HTW-DIL at Touché: Multimodal Dense Information Retrieval for Arguments

Notebook for the Touché Lab at CLEF 2024

Tamás Janusko[1], Aaron Kämpf[1,†], Denis Keiling[1,†], Jessica Knick[1,†], David Schäfer[1,†] and Maik Thiele[1]

[1]HTW Dresden, Friedrich-List-Platz 1, Dresden, 01069, Germany

Abstract

Retrieving images for arguments poses many of the problems of traditional information retrieval with the added challenge of being inherently multimodal. We adapt a dense retrieval approach to address this issue and acquire synthetic training data to fine-tune a multimodal model as part of our retriever. Furthermore we conduct ablation studies to examine the impact of different modalities and benchmark our approach against state-of-the-art methods. While the task itself is ambiguity-laden there appears to be a benefit of using only textual information for retrieving argumentative images.

Keywords

Machine Learning, Information Retrieval, Multimodal Retrieval, Image Retrieval

1. Introduction

In information retrieval tasks the key objective is to find the most relevant documents for a given query. This is also the case for the image retrieval for arguments as performed in task 3 of Touché@CLEF [1, 2] on the TIRA shared task platform [3]. Although methods like BM25 [4] are strong baselines, they are limited by their consideration of surface forms of the information-carrying elements in the document alone. While this leads to diminished recall when dealing with synonyms, for example, the case of multimodal retrieval poses the question of how to operate on the surface level at all.
Joint text and image embeddings solve both the challenges of deeper semantic understanding as well as that of relating textual to visual information. In the case of image retrieval for arguments we are confronted with an additional layer of implicitness which out-of-the-box embeddings are not equipped to handel, thus we propose a method inspired by Facebook Research's dense passage retrieval (DPR) System [5] where we fine-tune a multimodal large language model (MLLM) to maximize similarity scores of matching argument and image/text pairs.

2. Task Description

We are provided with 136 arguments consisting of topic, premise, claim, stance and type. The task is to retrieve supporting images from a web crawl of approx. 9.000 samples where each image is accompanied by additional information, among others the text content of the encompassing website and the search query used to obtain that image. A particularly difficult aspect of the task is that several arguments share a topic and have similar premises and claims while their differing stances are grounded in subtle deviations of the lines of reasoning.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

† Authors contributed equally

tamas.janusko@htw-dresden.de (T. Janusko); aaron.kaempf@stud.htw-dresden.de (A. Kämpf); denis.keiling@stud.htw-dresden.de (D. Keiling); jessica.knick@stud.htw-dresden.de (J. Knick); david.schaefer@stud.htw-dresden.de (D. Schäfer); maik.thiele@htw-dresden.de (M. Thiele)
0000000021665977X (M. Thiele)

3. Methodology

3.1. Multimodal DPR

As presented in the original DPR paper we also frame retrieval as a metric learning problem aiming to maximize dot product similarity of matching queries and targets. While DPR retrieves passages - chunks derived from larger documents - we treat each image and the accompanying textual information (website text summary and web search query) as a unit during training, although only the image is evaluated eventually.

The method employs an in-batch negative training scheme where for a given image a matching argument (positive) doubles as a negative example when paired with any other image within the batch, thus efficiently increasing the data size. Additionally, a randomly sampled argument is passed along with the positive argument to function as yet another negative for each image in the batch. This way we yield n positive pairings and n^2 negative pairings for a batch size of n. On this basis we compute the negative log-likelihood loss as implemented in the original DPR code[1] using the MLLM Moondream2[2] to facilitate operations in joint embedding space. Phi 1.5 [6] is the underlying LLM with SigLIP [7] providing the vision capabilities in Moondream2. This model choice is motivated by the favorable reported performance as well as its moderate size which is manageable with the hardware available to us.

We use a learning rate of 1e-5 with linear warm-up from 10% over the first 10% of the training and then decay back to 10% over the remaining run. With a batch size of 16 training is performed for two epochs (Ep2), with the exception of an approach with image and text input that we train for three epochs (Ep3) to probe the onset of possible overfitting.

The fine-tuned model is then used to embed the query argument as well as the image/text pairs. FAISS [8] indices are computed for all embedded image/text pairs and the top-k most similar instances for each argument embedding are retrieved.

Note that the final evaluation is only based on the retrieved image, rendering the textual website content as merely supportive context information for the retrieval task which is not directly considered by the judges.

3.2. Data Pre-Processsing

Since the input length of any large language model (LLM) is finite we perform several pre-selection and pre-processing steps. Firstly, we use only the image, website content text and query string to represent an image and its website context. Additionally, images are scaled to 256 pixels at their largest dimension, and content text is summarized with BART fine-tuned on CNN Daily Mail[3] for summarization [9]. Inputs too large for summarization models are chunked in suitable sizes and separately condensed and then re-concatenated. If a website's content consists mostly of structured text such as lists, no summarization is performed.

Since arguments are given in XML-format we join the argument elements and topic into a concise natural sentence, but without the type information.

3.3. Synthetic Train Data

In order to train a model in the first place we need a training set for the task at hand. Using the multimodal capabilities of OpenAIs GPT-4 [10] we generate synthetic arguments by inferring plausible argument elements from available image/website data. Each image/summary is used to derive one argument topic for which in turn a premise and claim are generated for pro and con stances. The

[1] https://github.com/facebookresearch/DPR/blob/main/dpr/models/biencoder.py#L254
[2] https://huggingface.co/vikhyatk/moondream2
[3] https://huggingface.co/facebook/bart-large-cnn

resulting argument is given in valid XML-format. We do not distinguish between *anecdotal* and *study* types as there were no examples of the latter at the time of development.

3.4. Ablative Approaches

In addition to using image and text data jointly (*Moondream Default, Ep2, Ep3*) we experiment with using the images (*Moondream Image*) and the textual information (*Moondream Text*) separately in order to examine whether an unimodal approach represents a feasible alternative within our DPR-like setup. For this purpose Moondream models are fine-tuned using only images or website content text (including the query string) with the same hyperparameters as the multimodal approach. Moreover, we employ *Ada*-embeddings[4] from OpenAI to represent the case of simple text-based retrieval with proven off-the-shelf technology. The rationale for this is that images found on websites are usually placed there deliberately by human authors with the intention of supporting the written content. Following that assumption and given robust text embeddings one can leverage this relation to obtain relevant images without taking them into account explicitly.

4. Manual Evaluation Results

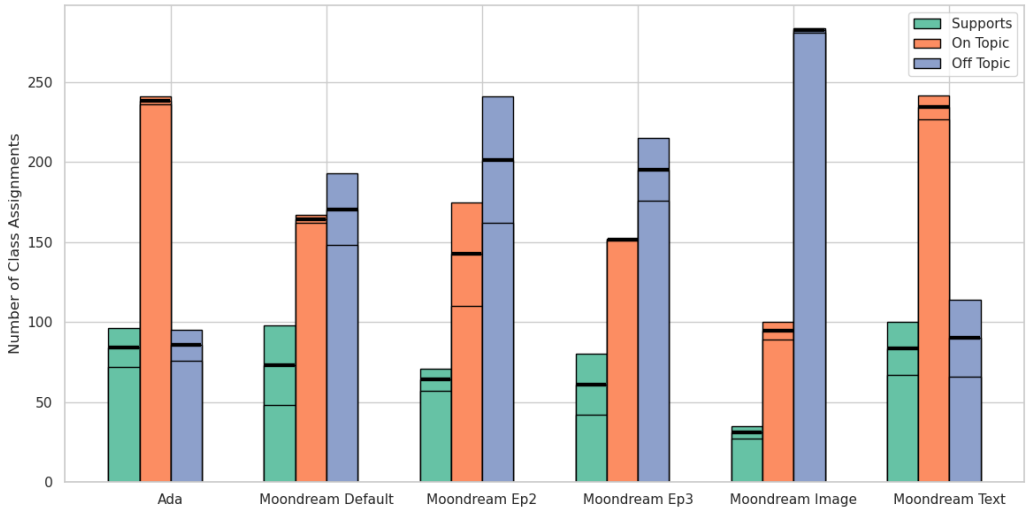

Figure 1: Class attributions per approach for retrieved images averaged over top-3 results

To obtain a better understanding of our approaches we manually evaluate the retrieval results. For each approach we examine the top-3 retrieved images and assign them to the mutually exclusive categories *supports argument*, *on-topic* and *off-topic* and compute the inter-annotator agreement using Cohen's kappa. Considering all 136 arguments, two annotators and top-3 results we yield 816 annotations per approach. In Figure 1 bars represent number of class instances from all top-3 runs and both annotators, with the bold marking representing the mean and the thinner upper and lower markings show min/max values found in annotations of single runs. We find the majority of retrieved images classified as not supportive for the corresponding argument. Only Ada and text-only Moondream show parity of supporting and off-topic classes with on-topic instances being the majority which can be interpreted as a trend of low performing image-only approaches to best performing text-only ones. This may be because only shallow semantics of images are captured by the model. It is

[4]https://platform.openai.com/docs/models/embeddings

also supported by kappa values found in Table 1 where we find the highest inter-annotator agreement for image-only Moondream, our worst performing model. In contrast, the best performing approaches, Ada and text-only Moondream, have the lowest kappa values. This can be interpreted as evidence of the inherent ambiguity of the task at hand and the many ways an image can support an argument.

Additionally, we compute the Jaccard index to quantify the similarity of results given by the different methods we employ. For this we use the top-10 results from all arguments and compare the IDs of the retrieved images. From Figure 2 we take that approaches differ substantially in their choice of relevant images, with Moondream-based approaches being fairly similar and Ada-based retrieval far behind with Moondream text-only as the closest approach. The findings from Figure 1 that text-based approaches differ significantly from the image-only Moondream approach are reaffirmed. But while the high similarity of Moondream image/text fine-tuned for two and three epochs is obvious, the non-similarity of Ada and text-only Moondream is surprising and again speaks for the high ambiguity of the challenge.

Method	Kappa
Ada	0.231
MD Std.	0.489
MD Ep2	0.439
MD Ep3	0.525
MD Image	**0.66**
MD Text	0.371

Table 1
Cohen's kappa inter-annotator agreement for two annotators and top-3 retrieved images

5. Conclusion

In this work we explored possibilities of multimodal dense image retrieval for arguments. We adapted the DPR-technique and fine-tuned a multimodal base model for different input modalities. Ablation studies suggest that text-only input is the most favorable input format and fine-tuning on images alone causes retrieval to stray off-topic. This is underlined by the similarly good results of both text-only approaches in contrast to their differences in size and purpose. However, it calls for further examination, preferably on a larger data set and with additional annotators since 136 samples and two annotators facilitate only moderately robust statistical analysis.
The main areas of interest are the contribution of information from text vs. image inputs, as well as the role and extent of ambiguity when mapping arguments to images. As a natural first step the possibility of bottlenecks caused by models that are too small and low image resolutions has to be ruled out.

References

[1] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. Garcıa Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

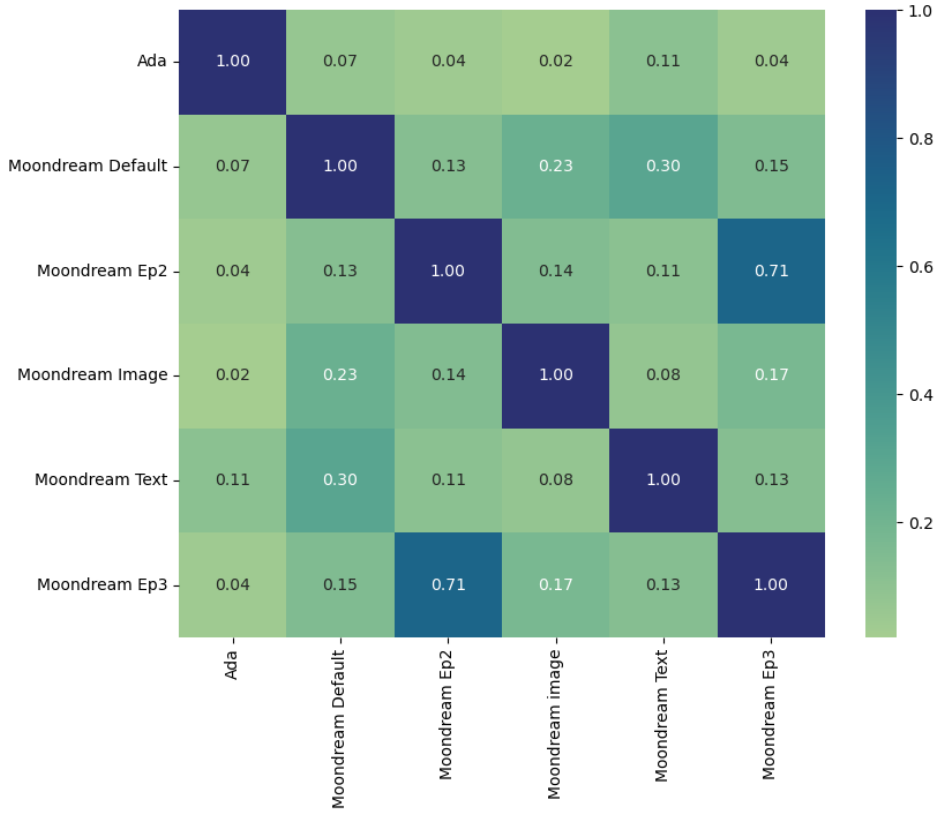

Figure 2: Jaccard indices quantifying similarities between approaches for top-10 retrieved images

[2] J. Kiesel, Ç. Çöltekin, M. Heinrich, M. Fröbe, M. Alshomary, B. D. Longueville, T. Erjavec, N. Handke, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, T. Reitis-Münstermann, M. Scharfbillig, N. Stefanovitch, H. Wachsmuth, M. Potthast, B. Stein, Overview of Touché 2024: Argumentation Systems, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[3] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[4] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Found. Trends Inf. Retr. (2009). URL: https://doi.org/10.1561/1500000019. doi:10.1561/1500000019.

[5] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 EMNLP, Association for Computational Linguistics, Online, 2020. doi:10.18653/v1/2020.emnlp-main.550.

[6] Y. Li, S. Bubeck, R. Eldan, A. D. Giorno, S. Gunasekar, Y. T. Lee, Textbooks are all you need ii: phi-1.5 technical report, 2023. URL: https://www.microsoft.com/en-us/research/publication/textbooks-are-all-you-need-ii-phi-1-5-technical-report/.

[7] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, 2023. arXiv:2303.15343.

[8] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, 2017. arXiv:1702.08734.

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461.

[10] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.