Boon: A Neural Search Engine for Cross-Modal Information Retrieval

Yan Gong
Department of Computer Science
Loughborough University
Loughborough, UK
y.gong2@lboro.ac.uk

Georgina Cosma
Department of Computer Science
Loughborough University
Loughborough, UK
g.cosma@lboro.ac.uk

ABSTRACT

Visual-Semantic Embedding (VSE) networks can help search engines understand the meaning behind visual content and associate it with relevant textual information, leading to accurate search results. VSE networks can be used in cross-modal search engines to embed image and textual descriptions in a shared space, enabling image-to-text and text-to-image retrieval tasks. However, the full potential of VSE networks for search engines has yet to be fully explored. This paper presents Boon, a novel cross-modal search engine that combines two state-of-the-art networks: the GPT-3.5-turbo large language model, and the VSE network VITR (VIsion Transformers with Relation-focused learning) to enhance the engine's capabilities in extracting and reasoning with regional relationships in images. VITR employs encoders from CLIP that were trained with 400 million image-description pairs and it was fine-turned on the RefCOCOg dataset. Boon's neural-based components serve as its main functionalities: 1) a 'cross-modal search engine' that enables end-users to perform image-to-text and text-to-image retrieval. 2) a 'multi-lingual conversational AI' component that enables the end-user to converse about one or more images selected by the end-user. Such a feature makes the search engine accessible to a wide audience, including those with visual impairments. 3) Boon is multi-lingual and can take queries and handle conversations about images in multiple languages. Boon was implemented using the Django and PyTorch frameworks. The interface and capabilities of the Boon search engine are demonstrated using the RefCOCOg dataset, and the engine's ability to search for multimedia through the web is facilitated by Google's API.

CCS CONCEPTS

• **Information systems** → **Web search engines**; **Image search**; **Web and social media search**; Chat; Search engine indexing; • **Computing methodologies** → *Visual content-based indexing and retrieval.*

KEYWORDS

cross-modal information retrieval; search engine; large language model; visual-semantic embedding; neural networks

MMIR '23, November 2, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0271-6/23/11.
https://doi.org/10.1145/3606040.3617440

ACM Reference Format:
Yan Gong and Georgina Cosma. 2023. Boon: A Neural Search Engine for Cross-Modal Information Retrieval. In *Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval (MMIR '23), November 2, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3606040.3617440

1 INTRODUCTION

Search engines have transformed the way people discover and access multimedia resources (such as texts, images, and videos) by providing fast and easy search capabilities [11, 29]. Traditional search engines typically rely on textual information such as metadata, tags, to identify and retrieve relevant images [23, 44]. Cross-modal information retrieval-based search engines enhance multimedia search experiences by leveraging advanced techniques like Natural Language Processing (NLP) and Computer Vision (CV) to bridge the gap between text and image modalities, allowing users to obtain relevant and accurate results [15, 38]. State-of-the-art neural networks for cross-modal information retrieval are Visual-Semantic Embedding (VSE) networks, which embed image-description pairs in a shared latent space and compute similarity scores for image-to-text and text-to-image retrieval tasks [13]. Li et al.[21] proposed Visual-Semantic Reasoning Network (VSRN++), which uses a Graph Convolutional Network (GCN) [18] to extract the relationships between image regions, resulting in high-level visual semantics. Chen et al. [3] presented a Variation of Visual-Semantic Embedding Network (VSE∞), which utilises a generalised pooling operator to uncover the optimal strategy for combining image and description representations. Radford et al. [31] proposed Contrastive Language-Image Pre-training network (CLIP), which enables efficient learning of visual concepts through natural language supervision using 400 million image-description pairs. Recently, Gong et al. [12] introduced VIsion Transformers with Relation-focused learning network (VITR) that enhances Vision Transformers (ViTs) by employing a local encoder to extract and reason about image region relations, combining reasoned results with pre-trained global knowledge (e.g. from CLIP) to predict similarity scores between images and descriptions. VITR outperformed various state-of-the-art networks, including CLIP, VSE∞, and VSRN++, in cross-modal information retrieval tasks, particularly in relation-focused cross-modal information retrieval [12]. As a result, this paper focuses on developing a search engine that incorporates VITR, improving user experience by emphasising information retrieval based on relations expressed in user queries and enhancing image-to-text and text-to-image retrieval performance.

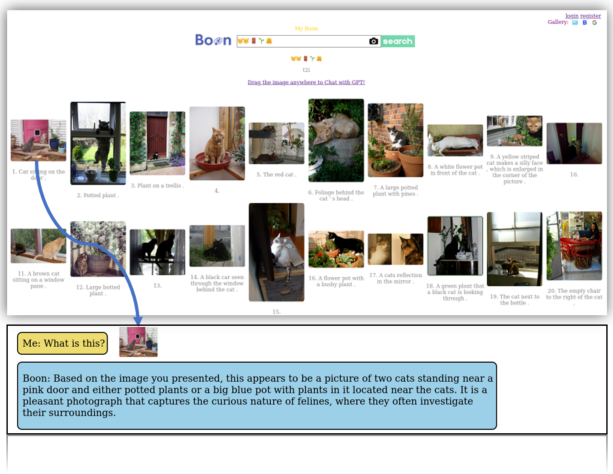


Figure 1: The proposed search engine, Boon, enables cross-modal information retrieval and facilitates conversations about images with users.

Additionally, recently developed Large Language Models (LLMs), such as ChatGPT [36], have exhibited exceptional capabilities in natural language understanding and generation [34], revolutionising various applications from conversational AI and content creation to sentiment analysis [17]. By integrating an LLM into a cross-modal information retrieval search engine, the engine can translate and summarise textual queries, addressing the constraints of existing VSE networks that support only brief English queries. On the other hand, LLMs, such as ChatGPT's 3.5 model, exhibit limitations in comprehending image modalities. However, these shortcomings can be mitigated by employing VSE networks to obtain the most relevant description for an image and utilising this description as a textual prompt for the LLM. Therefore, this paper presents Boon (shown in Figure 1), a novel cross-modal search engine that combines two state-of-the-art networks: ChatGPT (an LLM), and VITR (a VSE network) to enhance the engine's capabilities in extracting and reasoning with regional relationships in images. The contributions of this paper are as follows:

- The proposed Boon is a search engine that benefits from high cross-modal information retrieval performance due to its integration of VITR. Boon enables users to retrieve images using textual queries or to retrieve textual descriptions and their corresponding images using image queries from a gallery. Additionally, Boon re-ranks the results of Google's Programmable Search Engine API (Google's API) to make them more relevant to the query, and this improves search results, particularly for queries that contain relation-related content.
- The proposed search engine uses ChatGPT to support textual queries written in multiple languages. One of VITR's limitations is that it can only support textual queries in English. By combining the capabilities of VITR and ChatGPT, non-English queries detected using the Python LANGID library can be translated into English.
- ChatGPT's 3.5 model has limitations in its ability to comprehend image modalities. Boon can converse with end-users about images and this feature can ultimately enhance their experience while using the engine.

2 RELATED WORK

This section discusses related work on cross-modal information retrieval networks and large language models.

2.1 Cross-modal Information Retrieval Networks

Current works use VSE networks to embed image-description pairs in a shared latent space and calculate similarity scores for retrieval tasks [9, 13, 39, 40, 45, 46]. Faghri et al. [9] proposed an enhanced VSE architecture which employs a fully connected neural network and a Gated Recurrent Units (GRU) network [6] to embed image features (extracted by the Faster R-CNN [1, 32]) and descriptions as representations, respectively. Lee et al. [19] explored the full latent alignments between image regions and descriptive words to determine the similarity of image-description pairs. Li et al. [20, 21] enhanced image features with image region relations extracted by a GCN [18]. Chen et al. [3] proposed a variation of the VSE network that benefits from a generalised pooling operator, which uncovers the best strategy for pooling image and description representations.

The development of pre-trained networks for cross-modal information retrieval has advanced significantly in recent years [4, 5, 10, 24, 42]. Chen et al. [4] introduced a novel network that is an universal image-text representation learned through large-scale pre-training on four image-text datasets. Lu et al. [24] presented a novel collaborative two-stream vision-language pre-training approach for image-text retrieval that strengthens cross-modal interaction through instance-level alignment, token-level interaction, and task-level interaction. Radford et al. [31] proposed the pre-trained CLIP, which applies contrastive learning to align global visual representations and textual representations from a dataset containing 400 million image-description pairs. Recently, Gong et al. [12] proposed VITR, a novel network that augments the ViT by extracting and reasoning about image region relations. VITR addresses the limitations of ViT-based networks in relation-focused cross-modal information retrieval tasks [27, 48] and outperforms state-of-the-art networks such as CLIP in these tasks.

2.2 Large Language Models

LLMs have witnessed remarkable advancements in recent years, exhibiting exceptional performance across a wide range of NLP tasks [8, 16]. Early models, such as Word2Vec [7] and GloVe [30], paved the way by generating dense vector representations of words, while Recurrent Neural Networks (RNNs) [28] and Long Short-Term Memory (LSTM) [14] networks enabled sequential data processing. The advent of attention mechanisms and transformers, introduced by Vaswani et al. [37] further revolutionised the field of NLP. Building on these breakthroughs, more recent models such as BERT (Bidirectional Encoder Representations from Transformers) [8], OpenAI's GPT-3 [2] and ChatGPT [36] have harnessed the power of unsupervised pre-training and fine-tuning to achieve impressive

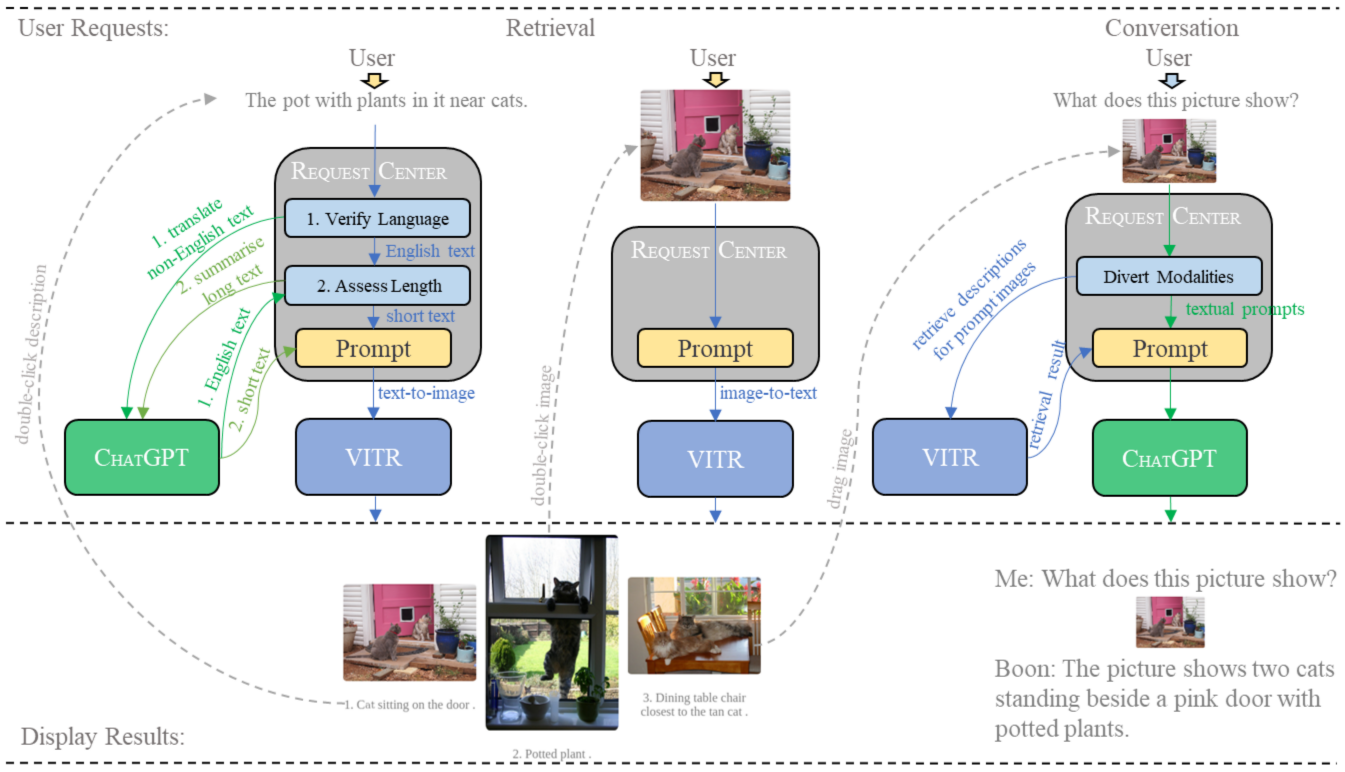


Figure 2: Flowchart of Boon's functionalities. For retrieval requests: 1) in text-to-image retrieval, users input a textual query to retrieve relevant images; and 2) in image-to-text retrieval, the image query is used to retrieve relevant descriptions and their corresponding images. For conversation requests, users input a textual query to converse with ChatGPT and can also upload images to enrich the discussion.

performance in various tasks, including natural language understanding and generation. Although these large-scale models have demonstrated unprecedented capabilities in NLP tasks, they have not been extensively employed in cross-modal search engines.

3 THE PROPOSED BOON CROSS-MODAL SEARCH ENGINE

This section presents the architecture of Boon along with a discussion of its retrieval and conversation modules, and a presentation of its front-end features and capabilities.

3.1 The Architecture of Boon

The proposed search engine, Boon, is illustrated in Figure 2. Boon caters to users' requests for multi-lingual cross-modal retrieval and conversation: 1) for retrieval, users can either input a textual query to search for relevant images or upload an image query to search for relevant descriptions and their corresponding images; and 2) for conversation, users can input textual prompts and upload prompt images to have a conversation about the image(s) with ChatGPT. Boon comprises three modules which are VITR, ChatGPT, and Request Center, and they function as follows.

The **VITR** module is ultilised for cross-modal information retrieval, and it aims to embed image-description pairs into a shared latent space, enabling the prediction of the pairs' similarity scores for the purpose of retrieval ranking [12]. VITR consists of: 1) a text encoder that encodes a description into a global representation and a set of local representations; and 2) a ViT encoder and a CNN-based local encoder encode an image and its regions into a global representation and a set of local representations, respectively. VITR can utilise the encoders from CLIP to obtain global and local representations of images and texts, which were trained on 400 million image-description pairs. VITR was fine-tuned on the RefCOCOg dataset [26] to learn reasoning relations and aggregate reasoned results from local representations, along with global knowledge to enhance relation-focused cross-modal information retrieval performance.

The **ChatGPT** module utilises the GPT-3.5-turbo model through the ChatGPT API for translation and summarisation of textual queries, and to generate sentences for conversations with users based on various prompts.

The **Request Center** module activates user requests to generate prompts for the VITR and ChatGPT modules. The details of how the Request Center activates different user requests will be introduced in sections 3.2 and 3.4.

3.2 Retrieval Requests

Text-to-Image Retrieval. For text-to-image retrieval requests, users can input a textual query to retrieve relevant images. The back-end Request Center processes the textual query as follows: 1) Verifies the language of the textual query using the Python langid library [25]. If the query is not written in English, the Request Center asks ChatGPT to translate it into English with the prompt: 'Translate the following text to English, provide the result directly without explanations: (the textual query).' The prompt for ChatGPT does not require the query to be in a specific language. 2) Assesses the length of the textual query. If a query exceeds the 77-token limit permitted by VITR, the Request Center requests ChatGPT to summarise the query in order to meet the length requirement. The processed textual query serves as the prompt for VITR, and then VITR returns the text-to-image retrieval results to be displayed.

Image-to-Text Retrieval. Image-to-text retrieval requests use image queries to retrieve relevant descriptions, which are then displayed along with their corresponding images. The back-end Request Center servers the image query uploaded by users as the prompt for VITR. VITR ranks the descriptions based on their relevance to the image query, and the descriptions and their corresponding images are displayed. Users can find interest in the displayed images.

Switch Modes to Access Various Galleries. Users can switch between the 'My Album', 'My Boon', and 'My Google' modes to access images from various galleries. Each mode accesses images from an independent gallery: 1) 'My Album' mode provides cross-modal information retrieval performance for managing users' pictures, and users can create an account and establish their personal gallery by uploading up to a fixed number of images (e.g., 500) themselves. Each user can search for and retrieve images that reside within their personal gallery. 2) 'My Boon' mode, creates a common gallery and for demonstration purposes it was populated using the RefCOCOg dataset [26], which contains 25 799 real-world images, with 21 899 of them having corresponding relevant descriptions. This allows all users to search for and retrieve images. 3) 'My Google' mode, enables users to search for and retrieve the relevant images related to their textual query from the web, using Google's API. Boon re-ranks the results returned by Google's API to provide a more accurate image retrieval service for users.

3.3 Database

The database is built for the common gallery of 'My Boon'. To optimise retrieval time, the representations for images and descriptions needed by VITR have been pre-encoded and stored in the database. As illustrated in Figure 3, the database files 'imGloRp.npy' (39.6MB), 'imLocRp.npy' (5.2GB), 'deGloRp.npy' (137.6MB), and 'deLocRp.npy' (10.6GB) store the global representations of images, the local representations of images, the global representations of descriptions, and the local representations of descriptions, respectively. By directly accessing the saved representation values from the database files, Boon eliminates the need for encoding images and descriptions during the retrieval process, resulting in faster retrieval.

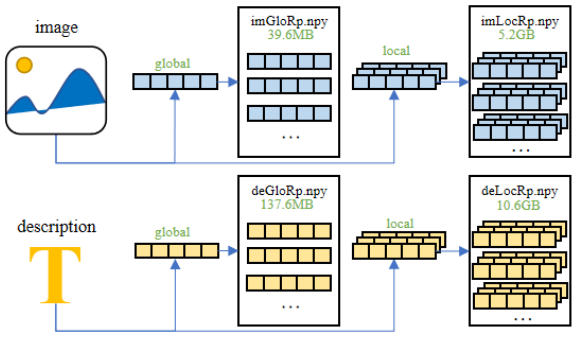


Figure 3: The encoded global and local representations for both images and descriptions generated by VITR have been stored in the '.npy' files to improve retrieval speed.

3.4 Conversation Requests

For conversation requests, users can input a textual prompt to engage in a conversation with ChatGPT, as well as upload images (multiple images are supported) for discussion. The back-end Request Center initially checks whether the user's query includes images. If it does, the Request Center prompts VITR to retrieve the most relevant descriptions from a created description pool for the uploaded images, and the retrieved descriptions will serve as prompts for ChatGPT. The description pool is derived from the MS-COCO dataset [22], which encompasses 634 083 diverse descriptions capable of accurately depicting real-world images.

The prompt for ChatGPT, as illustrated in Figure 4, incorporates the roles of the user, assistant (ChatGPT), and system. First, a series of conversation histories between the user and the assistant are input as the prompts for ChatGPT to ensure continuity in the conversation. Second, the system prompt instructs ChatGPT to pretend it can view the images while reminding it to avoid discussing images in its response if the user's question is unrelated to them. Finally, the retrieved descriptions for the user's uploaded images are combined with the user's current question to form the user's prompt for ChatGPT. Once ChatGPT receives the prompt, it generates a response, which is then displayed by Boon.

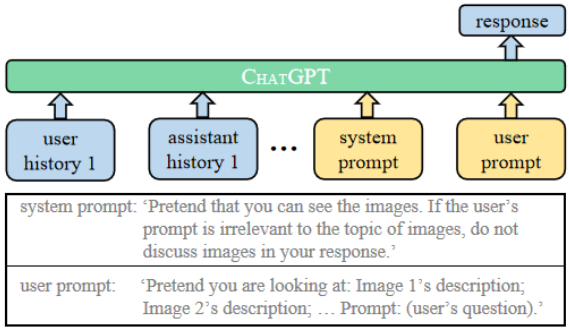


Figure 4: The method of facilitating a conversation about images with ChatGPT using the prompts of the roles of the user, assistant, and system.

3.5 The Front-End of Boon

Figure 5 illustrates the front-end components of Boon, which include (a) the navigation interface, (b) the retrieval interface, and (c) the conversation interface. In the navigation interface, users can input a textual query in the provided text box or upload an image query using the upload button. Once the search button is clicked, the retrieval results are displayed on the retrieval interface. Both the navigation and retrieval interfaces include a button for navigating to the conversation interface. Within the conversation interface, users can input textual prompts in the text box and upload images using the upload button. After clicking the send button, the conversation history appears on the conversation interface.

To enhance the user experience, Boon incorporates mouse actions.

- When a user double-clicks a displayed description, the clicked description becomes the textual query for a text-to-image retrieval request.
- When a user double-clicks a retrieved image, the clicked image becomes the image query for an image-to-text retrieval request.
- When a user drags a retrieved image, the dragged image is transferred to the conversation interface as the prompt image.

4 RESULTS

This section presents retrieval request results, including visuals of multi-lingual results, re-ranking for web images via Google's API, and quantitative findings. It also highlights Boon's image-related conversation requests with visuals and quantitative outcomes.

4.1 Implementation Details

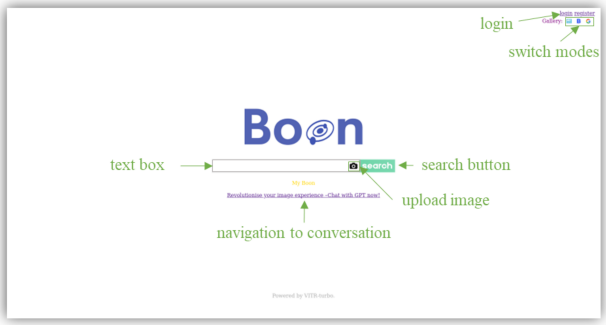
A high-performance PC with a single NVIDIA RTX 3080 graphics card and 64GB of memory can meet the minimum requirements for running Boon. The proposed Boon was implemented using the Django framework. For the VITR module, the model VITR$_L$ [12] utilising the encoder of 'ViT-L/14' from CLIP was employed, and the turbo setting N was set to 200. The code for VITR was implemented with the PyTorch framework.

4.2 Results of Retrieval Requests

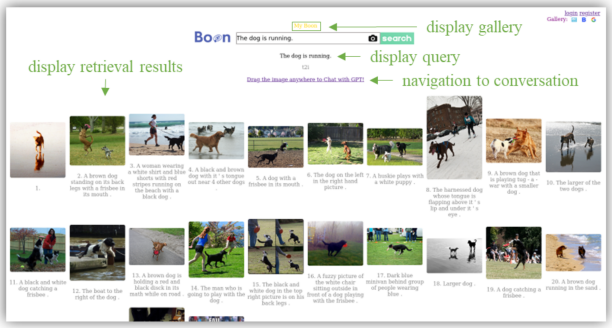
Presenting Examples of Retrieval Requests. Figure 6 visually presents several retrieval results using non-English, English, and long textual queries, as well as the re-ranking results for images on the web retrieved through Google's API.

Figure 6 (a) displays four examples of the top relevant result when employing non-English textual queries for retrieval. Four languages Chinese, Korean, Greek, and Emoji were tested. Each language was translated into English by the ChatGPT module of Boon before being used by the VITR module to search for relevant images. The ChatGPT supported over 100 different languages for translation.

Figure 6 (b) showcases two examples of the top two relevant results when using long textual queries (written in English) for retrieval. The first example's query was a story about two dogs, while the second example's query was a news article about horse riding.



(a) Navigation interface.



(b) Retrieval interface.

(c) Conversation interface.

Figure 5: The front-end components of Boon, include (a) the navigation interface, (b) the retrieval interface, and (c) the conversation interface.

The results for both examples were relevant to their respective queries.

Figure 6 (c) demonstrates examples of Boon re-ranking the retrieval results from Google's API. Considering that transferring images from Google's API to Boon takes time, and users' focus is typically the top retrieved results, Boon obtains 40 retrieval results for each query using Google's API. It then recalculates the relevance between these 40 retrieved results and the query to re-rank them. In Figure 6 (c), the top retrieved images by Google's API were irrelevant to the queries. Meanwhile, Boon re-ranked the retrieval results to position these images at lower rankings, and the top retrieved

(a) The top images retrieved using queries in Chinese, Korean, Greek, and Emoji languages.

(b) The top two images retrieved using long textual queries.

(c) Compare the rankings of retrieved images in response to queries between Google's API and Boon, where Boon corrects and improves Google's erroneous results.

Figure 6: The retrieval examples for Boon include: (a) translating non-English queries for retrieval; (b) using long textual queries for retrieval; and (c) re-ranking images to the queries retrieved by Google's API.

Me: What does the picture show?

Boon: The picture shows an adorable little girl who seems to be enjoying a delicious slice of pizza, which she's holding in her hand. The pizza is sitting on a plate in front of her, with some crumbs visible on the plate. The girl looks joyful with a big smile on her face, and she's wearing a cute dress with matching hair accessories. The background of the image is blurry, but it appears to be a kitchen or dining room. Overall, the picture is heartwarming, capturing the simple pleasures of childhood, such as enjoying a tasty slice of pizza.

Me: What are the relationships in the picture?

Boon: From Image 1, it appears that there are a few relationships depicted. The little girl is at the center of the image, and she seems to have a positive relationship with the pizza she is eating. The way she is holding the pizza slice suggests that she's enjoying every bite of it. Additionally, there is an implied relationship between the girl and the person who gave her the food. It's possible that the person who made or served the pizza has a caring relationship with the girl. Overall, the picture shows a moment of comfort and enjoyment for the little girl, with good relationships portrayed between her and the food she's eating, and possibly between her and someone who provided her with the food.

Me: What is the most popular pizza flavor?

Boon: Based on my knowledge, there are many popular pizza flavors that people enjoy, and each flavor varies depending on culture and preferences. However, some of the most popular pizza toppings include pepperoni, sausage, mushroom, onion, green pepper, and black olives. Additionally, another popular pizza topping is tomato sauce paired with fresh mozzarella cheese and basil, known as the Margherita pizza - a staple in Italian cuisine. Ultimately, the most popular pizza flavor could vary widely depending on the region or culture, so it's hard to give a definitive answer.

(a)

Me: Τι ζώα είναι στη φωτογραφία;
(Greek, means: What animal is in the picture?)

Boon: Βλέποντας την εικόνα, βλέπω ότι στη φωτογραφία υπάρχει μια ζέβρα που στέκεται σε ένα πεδίο με χόρτα και άνθη. Επομένως, το ζώο που είναι στη φωτογραφία είναι μια ζέβρα.
(Means: Looking at the picture, I see that in the photo there is a zebra standing in a field of greens and flowers. Therefore, the animal in the photo is a zebra.)

Me: 图片里有几只斑马?
(Chinese, means: How many zebras are in the picture?)

Boon: 我注意到这张照片中有一只斑马。我没有看到其他斑马的存在。
(Means: I noticed a zebra in this photo. I didn't see the existence of other zebras.)

Me: Von welchem Kontinent kommen Zebras?
(German, means: Which continent do zebras come from?)

Boon: Die meisten Arten von Zebras kommen in Afrika vor. Sie sind in verschiedenen afrikanischen Ländern wie Südafrika, Botswana, Simbabwe, Tansania und Kenia zu finden.
(Means: Most species of zebra are found in Africa. They can be found in various African countries such as South Africa, Botswana, Zimbabwe, Tanzania and Kenya.)

Me: Based on the picture, please generate the emoji.
(English, Boon's response should include the emoji.)

Boon: 🦓🌸🌾
I would suggest these three emojis for the picture: a zebra emoji to represent the zebra in the field, a flower emoji to represent the flowers in the field, and a grass emoji to represent the grass in the field. These emojis would help convey the peaceful and natural atmosphere of the picture.

(b)

Me: What is the difference between the three pictures?

Boon: The main difference between the three pictures is their settings and the people in them. Image 1 depicts a group of people enjoying their time on the deck of a yacht, while Image 2 shows a group of people eating in the woods next to a bench. On the other hand, Image 3 portrays a family sitting down to eat inside a house. Another difference is the ambiance of the images - Image 1 is outdoors, while Image 3 is indoors, and Image 2 is in a natural environment. Moreover, the people in the images are all different, with Image 1 showing a group of friends having a good time, Image 2 portraying a group of unknown people, and Image 3 featuring a family. Lastly, the table settings are different; Image 1 and 3 show a more formal environment, while Image 2 is more casual.

Me: What are the similarities between the three pictures?

Boon: The main similarity between the three pictures is that they all feature groups of people sharing a meal together. While the settings and people may differ, the concept of shared food and enjoyment of a meal is present in all three images. Another similarity is the fact that all of the images show a form of socializing - in Image 1, friends are enjoying each other's company on a boat, in Image 2, people are standing and eating together in nature, and in Image 3, a family is eating together in a formal dining room setting. Finally, all three images depict a moment of togetherness, whether it's friends, strangers, or family members sharing a meal together.

Me: Which picture is more likely to display an indoor environment?

Boon: Image 3, with a family sitting down to eat at a lighted dining room table is more likely to display an indoor environment. While Images 1 and 2 show people gathered outdoors, Image 3 is set in a dining room, indicating that the setting is inside a house or building. Therefore, Image 3 is more likely to display an indoor environment than the other two images.

(c)

Me: Based on the picture, write a story.

Boon: Once upon a time, there was a young boy named Alex, who had a love for football. He spent every spare moment on the school field practicing his moves, and he always felt happy when he had a ball at his feet. One day, after school, Alex hopped onto the school bus with his friends. They were chatting and joking around when they arrived at Alex's stop. He hopped off the bus and made his way down the familiar street towards his house. As he arrived, his dog, Buddy, was waiting to welcome him home. Alex and Buddy loved to play football together, and Buddy was always willing to get in on the action. Alex got changed into his kit, and the two of them went out into the garden. Alex had been working on his kicking, and he showed off his new moves to Buddy. He kicked the ball high into the air, and Buddy ran after it, his tail wagging. Alex and Buddy spent hours playing with the ball until the sun began to set. Alex's legs ached from all the playing, but he was grinning from ear to ear. He loved spending time with Buddy, and he couldn't wait to get out on the field again tomorrow. As they headed inside, Alex's mom greeted them with some snacks, and they all snuggled up together watching TV. Alex was happy with how his day had turned out, and he felt grateful for the simple things, like playing football with his best friend.

Me: What content in the story does not exist in the picture?

Boon: The third section of the story where Alex played football with his dog, Buddy, in the garden is not included in any of the images. Image 1 shows a boy kicking a soccer ball, but there is no mention of a garden in the picture. Image 2 displays several children exiting a bus, but there is no reference to Alex or his friends in the image. Image 3 shows a child playing with a dog, which is similar to the story, but there is no football or mention of a garden in the picture. Therefore, the content that does not exist in the picture is the specific scene of Alex playing football with his dog in the garden.

(d)

Figure 7: Examples of conversation requests about images include: (a) conversations about a single image, (b) multi-lingual conversations about an image, (c) conversations about multiple images, and (d) writing a story based on multiple images.

images by Boon were presented for comparison. For example, in the second scenario, the user was searching for a picture of a cat on top of an object. However, the top retrieved image by Google's API featured the movie 'Top Gun' with cats, ignoring the relation expressed in the user's query. Boon then re-ranked this image to the 19th position in terms of relevance and presented a more relevant image at the top.

Quantitative Results. The retrieval performance of Boon was quantitatively evaluated using VITR [12]. Table 1 and Table 2 compare the proposed Boon with baseline methods on the relation-focused dataset RefCOCOg [26] and the benchmark dataset Flickr30K [41], respectively, for both image-to-text and text-to-image retrieval [12]. The evaluation measure used is Recall at rank k (Recall@k), which is defined as the percentage of relevant items among the top k retrieved results [33]. In the RefCOCOg test set, as shown in Table 1, Boon achieved average Recall@1 values of 45.2% for image-to-text retrieval and 29.5% for text-to-image retrieval, outperforming CLIP by 2.8% and 4.3% respectively. In the Flickr30K test set, as shown in Table 2, Boon achieved average Recall@1 values of 94.7% for image-to-text retrieval and 82.5% for text-to-image retrieval, outperforming CLIP by 2.1% and 4.7% respectively.

Table 1: Results of cross-modal information retrieval networks on the RefCOCOg test set. Table shows average Recall@k (%) values.

Network	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
VSRN++ [12]	20.0	44.9	57.3	13.8	34.6	47.8
VSE∞ [12]	31.1	58.3	69.7	19.5	42.8	55.2
CLIP [12]	42.4	65.5	75.1	25.2	48.9	60.4
Boon	**45.2**	**71.1**	**80.5**	**29.5**	**55.1**	**66.8**

Table 2: Results of cross-modal information retrieval networks on the Flickr30K test set. Table shows average Recall@k (%) values.

Network	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
VSRN++ [21]	79.2	94.6	97.5	60.6	85.6	91.4
VSE∞ [3]	88.7	98.9	99.8	76.1	94.5	97.1
CLIP [12]	92.6	99.2	99.6	77.8	95.2	97.7
Boon	**94.7**	**99.7**	**99.9**	**82.5**	**96.7**	**98.3**

Assessing Retrieval Time for Each Query. The average retrieval time for each query using Boon was experimentally assessed. For image-to-text retrieval involving 21 899 textual descriptions, the average retrieval time was 0.19 seconds. Meanwhile, for text-to-image retrieval encompassing 25 799 images, the average retrieval time was 0.68 seconds.

4.3 Results of Conversation Requests

Presenting Examples of Conversation Requests about Images. Figure 7 showcases four examples of conversation requests about images in Boon. In Figure 7 (a), a series of conversations revolve around a picture of a girl eating pizza, such as those regarding relationships in the picture. In Figure 7 (b), Boon showcases its multi-lingual proficiency in conversation about images. A zebra picture was uploaded, and questions were asked in Greek, Chinese, German, and English. Boon accurately responded in the respective languages and generated the appropriate emoji as the response based on the image as requested. Figure 7 (c) demonstrates that Boon can support multiple prompt images and answer questions about their differences and similarities. In Figure 7 (d), Boon has written a story based on three related pictures. The story is about a boy who played football at school and then played with his dog after school. Unlike other visual question-answering networks [35, 43], Boon can continuously communicate with users and answer high-level semantic questions.

Quantitative Results. Ensuring that the retrieved descriptions by Boon accurately describe the prompt images is the key to fulfilling the conversation request regarding images. To evaluate this, the paper utilises the Flickr30K dataset [41], and the descriptions in the dataset were not included in Boon's description pool. Specifically, Boon retrieved relevant descriptions from its description pool for the 1000 images in the Flickr30K test set and compared them with the images' corresponding descriptions in the dataset. BertScore, which utilises contextual embeddings from BERT to compare the similarity between two pieces of text [47], was used as an evaluation measure with a maximum value of 1. According to the evaluation results, Boon's average BertScore on the Flickr30K test set was 0.91.

5 CONCLUSION

VSE networks improve search engine accuracy by associating visual content with relevant text. They can be used in cross-modal search engines to retrieve multimedia resources, by embedding image and textual descriptions in a shared latent space. This paper introduces a novel cross-modal search engine, Boon, which improves user experience in image-to-text and text-to-image retrieval tasks by incorporating the cutting-edge VSE network, VIsion Transformers with Relation-focused learning (VITR), and ChatGPT that is an advanced Large Language Model (LLM). Boon leverages VITR to emphasise information retrieval based on user query relations and enhance both image-to-text and text-to-image retrieval performance. Furthermore, it utilises ChatGPT to facilitate translations in multiple languages and enable conversations about images, broadening accessibility for various audiences, including visually impaired individuals. By supplying relevant image descriptions obtained from Boon's integrated VITR as input prompts, the limitations of ChatGPT's 3.5 model in comprehending images are overcome. The interface and capabilities of Boon's search engine are demonstrated using the RefCOCOg dataset, and its ability to search for multimedia online is facilitated by Google's API. Future developments for Boon will involve implementing text-video retrieval functions. Furthermore, a deep comparison between Boon and existing search engines like Google and Bing will be part of future work.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for

image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[3] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 15789–15798.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*. 104–120.

[5] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, et al. 2022. ViSTA: vision and scene text aggregation for cross-modal retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5184–5193.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1724–1734.

[7] Kenneth Ward Church. 2017. Word2Vec. *Natural Language Engineering* 23, 1 (2017), 155–162.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 4171–4186.

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*. 12.

[10] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems* 33 (2020), 6616–6628.

[11] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashionbert: text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the SIGIR on Research and Development in Information Retrieval*. 2251–2260.

[12] Yan Gong and Georgina Cosma. 2023. VITR: augmenting vision transformers with relation-focused learning for cross-modal information retrieval. *arXiv preprint arXiv:2302.06350* (2023).

[13] Yan Gong, Georgina Cosma, and Hui Fang. 2021. On the limitations of visual-semantic embedding networks for image-to-text information retrieval. *Journal of Imaging* 7, 8 (2021), 125.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[15] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[16] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274.

[17] Michael R King and ChatGPT. 2023. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cellular and Molecular Bioengineering* 16, 1 (2023), 1–2.

[18] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of International Conference on Learning Representations*.

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*. 201–216.

[20] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of International Conference on Computer Vision*. 4654–4662.

[21] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2022. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 641–656.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Proceedings of the European Conference on Computer Vision*. 740–755.

[23] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of International Conference on Computer Vision*, Vol. 2. Ieee, 1150–1157.

[24] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. 2022. COTS: collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 15692–15701.

[25] Marco Lui and Timothy Baldwin. 2012. langid.py: an off-the-shelf language identification tool. In *Proceedings of the ACL System Demonstrations*. 25–30.

[26] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 11–20.

[27] Mingyuan Mao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, Errui Ding, Baochang Zhang, Shumin Han, et al. 2021. Dual-stream network for visual recognition. *Advances in Neural Information Processing Systems* 34 (2021), 25346–25358.

[28] Larry R Medsker and LC Jain. 2001. Recurrent neural networks. *Design and Applications* 5, 64-67 (2001), 2.

[29] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 4 (2021), 1–23.

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning*. 8748–8763.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2016), 1137–1149.

[33] Tefko Saracevic. 1995. Evaluation of evaluation in information retrieval. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 138–146.

[34] Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. 2023. ChatGPT and other large language models are double-edged swords. *Radiology* 307, 2 (2023), e230163.

[35] Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4613–4621.

[36] OpenAI Team. 2022. ChatGPT: optimizing language models for dialogue.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017), 5998–6008.

[38] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[39] Yaxiong Wang, Hao Yang, Xiuxiu Bai, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2020. PFAN++: bi-directional image-text retrieval with position focused attention network. *IEEE Transactions on Multimedia* 23 (2020), 3362–3376.

[40] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 10941–10950.

[41] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[42] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViL: knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3208–3216.

[43] Jing Yu, Zihao Zhu, Yujing Wang, Weifeng Zhang, Yue Hu, and Jianlong Tan. 2020. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognition* 108 (2020), 107563.

[44] Lei Zhang and Yong Rui. 2013. Image search—from thousands to billions in 20 years. *ACM Transactions on Multimedia Computing, Communications, and Applications* 9, 1s (2013), 1–20.

[45] Peng-Fei Zhang, Yang Li, Zi Huang, and Xin-Shun Xu. 2021. Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Multimedia* 24 (2021), 466–479.

[46] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. 2020. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3536–3545.

[47] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[48] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. RegionCLIP: region-based language-image pretraining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 16793–16803.