Evaluating Image Similarity Using Contextual Information of Images with Pre-trained Models

Juyeon Kim[1], Sungwon Park[2,3], Byunghoon Park[3], and B. Sooyeon Shin[4(✉)]

[1] Sorbonne University, Paris, France
juyeon.kim@etu.sorbonne-universite.fr
[2] Korea University, Sejong, South Korea
apcm05@korea.ac.kr
[3] T3Q Co., Ltd., Seoul, South Korea
warmpark@t3q.com
[4] Center for Creative Convergence Education, Hanyang University, Seoul, South Korea
shinsy@hanyang.ac.kr

Abstract. This study proposes an integrated approach to image similarity measurement by extending traditional methods that concentrate on local features to incorporate global information. Global information, including background, colors, spatial representation, and object relations, can leverage the ability to distinguish similarity based on the overall context of an image using natural process techniques. We employ Video-LLaMA model to extract textual descriptions of images through question prompts, and apply cosine similarity metrics, BERTScore, to quantify image similarities. We conduct experiments on images of the same and different topics using various pre-trained language model configurations. To validate the coherence of the generated text descriptions with the actual theme of the image, we generate images using DALL-E 2 and evaluate them using human judgement. Key findings demonstrate the effectiveness of pre-trained language models in distinguishing between images depicting similar and different topics with a clear gap in similarity.

Keywords: Vision language · Image similarity · Natural language processing · Computer Vision · Vision Transformer · Large language model

1 Introduction

Vision language models integrate both vision and language modalities in a single model, enabling the processing of images and natural language texts. They find application in diverse tasks, including visual question answering and image captioning [2].

S. Bouzefrane et al. (Eds.): MSPN 2023, LNCS 14482, pp. 188–204, 2024.
https://doi.org/10.1007/978-3-031-52426-4_13

The field of image understanding and representation has witnessed significant advancements, with notable approaches including Convolutional Neural Networks (CNNs) and Vision Transformers (ViT). CNNs are known for their deep architecture, capable of detecting diverse image features through filter operations. ViT, alternatively, partitions images into patches and processes them as sequences, employing multi-layer transformers for efficient feature extraction. In the domain of large language models, Video-LLaMA merges language and vision understanding through its Vision-Language Branch, featuring a pre-trained image encoder and contextualising temporal information. BERT stands out for its contextual interpretation of text, leveraging transformers with self-attention to capture rich contextual information within the text.

Our research primarily aims to use the vision transformer's image understanding capabilities and combine them with text generated by language models to compute image similarities. This paper is organised as follows. We review some important studies on image understanding methods related to our methodology in Sect. 2. Section 3 presents our methods based on Video-LLaMA and BERTScore (Sect. 3.1) and our approach to generate image descriptions with contextual global information (Sect. 3.2) and measure image similarity scores (Sect. 3.3). Section 4 presents the experimental setup of the image dataset (Sect. 4.1) and its procedures (Sect. 4.2). The results of our experiment are presented in Sect. 5: images of the same topic in Sect. 5.1 and different topics in Sect. 5.2. A comparison of the pre-trained language models is discussed in Sect. 5.3. Finally, we conclude our study and provide prospects for future advancements in image similarity in Sect. 6.

2 Related Works

CNN. Convolutional neural networks (CNN) [3] consist of multiple layers, each dedicated to detecting distinct image features. Various filters with different resolutions are applied to the training images, and the output from each convolutional layer serves as the input for the subsequent layer. Filters are used to detect special patterns, such as brightness and boundaries, in an image by detecting changes in the intensity values of the image. They allow the network to learn the hierarchical representations of the input image. By stacking multiple convolutional layers with different filters, CNNs can learn increasingly complex features of the input image, leading to a higher accuracy in image recognition tasks.

ViT. In Vision Transformer (ViT) [4], input images are partitioned into fixed-size patches, with each patch undergoing tokenization. These tokenized patches are constructed as sequences and subjected to linear embedding before their incorporation into the transformer model. This contrasts with CNNs, which directly use raw pixel values. Within the ViT architecture, the transformer model comprises multiple encoder layers. Each encoder unit comprises multi-head self-attention and Multi-Layer Perceptron (MLP) blocks. Layer normalisation was uniformly applied to all the blocks, with residual connections introduced following each block. This design empowers ViT to discern the intricate hierarchical features within the input images. Upon traversing the transformer encoders, the initial vector of the encoder's terminal output plays a crucial role in classification. This vector is directed through an MLP Head that features a single hidden layer for classification tasks. During fine-tuning, it traverses a single linear layer, instead of the aforementioned MLP head. ViT's strategy of image partitioning into patches and subsequent treatment as sequential data has been shown as a computationally efficient and accurate approach compared to CNNs.

Video-LLaMA. In prior research efforts, Large Language Models (LLMs) have been trained with varying parameters, ranging from 7 billion to 65 billion. The Vision-Language Branch within the architecture of Video-LLaMA [6] is designed to allow LLMs to understand visual content. This branch consists of a frozen pre-trained image encoder, which is fixed and extracts features from video frames (see Fig. 1). In addition, a position embedding layer was employed to inject temporal information into the video frames. The video Q-former shares architectural similarities with the query transformer (Q-former) used in BLIP-2 [7]. This process results in a video embedding vector that encapsulates the representations of the input video. To ensure compatibility with the language model, a linear layer is introduced, which transforms the video embedding vector into a video query vector of the same dimensions as the LLM's text embedding. During the Forward Pass, a video soft prompt establishes a connection between the input text embedding and frozen LLM, guiding it to generate contextually relevant text based on the video content. In the implementation of the Vision-Language Branch, the visual components of the pre-trained BLIP-2 model serve as a frozen visual encoder. This includes ViT-G/14 from EVAPLIP and pre-trained Q-formers. Notably, the remaining components, such as the position embedding layers, video Q-formers, and linear layers, are initialized randomly and subsequently optimised to seamlessly integrate with the frozen LLM.

BERT Contextual Interpretation. BERT, or Bidirectional Encoder Representations from Transformers [8], relies on contextual embeddings to process input. Initially, input tokens are transformed into embedding vectors, where the dimensionality of these vectors, known as d_{model}, serves as an input for BERT. These embedding vectors are computed for each word in the input sequence following internal operations. The resulting output embeddings in BERT encapsulate contextual information that extends across the entire sentence, referencing all words within the context. This process, in which each word interacts with every other word, is achieved through a self-attention mechanism. In essence, BERT's use of contextual embeddings and the self-attention mechanism across its transformer encoder layers enables it to capture and represent the contextual nuances of input text comprehensively, yielding output embeddings that reflect the context of the entire textual input.

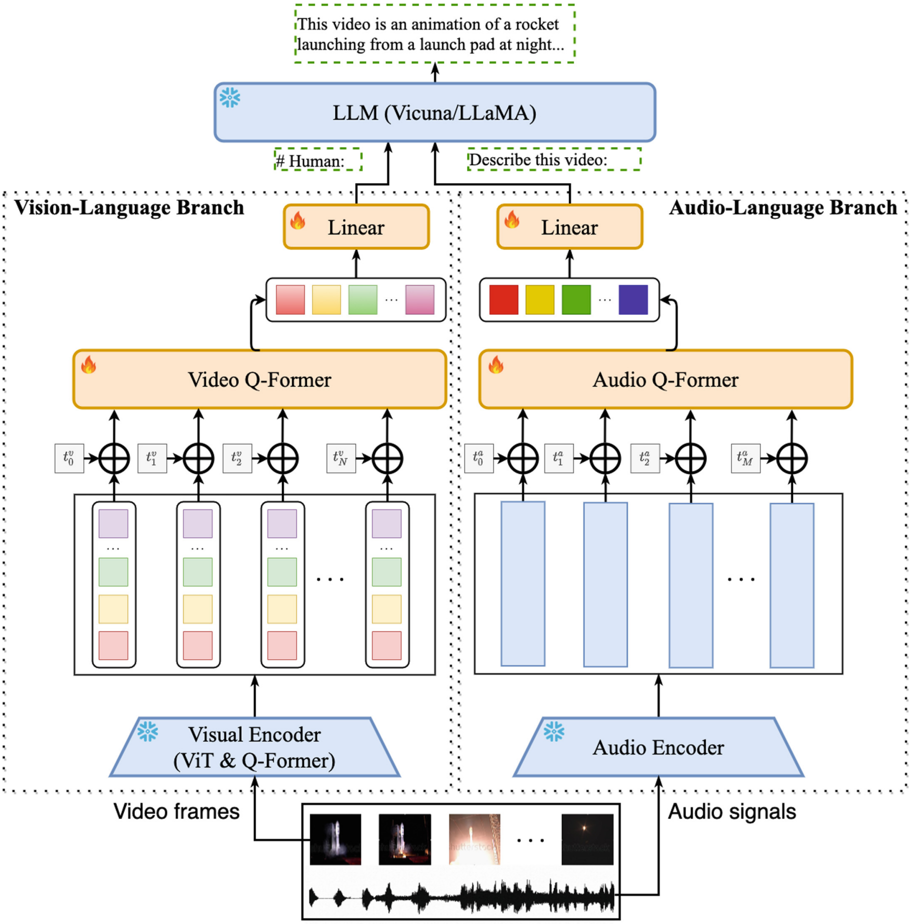


Fig. 1. Overall architecture of Video-LLaMA [6]

3 Methods

In this section, we detail our methodology, starting with an overview of our approach for generating textual image descriptions and computing image similarity scores.

3.1 Video-LLaMA and BERTScore Methodology

The Video-LLaMA [6] framework consists of two integral components: the Vision-Language and Audio-Language branches. For the scope of this paper, we focus solely on the elements within the Vision-Language (VL) branch. Our specific interest lies in comparing images and text. Video-LLaMA, as a multi-modal framework, is designed to enhance language models with the ability to understand both visual and auditory content in videos. The VL branch uses a pre-trained image encoder, specifically ViT-G/14 [5], to encode visual information from the input video or image.

In the context of text generation, traditional evaluation metrics such as BLEU [10] and ROUGE have been commonly employed [11]. We propose the use of automated metrics inspired by BERTScore to assess the quality of text generation. Notably, BERTScore exhibited superior performance in machine translation and image captioning when compared with conventional metrics such as BLEU [1].

We employ a pre-trained model to encode the reference and candidate input texts. In our experiments, we tested three transformer-based models, RoBERTa large model [13], XLM model [14], and BERT base model (uncased) [8]. These models have achieved state-of-the-art performance on NLP tasks, including text classification, question answering, and language modelling. The specificity of the models is included in Table 1.

Table 1. Pre-trained models used in our experiments [14,16]

Architecture	Shortcut name	Details of the model
BERT	bert-base-uncased	12-layer, 768-hidden, 12-heads, 110M parameters Trained on lower-cased English text
XLM	xlm-mlm-en-2048	12-layer, 2048-hidden, 16-heads, 665M parameters XLM English model
RoBERTa	roberta-large	24-layer, 1024-hidden, 16-heads, 355M parameters RoBERTa using the BERT-large architecture

3.2 Our Approach

The proposed system follows the architecture of Video-LLaMA as illustrated in Fig. 2. The procedure begins with the input of a reference image, which is transformed into vectors using the Video-LLaMA image encoder, resulting in representations of specific dimensions. In the initial stage, the user poses a single question prompt for all input images. Subsequently, the model generates a descriptive text output containing contextual information regarding the image. We repeat the same process for the candidate image using the contextual information generated from both the reference and the candidate images to calculate cosine similarity.

BERTScore, introduced by Zhang et al. (2020) [1], was initially designed to evaluate machine translation outputs and image captioning. The core concept in BERTScore involves encoding the candidate and reference separately using a BERT-based model, and assigning a score based on the similarity between individual encoded tokens. Each word in the reference sentence is compared for cosine similarity with all the words in the candidate sentence, constituting the 'reference and candidate token-pair' evaluation mechanism. BERTScore leverages BERT to generate contextualised word representations. Our approach generally aligns with that of BERTScore.

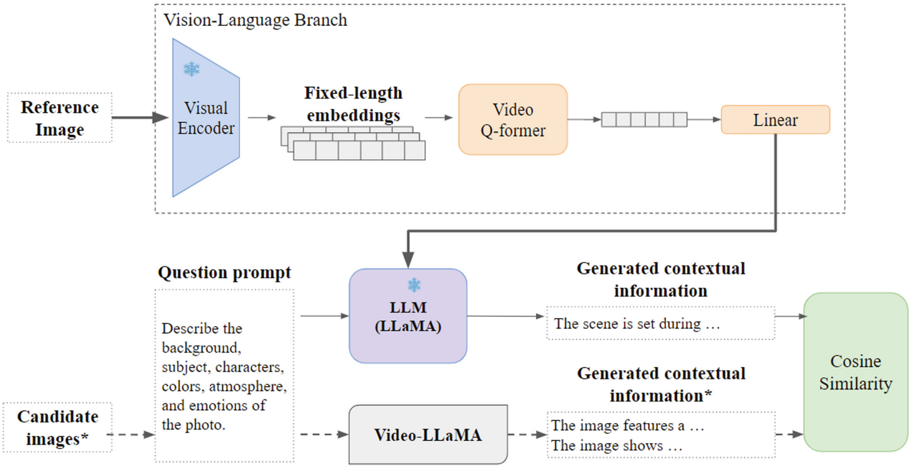


Fig. 2. Illustration of our approach.

3.3 Similarity Measure and Score

Given a reference sentence $x = \{x_1, \ldots, x_n\}$ and a candidate sentence $\hat{x} = \{\hat{x_1}, \ldots, \hat{x_m}\}$, we employ contextual embeddings to represent the tokens. We use different models and tokenizers according to each model. Matching is computed through cosine similarity as illustrated in Fig. 3.

The cosine similarity calculation between the encoded reference and encoded candidate tokens, as described by Zhang et al. (2020) [1], is expressed as: $sim(x_i, \hat{x_i}) = \frac{\mathbf{x}_i^\top \cdot \hat{\mathbf{x}_j}}{||\mathbf{x}_i|| \cdot ||\hat{\mathbf{x}_j}||}$. Although this involves a comparison of individual tokens x_i and $\hat{x_i}$, their vector representations $\mathbf{x_i}$ and $\hat{\mathbf{x}_j}$ encapsulate contextual information.

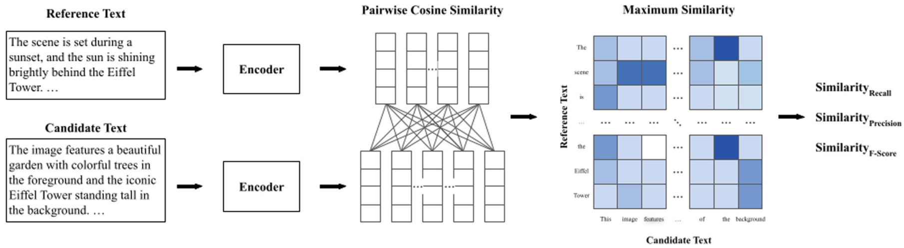


Fig. 3. Illustration of the similarity scores computation: when provided with the reference text x and candidate $\hat{x}$, we calculate BERT embeddings and pairwise cosine similarity. We proceed with greedy matching, choosing the maximum value from both rows and columns of the similarity matrix. [1].

To calculate the similarity score recall S_R, we match each token in the reference text x to its corresponding token in the candidate text $\hat{x}$ and to calculate the similarity score precision S_P, we match each token in the candidate text $\hat{x}$ to its corresponding token in the reference text x. We use a greedy matching approach, pairing each token with its most similar counterpart in another sentence. The F-score is then computed by combining precision and recall. We apply these techniques using the following equations:

$$S_R = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, S_P = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, S_F = 2\frac{S_P \cdot S_R}{S_P + S_R} \tag{1}$$

We observe that the computed scores are confined to a narrow range. To enhance the score readability, following Zhang et al. (2020) [1], we proceed with min-max normalisation. We use the experimental lower threshold b and upper threshold h to linearly rescale S_R. For example, the normalised value $\hat{S_R}$ of S_R is:

$$\hat{S_R} = \frac{S_R - b}{h - b} \tag{2}$$

Following this process, the scores typically fell within the range of 0 to 1. We apply the identical procedure to both S_P and S_F.

4 Experiments

For our experiment, we used Video-LLaMA Demo [6] to process the collected image dataset. By posing questions using the Video-LLaMA Demo prompt, we obtained text-based answers. We then employed BERTScore to assess the textual outputs in terms of their relevance to original image themes. In addition, we explored the possibility of transforming text-based descriptions back into images resembling the originals using the DALL-E 2 model. This enabled a comparative analysis of the original and generated images based on subjective assessment.

4.1 Data Collection and Preparation

Before data collection, we chose five image topics to compute the image similarity scores. We then sourced image data related to these five topics from Kaggle [15] and curated a dataset consisting of 25 images, which are provided in Table 2.

Table 2. Topics of image dataset.

Theme	Topic
1. Sculpture	Eiffel Tower
2. Animal	Butterfly
3. Human Activity	Eating Ice Cream
4. Landscape	Rainbow
5. History	Pyramid

4.2 Procedures

Question Prompt for Video-LLaMA. Upon feeding the images into the Video-LLaMA model, we presented a single question: "Describe the background, subject, characters, colors, atmosphere, and emotions of the photo.". This question was posed collectively for all images, aiming to capture the overall context of each image, including the typical elements found in such visuals. The formulation of this prompt aims to encapsulate contextual details from the visual attributes of the object or event, covering elements like color, and the overall description of the scene [12].

DALL-E 2 Text-to-Image Transformation. To validate the text generated by Video-LLaMA, we employed DALL-E 2, a text-to-image conversion model. Considering text length constraints (540 characters and 76 words), we used a segment of the original text output from Video-LLaMA to generate corresponding images.

Comparison of DALL-E 2 Generated Images with Originals. We compared the images generated by DALL-E 2 with the original images by using direct human judgement to evaluate the fidelity and accuracy of the generated images. The process is illustrated in Fig. 4.

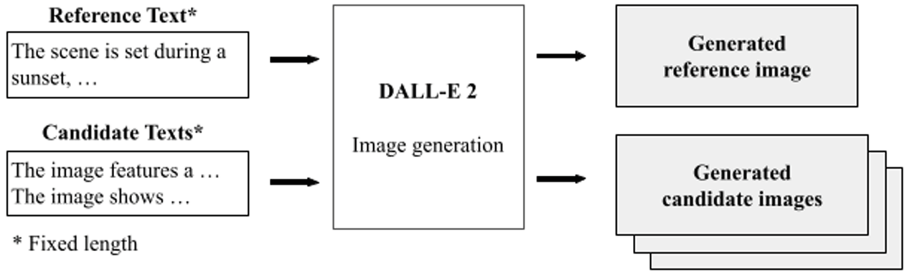


Fig. 4. Illustration of validation process using DALL-E 2

Cosine Similarity Analysis. The text outputs are compared to the corresponding image themes using cosine similarity, allowing us to quantify the alignment between the generated descriptions and the expected themes. This study investigates contextual embedding models that create embeddings at various layers within an encoder network. Prior research by Liu et al. (2019) [9] has highlighted the efficacy of intermediate layers for semantic tasks. We adopted the recommended representation layer for the similarity score, a selection made using the validation set outlined in prior studies by Zhang et al. (2020) [1], namely, the WMT16 dataset.

5 Results

The results of our experiments are described in detail in the following sections. Initially, we compared images within the same topic, followed by a comparison of images across different topics.

5.1 Images of the Same Topic

We show three sample images centered around the Eiffel Tower, as presented in Fig. 5.

Generated Image Descriptions

Question: Describe the background, subject, characters, colors, atmosphere, and emotions of the photo.

Reference Image 0 (Fig. 5a): The background of the image is the Eiffel Tower at night. The tower is lit up in the colors of the French flag, red, white, and blue. The fountains in front of the tower are also lit up in the same colors. The atmosphere of the image is romantic and beautiful. The emotions of the image are awe and wonder at the sight of the Eiffel Tower lit up in the colors of the French flag.

(a) Reference image 0 (b) Candidate image 1 (c) Candidate image 2

Fig. 5. Reference and candidate images of Eiffel Tower (Color figure online)

Candidate Image 1 (Fig. 5b): The scene is set during a sunset, and the sun is shining brightly behind the Eiffel Tower. The tower appears to be the main focus of the photo, towering over the landscape in front of it. There is a car parked in the foreground, with two cars visible to the left of the Eiffel Tower, and two cars visible to the right of it. Additionally, there are two bird silhouettes present in the scene, one on the right and one on the left of the tower. The overall atmosphere is serene and peaceful, with the golden sunset illuminating the Eiffel Tower and the surrounding area. There is no indication of any people or other objects in the scene, but one might imagine the tower attracting crowds of tourists and locals to admire its beauty.

Candidate Image 2 (Fig. 5c): The image features a beautiful garden with colorful trees in the foreground and the iconic Eiffel Tower standing tall in the background. The Eiffel Tower is an iron tower located in Paris, France, and it is an iconic landmark known throughout the world. It is a symbol of romance and Parisian culture, attracting millions of visitors annually. The garden is well-maintained with green grass, providing an ideal setting for a leisurely stroll. A path leads the viewer through the garden, making it easier to enjoy the scenery and take in the beauty of the Eiffel Tower. The trees are decorated with leaves of varying hues, adding to the vibrant and colorful atmosphere. The scene is peaceful, allowing visitors to relax and enjoy the surrounding beauty.

DALL-E 2 Input Text and Generated Images
We verified the context2image using images from the same topic.

Generated Image of Reference 0 (Fig. 6a): The background of the image is the Eiffel Tower at night. The tower is lit up in the colors of the French flag, red, white, and blue. The fountains in front of the tower are also lit up in the same colors. The atmosphere of the image is romantic and beautiful. The emotions of the image are awe and wonder at the sight of the Eiffel Tower lit up in the colors of the French flag.

Generated Image of Candidate 1 (Fig. 6b): The scene is set during a sunset, and the sun is shining brightly behind the Eiffel Tower. The tower appears to be the main focus of the photo, towering over the landscape in front of it. There is a car parked in the foreground, with two cars visible to the left of the Eiffel Tower, and two cars visible to the right of it. Additionally, there are two bird silhouettes present in the scene, one on the right and one on the left of the tower.

Generated Image of Candidate 2 (Fig. 6c): The image features a beautiful garden with colorful trees in the foreground and the iconic Eiffel Tower standing tall in the background. The Eiffel Tower is an iron tower located in Paris, France, and it is an iconic landmark known throughout the world. It is a symbol of romance and Parisian culture, attracting millions of visitors annually. The garden is well-maintained with green grass, providing an ideal setting for a leisurely stroll. A path leads the viewer through.

(a) New Eiffel Tower 1-1 (b) New Eiffel Tower 1-2 (c) New Eiffel Tower 1-3

Fig. 6. Generated image outputs from DALL-E 2 (Color figure online)

5.2 Images of Different Topics

In this context, we illustrate our comparative analysis by juxtaposing images Fig. 7) from four different topics with the reference image of the Eiffel Tower.

Generated Image Descriptions

Question: Describe the background, subject, characters, colors, atmosphere, and emotions of the photo.

Butterfly (Fig. 7a): The background of the image is a pink flower garden with white and purple pink flowers. The subject of the image is a black butterfly sitting on a pink flower with its wings open, showcasing its vibrant colors. The colors used in the image are primarily pink and purple, with the butterfly's wings and the flowers being the most prominent. The atmosphere of the image is peaceful and serene, with the butterfly appearing to be in a state of contemplation as it rests on the flower. The emotions conveyed in the image are serenity and tranquility.

Eating Ice Cream (Fig. 7b): The scene features a young girl wearing a white dress, standing in a park, eating a green and white ice cream cone. The ice

(a) Butterfly

(b) Eating ice cream

(c) Rainbow

(d) Pyramid of Djoser

Fig. 7. Images of different topics: Butterfly, Eating ice-cream, Rainbow, Pyramid (Color figure online)

cream cone stands out from her dress due to its bright color, and the contrast is visually appealing. Her expression shows she is delighted, possibly tasting a flavor she enjoys. In the background, there is a building or structure, possibly an amusement park, providing an entertaining setting for the girl to enjoy her ice cream.

Rainbow (Fig. 7c): The scene features an incredible double rainbow over a beach, accompanied by some trees and sand. The rainbow seems to have started just a few seconds ago, with a beautiful arc of colors spanning the horizon. It appears that the sunset has started, creating a warm and serene atmosphere in the scene. The rainbow is visibly captivating the beachgoers' attention, bringing a sense of wonder and delight to the picture. The colors of the rainbow, combined with the sunset, make the scene very pleasant, making it a momentous and memorable sight for those present.

Pyramid (Fig. 7d): The image shows the ruins of a large stone structure, which appears to be a pyramid. The structure is made of large, rough stone blocks and is surrounded by sand dunes in the desert. The pyramid has several chambers and entrances, and it appears to be in a state of disrepair. There are no people or animals in the image. The atmosphere is arid and desolate, with the only vegetation in the image being the small shrubs and cacti that can be seen growing around the base of the pyramid. The emotions conveyed in the image are of awe and reverence for the ancient structure.

DALL-E 2 Input Text and Generated Images
We verified the context2image using images from the different topics.

(a) Generated Butterfly

(b) Generated Eating ice cream

(c) Generated Rainbow

(d) Generated Pyramid of Djoser

Fig. 8. Generated image outputs from DALL-E 2 (Images of different topics): Butterfly, Eating ice-cream, Rainbow, Pyramid (Color figure online)

Generated Image of Butterfly (Fig. 8a): The background of the image is a pink flower garden with white and purple pink flowers. The subject of the image is a black butterfly sitting on a pink flower with its wings open, showcasing its vibrant colors. The colors used in the image are primarily pink and purple, with the butterfly's wings and the flowers being the most prominent. The atmosphere of the image is peaceful and serene, with the butterfly appearing to be in a state of contemplation as it rests on the flower.

Generated Image of Eating Ice Cream (Fig. 8b): The scene features a young girl wearing a white dress, standing in a park, eating a green and white ice cream cone. The ice cream cone stands out from her dress due to its bright color, and the contrast is visually appealing. Her expression shows she is delighted, possibly tasting a flavor she enjoys. In the background, there is a building or structure, possibly an amusement park, providing an entertaining setting for the girl to enjoy her ice cream.

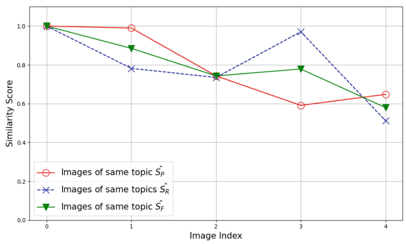
Generated Image of Rainbow (Fig. 8c): The scene features an incredible double rainbow over a beach, accompanied by some trees and sand. The rainbow seems to have started just a few seconds ago, with a beautiful arc of colors spanning the horizon. It appears that the sunset has started, creating a warm and serene atmosphere in the scene. The rainbow is visibly captivating the beachgoers' attention, bringing a sense of wonder and delight to the picture. The colors of the rainbow, combined with the sunset.

Generated Image of Pyramid (Fig. 8d): The image shows the ruins of a large stone structure, which appears to be a pyramid. The structure is made of large, rough stone blocks and is surrounded by sand dunes in the desert. The pyramid has several chambers and entrances, and it appears to be in a state of disrepair. There are no people or animals in the image. The atmosphere is arid and desolate, with the only vegetation in the image being the small shrubs and cacti that can be seen growing around the base of the pyram.

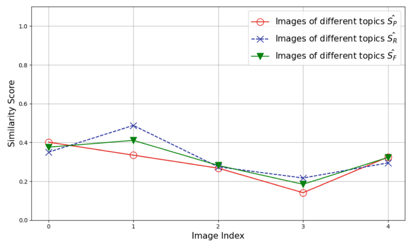
5.3 Comparison of Language Models

Our experiment involves two sets of images for the comparison of language models: Eiffel Tower images, representing similarity, and images depicting the act of eating ice cream, representing dissimilarity to the Eiffel Tower. To assess the image similarity, we leveraged specific pre-trained models, which are detailed in Table 1.

For English text generation evaluation, the recommended choice is the 24-layer RoBERTa-large model for computing BERTSCORE [1]. BERT-base-uncased is the base model of the BERT language model, and is learned with little data [8]. Xlm-mlm-en-2048 performs well in cross-lingual classification, and supervised and unsupervised machine translation. [14] We computed similarity scores of the images of the same topic and different topic scores in three BERT language models.



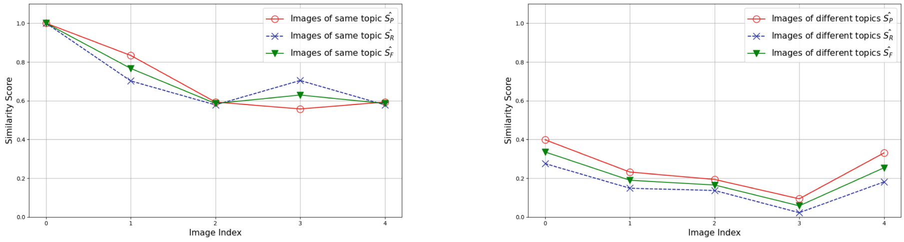
(a) Images of the same topic: five images of Eiffel Tower compared to the reference image of Eifel Tower with image index 0



(b) Images of different topics: five candidate images of Eating ice cream compared to the reference image of Eiffel Tower

Fig. 9. Rescaled similarity scores with RoBERTa-large. Each image index corresponds to an individual image within the dataset.

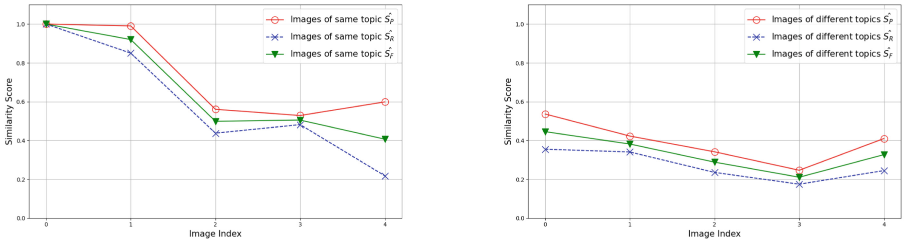
For RoBERTa-large (Fig. 9), images of the same topic showed a score of 0.5 or higher, and 0.4 or lower for images of different topics. Using BERT-base-uncased,



(a) Images of same topic with BERT-base-uncased

(b) Images of different topics with BERT-base-uncased

Fig. 10. Rescaled similarity scores with BERT-base-uncased



(a) Images of same topic with xlm-mlm-en-2048

(b) Images of different topics with xlm-mlm-en-2048

Fig. 11. Rescaled similarity scores with xlm-mlm-en-2048

we found that the similarity scores were above 0.6 for images considered as the same topic, and below 0.4 for images from various topics after rescaling, as shown in Fig. 10. For xlm-mlm-en-2048 shown in Fig. 11, scores for images of the same topic range from 0.2 to 1.0, whereas images of different topics receive scores of 0.6 or lower. Unlike other models, the score data for both the same topic and different topic images make it challenging to discern the relationship between the images. The language model that exhibits the most distinguishable image similarity scores among the experimental models is bert-base-uncased.

5.4 Comparison of Topics

With bert-base-uncased, we compared the F1 Similarity Score $\hat{S_F}$ for images of the same topic and different topics, as illustrated in Fig. 12. The comparison involved 25 images across five topics from our dataset, with each image index corresponding to a unique image of its topic. The Eiffel Tower, represented as a hollow circle ○, serves as the reference at an image index of 0. Similarity scores between the reference image and other Eiffel Tower images are plotted along the solid line. We repeated the comparison using the same Eiffel Tower reference image with images of different topics as candidate images, for example, the Butterfly topic represented as 'x' along the dashed lines, the Eating ice cream topic in ●, Rainbow in ▽, Pyramid in △.

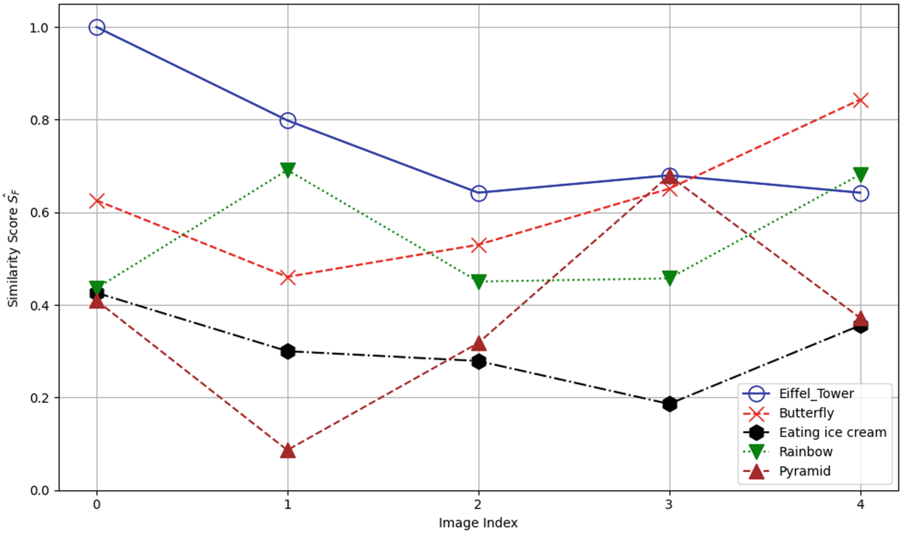


Fig. 12. Comparison of different topics (bert-base-uncased) with Fig. 5a as the reference image represented at image index 0 in ○.

We observed a cosine similarity score of 0.6 or higher for images categorised as the same topic, whereas scores below 0.6 indicated different topics. However, outliers exist where images of different topics score higher than 0.6. At image index 3, where a consistent similarity score is observed, we noticed shared information among images of the Eiffel Tower and the Pyramid. This shared information includes details about the blue sky, architectural structures, and a variety of colors depicted in all three images of the Eiffel Tower, the Butterfly, and the Pyramid. Further experiments using larger image datasets and varied rescaling techniques with different models are essential to comprehensively examine and address these cases.

6 Conclusion

We introduced an integrated approach that efficiently calculates image similarities by harnessing the capabilities of vision transformers and language models. Using BERTScore and Video-LLaMA, we extracted textual descriptions through question prompts, experimented with diverse pre-trained language model configurations, and quantified similarities with cosine similarity metrics. Our results highlight the effectiveness of specific pre-trained language models in distinguishing between images of the same and different topics, revealing a clear gap in similarity scores. We verified the coherence of the generated text by generating images using DALL-E 2 from the original image descriptions and subsequently comparing the generated images with the originals. However, this study is limited by the size of the image dataset used in the experiments, and a larger variety of image categories should be included to strengthen the demonstration of image similarity score. To further enhance our approach, we will explore various question prompts and diverse BERTScore models, optimise layers within each model, and integrate the similarity computations within Video-LLaMA.

References

1. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: evaluating text generation with BERT (2019)
2. Huggingface blog. https://huggingface.co/blog/vision_language_pretraining. Accessed 8 Sept 2023
3. Kim, P.: Convolutional neural network. In: Kim, P. (ed.) MATLAB Deep Learning, pp. 121–147. Apress, Berkeley (2017). https://doi.org/10.1007/978-1-4842-2845-6_6
4. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale (2020)
5. Zhai, X., et al.: Scaling vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
6. Zhang, H., Li, X., Bing, L.: Video-llama: an instruction-tuned audio-visual language model for video understanding (2023)
7. Li, J., et al.: Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models (2023). https://doi.org/10.48550/arXiv.2301.12597
8. Devlin, J., et al. BERT: pre-training of deep bidirectional transformers for language understanding (2018)
9. Liu, N.F., et al.: Linguistic knowledge and transferability of contextual representations (2019). https://doi.org/10.18653/v1/N19-1112
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. Association for Computational Linguistics (2002). https://doi.org/10.3115/1073083.1073135
11. Eddine, M.K., et al.: FrugalScore: learning cheaper, lighter, and faster evaluation metrics for automatic text generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2022)
12. Wang, X., Zhu, Z.: Context understanding in computer vision: a survey. Comput. Vis. Image Understanding **229** (2023)
13. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach (2019)
14. Lample, G., Conneau, A.: Cross-lingual language model pretraining (2019)
15. Griffin, G., Holub, A.D., Perona, P.: Caltech 256 Image Dataset
16. Huggingface docs, pre-trained models. https://huggingface.co/transformers/v3.4.0/pretrained_models.html