AI-Powered Image Tagging using Generative & Vision-Language Models

INTRODUCTION

This presentation explores an innovative approach to automated image tagging, leveraging the power of generative and vision-language models. Our method combines BLIP, KeyBERT, and CLIP to achieve highly accurate, context-aware image tagging, enabling various applications such as content moderation, smart photo albums, and automated social media tagging.

A GenAI project using BLIP. KeyBERT, and CLIP for automatic image tagging. This layout showcases the problem statement, solution overview, tools and technologies, and implementation details.

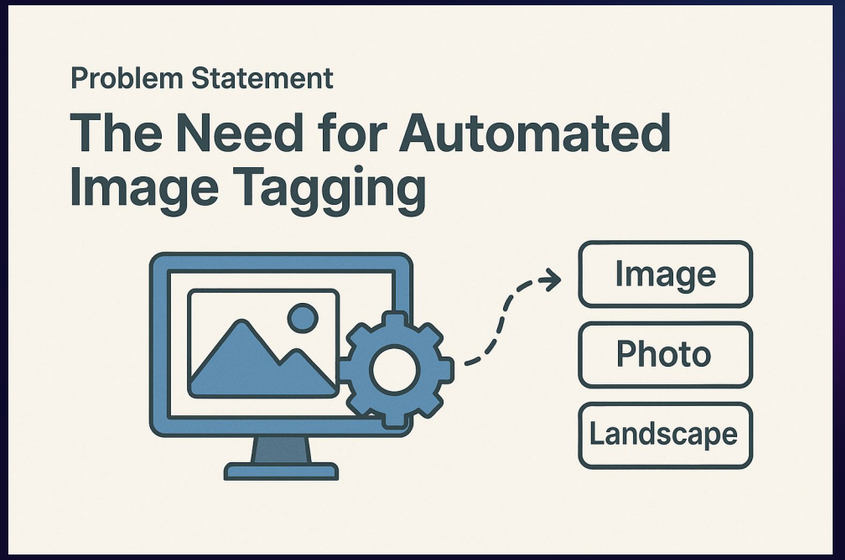
Problem Statement: The Need for Automated Image Tagging

Manual image tagging is a labor-intensive, time-consuming, and inconsistent process. The need for an automated, high-quality, and context-aware image tagging solution is paramount. Automatic image tagging has the potential to revolutionize various industries, enhancing content moderation, streamlining image search, and facilitating the creation of smart photo albums.

- Tedious and time-consuming
- Inconsistent and subjective
- Not scalable to large image datasets

Use Case Examples

- Instagram/Pinterest-style auto-tags
- Content moderation and discovery
- Smart photo albums



Solution Overview: BLIP + KeyBERT + CLIP Stack

1

Image Input

Raw image as input.

2

BLIP Captioning

Generates a context-aware caption using BLIP.

3

KeyBERT Tag Extraction

Extracts keywords using KeyBERT from the generated caption.

4

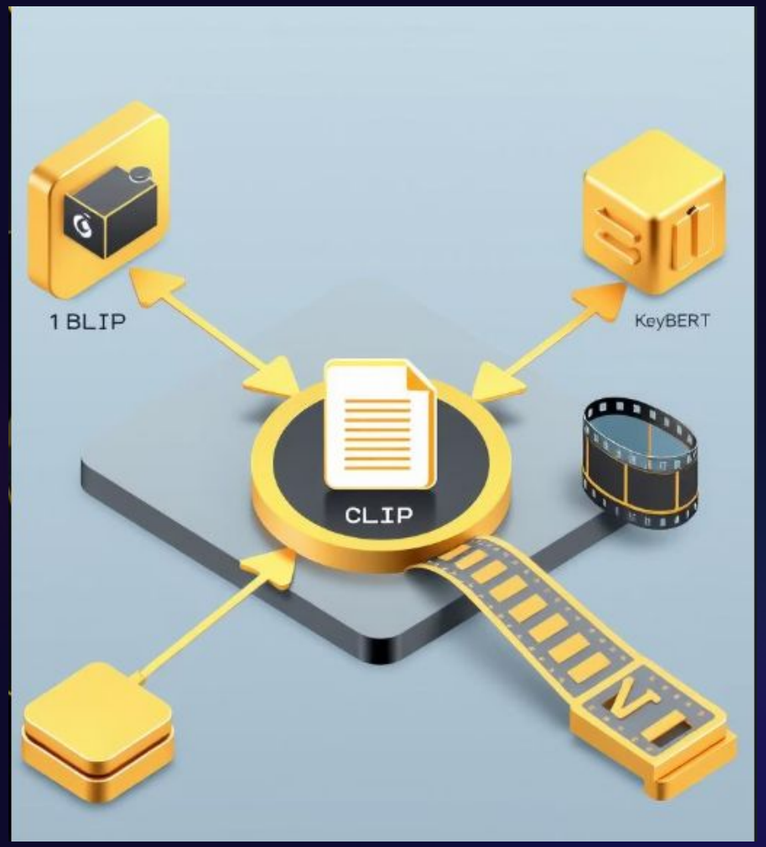
CLIP Relevance Scoring

Uses CLIP to filter and retain only visually relevant tags.

5

Final Tags

Outputs the refined set of tags.



Our solution leverages the strengths of three powerful models: BLIP, KeyBERT, and CLIP. The process begins with generating a descriptive caption of the input image using BLIP, followed by extracting relevant keywords from the caption using KeyBERT. Finally, CLIP is employed to assess the visual relevance of each tag, ensuring high-precision image tagging.

Tools & Technologies: The Building Blocks

BLIP

Vision-Language Captioning (Salesforce)

KeyBERT

Keyword Extraction (BERT embeddings)

CLIP

Image-Text Similarity (OpenAI)

PyTorch

Deep Learning Framework

Our project utilizes a combination of cutting-edge tools and technologies to achieve accurate and efficient image tagging. We leverage the power of BLIP for vision-language captioning, KeyBERT for keyword extraction, and CLIP for image-text similarity. These tools are built upon the foundation of Transformers, Sentence-Transformers, PyTorch, and are implemented within a Jupyter Notebook environment.

How BLIP Works: Generating Context-Aware Captions

BLIP (Bootstrapping Language-Image Pre-training) is a vision-language model developed by Salesforce. It is pre-trained on large image-text datasets to generate descriptive and context-aware captions for input images. BLIP excels at capturing the nuances and details of an image, providing a rich textual representation for downstream tasks.

Input: Raw Image

Takes an image as input and generates a context-aware caption.

Output: Caption

The output is a descriptive sentence about the image.

Pretrained Model

BLIP is pretrained on large image-text datasets.

Example: Image: [image of a dog running in a field].

Caption: “A dog running across a grassy field on a sunny day”.

How KeyBERT Works: Extracting Relevant Keywords

KeyBERT is a keyword extraction technique that leverages BERT embeddings to identify the most relevant keywords or key phrases from a given text. By calculating the cosine similarity between the text embedding and the embeddings of candidate keywords, KeyBERT effectively extracts the terms that best represent the content.

Input: Caption Text

Takes the caption generated by BLIP as input.

Output: Keyphrases

Returns the top N keywords/keyphrases from the text.

BERT Embeddings

Uses BERT embeddings and cosine similarity for extraction.

Example: Caption: "A group of friends hiking in the mountains during sunset".

Tags: [“hiking”, “mountains”, “friends”, “sunset”].

How CLIP Filters Tags: Ensuring Visual Relevance

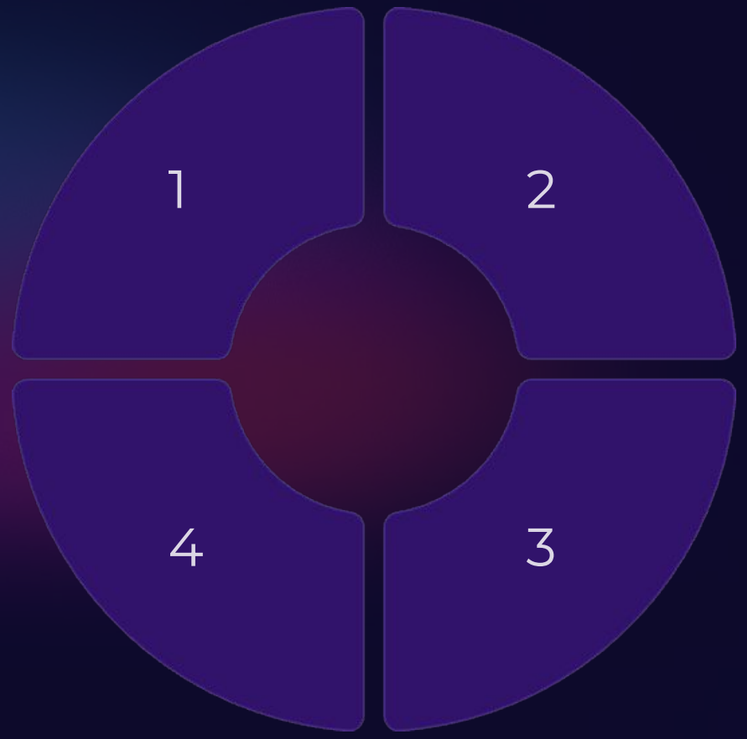


Image Input

Input Image

Tag Input

Extracted Keyword Tag

Similarity Score

Cosine similarity > 0.25

CLIP Embedding

Embed both image and tag.

CLIP (Contrastive Language-Image Pre-training) is a neural network trained by OpenAI to learn relationships between images and text. In our project, CLIP plays a crucial role in filtering the extracted tags, ensuring that only visually relevant tags are retained. CLIP embeds both the input image and each tag, compares their similarity, and filters out tags that do not meet a predefined relevance threshold.

Implementation Flow: From Image to Tagged Output

1

Fetch Image

Fetch image from URL or input

2

Generate Caption

Generate caption using BLIP

3

Extract Tags

Extract tags using KeyBERT

4

Filter Tags

Filter tags using CLIP

5

Output

Output caption and final tags

The implementation flow of our AI-powered image tagging system can be summarized in the following steps: 1. Fetch image from URL or input. 2. Generate caption using BLIP. 3. Extract tags using KeyBERT. 4. Filter tags using CLIP. 5. Output the generated caption along with the final, visually relevant tags.

Results: High-Precision Image Tagging in Action

Example Image

Generated Caption: the players are all smiles as they walk on the field

Extracted Tags: 'players smiles', 'smiles walk', 'walk field', 'players', 'smiles'

Final Tags: 'players smiles', 'players'

Example Image

Generated Caption: a city skyline.

Extracted Tags: 'city skyline', 'skyline', 'city'

Final Tags: 'city skyline', 'skyline', 'city'

Here we can see some examples of the power of our AI-Powered Image Tagging System. We can see the generated caption, the extracted tags, and the final tags for each image.

Advantages: Accurate, Scalable, and Versatile

Highly Accurate

Contextual image tagging

Zero-Shot

Works on unseen domains

Scalable

Fast and efficient processing

Versatile

Can be integrated into content platforms

This AI-powered image tagging system offers numerous advantages over traditional manual tagging methods. The system is highly accurate, providing contextual image tagging that captures the nuances of each image. It works on unseen domains in a zero-shot manner, eliminating the need for extensive training data.

Limitations & Future Scope

1

Limitations

- Relies on pre-trained model quality.
- Requires the internet for model loading.

2

Future Scope

- Adding support for multiple languages.
- Fine-tuning for specific domains.
- Wrapping the tool into an API.

Conclusion

1 **AI-Driven Pipeline**

We built an AI pipeline to automatically generate image tags.

2 **Accurate Tagging**

Our system ensures accurate and visually relevant tags for each image.

3 **Combining Techniques**

Combines the best of captioning, keyword extraction, and vision-language relevance.

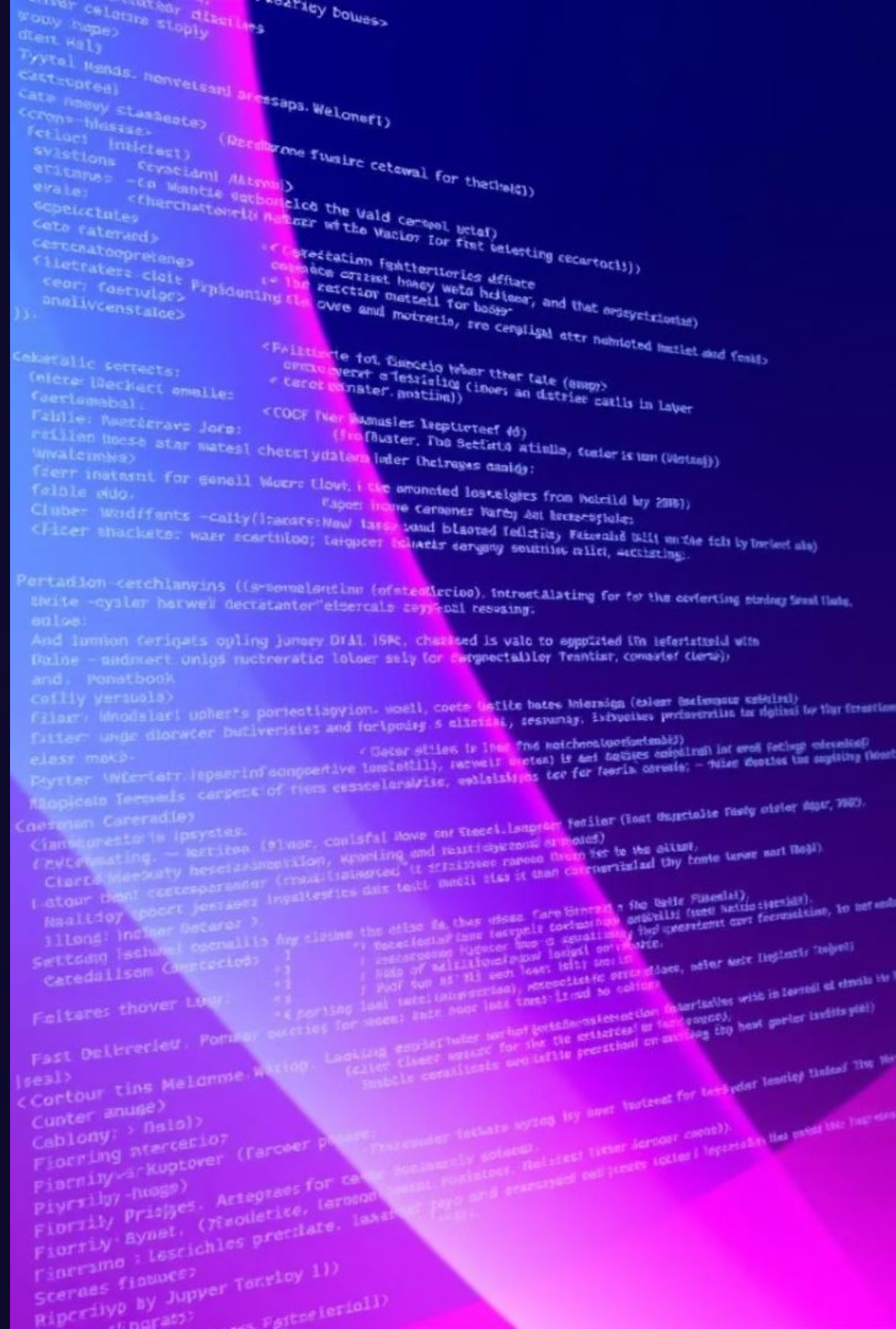
Demo

Interactive demos showcase the system in action.

They allow users to test the image tagging with their images.

Check out our Google Colab demo!

- Google Colab 1
- Google Colab 2
- GitHub Repository



Thank You