Journal of Innovations in Computer Science and Engineering
Shahid Beheshti Unversity

May 2023, Volume 1, Number 1

A Weighted Multi-Criteria Decision Making Approach for Image Captioning

Hassan Maleki Golandouz [a]✉ Code Orcid: 0009-0009-2836-4395,
Mohsen Ebrahimi Moghaddam [b] Code Orcid: 0000-0002-7391-508X,
Mehrnoush Shamsfard [c] Code Orcid: 0000-0002-7027-7529
a Faculty of Computer Science and Engineering, Shahid Beheshti University G.C, Tehran, Iran, hassanmalekii92@gmail.com
b Faculty of Computer Science and Engineering, Shahid Beheshti University G.C, Tehran, Iran, m_moghaddam@sbu.ac.ir
c Faculty of Computer Science and Engineering, Shahid Beheshti University G.C, Tehran, Iran, m-shams@sbu.ac.ir

ABSTRACT
Image captioning aims to automatically generate descriptions of images in natural language. This challenging problem in the field of artificial intelligence has recently gained significant attention in computer vision and natural language processing. Among the existing approaches, visual retrieval-based methods have shown high effectiveness. These approaches search for similar images and then construct a caption for the query image based on the captions of the retrieved images. In this study, we propose a method for visual retrieval-based image captioning. We utilize a multi-criteria decision-making algorithm to effectively combine multiple criteria with proportional impact weights in order to retrieve the most relevant caption for the query image. The main idea of our approach is to design a mechanism that retrieves semantically relevant captions for the query image and then selects the most appropriate caption by imitating human decision-making using a weighted multi-criteria decision-making algorithm. Experimental results on the MS COCO benchmark dataset demonstrate that our proposed method outperforms state-of-the-art models, providing more effective results.

KEYWORDS
Image Captioning, Machine Vision, Natural Language Processing, Multi-Criteria Decision Making, Transfer-based Approach

1. INTRODUCTION

Image captioning aims at automatically generating natural language descriptions for images. It has garnered considerable interest in the field of computer vision due to its wide range of applications, including facilitating human-machine interaction and enabling image retrieval through verbal communication. This area of research bridges two key domains in artificial intelligence: computer vision and natural language processing [1]. The generated descriptions aim to capture various visual aspects of the image, such as objects and their characteristics, scene features (e.g., indoor or outdoor), and verbalize interactions among individuals and objects within the scene [2].

Image captioning models are designed to process images and produce descriptive text that highlights the most important elements of the visual content. These models can be broadly categorized into two groups. The first group focuses on generating captions directly from images, utilizing techniques such as object detection, attribute prediction, scene classification, and action recognition, based on visual features [3]-[5]. These visual cues are then utilized to generate the caption through surface realization. In recent years, a particular set of generative approaches involves combining convolutional neural networks (CNNs) with recurrent neural networks (RNNs). Initially, high-level features are extracted from a CNN trained on image classification tasks, followed by a recurrent model that generates subsequent words conditioned on the image features and previously predicted words [6]-[9].

The second group of approaches in image captioning, known as data-driven approaches, treat the task as a retrieval problem [2]. These methods generate descriptions for a query image by searching for similar images and using the captions of the retrieved images. The description for the query image can be achieved by reusing the caption of the most similar retrieved image (transfer) or synthesizing a new caption based on the retrieved captions. Data-driven methods face two main challenges. Firstly, the retrieval set needs to be of high quality, ensuring relevant and informative captions. Secondly, a suitable similarity metric is required to measure the similarity between the query image and the retrieved captions. An effective similarity metric should consider the comparison between the query image and candidate captions, which belong to different modalities, and dynamically assign different weights to various criteria.

For instance, in Fig. 1-b, it can be observed that the query image and retrieved captions match well in terms of objects and attributes, but the crucial criterion for consideration is the matching of actions. On the other hand, in Fig. 1-a, object matching is more determinative, while in Fig. 1-c, attribute matching carries more weight. Therefore, a suitable similarity metric should dynamically assign appropriate weightings to different criteria (e.g., object, attribute, action) based on the specific image and its candidate captions, enabling the selection of the most appropriate caption for the query image.

Submit Date: 2022-08-23
Revise Date: 2023-05-01
Accept Date: 2023-05-20
✉ Corresponding author

a
a living **room** has a couch a **table** and a small **television**
a **room** with **television** and **lights** turned on and a **screen**
a **television** and some **books** in a **room**
shelves with **books** a **television** set a **table** a **cat** and a **clock**

b
an adult giraffe **standing** in a lush green tree filled field
a giraffe is **walking** around some bushes and trees
a very tall adult giraffe **eating** from a basket
a giraffe **looking** at the camera with trees in the background

c
a man sitting next to a **large brown** dog
a dog standing on a bench next to a man with **green** shirt
a dog is sitting on a bench beside a **young** man
a **large white** dog sits on a bench with people next to a path

a, b, c

Figure 1: Here are some examples that highlight the importance of weight allocation for each criterion in selecting the appropriate caption. Among the retrieved captions in scenarios (a), (b), and (c), **objects**, **actions** and **attributes** play a crucial role in selecting the appropriate caption, respectively.

This paper introduces a novel similarity metric between images and descriptions to enhance data-driven approaches in image captioning. The metric is designed based on the understanding that individuals consider different criteria when selecting an appropriate caption among retrieved captions. The similarity is measured by assigning impact weights to various criteria, including object matching, attribute matching, and action matching, which play significant roles in comparing the image and description. Firstly, semantically relevant descriptions are retrieved for the query image. Then, the matching score between the query image and retrieved captions is calculated using the proposed similarity metric. Finally, the most appropriate caption is selected.

The main contributions of this paper are twofold. Firstly, it introduces a novel step in multi-criteria decision making, where impact weights are assigned to each criterion to improve the overall results. Secondly, it presents a new similarity metric that considers multiple criteria and automatically determines their weights to measure the similarity between images and descriptions.

In general, data-driven image captioning methods can be classified into three categories: transfer-based approaches using visual information, which select and transfer a caption based solely on visual information [10, 11]; transfer-based approaches using both visual and textual information, which select and transfer a caption based on a combination of visual and textual information [12, 13]; and generation-based approaches, which generate a novel description by combining fragments of the retrieved descriptions [10, 14].

Transfer-based approaches using visual information. This category involves finding the visually closest image from a large set of images and transferring its caption as the final description. The Im2Text model [10] is one of the pioneering works in this category, proposing a two-step retrieval process for caption transfer. In the first step, global image features such as GIST [15] and Tiny Image [16] descriptors are used to find visually similar images. This step serves as a baseline for most retrieval-based models. In the second step (re-ranking), a range of detectors and scene classifiers, including object, stuff, pedestrian, and action detectors, are applied to the images based on the entities mentioned in the candidate captions. This semantic representation is then used to re-rank the associated captions in the final step.

In general, the proposed method shares similarities with the Im2Text model, but it also exhibits several notable differences:

- The proposed method utilizes both visual and textual information, whereas the Im2Text model relies solely on visual information.
- The Im2Text model requires training a classifier to select the appropriate caption from the retrieved captions. However, this approach may not always achieve the desired accuracy, and to improve accuracy, a large and diverse training dataset is necessary. In contrast, the proposed approach does not rely on training a classifier. Instead, it utilizes statistics and probabilistic methods to determine the most appropriate caption. This eliminates the need for extensive training data and offers a more efficient and flexible solution.
- In the Im2Text model, all criteria are assigned equal weight. In contrast, the proposed method dynamically and automatically determines the impact weight of each criterion based on the specific image and its content.

Another method in this group is proposed by Patterson et al. [11]. They introduce a large-scale scene attribute dataset to the computer vision community, which is a novel contribution. Using this dataset, they train attribute classifiers and demonstrate that the responses of these classifiers can serve as a more effective global image descriptor compared to traditional descriptors like GIST. They enhance the baseline model by substituting the global features with automatically extracted scene attributes, resulting in improved caption transfer performance.

Transfer-based approaches using visual and textual information. In this category, Mason and Charniak [12] adopt a caption transfer approach that treats it as an extractive summarization problem. They specifically focus on textual information during the final re-ranking step. They utilize the scene attributes descriptor proposed in [11] to represent images. First, visually similar images are retrieved from the training set, and then the conditional probabilities of observing a word in the query image's caption are estimated using non-parametric density estimation based on the captions of the retrieved images. The final output caption is obtained through two extractive summarization techniques: the SumBasic model [17] and Kullback-Leibler divergence between word distributions.

Another method proposed by [13] adopts an average query expansion approach based on compositional distributed semantics. They use features extracted from Visual Geometry Group Convolutional Neural Network model (VGG-CNN; [18]), trained on ImageNet to represent images. Visually similar images are retrieved, and a new query is constructed based on the average of distributed representations of the retrieved captions, weighted by their similarity to the input image. The method presented in [19] also employs CNN activations to represent images and performs k-nearest neighbor retrieval to find visually similar images from the training set. Similar to [12] and [13], it selects a caption from the candidate captions associated with the retrieved images that best describes the query image. However, it differs in the similarity representation between captions and the approach for choosing the best candidate. For each retrieved caption, the authors compute the n-gram overlap F-score with the other retrieved captions and define the consensus caption as the one with the highest mean n-gram overlap. They propose selecting a single caption corresponding to the caption with the highest mean n-gram overlap among the other candidate captions, estimated using an n-gram similarity measure.

Generation-based approaches. In the generation-based approach proposed by Kuznetsova et al. [14], their model follows a multi-step process to generate captions for query images. Firstly, the detectors and classifiers used in the re-ranking step of the Im2Text model are applied to the query image to extract and represent its semantic content. Instead of performing a single retrieval, their model conducts separate image retrieval steps for each detected visual element in the query image to gather relevant phrases from the retrieved captions. This step involves collecting three types of phrases: noun and verb phrases, prepositional phrases for stuff detections, and additional prepositional phrases for scene context detections.

Noun and verb phrases are extracted from the captions in the training set by considering visual similarity among object regions detected in both the training images and the query image. Prepositional phrases for stuff detections are collected by measuring the visual similarity of appearance and geometric arrangements between the detections in the query and training images. Similarly, prepositional phrases for scene context detections are obtained by calculating the global scene similarity between the query and training images. Finally, the collected phrases for each detected object are used in an integer linear programming (ILP) framework that considers factors such as word ordering and redundancy. This ILP process generates the output caption for the query image based on the collected phrases and their relationships.

Similarly, Gupta et al. [10] propose a model that uses hand-crafted visual features to retrieve visually similar images. They extract linguistically motivated phrases from the descriptions of these retrieved images and employ a joint probability model that considers image similarity and Google search counts to measure the relevance of these phrases. The final sentences are generated using the SimpleNLG tool [11]. Mert Kilickaya et al. [20] also propose a mechanism that combines deep features with an object-specific image representation to select relevant images from a large set. They employ a local image analysis method to explore similarities among local image regions, which are then used to collect phrases from the relevant set of images. Finally, sentences are generated using language models based on these collected phrases.

2. The proposed method

The proposed approach consists of two main parts: Part one focuses on retrieving semantically more relevant captions using the query image (as shown in Fig. 2). Part two involves selecting the most appropriate caption from the retrieved captions, based on a novel similarity metric (as illustrated in Fig. 5). In the following, we provide a detailed description of each part.

3. Part one: Retrieve semantically more relevant captions with the query image

3.1.1. Retrieving visually similar images

In data-driven approaches, the quality of the initial retrieval is crucial, and having accurate visual features is paramount [13]. To represent images, we utilize the top-layer features of a pre-trained CNN [18], which results in a 4096-dimensional feature vector. Our first objective is to identify a set of k nearest training images for each query image based on visual similarity. It is important to ensure that there are no outliers, as this greatly affects the effectiveness of the approach. Instead of using a fixed neighborhood, we employ an adaptive strategy to select the initial candidate set of image-caption pairs $\{(I_i, C_i)\}$ similar to [13]. For a given query image Iq, we employ a ratio test and only consider candidates that fall within a radius determined by the distance score between the query image and the nearest training image $I_{closest}$, as described in [13].

$$N(I_q) = \{(I_i, c_i) |\ dist(I_q, I_i) \leq (1+\varepsilon)\ dist(I_q, I_{closest}), I_{closest} = \arg\min dist(I_q, I_i),\ I_i \in T\} \quad (1)$$

where "dist" represents the Euclidean distance between two feature vectors. The set N denotes the candidate set based on the adaptive neighborhood, and T represents the training set. The parameter ε is the adaptive neighborhood parameter, which is a positive scalar value and is determined empirically.

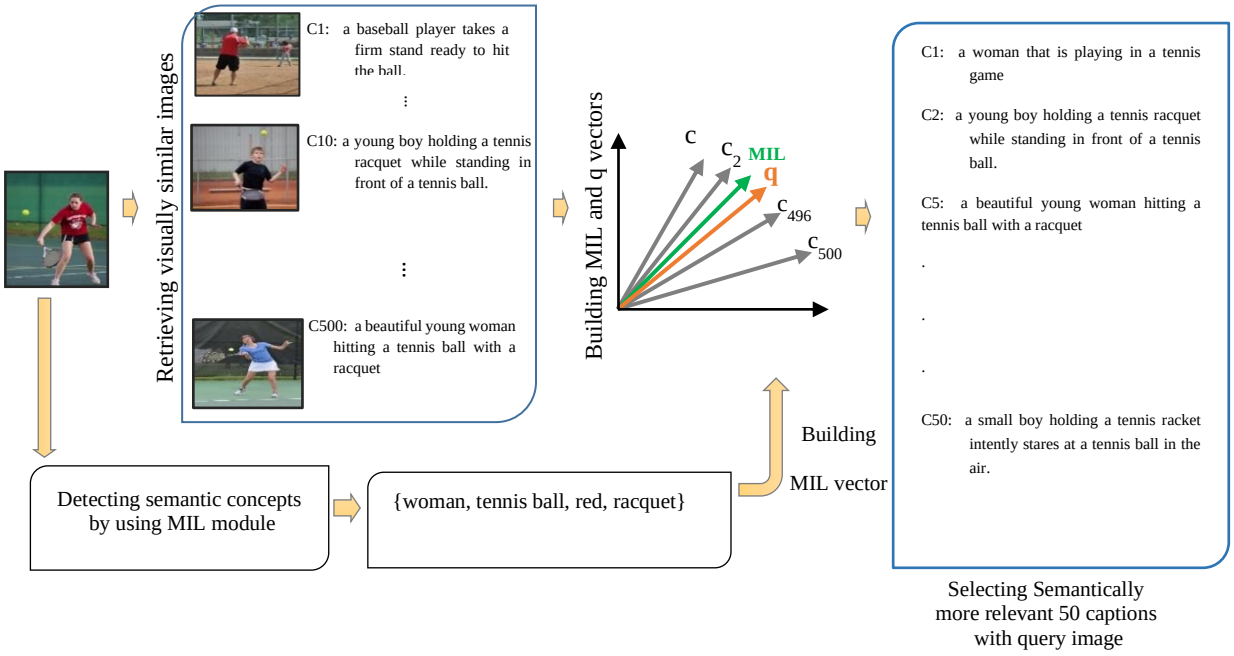


Figure 2. Part one of the proposed approach aims to retrieve semantically more relevant captions for the query image by leveraging both visual and textual information.

4. Detecting semantic concepts

For a query image Iq, the first step in part one is to retrieve visually similar images from a large collection of captioned images. Then, a pre-trained Multiple Instance Learning (MIL) model is used to detect a set of semantic concepts, or tags, that are likely to be part of the image's caption (Fig. 3). The method described in [21] is employed for this purpose. To detect these semantic concepts, a set of tags is selected from the caption text in the training set, and the k most common words in the training captions are used to determine the vocabulary of tags, following the approach outlined in [21] and [22].

Treating the prediction of semantic concepts as a multi-label classification task, the problem can be formulated as follows: Let there be N training examples, and let $y_i = [y_{i1}, \ldots, y_{ik}] \in \{0,1\}^k$ be the label vector for the i^{th} image, where $y_{ik} = 1$ if the image is annotated with tag k, and $y_{ik} = 0$ otherwise. Denoting v_i and s_i as the image feature vector and the semantic feature vector for the i^{th} image, respectively, the cost function to be minimized is [21]:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M}(y_{ik}\log s_{ik} + (1-y_{ik})\log(1-s_{ik})) \quad (2)$$

where $s_i = \sigma(f(v_i))$ is a K-dimensional vector with $s_i = [s_{i1}, \ldots, s_{ik}]$, σ(·) is the logistic sigmoid function and f(·) is implemented as a multilayer perceptron (MLP). During the testing phase, for each input image, a semantic concept vector s is computed. This vector represents the probabilities of all tags and is generated by the semantic-concept detection model.

5. Building MIL and q vector

In this study, the meaning of a word is represented by a 500-dimensional vector, which captures the context in which the word appears in a corpus. These word vectors are trained using the caption data from the MS COCO dataset, which consists of 620K captions. The word2vec model, specifically the predict-based model described in [24], is utilized for training the word vectors. During training, the minimum word count and the window size are set to 5 and 10, respectively.

Similar to the approach described in [25], we generate the vector representation of a caption by first removing stop words and then summing the vectors of the remaining words in the caption. Once the image tags are predicted, their corresponding vectors are obtained using word2vec. These tag vectors are then combined to form an MIL vector, which serves as an approximation of the ideal caption vector for the input image (Iq). The retrieved captions are subsequently re-ranked based on the cosine distance between their vectors and the MIL vector. Finally, we select the top n captions that are closest to the MIL vector as the candidate descriptions for the input image. The process of obtaining the n captions close to the MIL vector is outlined in Algorithm 1.

However, there are instances where the MIL model fails to predict any appropriate tags for the query image. In such cases, we adopt an alternative approach inspired by [10]. Instead of using the MIL vector, we create a vector called the q vector, which is obtained by taking the weighted average of the vectors of the retrieved captions. The calculation of the q vector is performed as follows [13]:

$$q = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} sim(I_q, I_i).c_i^j \qquad (3)$$

where N and M denote the total number of image-caption pairs in the candidate set N and the number of reference captions associated with each training image, respectively. The term $sim(I_q, I_i)$ refers to the visual similarity score of the image I_i to the query image I_q which is used to give more significance to the captions of images that visually are closer to the query image. $sim(I_q, I_i)$ is defined by the equation (4) as follows[13]:

$$sim(I_q, I_i) = 1 - dist\,(I_q, I_i)/Z^1 \qquad (4)$$

In cases where the MIL vector is constructed, we select the n nearest neighbors. Otherwise, we create the q vector and also choose the n nearest neighbors for further investigation. The number of retrieved captions that are close to the MIL vector or the q vector varies based on the number of images retrieved within a given radius, as specified by equation (1).
The maximum value of n is set to 50 because when retrieving visually similar images to the query image, a maximum of 100 images are retrieved, and each image has 5 captions. This results in a total of 500 captions being retrieved. By selecting 50 captions that are close to the MIL vector or the q vector, in the worst-case scenario, only 10=50/5 different types of descriptions may be obtained, each expressed in five different forms. Therefore, setting the number of neighbors to 50 ensures obtaining at least ten distinct caption types.

Algorithm1 obtain the n captions close to the MIL vector

Input: Query image and retrieved captions
Output: n captions close to the MIL vector
Begin
1. Predict query image tags, T={t_1, t_2, …, t_m }.
2. Obtain vectors of the tags by using word2vec model, V={ v_1, v_2, …, v_m }
3. Compute MIL vector, MIL= $\sum_{i=1}^{m} V_i$
4. Compute the cosine distance between the vectors of captions (C_i) and the MIL vector:
   D_i= Cosine distance(C_i . MIL_i) for i=1, … ,N' (number of retrieved captions)
5. Sort D_i descending and select top n as the n captions close to the MIL vector

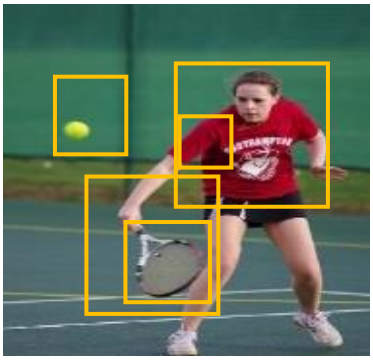
End

In the next step, the candidate captions are subjected to a more detailed comparison with the query image based on predefined criteria, including object matching, attribute matching, and action matching. This comparison process aims to identify the most appropriate caption by considering the aforementioned criteria and employing a multi-criteria decision-making algorithm. To facilitate the matching process, a prepared list (referred to as prepared list_1 in Fig. 4) is utilized. This list is constructed based on the captions from the MS COCO training dataset and consists of the 1000 most frequently occurring words along with their corresponding part-of-speech (POS) tags. In cases where a word has multiple POS tags, the most frequently observed POS tag is considered for matching purposes.

6. Part two: Selecting the most appropriate caption based on similarity metric

7. How to perform matches

Object and action matching. Matching objects in the query image with the objects (nouns) in the candidate captions is carried



{woman, tennis ball, red, racquet, hit}

Figure 3: A query image and predicted semantic concepts for it

out as follows:

$$The\ matching\ score\ = \frac{M*(+1)+[min(B,Q)-M]*(-1)-[|(Q-B)|*P]}{Q} \qquad (5)$$

[1] Z is a normalization constant

where Q and B denote the number of objects in query image and the candidate captions, respectively. M is the number of matches between query image objects and candidate captions objects, and P refers to the amount of penalty for non-matching which calculated as follows:

$$P = \frac{1}{2} \quad if Q > B \qquad (6)$$
$$P = \frac{1}{3} \quad if Q < B$$

The matching score of objects in the query image with the objects mentioned in the candidate caption is measured as follows: According to equation (5), first the number of object matching is obtained (M). For matches, the score of "+1" ($M * (+1)$), and for non-matching, the score of "-1" ($[\min(B, Q) - M] * (-1)$) are considered. If the number of objects in the query image (Q) and the candidate caption (B) are not equal, a penalty is imposed: ([| (Q-B) | * P]). But how can we measure the matching score between the objects in the query image and the objects mentioned in the candidate caption?
Consider the objects "car" and "truck". Although they may not have similar visual characteristics, they are semantically similar as they both represent vehicles used for road transportation. Hence, it is important to use a method that can measure the semantic similarity between two words rather than their apparent similarity. There are two methods that can be considered for measuring the matching score of objects in the query image with the objects in the candidate caption: a) Using WordNet to measure the similarity of two words. b) Using the word2vec vectors of two words and calculating the cosine similarity between them.
In this study, the second method is employed. The object matching score is obtained by calculating the cosine similarity between the word2vec vectors using equation (5). If the cosine similarity between two vectors is greater than or equal to the threshold H[2], it indicates a match with a corresponding similarity score. On the other hand, if the cosine similarity is below the threshold, it implies a non-match. The action matching process follows a similar procedure as the object matching, utilizing equation (5) and considering the aforementioned aspects.

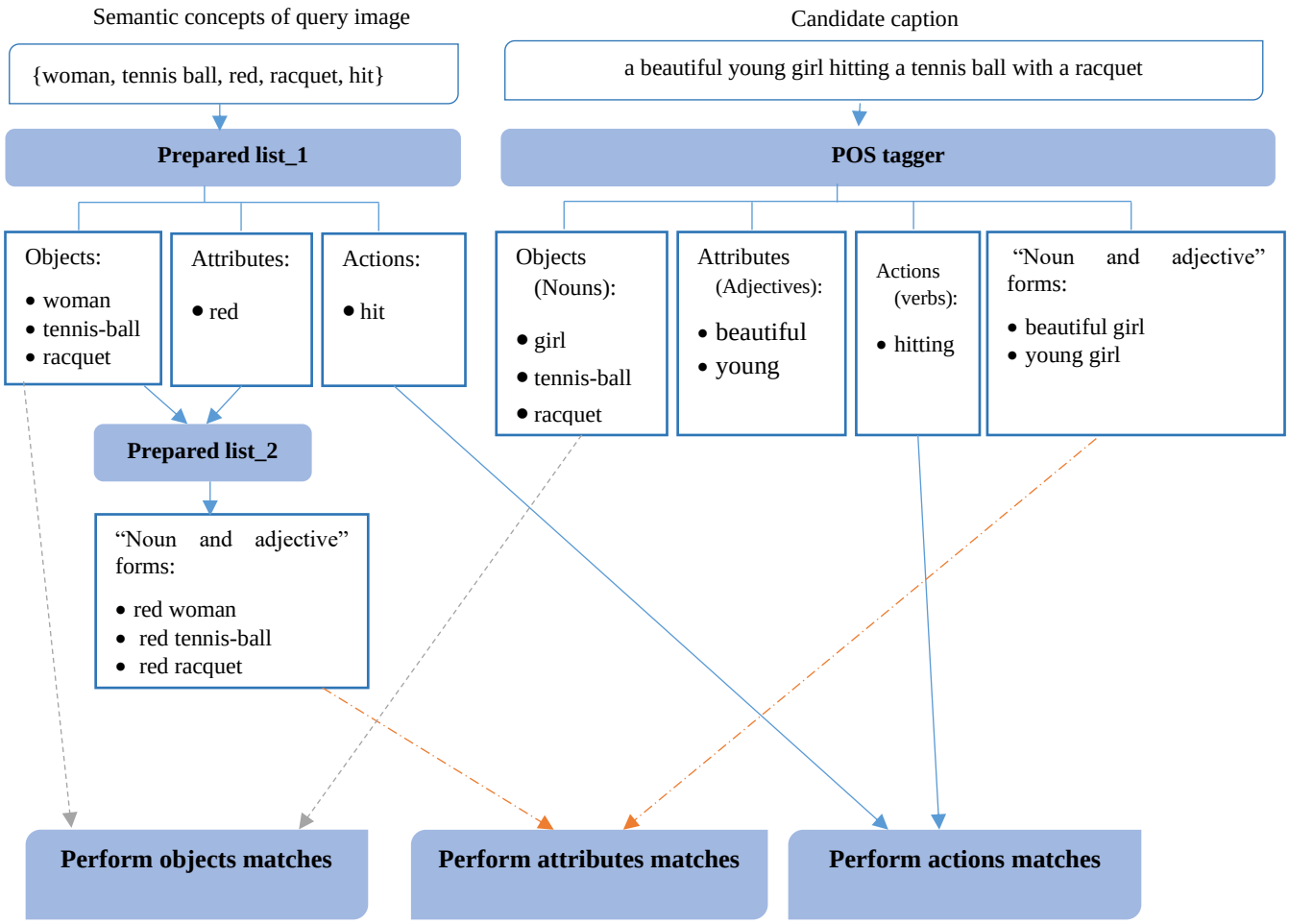


Figure 4: Diagram of how to perform matches

[2] The parameter H is a positive real value.

Attribute matching. The attribute matching process involves assigning each word identified as an attribute from the MIL output to a corresponding noun before performing the matching. To accomplish this, a prepared list (prepared list_2 in Fig. 4) is utilized. This list is created based on the captions of the MS COCO training dataset and contains all the "noun and adjective" forms found in the training set.
In the candidate captions, the "noun and adjective" forms are obtained using a POS tagger tool. The "noun (object) and adjective (attribute)" forms of the query image are then matched with the "noun and adjective" forms of the candidate captions using word2vec vectors. The vector representation of the "noun and adjective" form is constructed by summing up the vectors of its constituent words. Similar to objects and actions, attributes are also matched based on the cosine similarity between their word2vec vectors. If the cosine similarity between two attribute vectors is equal to or greater than the threshold H, they are considered as matches. In cases where a complete match occurs (i.e., the cosine similarity is one), a score of one is assigned. For other cases, the similarity score is used to represent the degree of similarity between the attributes. In Fig. 4, a diagram is presented illustrating how matches are performed based on the criteria discussed above. The caption with the highest overall score based on these criteria is selected as the most appropriate caption for the query image.

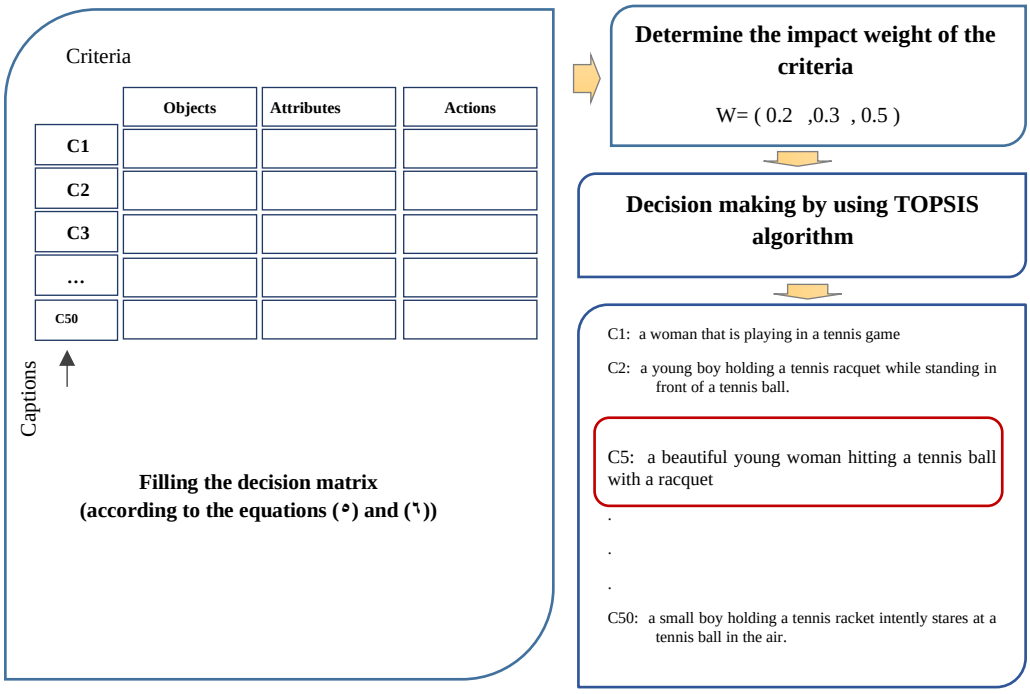


Figure 5: Part two of the proposed method: Selecting the most appropriate caption for the query image based on new similarity metric.

8. Filling the decision matrix

In many situations, it is desirable to make decisions that consider multiple criteria. Multiple Criteria Decision Making (MCDM) is a field that deals with the process of making decisions when there are multiple criteria involved, and these criteria often conflict with each other [26]. There are two main types of MCDM problems: one involves a finite number of alternative solutions and is known as multiple criteria decision, while the other deals with an infinite number of solutions and is called multiple objective optimization [26]. In the current study, we are dealing with a multiple criteria decision problem.
To describe a MCDM problem, a decision matrix is commonly used. Suppose we have m alternatives (captions) to be evaluated based on n criteria. The decision matrix is an m × n matrix, where each element Y_{ij} represents the value of the j_{th} criterion for the i_{th} alternative. In our case, the elements of the decision matrix are determined based on equations (5) and (6). Once the decision matrix is constructed, we can proceed with making decisions. However, an important step in the process is to determine the impact weights of the evaluation criteria.

9. Determine the impact weight of the criteria

In the current step, the impact weights of the criteria are determined using Shannon's entropy algorithm. Shannon's entropy is widely recognized and utilized particularly when it is not feasible to obtain appropriate weights based on preferences or conducting experiments with decision-makers [27]. Shannon's entropy is a fundamental concept in information theory, providing a means to quantify the level of uncertainty or randomness within a system [27]. By applying Shannon's entropy algorithm, the uncertainty associated with each criterion can be assessed, enabling the determination of their respective weights. The steps involved in Shannon's entropy algorithm are as follows:
1) Normalize the decision matrix.

$$Pij = \frac{r_{ij}}{\sum_{i=1}^{m} r_{ij}} \forall_{ij} \quad \text{for i = 1, . . . ,m and j=1, . . . , n. (7)}$$

The raw data are normalized to eliminate anomalies with different measurement units and scales. This process transforms different scales and units among various criteria into common measurable units to allow for comparisons of different criteria.

2) Compute entropy

$$E_j = -\frac{1}{\text{Ln m}} \sum_{i=1}^{m} P_{ij}\, Ln\left(P_{ij}\right) \text{ , For j=1, . . . , n} \quad (8)$$

3) Set the degree of diversification as:

$$d_j = 1 - E_j \text{ , For j=1, . . . , n} \quad (9)$$

4) Set the importance degree of attribute j:

$$W_j = \frac{d_j}{\sum_{s=1}^{n} d_s} \text{ , For j=1, . . . , n} \quad (10)$$

In the next step, Multi-criteria decision making is done by using TOPSIS (technique for order preference by similarity to an ideal solution) algorithm.

10. Decision making by using TOPSIS Algorithm

Selecting the most appropriate caption for the query image. Among the different methods of decision making with multiple criteria, the TOPSIS method was chosen for this study due to its advantages over other methods, such as the ability to incorporate both quantitative and qualitative criteria simultaneously. The TOPSIS method, introduced in [28] and referenced in [29], is a powerful approach for handling multiple criteria and identifying solutions from a finite set of alternatives. The underlying principle of the TOPSIS method is to select an alternative that has the shortest distance to the ideal solution and the farthest distance from the negative-ideal solution. The TOPSIS method consists of the following steps in its procedure:

1) Calculate the normalized decision matrix. The normalized value n_{ij} is calculated as

$$n_{ij} = \frac{r_{ij}}{\sqrt{\sum_{i=1}^{m} r_{ij}^2}} \quad \text{for i = 1, . . . , m and j=1, . . . , n.} \quad (11)$$

2) Calculate the weighted normalized decision matrix. The weighted normalized value v_{ij} is calculated as

$$v_{ij} = w_j n_{ij} \text{ for i = 1, . . ., m and j=1, . . . , n} \quad (12)$$

where w_j is the weight of the jth criterion, and $\sum_{j=1}^{n} w_j$=1 . These weights are obtained using the Shannon's Entropy Algorithm

3) Determine the positive-ideal and negative-ideal solution

$$A^+ = \{(v_1^+, v_2^+, \ldots, v_n^+)\} = \{(\max v_{ij} | i \in O), (\min v_{ij} | i \in I)\} \quad (13)$$
$$A^- = \{(v_1^-, v_2^-, \ldots, v_n^-)\} = \{(\min v_{ij} | i \in O), (\max v_{ij} | i \in I)\} \quad (14)$$

where O is associated with benefit criteria, and I is associated with cost criteria.

4) Calculate the separation measures, using the n-dimensional Euclidean distance. The separation of each alternative from the ideal solution is given as

$$d_i^+ = \left\{\sum_{i=1}^{m} \left(V_{ij} - V_i^+\right)\right\}^{\frac{1}{2}} \ \forall i. \quad (15)$$

Similarly, the separation from the negative-ideal solution is given as

$$d_i^- = \left\{\sum_{j=1}^{m} \left(V_{ij} - V_i^-\right)\right\}^{\frac{1}{2}} \ \forall i. \quad (16)$$

5) Calculate the relative closeness to the ideal solution. The relative closeness of the alternative A_i with respect to A^+ is defined as

$$cl_i = \frac{d_i^-}{d_i^+ + d_i^-} \text{ for i = 1, . . ., m. Since } d_i^- \geq 0 \text{ and } d_i^+ \geq 0, \text{ then clearly } cl_i \in [0, 1]. \quad (17)$$

6) Rank the preference order. ranking alternatives using this index in decreasing order.

11. EXPERIMENTAL RESULTS

Details about experimental setup, are given below.

12. Corpus

The representation of words in our approach is derived from the captions found in the MS COCO dataset, which consists of a collection of 620K captions. As a preprocessing step, all captions in the corpus are converted to lowercase, and punctuation is removed. Following a similar methodology as described in [13], we employ 500-dimensional vectors that are trained using the word2vec model [24].

13. Dataset and Settings.

We conducted our experiments on the widely used MS COCO [23] dataset, which is a large-scale dataset consisting of 123K images. The dataset comprises 82,783 training images and 40,504 validation images. Each image in the dataset is accompanied by five reference captions, which are annotated by different individuals. The MS COCO dataset is known for its diversity, containing images with multiple objects and rich contextual information. Due to these characteristics, it serves as a challenging testbed for image captioning and has been widely utilized in recent research on automatic image captioning. To ensure a fair comparison with previous works, we followed the train, validation, and test splits provided by [6]. Specifically, we utilized all 82,783 images from the training set for training our model. For validation purposes, we used 5K images, and for testing, we employed another 5K images.

In our experiments, we employed the validation split of the MS COCO dataset as a tuning set to optimize the hyperparameters of our method. This involved fine-tuning and selecting the optimal values for various parameters to enhance the performance of our approach. The test split of the dataset was then used for evaluation and reporting the results. To build our knowledge base, we considered all the image-caption pairs from both the training and validation splits. One important parameter in our method is H, as mentioned in section 2.2.1. We empirically set the value of H to 0.85. The MS COCO dataset is continuously being updated and improved. For the purpose of our study, we relied on version 1.0 of the dataset to ensure consistency in our results. The publicly available code [6] was followed to preprocess the captions from the dataset. This preprocessing step resulted in a vocabulary size of 8791 for the MS COCO dataset.

14. Metrics

The proposed approach is compared with several baseline models and existing approaches in the field. These include the adapted baseline model VC from Im2Text [10], which selects the caption of the visually closest image as the description. Additionally, the word frequency-based approaches MC-SB and MCKL from [12] are used for comparison, as well as the model QE presented by [13], which utilizes an average query expansion approach based on compositional distributed semantics.

To ensure a fair comparison with the mentioned models, we employ the same similarity metric and training splits for retrieving visually similar images across all models. The evaluation of caption quality is conducted using a range of established metrics, which are extensively discussed in [30-31]. These metrics include BLEU [32], METEOR [33], and ROUGE-L [31]. Each metric assesses the agreement between the generated captions from automatic systems and the ground-truth captions. To facilitate this evaluation, we utilize the publicly available Python evaluation API provided by the MSCOCO evaluation server.

Table1. Quantitative results. In all columns, the higher numbers indicate a better performance.

	B-1	B-2	B-3	B-4	METEOR	ROUGE
PROPOSED	**50.0**	**30.1**	**18.3**	**11.4**	**17.6**	**37.3**
****QE_2**	44.7	24.9	13.9	7.9	14.2	32.9
***QE**	-	-	-	5.36	13.17	-
MC-KL	-	-	-	4.04	12.56	-
MC-SB	-	-	-	5.02	11.78	-
VC	-	-	-	3.71	10.07	-

*QE is the result reported by [13].
**QE_2 is the same as the QE approach, but authors obtained new results.

15. Quantitative evaluation results

Table 1 presents the quantitative results based on evaluation metrics. The proposed approach demonstrates superior performance compared to the VC, MC-SB, MC-KL, and QE models.

16. Qualitative evaluation results

Figure I-2 in Appendix I, showcases example results obtained with the proposed method on the MS COCO benchmark dataset. To facilitate comparison, the figure includes ground truth human descriptions as well as a match graph displaying the similarity between the retrieved caption and the 5 reference captions of the query image. Based on this figure, it is evident that the proposed method, employing the multi-criteria decision-making mechanism, successfully selects captions that outperform those generated by other methods.

In Fig. I-3 in Appendix I, there are instances where the proposed approach exhibits limitations. Although the system may not always produce the most desirable results in these cases, it is capable of capturing some semantic relations accurately. However, errors in the MIL outputs can impact the selection of the final caption. For instance, in Fig. I-3-a, the MIL model predicts the word "female" with a probability of 0.18 for the image. This prediction influences the subsequent steps, leading to the selection of captions containing the attribute "female" as the final caption, resulting in a BLEU-4 score of zero. Similarly, in Fig. I-3-b, the MIL model predicts the word "sandwich" with a probability of 0.16, which leads to the selection of an incorrect caption in the following steps.

In other cases, the utilization of all outputs from the MIL model instead of focusing on the words that represent the main subject of the image has resulted in the selection of inappropriate captions. For instance, in Fig. I-3-d, the MIL model predicts the words "paper" and "person" with probabilities of 0.9 and 0.97, respectively, which are then used in the subsequent steps. However, in the image, although "paper" and "person" are present to some extent, they are not the main objects of focus.
Furthermore, the limitations of the Word2vec model in accurately representing words, where the distinction between two words cannot be properly captured by their vectors, have led to the selection of incorrect captions. For example, in Fig. I-3-c, the MIL model predicts the words "red" and "black" with probabilities of 0.30 and 0.25, respectively. However, the retrieved caption mentions the color "blue" ("blue jacket") instead. This discrepancy arises due to the cosine similarity between "red" and "blue" in the Word2vec model trained on the MS COCO dataset, which is 0.73. Additionally, the cosine similarity between the terms "blue jacket" (extracted from the candidate caption) and "red jacket" (obtained from the MIL output) is 0.88, exceeding the threshold value of 0.85. Consequently, this incorrect matching leads to the selection of an erroneous caption for the image.

17. CONCLUSION

One limitation of this work is the incorrect prediction or the absence of word prediction by the MIL model. As depicted in Fig. I-1-a and Fig. I-1-b, the proposed method comprises two parts, with the MIL outputs playing a crucial role in both. However, the MIL model may only detect a subset of the objects, attributes, and actions in the query image or it may identify only a few of them, resulting in incomplete word predictions. In some cases, the MIL model may fail to identify any words at all.
The words predicted by the MIL model are utilized in two key aspects of the proposed method. Firstly, in the initial part, the MIL outputs are used to create the MIL vector, which in turn helps in selecting the most relevant captions. Secondly, in the subsequent part, the MIL output is employed to assess the level of similarity between candidate captions and the query image. Consequently, any errors or inaccuracies in the MIL outputs directly impact the performance of both parts. It would be beneficial to verify the relevance of the MIL output words before utilizing them in subsequent steps. This validation process would ensure that the selected captions align closely with the main objective of the image, enhancing the retrieval of descriptions that are more closely aligned with the image's intended meaning.
Another limitation of this work is the word2vec model, which fails to generate accurate vectors for certain words. As a result, the vector representations of these words do not effectively capture the differences between them.
Therefore, one of our future plans is to enhance both the word2vec model and the MIL model in relation to the aforementioned issues. Addressing these concerns has the potential to significantly improve the performance of the proposed method. Another future plan involves expanding the number of criteria in the decision-making process. This can be achieved by conducting specific analyses on the query image and the retrieved captions, allowing for a more comprehensive evaluation and selection process.
In conclusion, we have introduced a framework for visual retrieval-based image captioning that leverages a multi-criteria decision-making algorithm. This approach effectively combines multiple criteria with proportional impact weights to retrieve the most relevant caption for a given query image. Experimental evaluations conducted on the MS COCO benchmark dataset have demonstrated the superiority of our framework compared to other existing approaches. The utilization of criteria with proportional impact weights has contributed to significantly improved results in terms of caption relevance and effectiveness.

18. ACKNOWLEDGMENT

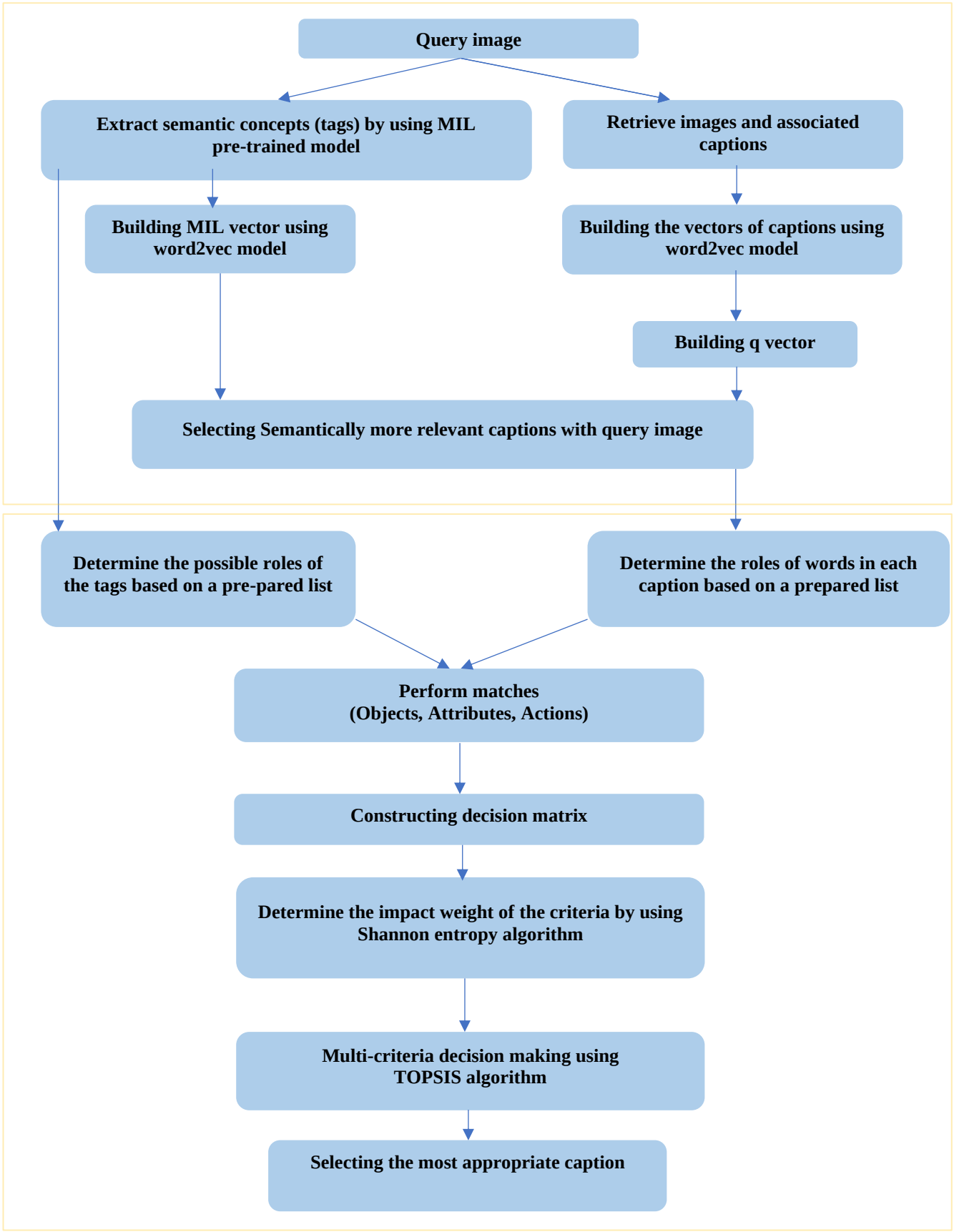
19. REFERENCES

[1] X. Li, X. Song, L. Herranz, Y. Zhu and et al., “Image captioning with both object and scene information,” in *Proceedings of the 2016 ACM on Multimedia Conference*, 2016, pp. 1107–1110.
[2] R. Bernardi, R. Cakici, D. Elliott and et al., “Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures.,” *J. Artif. Intell. Res.(JAIR)*, vol. 55, pp. 409–442, 2016.
[3] A. Farhadi, M. Hejrati, M. A. Sadeghi and et al., “Every picture tells a story: Generating sentences from images,” in *European conference on computer vision*, 2010, pp. 15–29.
[4] G. Kulkarni, V. Premraj, V. Ordonez and et al., “Babytalk: Understanding and generating simple image descriptions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, 2013.
[5] M. Mitchell, X. Han, J. Dodge and et al., “Midge: Generating image descriptions from computer vision detections,” in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 747–756.
[6] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
[7] K. Xu, J. Ba, R. Kiros and et al., “Show, attend and tell: Neural image caption generation with visual attention,” in *International Conference on Machine Learning*, 2015, pp. 2048–2057.
[8] X. Chen and C. Lawrence Zitnick, “Mind’s eye: A recurrent visual representation for image caption generation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2422–2431.
[9] F. Xiao, W. Xue, Y. Shen, and X. Gao, “A new attention-based LSTM for image captioning,” *Neural Processing Letters*, vol. 54, no. 4, pp. 3157–3171, 2022.
[10] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2text: Describing images using 1 million captioned photographs,” in *Advances in Neural Information Processing Systems*, 2011, pp. 1143–1151.
[11] G. Patterson, C. Xu, H. Su, and J. Hays, “The sun attribute database: Beyond categories for deeper scene understanding,” *Int. J. Comput. Vis.*, vol. 108, no. 1–2, pp. 59–81, 2014.
[12] R. Mason and E. Charniak, “Nonparametric Method for Data-driven Image Captioning.,” in *ACL (2)*, 2014, pp. 592–

598.
[13] S. Yagcioglu, E. Erdem, A. Erdem and et al., "A Distributed Representation Based Query Expansion Approach for Image Captioning.," in *ACL (2)*, 2015, pp. 106–111.
[14] P. Kuznetsova, V. Ordonez, A. C. Berg and et al., "Collective generation of natural image descriptions," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 2012, pp. 359–368.
[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[16] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, 2008.
[17] A. Nenkova and L. Vanderwende, "The impact of frequency on summarization," *Microsoft Res. Redmond, Washington, Tech. Rep. MSR-TR-2005*, vol. 101, 2005.
[18] K. Chatfield, K. Simonyan, A. Vedaldi and et al., "Return of the devil in the details: Delving deep into convolutional nets," *arXiv Prepr. arXiv1405.3531*, 2014.
[19] J. Devlin, H. Cheng, H. Fang and et al., "Language models for image captioning: The quirks and what works," *arXiv Prepr. arXiv1505.01809*, 2015.
[20] Kilickaya, Mert, et al. "Data-driven image captioning via salient region discovery." IET Computer Vision 11.6 , 2017, 398-406.
[21] H. Fang, S. Gupta, F. Iandola and et al., "From captions to visual concepts and back," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1473–1482.
[22] Z. Gan, C. Gan, X. He and et al., "Semantic compositional networks for visual captioning," *arXiv Prepr. arXiv1611.08002*, 2016.
[23] T.-Y. Lin, M. Maire, S. Belongie and et al., "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.
[24] T. Mikolov, I. Sutskever, K. Chen and et al., "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
[25] W. Blacoe and M. Lapata, "A comparison of vector-based representations for semantic composition," in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 2012, pp. 546–556.
[26] L. Xu and J.-B. Yang, *Introduction to multi-criteria decision making and the evidential reasoning approach*. Manchester School of Management Manchester, 2001.
[27] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.
[28] D. L. Olson, "Comparison of weights in TOPSIS models. Mathematical and Computer Modeling, 40 (7-8), 721–727." 2004.
[29] K. Yoon and C.-L. Hwang, *Multiple attribute decision making: methods and applications*. SPRINGER-VERLAG BERLIN AN, 1981.
[30] D. Elliott and F. Keller, "Comparing automatic evaluation measures for image description," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers*, 2014, vol. 452, no. 457, p. 457.
[31] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
[32] K. Papineni, S. Roukos, T. Ward and et al. , "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002, pp. 311–318.
[33] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 376–380.

20. APPENDIX I

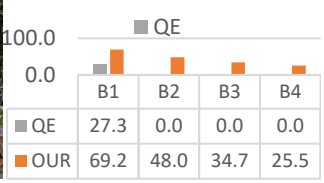
Fig. 1 illustrates the conceptual diagram of our proposed approach for image captioning. It provides a representation of the key components and their interactions within the framework. Fig. 2 and Fig. 3 also showcase examples of input images and the corresponding generated descriptions using our proposed method. Fig. 2 demonstrates instances where our approach has successfully produced high-quality output. Conversely, Fig. 3 presents examples where our approach has generated captions that are deemed less accurate or suboptimal compared to other methods.



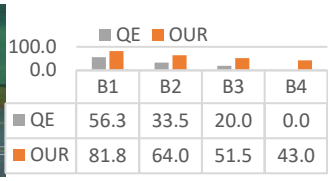
a

b

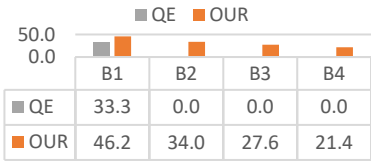
Figure 1: The conceptual diagram of our proposed approach for image captioning which consists of two parts; part one (a): retrieve semantically more relevant captions with the query image, part two (b): selecting the most appropriate caption among the candidate captions.



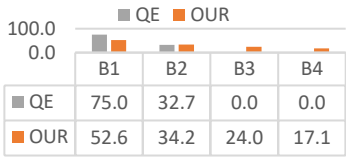
QE	the flags of many nations flying by big ben in london
OUR	landscape of a clock tower attached to a large building in a city
HUMAN	a large crowd is attending a community fair
	a crowd of people walking in an outdoor fair
	a crowd of people at a festival type event in front of a clock tower
	the building has a clock displayed on the front of it
	a festival with people and tents outside a clock tower



QE	a small boy holding a tennis racket intently stares at a tennis ball in the air
OUR	a beautiful young woman hitting a tennis ball with a racquet
HUMAN	a woman hitting a tennis ball on a court
	a woman swinging a tennis racquet towards a tennis ball
	a female tennis player finishes her swing after hitting the ball
	a woman bending slightly to hit a tennis all with a racket
	a female in a red shirt is playing tennis



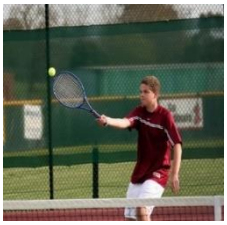
QE	a man and boy blow out a candle on a birthday cake
OUR	an older woman sits in front of a cake near a young woman
HUMAN	a woman standing over a pan filled with food in a kitchen
	a woman smiling while she prepares a plate of food
	a smiling woman standing next to a plate of food she made
	a woman in a bright pink summer shirt smiles and displays a party platter she has made
	a person standing in front of a counter top and a tall pile of food

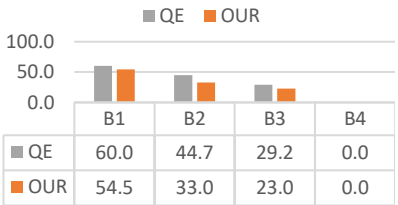


QE	a chair and a table in a room
OUR	modern living room with a ceiling fan two couches a coffee table a fireplace and a large screen tv
HUMAN	a little room and dining room area with furniture
	a living room with a big table next to a book shelf
	a living room decorated with a modern theme
	a living room with wooden floors and furniture
	the large room has a wooden table with chairs and a couch

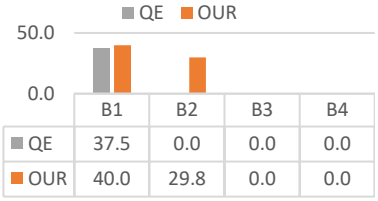
a b c d

Figure 2: Some example input images and the generated descriptions, that the proposed method has produced a good output compared to other methods

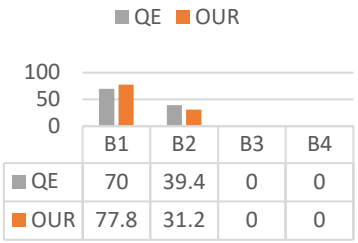




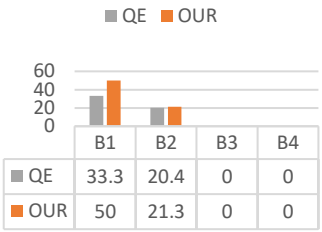
QE	a tennis player swings his racket at a tennis ball
OUR	a female tennis player lunges forward to return the tennis ball
HUMAN	a guy in a maroon shirt is holding a tennis racket out to hit a tennis ball
	a man on a tennis court that has a racquet
	a boy hitting a tennis ball on the tennis court
	a person hitting a tennis ball with a tennis racket
	a boy attempts to hit the tennis ball with the racquet



QE	many different types of vegetables on wooden table
OUR	a sub sandwich on a wooden tray on a table
HUMAN	a wooden cutting board with cheese bread and a knife on it
	a cutting board topped with cheese bread and a knife
	a cutting board with carrots and thin breading
	sliced bread and cheese sits on a cutting board with a sharp knife
	carrots bread and knife on top of cutting board



QE	a man wearing skis at the bottom of a slope
OUR	a man in a blue jacket on snow skis
HUMAN	a man on skis is posing on a ski slope
	a person on a ski mountain posing for the camera
	a man in a red coat stands on the snow on skis
	a man riding skis on top of a snow covered slope
	a lady is in her ski gear in the snow



QE	a couple slices of pizza on a cardboard box
OUR	a person is holding a large paper box with food in it
HUMAN	hot dog on a roll with cheese onions and herbs
	a sandwich has cilantro carrots and other vegetables
	a hotdog completely loaded with onions and leaves
	a hand holding a hot dog on a bun in a wrapper
	the hotdog bun is filled with carrots and greens

a b c d **Figure 3**: Some example input images and the generated descriptions, that the proposed method has produced a bad output compared to other methods