

Jean-Luc Picard at Touché 2023: Comparing Image Generation, Stance Detection and Feature Matching for Image Retrieval for Arguments

Notebook for the Touché Lab on Argument and Causal Retrieval at CLEF 2023

Max Moebius^{1,*†}, Maximilian Enderling^{1,†} and Sarah T. Bachinger^{1,†}

¹Friedrich-Schiller-University Jena, 07737 Jena, Germany

Abstract

Participating in the shared task "Image Retrieval for arguments", we used different pipelines for image retrieval containing Image Generation, Stance Detection, Preselection and Feature Matching. We submitted four different runs with different pipeline layout and compare them to a given baseline. Our pipelines perform similarly to the baseline.

Keywords

Image Retrieval, Image Generation, Feature Matching

1. Introduction

As the saying goes, "a picture is worth a thousand words". A convincing argument in writing should be accompanied by an equally convincing image. There are no perfect out-of-the-box solutions so far, which is why we participated in the shared task "Image Retrieval for arguments"[1] from Touché [2].

2. Background

In the following section, related work covering image generation with Stable Diffusion and Feature Matching is reviewed.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

†These authors contributed equally.


✉ max.moebius@uni-jena.de (M. Moebius); maximilian.enderling@uni-jena.de (M. Enderling); sarah.bachinger@uni-jena.de (S. T. Bachinger)

🌐 <https://github.com/ArcticF0x99> (M. Moebius); <https://github.com/BMI24> (M. Enderling); <https://github.com/stbachinger> (S. T. Bachinger)

🆔 0009-0005-5422-2164 (S. T. Bachinger)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2.1. Stable Diffusion

”Stable Diffusion is a latent text-to-image diffusion model capable of generating photo-realistic images given any text input.”[3]

Stable Diffusion is a neural network that generates an image corresponding to a given text input (so-called **prompt**). It is possible to specify a certain image style (e.g. ”comic”) by including it in the prompt.

For our approach, the version `stable-diffusion-v1-4` was used, which was created by resuming training from `stable-diffusion-v1-2` with 225,000 steps at a resolution of 512×512 pixels[3].

2.2. Feature Matching

”Feature matching refers to the act of recognizing features of the same object across images with slightly different viewpoints”[4]. In this context, features are defined by keypoints, which causes their name `feature keypoints`. These feature keypoints mark an area that is particularly interesting or defining in an image.

SIFT is a feature descriptor used to detect, describe and match local features of images. For that, the descriptor uses a database of images to compare with. Every feature of the new image is compared to the database with Euclidean distance of the feature vectors to recognize objects[5]. Broadly speaking, SIFT extracts feature keypoints and feature descriptor from an image. The descriptors contain the visual description of images and are commonly used to determine the similarity between images.

FLANN stands for `Fast Library for Approximate Nearest Neighbors` and is used for fast nearest neighbor search in large datasets or images. In general, a matcher takes the descriptors of two images, builds pairs of features with one feature from each image, and calculates for every pair a distance. The smaller the distance, the more similar the features are to each other. With clustering and search in multidimensional spaces, the matching by FLANN is more efficient for larger datasets compared to the often used `BFMatcher`[6].

Using a threshold, the feature matches are filtered. Every match has a distance value while the threshold is also a value. All matches with a distance under the threshold value are accepted and determined as good. The smaller the accepted distance values, the more similar are the matches between two images. That means the similar images are to each other.

Additionally, homography is used to determine the transformation between points in an image and projects them on to an image plane with a normalized camera. That means objects are viewed usually from different angles in two images, but they show the same objects. Therefore, the images have a different perspective. With homography, the objects in the image are made comparable by bringing the images into the same perspective. (Compare [7])

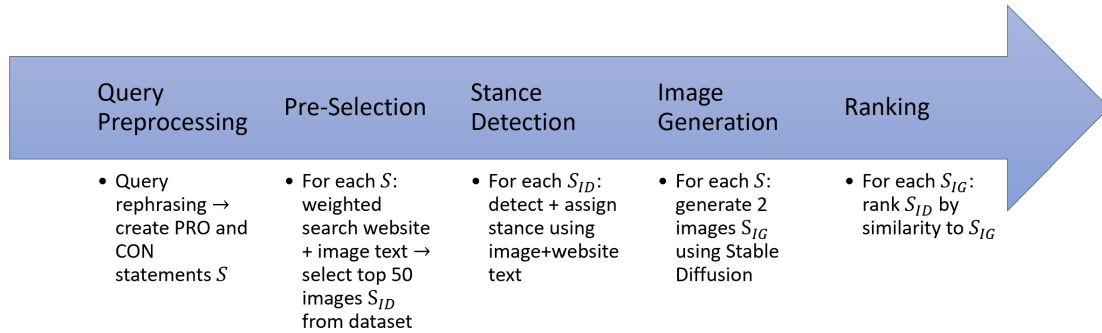


Figure 1: The whole pipeline

3. Approach

Our implementation is available on GitHub¹. The complete pipeline containing all steps is shown in Figure 1. In its current state, it's a full-rank pipeline.

3.1. Query Preprocessing

To isolate possible important terms for the following queries, the spaCy[8] library was used to parse the topic questions and generate a parse tree.

Then, all tokens that were identified as punctuation and the root word (a verb) were excluded. From the remaining, only words that had a lower Zipf frequency[9] (log base 10 of times the word appears per billion words) than 5.6 were kept. This threshold was chosen because it was the average Zipf word frequency in the given corpus. After this, it was assumed that only relevant words were kept. To support the query, the root word was placed in front of the remaining words, which kept their order.

For example, the phrase "Do we need sex education in schools?" was decomposed into the root word "need" and the remaining words "sex", "education", and "schools". The resulting intermediate query would be "need sex education schools".

For PRO queries, the intermediate query was used as it is. For the CON queries, the negation "not" was put in front of the intermediate query.

3.2. Image Preselection

For the initial image preselection, an index was constructed with PyTerrier[10] containing the ID of a document and its text content. Here, at most 4096 characters were used. Then, BM25[11] was used to retrieve the best 50 images for a given query, generated as described in 3.1.

¹https://github.com/ArcticF0x99/ir_image_retrieval

3.3. Stance Detection

For the stance detection, we used a Hugging Face pipeline² that implements the work from Wenpeng Yin and Roth [12] for zero-shot classification with the BART model after being trained on the MultiNLI data set. The pipeline was freely available and can be loaded with the "zero-shot-classification" pipeline from hugging face³.

The model receives a text and different labels and outputs the probability of each label being a good descriptor for the given text. Hence, either "contra", "pro" or "neutral" was added in front of the given query.

The highest probability was assumed to show the stance of the image. They were sorted according to the score and the image IDs returned.

3.4. Image Generation

The image generation is used to generate images with given queries/phrases. For that, we used `Stable Diffusion`.

A generated image for a query displays the information of the query visually. With the use of image comparison methods (like `Feature Matching`), every image of a dataset is compared to the generated image. The more similar an image is to the generated one, the better the image to the query. The next section explains that in more detail.

3.5. Image Ranking (Feature Matching)

For a given query, a series of images are ranked using `Feature Matching`. The set of images may include a variety of styles, therefore feature matching with generated images of various styles can improve query matching overall. As a result, cartoonish and photorealistic visuals are produced. For this, we concatenate a style prompt ("a photograph about the topic:" or "an image in comic style about the topic:") and the query, which is then processed by `Stable Diffusion` to generate an image.

Matching features between a pair of images are detected using feature matching, always between one generated image and one from the queried dataset (see section 2.2). As a result, for each image, the number of matches above the threshold is returned. The assumption is that an image with many good matches is a desirable result for the query. The images are rated according to the numbers of matches, and the top ten images are returned.

4. Results

4.1. Submission

We submitted 5 runs through Tira [13] with different combinations of the approaches described in Section 3, namely:

²<https://huggingface.co/facebook/bart-large-mnli>

³`classifier = pipeline("zero-shot-classification", model="facebook/bart-large-mnli")`

Table 1

Precision@10, precision@1, mean average precision (MAP), and p-value for student’s test for the pipelines with the curated data corpus

Pipeline ID	Precision@10	Precision@1	MAP	p-value
-1	0.146	0.130	0.122	
0	0.155	0.160	0.111	0.699
1	0.131	0.150	0.092	0.271
2	0.115	0.181	0.146	0.491
3	0.149	0.120	0.109	0.658

1. **Pipeline -1:** Baseline provided by the Touché Team [2]
2. **Pipeline 0:** Query Preprocessing and Image Preselection only
3. **Pipeline 1:** Query Preprocessing and Image Preselection with Stance Detection on text
4. **Pipeline 2:** Query Preprocessing and Image Preselection with Stance Detection on image text
5. **Pipeline 3:** Query Preprocessing and Image Preselection with Stance Detection on text and image text

In every approach, Image Generation and Image Ranking was used to determine the final ranked results.

4.2. Relevance evaluation

From the 5 pipelines, we collected the top 10 images returned by the pro and con queries for each of the 50 topics, gathering a total of 5000 images. After duplicate removal, 2091 images remained and were independently judged by three annotators with the labels *off-topic*, *pro*, *con* and *neutral*. The Fleiss-Kappa for the three annotators was 0.38. For the evaluation of the different approaches, the image judgments were curated as follows:

- if two annotators agree on a label, this label is chosen
- if there is no majority agreement and the image is labeled by more than one as *on topic*, its label will be *neutral*

4.3. Pipeline Evaluation

We evaluated the pipelines with the curated data as described above and with the judgments from the Touché Team [2]. Precision@10, precision@1 and mean average precision were calculated. Furthermore, we used a student’s t-test to find out whether the difference in values for average precision of the individual runs is significant compared to the -1 (baseline) run. Table 1 shows the results for our curated judgements, Table 2 for the Touché judgements.

5. Discussion

The results above indicate that even though the precision values vary across the different pipelines, none are statistically significant different from the baseline.

Table 2

Precision@10, precision@1, mean average precision (MAP), and p-value for student’s test for the pipelines given by the Touché Team

Pipeline ID	Precision@10	Precision@1	MAP	p-value
-1	0.127	0.130	0.097	
0	0.141	0.170	0.100	0.911
1	0.115	0.120	0.084	0.602
2	0.090	0.106	0.113	0.621
3	0.122	0.110	0.096	0.956

We observed relatively low values for precision on average. This may be partially due to missing relevant pictures in the corpus. Judging by the low Fleiss-Kappa, the task of evaluating the stance and relevance of an image for a certain topic was ambiguous, which was confirmed by the annotators. For future applications, an annotation guideline should be given to the annotators to avoid confusion. Overall, the results with the different judgements seem to agree.

We see that the inclusion of stance detection on the website text seems to be not beneficial in our case. Future work could determine if other stance detection models also suffer from this. Furthermore, selecting a higher amount of pictures than the current number of 50 during pre-selection could lead to different results. This is supported by the fact that out of the 5000 results for the 5 pipelines, actually only 1938 were unique. Since all of our results are chosen from the same $50 (\# \text{ topics}) \cdot 2 (\# \text{ of stances}) \cdot 50 (\# \text{ of pre-selected images per query via BM25, see section 3.2}) = 5000$ images, this may be part of the poor results.

6. Conclusion

As seen in Section 4, our pipelines perform not significantly better or worse than the baseline. The assumption was that with the inclusion of more information like stance detection the pipeline would perform better, namely pipeline 3 would perform best. We could not find evidence for that with our approach.

In the experimental setting, the pipeline 0 was the best in precision@10 for both judged images and pipeline 2 the best MAP for both judged images. For precision@1, pipeline 0 was best with our curated data, for the given data pipeline 2 performed best.

Different things could be changed in order to get better results like testing other variations of stance detection and image generation with image ranking. Other, untested approaches and a change in the approach combination used may lead to different results. As mentioned in the section 5, maybe the corpus is at fault and more pictures with a clearer stance may lead to better results. Increasing the number of pre-selected images could also serve as a topic for further research.

References

- [1] J. Kiesel, N. Reichenbach, B. Stein, M. Potthast, Image Retrieval for Arguments Using Stance-Aware Query Expansion, in: K. Al-Khatib, Y. Hou, M. Stede (Eds.), 8th Workshop on Argument Mining (ArgMining 2021) at EMNLP, Association for Computational Linguistics, 2021, pp. 36–45. URL: <https://aclanthology.org/2021.argmining-1.4/>. doi:10.18653/v1/2021.argmining-1.4.
- [2] A. Bondarenko, M. Fröbe, J. Kiesel, F. Schlatt, V. Barriere, B. Ravenet, L. Hemamou, S. Luck, J. Reimer, B. Stein, M. Potthast, M. Hagen, Overview of Touché 2023: Argument and Causal Retrieval, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. 14th International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, p. to appear.
- [3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.
- [4] R. Roelke, Local feature matching, 2023. URL: <https://cs.brown.edu/courses/cs143/2013/results/proj2/rroelke/>, visited on 2023-03-08.
- [5] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International journal of computer vision* 60 (2004) 91–110.
- [6] G. Bradski, *The OpenCV Library*, 2000.
- [7] O. Chum, T. Pajdla, P. Sturm, The geometric error for homographies, *Computer Vision and Image Understanding* 97 (2005) 86–102.
- [8] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, *spaCy: Industrial-strength Natural Language Processing in Python* (2020). doi:10.5281/zenodo.1212303.
- [9] G. K. Zipf, *The psycho-biology of language: An introduction to dynamic philology*, volume 21, Psychology Press, 1999.
- [10] C. Macdonald, N. Tonello, Declarative experimentation in information retrieval using pyterrier, in: *Proceedings of ICTIR 2020*, 2020.
- [11] S. E. Robertson, S. Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in: *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, organised by Dublin City University, Springer, 1994, pp. 232–241.
- [12] J. H. Wenpeng Yin, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, in: *EMNLP*, 2019. URL: <https://arxiv.org/abs/1909.00161>.
- [13] 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. Goeriot, 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. URL: https://doi.org/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

Acknowledgments

We thank the Touché organization team and Prof. Hagen and his chair for their continued helpful suggestions and support.