An Empirical Exploration of Perceived Similarity between News Article Texts and Images

Lucien Heitz[1,2,*], Abraham Bernstein[1] and Luca Rossetto[1]

[1] University of Zurich, Switzerland

[1] UZH - Digital Society Initiative, Switzerland

Abstract

The NewsImages task at MediaEval implicitly assumes that there is a one-to-one mapping between news articles and images, given that there is exactly *one* image that is considered a fit in the evaluation phase. In this quest for insight, we empirically explore this assumption. We conduct a user study where we show participants images from different sources and ask how well the image fits a given article from the NewsImages task. We find that 1.) there can be multiple images per article that are considered equally fitting, 2.) images from within the task dataset can beat the ground truth images for certain articles, and 3.) AI-generated articles underperform in comparison with editorially selected images. Based on our insights, we suggest an alternative evaluation strategy for the task and a clear separation of editorial images and AI-generated content.

1. Introduction

The MediaEval NewsImages benchmark [1] aims at deepening the understanding of the relation between news articles and editorially selected images. The task participants are asked to come up with computational means to find the correct mapping between a set of articles and images in a test set. The quality of the restored mapping is assessed using measures such as Mean Reciprocal Rank and Hits@k. These evaluation metrics assume—at least implicitly—that only one image matches an article; all other options are deemed equally incorrect. Given a non-literal relationship between the content of an article and its image, the assumption that there could only ever be exactly one matching image appears to be overly restrictive.

In this Quest for Insight, we empirically investigate the perceived fit between a given news headline and teaser with several images. We sampled a subset of the task dataset and paired every article with four images. We then asked participants to rate how fitting the images are. The evaluated images include 1.) the ground truth/baseline provided by the dataset, 2.) an image retrieved from an external stock image platform, 3.) an AI-generated image, and 4.) an alternative image from the task dataset. We show that the NewsImage evaluation procedure can lead to a situation where the task formulation promotes sub-optimal image selection. We, therefore, suggest and discuss alternative evaluation strategies.

Our motivation for doing so is to increase the external validity and practical applicability of the insights gained from the NewsImages tasks. There is still a need for a better understanding of writing headlines, assigning images to stories, as well as the tools supporting news editors. The focus on perceived image fit is, therefore, only a first step in assessing the practical relevance of the text-image matching strategies in this task.

MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online

[*] Corresponding author.

heitz@ifi.uzh.ch (L. Heitz); bernstein@ifi.uzh.ch (A. Bernstein); rossetto@ifi.uzh.ch (L. Rossetto)

0000-0001-7987-8446 (L. Heitz); 0000-0002-0128-4602 (A. Bernstein); 0000-0002-5389-9465 (L. Rossetto)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Approach

We implement our experiment as an online survey on Qualtrics and recruited participants via Prolific.[1] We chose a within-subject design experimental setup with repeated measurements for the user survey. We recruited a total of $N = 73$ users. Participation required users to be fluent English speakers. No additional requirements were imposed.

Each participant rates the perceived fit between the title plus the lead of an article and different images. The articles were randomly chosen from the GDELT2 training dataset. For each article headline and lead, users are shown a selection of four different images. This image selection was created using the following methods:

Ground truth As a baseline, we use the image defined as the corresponding fit in the GDELT2 training set. For a given article, the dataset includes either an editorially selected image from the respective news outlet or an AI-generated image from the task organizers. We use the GDELT2 dataset, as this is the only task dataset containing news in English with both editorially selected and AI-generated images.

Stock image In order to be able to compare both within and outside of the dataset, we included an image obtained from the free stock photo platform Unsplash.[2] We retrieve images using the article headline and lead as search terms.[3] The first image from the search results was selected without any further manual curation. We added stock images to see if outside-domain sources serves as an alternative image pool.

AI generated The AI image is generated with Stable Diffusion [2] using the Realistic Vision V6 model.[4] The article headline and lead are used as positive prompts, together with the recommended negative prompts of the model, and DPM++ SDE Karras (25 steps) as sampler with a randomized seed for each picture. We include an additional source of AI images, as this enables us to tell whether people like/dislike AI-generated images in general, or if there exist model-specific preferences.

CLIP-based retrieval We include an additional image from the dataset that was not considered a match. We used the image with the best rank that was different from the ground truth, retrieved using an OpenCLIP model [3] that was pre-trained on the LAION-5B [4] dataset.[5] For more details on our retrieval approach, please see our Working Notes paper [5]. We feature alternative images from the task dataset to 1.) better assess the performance of our OpenCLIP approach and 2.) look at the performance of inside-domain images.

Each user was given five article units. One such unit consists of four questions where one article title and lead are paired with each of the four image choices. Participants had to rate how well each of the four image choices fit the given article headline and lead. Overall, we showed 20 different news article-image pairs to each user.

A total of 20 news articles were randomly selected from the task dataset. Between 16 and 19 user-ratings were recorded for each question unit of these news articles. Question units were randomly assigned to users, as was the ordering of article-image pairings within each unit.

[1] Official website of the Qualtrics: https://www.qualtrics.com, and Prolific: https://www.prolific.com/

[2] Official website of Unsplash: https://www.unsplash.com/

[3] As GDELT2 presents a collection of articles from different news outlets, not every article includes a lead. In case no lead is available, we instead selected the article's first sentence.

[4] Available online: https://www.huggingface.co/SG161222/Realistic_Vision_V6.0_B1_noVAE

[5] Available online: https://www.huggingface.co/laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k

We asked users the following question to assess the fit for each of the four images: *"Please indicate to what extent you agree with the following statement: I think this image fits the news article."* The agreement for each article-image pair was expressed on a Likert scale from 1 (Strongly disagree) to 7 (Strongly agree).

3. Results and Analysis

Figure 1 shows the agreement on the fit of an image with a given news article, grouped by the four different image sources. Ratings in green express the perceived fit for natural images, i.e., editorially selected images. Ratings in orange express the perceived fit for images that were generated. In our analysis in this section, we focus on 1.) participants' disagreement in terms of fit across image sources and 2.) the model selection for image generation together with the impact of AI-generated content.

The perceived fit of an image ranges from "Strongly agree" to "Strongly disagree" across all four groups. This range of ratings is an indicator for us that there is a certain degree of disagreement in terms of image fit among users. Not all users agree on what a good fit is. Multiple images per article can be perceived to be equally good in terms of fit. This is true across all image sources, applying to both natural as well as generated images. If there truly was only one fitting image—as the evaluation process of the task assumes—we would expect a predominantly disagreeing rating for all but the ground truth. This is, however, not what we observe in our results. Both our generated AI images and CLIP results can match the user ratings of the ground truth. One source that cannot match the ground truth is stock images. We see that outside-domain images seem to provide a worse perceived fit than inside-domain images retrieved via CLIP.

Regarding the added AI-generated content, we see a substantial difference in the rating distributions when comparing the natural and generated images in the ground truth. And while our own AI-generation pipeline manages to achieve a more favorable rating distribution, more closely mirroring the ratings of natural images in the ground truth, it nevertheless fails to provide a high count of results with a "Strongly agree" rating. While leveraging CLIP's capacity for content curation does shift the rating of the ground truth AI images to more favorable ratings, it still falls short of providing results comparable with natural images. We, therefore, see a clear qualitative difference between the two image types, one that seemingly cannot be alleviated by the image selection/retrieval technique. At the same time, however, we find that there is no general difference in perceived fit when it comes to AI-generated content. The fitness of the images has less to do with the nature of the image and is more dependent on the model used for generating the images.

4. Discussion and Outlook

Based on our survey results, we would like to suggest two potential changes for future News-Image tasks. Our first discussion point centers around the evaluation procedure. The implicit assumption of the evaluation of the retrieval task is that there is exactly one fitting image for each article. If this one-to-one relation between articles and images were to exist, we would expect Figure 1 to mirror this assumption. We would need to see only top ratings in the ground truth group and high disagreement ratings across all other image sources. However, this is not the case. Hence, the task evaluation procedure fails to account for the fact that there are multiple images—even within the same task dataset—that fit an article equally well, if not better.

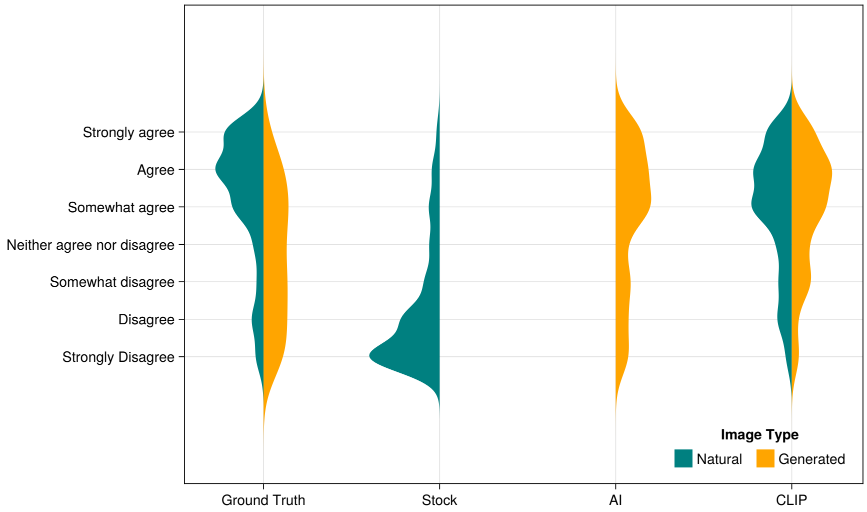

Figure 1: Distribution of participants' agreement with the statement "I think this image fits the news article." with respect to the four different image sources.

Therefore, we think the evaluation procedure should be adjusted, as the implicit assumption of exactly one relevant image per news article does not hold. It might be necessary to switch from an ex-ante relevance assessment to an ex-post one. A possible inspiration is the Inferred Mean Average Precision [6] metric, used by the TRECVid Ad-Hoc Video Search task.

The second point of the discussion focuses on the inclusion of AI-generated content. Our results show that AI images tend to receive lower overall ratings. As a consequence, we see that including both natural and generated images creates a certain tension in the task's goals/aspiration vs. its execution. Suppose finding a fitting image is the primary objective of this task, and AI-generated content is part of the ground truth. In that case, this leads to the creation of retrieval techniques that select images that are empirically shown to be perceived as less fitting than comparable natural ones. Competing systems would be required to focus on re-creating/guessing the image generation pipeline of AI content.

Shifting the task focus from finding fitting images to trying to recreate image-generation pipelines should be avoided, as this is in conflict with the goal of the task. AI-content should, therefore, be treated differently from the editorial one. One possible solution is to have two dedicated sub-tasks with two different datasets, one with and one without AI images. Furthermore, the decision not to communicate the image generation pipeline needs careful reconsideration, as this has significant implications for the system design.

Using our survey findings, we would like to conclude our paper by giving an outlook for additional real-world challenges regarding text-image matching for news articles. Among the most important questions to investigate is the process of how editors select images. Do they primarily focus on the title, the lead, of the article text? The practical implication of this question is that the task organizers might need to provide additional information, as the dataset currently does not feature full articles. Related aspects that are worth investigating for building retrieval pipelines are the intentions of editors when selecting images, the tools at their disposal, domain-specific requirements, together with user expectations, and varying preferences.

Acknowledgments This work was partially funded by the Digital Society Initiative (DSI) of the University of Zurich under a grant of the DSI Excellence Program and the Swiss National Science Foundation through project MediaGraph (contract no. 202125).

References

[1] A. Lommatzsch, B. Kille, Özlem Özgöbek, M. Elahi, D.-T. Dang-Nguyen, News Images in MediaEval 2023, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024, p. 4. URL: https://irml.dailab.de/wp-content/uploads/2023/11/NewsImages2023-LabOverview-v20231101.pdf.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 10674–10685. URL: https://doi.org/10.1109/CVPR52688.2022.01042. doi:10.1109/CVPR52688.2022.01042.

[3] 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, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, 2023, pp. 2818–2829. URL: https://doi.org/10.1109/CVPR52729.2023.00276. doi:10.1109/CVPR52729.2023.00276.

[4] 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, in: NeurIPS, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

[5] L. Heitz, Y. K. Chan, H. Li, K. Zeng, A. Bernstein, L. Rossetto, Prompt-based alignment of headlines and images using openclip, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[6] E. Yilmaz, J. A. Aslam, Estimating average precision when judgments are incomplete, Knowl. Inf. Syst. 16 (2008) 173–211. URL: https://doi.org/10.1007/s10115-007-0101-7. doi:10.1007/S10115-007-0101-7.