CEDNAV–UTB: Efficient Image Retrieval for Arguments with CLIP

Notebook for the Touché Lab at CLEF 2025

Diego Alberto Guevara Amaya[1], Jairo Enrique Serrano Castañeda[2], Juan C. Martínez-Santos[2] and Edwin Puertas[2]

[1]Naval Technological Development Center, Colombian Navy, Cartagena, Colombia
[2]School of Digital Transformation, Universidad Tecnológica de Bolívar, Cartagena, Colombia

Abstract

This paper introduces an efficient and reproducible system for argumentative image retrieval developed by the UTB–CEDNAV team for the 2025 edition of the Image Retrieval for Arguments challenge at Touché@CLEF. The system leverages the CLIP model (ViT-B/32) to represent textual arguments through images. Unlike previous approaches that rely heavily on complex text processing, image generation models, or multi-stage architectures, this solution focuses on computational simplicity. It significantly reduces energy consumption by reusing embeddings, enabling parallel processing, and eliminating redundant steps. According to measurements made using the CodeCarbon tool, this strategy resulted in an energy consumption reduction of over 85% in subsequent runs. The implementation is easy to deploy in environments like Google Colab and adheres to all Touché evaluation standards. This work provides a strong baseline for developing sustainable and scalable multimodal retrieval systems.

Keywords

Sustainable AI, Computational efficiency, Image retrieval, CLIP, Multimodal modeling

1. Introduction

Image Retrieval for Arguments is a task that leverages natural language processing and computer vision to enhance the analysis, generation, and presentation of complex ideas. This task is part of the Touché Lab at CLEF 2025 challenge [1], organized in collaboration with ImageCLEF [2], and evaluates systems capable of retrieving or generating images relevant to a textual argument. Each image should help convey the argument by illustrating it, providing examples, or evoking an emotional response, as shown in Figure 1. The dataset includes 200 arguments and a collection of over 1,000 images per argument.

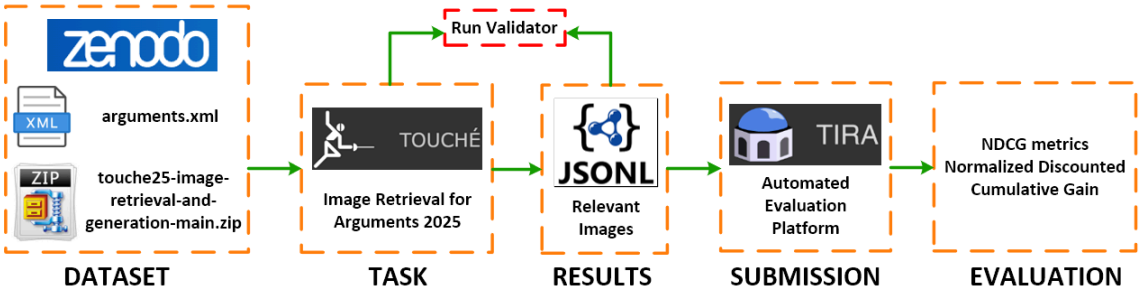

Figure 1: Retrieval for Arguments Touché-CLEF

CLEF 2025 Working Notes, September 9 – 12 September 2025, Madrid, Spain
guevarad@utb.edu.co (D. A. G. Amaya); jserrano@utb.edu.co (J. E. S. Castañeda); jcmartinesz@utb.edu.co (J. C. Martínez-Santos); epuerta@utb.edu.co (E. Puertas)
0009-0003-3192-0328 (D. A. G. Amaya); 0000-0001-8165-7343 (J. E. S. Castañeda); 0000-0003-2755-0718 (J. C. Martínez-Santos); 0000-0002-0758-1851 (E. Puertas)

To accomplish the task, we used the CLIP model (Contrastive Language–Image Pretraining) [3]. This model encodes both text and images into vector representations, enabling direct comparison and measurement of their semantic relatedness. CLIP has demonstrated strong performance in multimodal tasks and requires no additional training when used directly as a retrieval engine.

Current systems often prioritize improving retrieval accuracy without considering the computational efficiency of the process. Recent studies, such as Anthony et al. (2021) [4], highlight the importance of measuring and minimizing the carbon footprint during the training and execution of models, encouraging the use of tools that effectively track and optimize energy consumption. However, moving toward more sustainable artificial intelligence requires addressing the increasing energy demands of modern models. Canales (2024) [5] discusses several strategies to reduce the environmental impact of AI systems. In response to this need, the present research proposes a solution that optimizes the use of cloud infrastructure, executes processes in parallel, and reuses intermediate results—such as embeddings and rankings—to reduce energy consumption without compromising task performance.

The task of Image Retrieval for Arguments has practical applications in education, digital media, and language assistance systems. Images reinforce textual content, enhance understanding of technical or abstract concepts, and support visual assessment, thereby reducing bias and misinterpretation. Moreover, integrating relevant images into automatic argument generation and analysis pipelines contributes to the development of more valuable and accessible multimodal systems.

This work presents a functional, easy-to-understand, and optimized baseline that solves the task using CLIP without requiring additional training. The system delivers reproducible and reliable results with minimal manual intervention. Key contributions include:

- A multimodal pipeline for image retrieval using CLIP.
- A computational efficiency strategy that minimizes unnecessary resource usage.
- A validated baseline on the dataset provided by Touché 2025.

We organized the remainder of the document as follows: Section 2 presents the previous approaches used in earlier editions of the challenge and compares them with the proposed methodology. Section 3 describes the general architecture of the system and its workflow. Section 4 details the validation and preliminary evaluation process. Section 5 offers a critical discussion of the results obtained. Finally, Section 6 proposes possible lines of future work.

2. Background

This section provides context for the study by reviewing prior approaches, the CLIP model used, and the criteria applied for data selection in the UTB–CEDNAV System. It addresses four key aspects: (i) related work in previous argumentative image retrieval tasks, (ii) the text–image matching model that serves as the core of the system, (iii) the data selection strategy designed to ensure both efficiency and relevance and (iv) a quantitative evaluation of the system's environmental impact, offering insight into its computational sustainability compared to more resource-intensive methods. This review situates the proposed approach within the current state of the art and justifies the methodological decisions made.

2.1. Related Work

The task of Image Retrieval for Arguments has been in previous editions of the Touché challenge through various methods. Brummerloh et al. (2022)[6] employed sentiment analysis with BERT, optical character recognition (OCR) with Tesseract, and image clustering, which improved stance classification but relied heavily on text processing and manual validation. Elaina et al. (2023)[7] incorporated ChatGPT-generated arguments and combined CLIP with IBM Debater as a re-ranker, which introduced generative biases and reduced accuracy. Ostrower et al. (2024)[8] proposed generating reference images using TinyLLaMA and Stable Diffusion to compare with the corpus via CLIP. Still, the high computational cost prevented surpassing the traditional baseline. In contrast, the UTB-CEDNAV System avoids the

use of OCR, sentiment analysis, artificial generation, and external services. It relies solely on real data (image captions), significantly reducing computational load, bias, and ambiguity.

2.2. CLIP (ViT-B/32)

Developed by OpenAI and introduced by Radford et al. (2021) [9], CLIP is a multimodal learning model designed to associate images and text within a shared vector space. Although primarily built for image-text matching, its architecture supports comparisons across different modalities—text-to-text, image-to-image, and text-to-image—while preserving semantic consistency. This flexibility makes it especially effective for tasks such as argumentative image retrieval, where semantic similarity between claims and captions is crucial. In the UTB-CEDNAV System, CLIP is used precisely for this purpose, leveraging its ability to represent complex concepts in a unified space without requiring OCR or sentiment analysis.

2.3. Select Dataset Features

To optimize task performance, the work draws on findings by Theng and Bhoyar (2024) [10], who emphasize that the quality and relevance of data directly impact model performance. Based on this, they established three criteria to guide dataset selection:

- **Computational efficiency**: Reducing unnecessary data lowers processing time and resource usage.
- **Direct semantic relevance**: Prioritizing elements closely tied to the task objective enhances model interoperability.
- **Reduction of non-informative textual noise**: Eliminating irrelevant or redundant content prevents the model from learning spurious patterns, as described by Maheronnaghsh et al. (2024) [11]

2.4. Environmental Impact Assessment

To evaluate the environmental impact of the UTB–CEDNAV System, the team employed CodeCarbon [12], an open-source library developed by MLCO2, to estimate the carbon footprint associated with the computational load of running machine learning models. This tool tracks the energy consumption of Python scripts. It translates it into estimated CO_2 emissions, taking into account factors such as hardware type, geographical location, and runtime duration.

The integration of CodeCarbon reflects a growing need to develop AI systems that are both sustainable and transparent regarding their environmental cost. Unlike previous approaches to argumentative image retrieval, this work not only avoids computationally intensive techniques like OCR or synthetic image generation but also quantifies its efficiency using objective environmental metrics.

The values obtained through CodeCarbon support the system's minimalist design, demonstrating that we achieved strong performance while maintaining low energy consumption, thereby reinforcing the feasibility of sustainable solutions in real-world scenarios.

3. System Overview

Building on the outlined context, we designed the system to address the task of Image Retrieval for Arguments by adhering to two core principles: *processing efficiency* and *sustainable use of computational resources.*

3.1. Selecting Dataset Elements

The system begins with an analysis of the official Touché 2025 dataset, published on Zenodo [13], which comprises 32,339 images associated with 128 claims across 27 argument topics.

After reviewing the dataset's structure and content, we selected the following components:

- *arguments.xml*: Contains the textual arguments, particularly the claims that define the core of each argument (see Figure 2)

```
▼<argument>
   <id>1-1</id>
   <topic>Automation in the Workforce</topic>
   <claim>Automation increases work efficiency</claim>
 </argument>
```

Figure 2: Example argument structure [13]

- *touche25-image-retrieval-and-generation-main.zip*: A 500+ GB archive that includes images, HTML files, captions, and metadata. The key component is image-caption.txt, which provides precise and efficient image descriptions (see Figure 3)

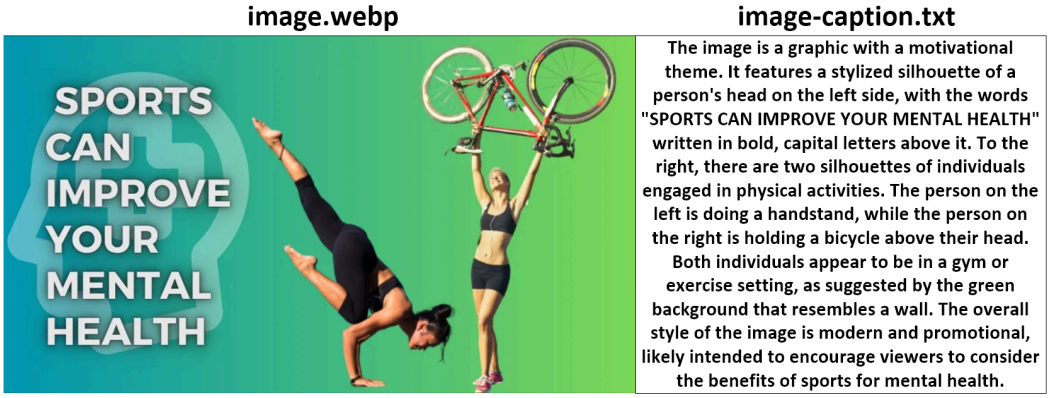

Figure 3: Example "image.web" and "caption.txt" [13]

After analyzing the metadata provided by the organizers, we observed that each image has a corresponding caption, offering a precise and concise description. Given this consistency, and in line with our objective of minimizing computational cost, we opted to compare textual embeddings between the claims of the arguments and the captions of the images. This approach allowed us to avoid direct image processing while preserving semantic alignment throughout the retrieval process.

3.2. Embeddings Pipeline

Once we identified the relevant dataset elements, the system loads captions from image-caption.txt files in parallel using ThreadPoolExecutor, as described by Sreedeep S. (2024) [14]. Each caption is then linked to its corresponding image_id and stored in a dictionary-like structure for easy retrieval.

Both the claims and captions within the system's data structure are transformed into normalized vector representations, known as embeddings, using the CLIP model (ViT-B/32). This process converts each textual input into a feature vector that captures its semantic meaning within a shared multidimensional space. Once generated, these embeddings are stored in organized, separate files, enabling efficient reuse in future system runs. As illustrated in Figure 4, before processing new data, the system checks for existing precomputed representations to avoid unnecessary recomputation. This strategy optimizes both execution time and computational resource usage, aligning with the system's commitment to sustainability. The resulting reduction in carbon footprint was quantified using the CodeCarbon tool, allowing the team to assess the system's positive environmental impact. This practice not only enhances overall system performance but also supports scalability and ensures the reproducibility of experiments.

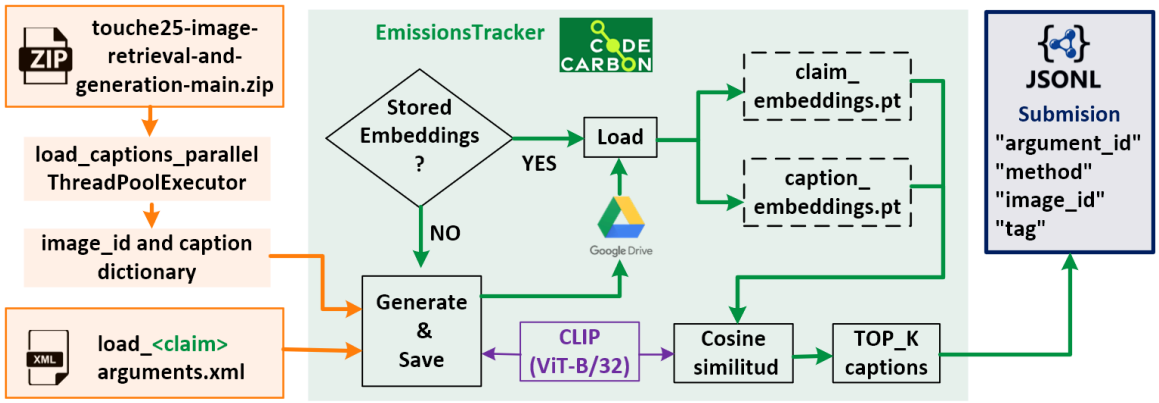

Figure 4: Pipeline CEDNAV-UTB System

3.3. Retrieval

Considering that the captions reliably describe the content of the images and are available for all entries, we decided to compute semantic similarity exclusively between claim and caption embeddings. This text–text comparison ensures alignment with the task goals while significantly reducing the computational cost associated with direct text–image processing.

To identify the most relevant images for each argument, the system computes cosine similarity between embeddings generated for each claim and all captions. Cosine similarity measures the angle between vectors in a multidimensional space, reflecting semantic alignment regardless of vector magnitude. Values closer to 1 indicate higher semantic alignment between the claim and caption, signaling more relevant associated images.

Once we calculate all cosine similarities, the system retrieves the images whose captions rank among the TOP_K most similar for each claim. It ensures that only the pictures with the most semantically aligned descriptions are selected. The final results are organized and formatted according to Touché 2025 specifications, producing an output file in submission.jsonl format.

3.4. Implementation

The system was implemented in Google Colab, providing a suitable balance between performance, simplicity, and scalability. The reuse of embeddings and conditional file downloading help minimize memory and storage usage.

To assess the environmental impact of the system, we employed the *CodeCarbon* tool to estimate energy consumption and the associated CO_2 emissions generated during pipeline execution. The initial run, which involved generating all embeddings from the *captions* and *claim* indices, consumed approximately **0.00349 kWh** of electricity, resulting in **0.00093 kg of CO_2** emissions. In contrast, subsequent runs, which reused the stored embeddings, demonstrated significantly lower energy usage, averaging just **0.00013 kg of CO_2** per run, as illustrated in Figure 5. It highlights the effectiveness of the reuse strategy as a means to reduce the system's carbon footprint.

4. System Validation

The source code of the UTB–CEDNAV System is openly available on GitHub [15], facilitating review, reuse, and extension by other researchers.

To ensure compliance with the Touché 2025 challenge rules, the system output was validated using the Official Validator, provided by the organizers. Upon uploading the generated jsonl file, the validation log reported:

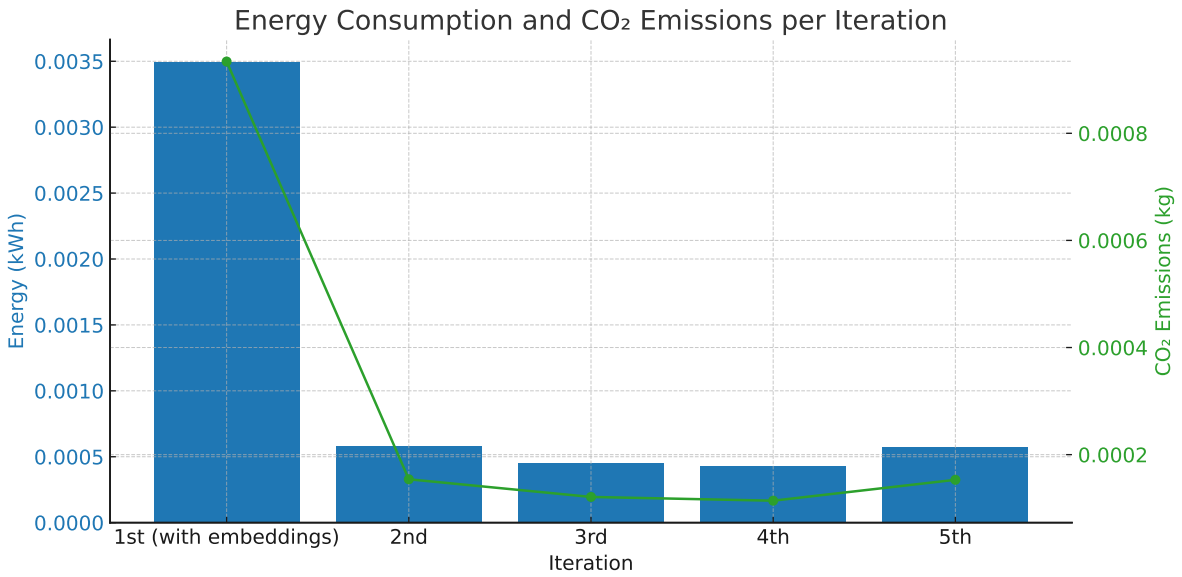

Figure 5: Energy emissions iterations

Note: The second, third, fourth, and fifth iterations reused the embeddings generated during the first iteration, optimizing execution time and reducing the system's overall energy consumption.

INFO: Starting run file validation on Monday, May 26, 2025, 15:09:59 GMT-0500 (Standard hour in Colombia) SUCCESS: Run file valid

This result confirms that the file meets the technical requirements of the challenge and can be used in the official evaluation, strengthening the traceability, validity, and reproducibility of the reported results. Additionally, for each argument, the system presents the most relevant images retrieved by the algorithm, as illustrated in Figure 6.

Figure 6: Retrieval Validation Successful

We registered the UTB–CEDNAV System on the challenge's official platform through TIRA [16], which successfully validated its output format and structure. However, at the time of this submission, the complete automatic evaluation remains unavailable as the organizers are finalizing the relevance judgments.

According to direct communication with the organizing team, the current evaluation is preliminary. It serves to identify potential issues, such as missing image coverage per query or invalid retrieved image_ids. The official results will be available after the submission deadline once a valid judgment pool system has been output. This approach ensures a fair and consistent evaluation across all participating systems.

5. Discussion

The UTB-CEDNAV system stands out from previous work by adopting a direct, reproducible, and computationally efficient approach. Unlike earlier proposals that rely on intensive text processing—such as OCR and sentiment analysis [6], re-ranking with external argumentation models [7], or synthetic visual generation through diffusion models [8]—this system operates solely on real data provided in the official dataset (claims and captions).

This design avoids unnecessary technological dependencies such as re-rankers, additional classifiers, or external APIs. Not only does this simplify the pipeline, but it also significantly reduces computational resource consumption. While approaches using OCR or image generation may require 4 to 10 times more operations per argument (due to the use of heavy models like Tesseract, LLaMA, or Stable Diffusion), UTB–CEDNAV System limits itself to direct transformation with CLIP and similarity comparison. This results in shorter execution times and lower energy consumption.

A key optimization implemented in the system is the pre-check for already computed embeddings. Before processing new data, the system checks if it has already stored embeddings to avoid redundant computations. This technique significantly reduces both runtime and resource usage by promoting intelligent reuse of intermediate results.

As a result, we minimized the system's carbon footprint in line with the principles of sustainable artificial intelligence. This reduction was quantified using the CodeCarbon tool, which enabled an accurate estimation of the system's positive environmental impact compared to more computationally intensive alternatives.

The deliberate exclusion of HTML parsing, generative models, or synthetic data reflects a commitment to algorithmic sustainability and methodological traceability. Furthermore, the successful validation of the result file using the official challenge tool confirms strict compliance with the task rules, ensuring both reproducibility and reliability.

6. Conclusions

This work presents a robust, reproducible, and environmentally responsible system for argumentative image retrieval in the Touché 2025 challenge. Its design is grounded in three core principles:

- The exclusive use of the CLIP model (ViT-B/32) to transform text into embeddings within a shared vector space.
- Efficient batch processing with reuse of previously generated resources.
- A mindful data selection strategy that avoids redundant operations and reduces computational load.

Unlike other approaches that integrate generative models, synthetic visual analysis, or additional neural networks for classification or re-ranking, this system minimizes technical complexity and energy consumption. It makes it particularly well-suited for resource-constrained environments or institutions committed to digital sustainability.

Additionally, the strategy of reusing previously stored representations proved highly effective: after the initial run, which required whole embedding generation, subsequent executions showed an **energy consumption reduction of over 85%**, with average emissions as low as **0.00013 kg of CO_2** per run. This measurable difference highlights the positive impact of avoiding unnecessary recomputation. It reinforces the importance of designing optimized pipelines that prioritize both computational efficiency and environmental sustainability in resource-intensive AI tasks.

7. Future Work

Future directions for the system include:

- Integrating lightweight models for semantic stance classification, enabling the system not only to assess image-argument relevance but also to determine whether an image supports or opposes a given argument
- Evaluating low-impact *visual question answering* techniques for re-ranking previously retrieved results, using simple model queries such as "Does this image support the argument?" to improve result ordering with minimal computational cost.
- Exploring hybrid embeddings that combine efficiency with lightweight generative capabilities, blending CLIP with small models that better capture argumentative context without adding latency or complexity

Together, these enhancements position the UTB–CEDNAV System as a viable path toward more sustainable multimodal artificial intelligence without compromising performance or coherence in the task of argumentative retrieval.

Acknowledgments

We thank the Integral Naval Education Command of the Colombian Navy for providing the necessary resources and the Naval Technological Development Center for offering a suitable environment to conduct this research. We thank the team of the Artificial Intelligence Laboratory VerbaNex [1], affiliated with the UTB, for their contributions to this project.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 to:

- Write and structure the scientific article.
- Synthesize comparisons between previous approaches.
- Verify grammatical consistency and argumentative clarity.

After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

CRediT Author Statement

Diego Alberto Guevara Amaya: Conceptualization, Methodology, Software, Formal Analysis, Writing – Original Draft, Visualization.

Jairo Enrique Serrano Castañeda: Supervision, Writing – Review & Editing, Project Administration.

Juan C. Martinez-Santos: Resources, Validation, Writing – Review & Editing.

Edwin Alexander Puertas Del Castillo: Funding Acquisition, Institutional Support, Writing – Review & Editing.

References

[1] J. Kiesel, Ç. Çöltekin, M. Gohsen, S. Heineking, M. Heinrich, M. Fröbe, T. Hagen, M. Aliannejadi, T. Erjavec, M. Hagen, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, H. Scells, I. Zelch, M. Potthast, B. Stein, Overview of Touché 2025: Argumentation Systems, 2025. URL: https://link.springer.com/chapter/10.1007/978-3-031-88720-8_67. doi:10.1007/978-3-031-88720-8_67, accessed May 26, 2025.

[1] https://github.com/VerbaNexAI

[2] M. Heinrich, J. Kiesel, M. Wolter, M. Potthast, B. Stein, Touché-argument-images | ImageCLEF / LifeCLEF - multimedia retrieval in CLEF, 2025. URL: https://www.imageclef.org/2025/argument-images, accessed May 26, 2025.
[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. arXiv:2103.00020, accessed May 26, 2025.
[4] L. F. W. Anthony, B. Kanding, R. Selvan, Carbontracker: Tracking and predicting the carbon footprint of training deep learning models, 2020. URL: https://arxiv.org/abs/2007.03051. arXiv:2007.03051, accessed May 26, 2025.
[5] J. C. Luna, Sustainable ai: How can ai reduce its environmental footprint?, 2024. URL: https://www.datacamp.com/es/blog/sustainable-ai, accessed May 26, 2025.
[6] T. Brummerloh, M. L. Carnot, S. Lange, G. Pfänder, Boromir at Touché 2022: Combining Natural Language Processing and Machine Learning Techniques for Image Retrieval for Arguments, 2022. URL: http://ceur-ws.org, cLEF 2022, September 5–8.
[7] D. Elagina, B.-A. Heizmann, M. Koch, G. Lahmann, C. Ortlepp, Neville Longbottom at Touché 2023: Image Retrieval for Arguments using ChatGPT, CLIP and IBM Debater, 2023. URL: http://ceur-ws.org, cLEF 2023, September 18–21.
[8] B. Ostrower, P. Aphiwetsa, DS@GT at Touché: Image Search and Ranking via CLIP and Image Generation, 2024. URL: http://ceur-ws.org, cLEF 2024, September 09–12.
[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. Accessed May 26, 2025.
[10] D. Theng, K. K. Bhoyar, Feature selection techniques for machine learning: a survey of more than two decades of research, 2025. URL: https://link.springer.com/10.1007/s10115-023-02010-5. doi:10.1007/s10115-023-02010-5, accessed May 26, 2025.
[11] M. J. Maheronnaghsh, T. Akbari Alvanagh, Robustness to spurious correlation: A comprehensive review, 2024. Accessed May 26, 2025.
[12] A. Lacoste, S. Luccioni, V. Schmidt, T. Dandres, Codecarbon: Estimate the carbon footprint of your compute usage, 2021. URL: https://github.com/mlco2/codecarbon. doi:10.5281/zenodo.5105071, accessed May 26, 2025.
[13] M. Heinrich, J. Kiesel, M. Wolter, M. Potthast, B. Stein, Touché25-image-retrieval-and-generation-for-arguments, 2025. URL: https://doi.org/10.5281/zenodo.15123526. doi:10.5281/zenodo.15123526, accessed May 26, 2025.
[14] S. S., Parallel processing in python with ThreadPoolExecutor, 2024. URL: https://www.linkedin.com/pulse/parallel-processing-python-threadpoolexecutor-sreedeep-surendran-hsbhc, accessed May 26, 2025.
[15] HIPERDAGA, cleaf25-image-retrieval-for-arguments, https://github.com/HIPERDAGA/cleaf25-image-retrieval-for-arguments, 2025. Accedido el 26 de mayo de 2025.
[16] 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, 2023. doi:10.1007/978-3-031-28241-6_20, accessed May 26, 2025.