: *C. C. HURLEY, Hardware, Agricultural Implements, Paints OILS, GLASS, CUTLERY, GUNS, ETC*
2nd prediction: *HOUSEHOLD WARE*
3rd prediction: *PORTABLE MILLS For Corn Meal STRAUR & CO., P. O. Box 1430, Cincinnati.*

(d) **Correct caption:** *FOR MEN ONLYYOUNG MEN OLD MEN OUR NEW BOOK*
1st prediction: *BEWARE OF THE RANGE PEDLER! THE MALLEABLE RANGE MADE IN SOUTH BEND*
2nd prediction: *HOSTETTER'S CELEBRATED STOMACH BITTERS*
3rd prediction: *Doctor Henderson OVER 27 YEARS OF SPECIAL PRACTICE Seminal Weakness & Sexual Debility,Syphilis,Book Free Museum of Anatomy,Stricture Rheumatism*

Fig. 1: The figure presents four images from the dev-0 set in which the model achieved the worst results in predicting the correct caption

[15] A. Radford et al. "Learning Transferable Visual Models from Natural Language Supervision". In: Int. Conf. Mach. Learn. PMLR, 2021, pp. 8748–8763.

[16] Pokrywka, J., Gralinski, F., Jassem, K., Kaczmarek, K., Jurkiewicz, K., & Wierzchoń, P. (2022, July). Challenging America: Modeling language in longer time scales. In Findings of the Association for Computational Linguistics: NAACL 2022 (pp. 737-749).

[17] Lee, B. C. G., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., & Weld, D. S. (2020). The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America. Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 3055–3062.

[18] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, J. Jitsev, Reproducible Scaling Laws for Contrastive Language-Image Learning , Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2818-2829

[19] 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.

[20] You, Y., Li, J., Reddi, S. J., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., & Hsieh, C.-J. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.

[21] Li, Y., Fan, H., Hu, R., Feichtenhofer, C., & He, K., "Scaling Language-Image Pre-Training via Masking." CVPR, 2023.

[22] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. 2022.

[23] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D. & Li, L.-J. (2016). YFCC100M: the new data in multimedia research.. Commun. ACM, 59, 64-73.

[24] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, J. Jitsev, LAION-5B: An open large-scale dataset for training next generation image-text models, NIPS 2022, pp. 25278-25294.