Enhancing Domain-Specific Web Crawling with CLIP

Team #15

Aleksandra Pshenova 20244673 | Julien Dupont 20256086 | Nuradil Zhambyl 20220862 | Matthew Heffernan 20256098

Abstract

We propose a novel method of targeted Web crawling and classifying pages using both textual content and images from each webpage. Our proposed method is based on 'BERT Based Topic-Specific Crawler' [1], and we replace SBERT textual embeddings with CLIP for both text and image.

- **Problematic:** Web page classification only considered the scraped textual content, ignoring the images scraped from each website.
- **Proposed Solution(s):** add CLIP features to the pipeline to process image output and adapted the pipeline for live Web search.

Motivation

By enhancing the base SBERT-based model with keyword-driven Web search, we aim to create a useful tool with specific application in domains where precise categorization of the Web search results is required.

I. Contributions

- Enabled **visual embedding** via CLIP
- Switched from **SBERT to CLIP** textual embeddings
- Implemented **image classification** and **filtering** system for CLIP
- Advanced existing system to **keyword-guided live Web search**, instead of hardcoded URL list
  → created a useful topic-specific search tool
- **Evaluated** the modified model using chunks embedding and LLM aided text summarization
- **Optimized and refactored** the original code

III. Related Work & Method

This project is based on 'BERT Based Topic-Specific Crawler [Tawil, Alqaraleh, in Innovations in Intelligent Systems and Applications Conference (ASYU), 2021].

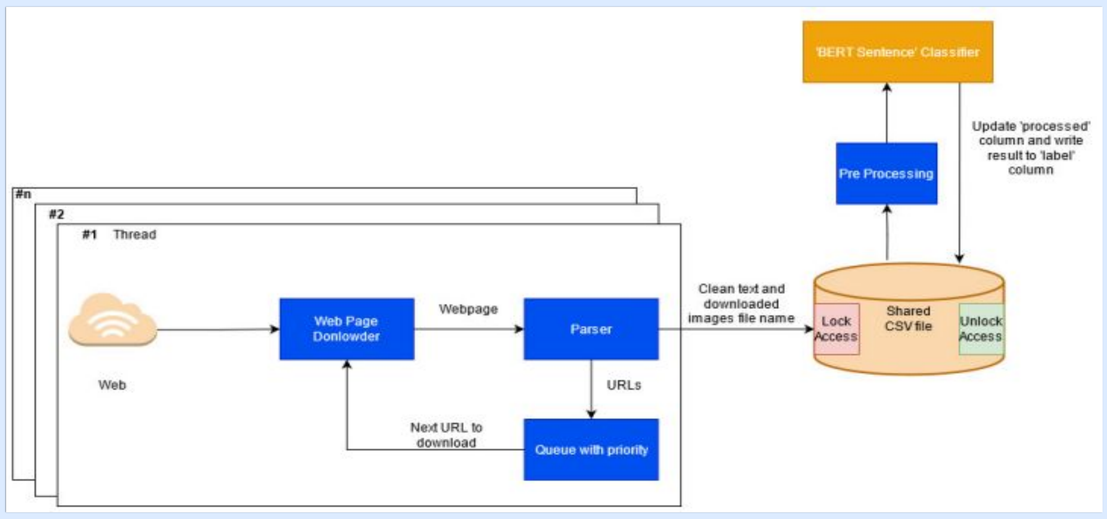

Fig. 1. The original web crawling pipeline proposed in [1]

$$\hat{c} = \arg\max_{c \in C} \cos\left(\Phi_{\text{sent}}(x), \Phi_{\text{sent}}(c)\right) \quad (1)$$

Where: Φ is the embedding model. C is a set of the predefined classes. x is the input text.

Fig. 2. Cosine similarity calculation between content and label embeddings used in [1]

The original system used recursive scraping to retrieve and download pages from a list of predefined URLs, and then classified the pages into one of predefined categories based on semantic similarity between page text content and label embeddings.

According to the authors, system is capable to download and index images in order to be used in **future work** to classify the web pages using **both their text and images**.

In order to enable image encoding in the current model, we added an image encoder from CLIP. We also changed the text encoder to CLIP for uniformity.

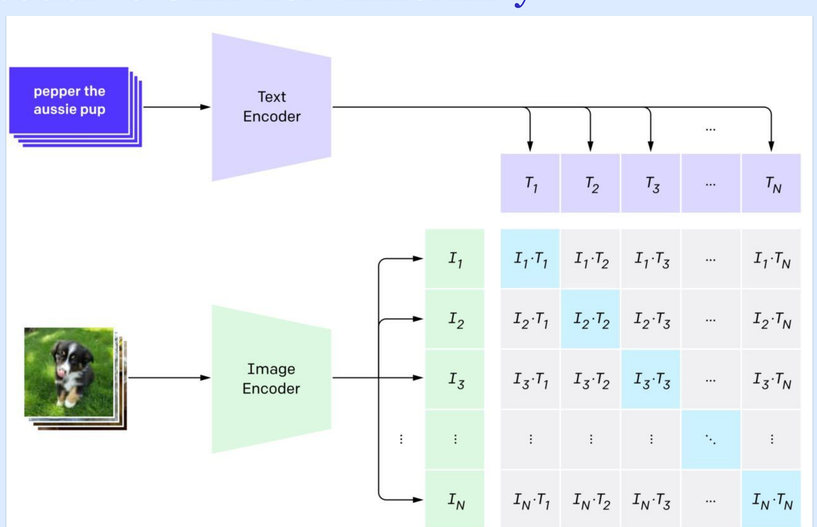

Fig. 3. CLIP Contrastive pre-training approach

To shape the system into a tool with real-world application, we swapped static URL database input with a keyword-based query, scraped the top relevant links, and classifies them. We also enabled an image filter for CLIP to dedupe and filter images irrelevant to the query.

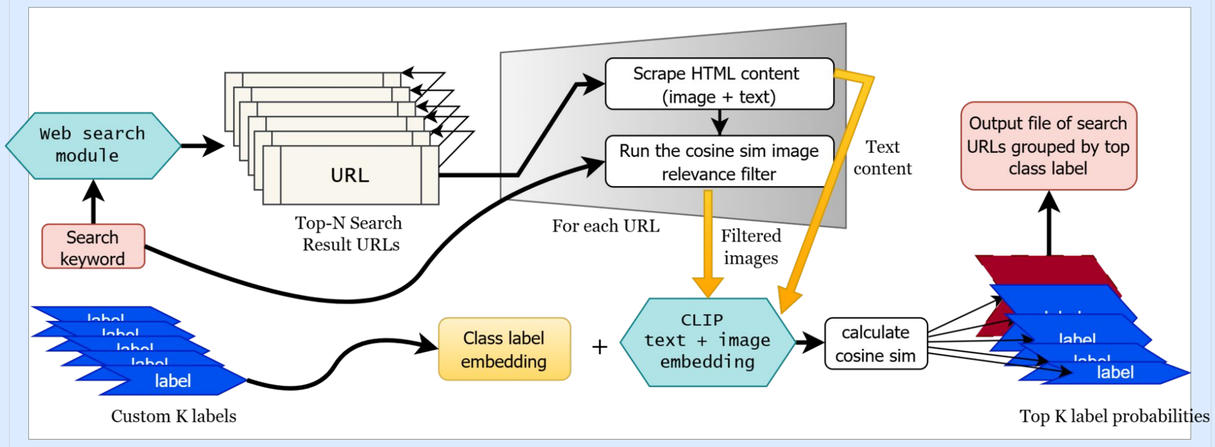

Fig. 4. Our proposed keyword-driven Web search pipeline

IV. Evaluation & Results

We compare our proposed system results with the baseline model:

Total predictions: 4460	Tech	History	Business	Politics
Match	1230	334	200	133
Mismatch	654	516	21	38

Table 1. Number of matching and mismatching samples in the baseline system

Total predictions: 451	Tech	History	Business	Politics
Match	66	147	4	26
Mismatch	94	50	47	10

Table 2. Results for our system using all page content embedded in chunks

Total predictions: 579	Tech	History	Business	Politics
Match	189	130	11	38
Mismatch	37	106	53	2

Table 3. Results for our system using summarized page content

Evaluation results show that adding CLIP alone does not improve base model accuracy due to its limited context window and noisy image inputs. This creates more space for improvement and thus we propose our modified pipeline

VI. Limitations & Future works

- **Current system could benefit from domain-specific fine-tuning.** In this project, we only utilized pre-trained models without any fine-tuning.
- Enabling more filters using cosine sim and rule-based systems could make the system more robust and accurate.

References:

[1] Y. Tawil and S. Alqaraleh, "BERT based Topic-Specific Crawler," *IEEE Conference Publication | IEEE Xplore*, Oct. 06, 2021.

[2] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," *arXiv.org*, Feb. 26, 2021.