Product

SafeBrowse: Protection from Not Suitable for Work (NSFW) Images

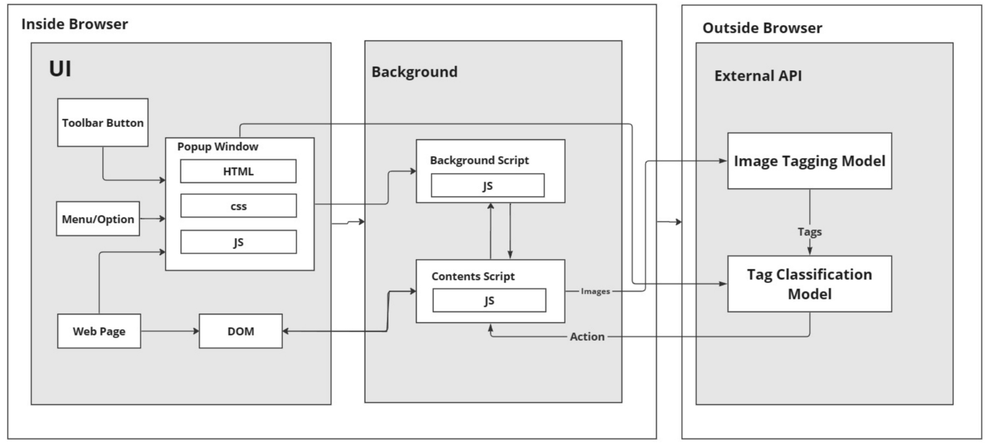
Jasmit Rathod , Ravishankar Pandey , Simar kaur
Department of Information Technology, The Bombay Salesian Society's
Don Bosco Institute of Technology Mumbai - 400070

2023 - 2024

Abstract

In today's online world, the increasing presence of Not Suitable for Work (NSFW) content raises concerns about its potential detrimental impact to mental well-being. Amidst these concerns, this project introduces "SafeBrowse", a novel browser extension meticulously designed to enhance online safety by preventing the display of NSFW images. What sets SafeBrowse apart is its unique approach, departing from conventional methods relying on simplistic keyword lists and image recognition algorithms. The extension employs a sophisticated process, beginning with the extraction of image URLs from web pages. It then utilizes advanced image-to-text models, specifically GIT and BLIP, to generate detailed textual descriptions of images. The all-MiniLM-L6-v2 sentence transformer model is used to calculate the cosine similarity score between the captions generated by GIT and BLIP using a threshold of 0.65 to determine selection of more accurate caption amongst the two. Followed by which, a finely-tuned deBERTa model subsequently classifies captions into categories such as violent, sexually explicit, and neutral, harnessing both visual and textual data to enrich content evaluation. SafeBrowse also boasts a dynamic blocking mechanism, enabling users to customize content filtering based on their sensitivity levels, offering a personalized and adaptive safeguard. While initial pilot testing emphasizes the effectiveness of the ensemble approach rather than specific accuracy metrics, the paper outlines a forward-looking agenda, prioritizing enhancements in user experience. Users can access the extension's user interface through a convenient popup window, triggered by clicking on the extension icon. This project is poised to make a significant impact in the online ecosystem by enhancing user safety, reducing unwanted content exposure, and granting individuals greater autonomy over their online experience. With coanted content exposure and granting greater autonomy, it contributes to a safer digital environment for all users.

System Architecture



Results

Fig 1. Before SafeBrowse Fig 2. After SafeBrowse

1) Fig 1. displays a web page with 3 images in which two of them are violent/nsfw.
2) In Fig 2. the violent images are blocked by SafeBrowse's blocking mechanism showing a message.

A noteworthy pattern emerges as processing time increases non-linearly with larger image sets, suggesting potential performance bottlenecks. A linear regression model as shown in Figure effectively captures this relationship, suggesting a 0.155 minute increase in processing time for each additional 9 image URLs. This non-linear growth in processing time indicates a time complexity likely greater than linear, potentially approaching quadratic or higher.

Literature Review

Model	Accuracy	Focus	Method	Limitations
DNN-based solution (Bicho et al.)	94 %	Web archiving	Transfer learning with Arquivo.pt dataset	Computationally expensive feature extraction
ChildNet (Zhuravlev et al.)	Superior to classical CNNs	Pixel-based nudity detection	21-layer deep neural network with reduced filter size	Limited to pixel-based features
CNNs (Smith et al.)	97%	NSFW content detection	Convolutional Neural Networks	May struggle with grayscale images
pLSA (Lienhart & Hauke)	92.7% correct positive rate, 1.9% false positive rate	Protecting children online	Probabilistic Latent Semantic Analysis	Lower accuracy compared to some CNN models
SafeBrowse	Not specified (pilot testing focused on ensemble approach)	Secure browsing experience	Image URL extraction, image-to-text models (GIT, BLIP), deBERTa classification, dynamic blocking	Real time user feedback not considered

Algorithm

```
console.log("Welcome To SAFEBROWSE's console");
// Function to extract image URLs from the page
function extractImageUrls() {
  const images =
Array.from(document.querySelectorAll("img"));
  console.log("image array : ",images);
  return images.map((img) => img.src);
}
// Store original image URLs before hiding them
const originalImageUrls = {};
```

1) Extension is initialised with a welcome message.
2) Image urls from src are extranted and stored in an arrray.

```
# Generate captions from both
caption_git = git_coco_pipeline(image,
max_new_tokens=50)[0]['generated_text']
            print("Caption generated by GIT:" , caption_git)
            caption_blip = blip_pipeline(image,
max_new_tokens=50)[0]['generated_text']
            print("Caption generated by BLIP:" , caption_blip)
            # Select the best caption
            best_caption = select_best_caption(caption_git,
caption_blip)
```

3) Git and Blip caption is generated from extracted url.
4) Best caption is selected and sent to server.

```
# Classify the selected caption and send response to server
 for text in texts:
    result = classifier(text, labels , max_new_tokens=50)
    results.append({
       "text": text,
       "category": result["labels"][0],
       "scores": result["scores"][0]
    })
```

5) Zero-shot classification is employed to categorize the selected caption into one of three categories: "Neutral," "Sexually Explicit," and "Violent".

6) Subsequently, the image is blocked based on its classified category.

```
if ((category === "Sexually explicit" || category === "Violent") && score > 0.5) {
    console.log(`Blocking image with URL: ${imageUrl}`);
    // Load default image instead of blocking
    const originalSrc = originalImageUrls[imageUrl];
    if (originalSrc) {
      const images = document.querySelectorAll("img");
      images.forEach(img => {
        if (img.src === originalSrc) {
          img.src =
'https://placehold.jp/20/dedede/433d3d/250x250.png?text=Content%20is%20blocked%20%
0Aby%20SafeBrowse'; //Load placeholder
          img.style.visibility = "visible"; // Make the image visible
        }
      });
    }
}
```

References

[1] SciSpace - Question. "What are the effects of NSFW content on mental health?: 5 answers from research papers." (n.d.). [Online]. Available: https://typeset.io/questions/what-are-the-effects-of-nsfw-content-on-mental-health-337gyhraoo.

[2] Alguliyev, R.M., Abdullayeva, F.J., & Ojagverdiyeva, S.S. (2022). "A Deep Learning Approach to Identify Not Suitable for Work Images." Journal of Information Security and Applications, 65(C), 103123. DOI: 10.1016/j.jisa.2022.103123.

[3] Zhelonkin, Dmirty & Karpov, Nikolay. (2020). Training Effective Model for Real-Time Detection of NSFW Photos and Drawings. 10.1007/978-3-030-39575-9-31.

[4] Lienhart, Rainer & Hauke, Rudolf. (2009). Filtering adult image content with topic models. 1472 - 1475. 10.1109/ICME.2009.5202781.

[5] S. Takkar, A. Jain, and P. Adlakha, "Comparative study of different image captioning models," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1366-1371, doi: 10.1109/ICCMC51019.2021.9418451.

[6] J. Wang, Z. Yang, X. Hu, & L. Li, "GIT: A Generative Image-to-Text Transformer for Vision and Language," arXiv:2205.14100 [cs.CV], Dec 2022. [Online]. Available: https://arxiv.org/abs/2205.14100.

[7] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," arXiv:2201.12086v2 [cs.CV], 15 Feb 2022.

[8] Michel, P., Bailleul, A., & Le Cun, Y. (2021). MiniLM: Efficient Pre-training of Language Models with Large Masked Language Models. arXiv preprint arXiv:2212.07617: https://arxiv.org/abs/2212.07617

[9] He, P., Liu, X., Gao, J., & Chen, W. (2021). "DeBERTa: Decoding-enhanced BERT with disentangled attention," arXiv:2006.03654v6 [cs.CL], 6 Oct 2021.