An image selection framework for automatic report generation

Changhun Hyun 1 · Chan Hur 1 · Hyeyoung Park 1

Received: 15 July 2021 / Revised: 23 February 2022 / Accepted: 10 April 2022 /
Published online: 18 May 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
The development of IoT technologies and social network services (SNS) are contributing to the growth of big data. However, the vast amount of data makes it difficult for users to find the information they need, and as a result, the demand for a system that provides the desired information in a well-organized form is increasing. Many studies are being conducted to extract desired information from data, and application studies such as automatic report generation are also being conducted. To generate a report for a given topic, a report generation system is required to extract essential information from big data and re-organize it in a compact form. Image selection system also plays an important role in automatic report generation as insertion of appropriate images can increase the completeness and readability of the report. In this study, we propose an image selection framework for recommending an appropriate image for a part of a report by combining textual information used in text-based image retrieval and visual features used in content-based image retrieval. In addition, the proposed image selection framework adopts an image filtering module that is specially designed for filtering out some images that are not suitable for use in reports. Through experiments on two datasets and comparative experiment with state-of-the-art work, we confirmed that our proposed method recommends images that fit the user’s intention, and its practical applicability.

Keywords Automatic report generation · Automatic image selection · Image retrieval · Image re-ranking · Image filtering

✉ Hyeyoung Park
hypark@knu.ac.kr

Changhun Hyun
chhyun@knu.ac.kr

Chan Hur
chanhur94@gmail.com

1 Computer Science Engineering, Kyungpook National University, Daegu, South Korea
1 Introduction

Over the past few years, IoT technology has developed significantly and has been applied to various fields of human life such as smart homes, wearables, and healthcare, making our lives more convenient and enriching. Consequently, it has become possible to collect tremendous amounts of information from IoT devices along with the Internet and social network service (SNS), which makes a significant contribution to the collection of big data. However, we cannot help but wonder whether the vast amount of data is trustworthy. IBM newly defined veracity as a fourth characteristic of big data representing the reliability of data. According to statistics measured by IBM in 2016, poor quality data costs $3.1 trillion every year. Similarly, because it has become difficult to obtain compact and well-organized information for users’ requests from vast amounts of data, the demand for data analysis technology has naturally increased.

Owing to such demand, methodological studies on extracting core and meaningful information from data and application studies such as automatic report generation using various types of data are being conducted. Automatic report generation is a technology that automatically generates report-type documents by organizing and summarizing various types of data on a given topic. Although reports are mainly written based on text, when appropriate images are inserted that well represent paragraphs or phrases, the readability and completeness of the report can be improved. Image search engines such as Google and Bing are commonly used to find images to be inserted in reports. However, it is costly and tiresome for users to look over all the numerous images in search results. In this regard, we propose a framework that automatically selects an appropriate image for a given part of the report by ranking candidate images. Although there have been several studies on image retrieval, the image selection method proposed in this paper is different in that it proposes an overall framework for the specific purpose of automatic report generation.

Report generation in this paper means creating a text summary of numerous news articles and inserting an image that represents the given summary sentence. Figure 1 shows an example of automatically generated report and our proposed image selection module applied to it. When a user inputs a query, news articles related to the query are obtained from the database, and the text part of the report is created by extracting the core contents of the articles through the text summarization process. The goal of our proposed image selection framework is to select the most appropriate image for the given query from the numerous candidate images, which are included in the related news articles used in the text summarization process. To achieve the specific purpose of selecting images for reports, the proposed image selection framework comprises three modules: candidate image filtering, image scoring, and ranking modules. The candidate image filtering module primarily filters out unsuitable images for use in reports. To achieve this, primary candidate images obtained from a database are filtered through an image classifier based on a deep neural network that has been trained to meet the purpose of our study. Subsequently, in the image scoring module, each candidate image is matched with a part of the report presented as phrases, using both textual and visual features. Finally, in the ranking module, the final score for each image is obtained by combining the text-text and image-text matching scores, and the candidate with the highest score is selected as the most suitable image for the given part of the report.

The remainder of this paper is organized as follows. The following section reviews related works on image retrieval, which is the core subject of this study. Section 3 presents detailed...
descriptions of each module along with the overall structural description of the proposed image selection framework. In Section 4, the experimental results of the proposed method using various queries and data and comparative experimental results with state-of-the-art work are presented. Finally, we conclude our work and present the direction of future work in Section 5.

2 Related works

Text-based image retrieval [15, 34, 37], which recommends an image by matching the text information such as title and tag annotated from the candidate images when a query is provided in text, is applied to image search engines such as Google and Flickr. Chaudhary et al. proposed a text-based image retrieval framework for three categories of complex queries: long, ambiguous, and abstract by comprising knowledge base construction and an image retrieval module [3]. Qian et al. proposed a tag-based image retrieval system with a topic diverse re-ranking method [25]. Text-based image retrieval is more user-friendly than content-based image retrieval, where image is queried, and fit more general image search situations. Nonetheless, this scheme has several drawbacks. One is that it retrieves numerous candidate images from the database that do not fit the user’s intention. This problem arises from the semantic difference between the text information annotated in the image and the query entered by the user [27]. This often occurs when the query is out of the range of the annotated image in the database, or the query is not detailed. The other is that it retrieves images only by relying on tagged text information, and that additional work is required after text-based image search in order to utilize advanced search options such as color and resolution provided by various image search engines. Therefore, further refinements are required for more reliable image retrieval.
To address the above-mentioned problem, content-based image retrieval methods have been proposed. Content-based image retrieval [7, 13, 18, 32] is a technology that selects an appropriate image when a query is provided as an image. To this end, after analyzing the content information included in the image and representing it as a feature vector, the similarities between the two feature vectors of the query image and the candidate images from the database are calculated, and an image with high similarity is recommended. Related studies include a method of converting images into feature vectors using deep learning techniques and calculating similarity [32]. Su et al. proposed three types of content-based image retrieval strategies to achieve high-quality image retrieval performance in terms of unsupervised, semi-supervised and supervised using conceptual features [29]. However, these existing studies depend on the performance of the feature extractor and have the disadvantage of not being able to extract detailed information such as specific people and places. Therefore, hybrid approaches for text-based and content-based image retrieval are necessary.

Existing hybrid approaches [8, 10, 16, 19, 20, 31] that utilize both visual and textual features in the field of image retrieval have been used for specific purposes, such as improving the performance of image retrieval [31], and visual question answering (VQA) [16]. Vu et al. [31] proposed a unified framework for clustering and re-ranking images. Their approach aims to improve the general performance of text-based image retrieval for queries composed of simple words rather than sentence queries used in our study. Liang et al. [16] proposed a focal visual-text attention (FVTA) network for VQA. Their research uses both visual and textual features for finding images that can answer a given question.

Similar to our study, there are studies with the goal of retrieving images suitable for a text query [9, 19]. However, Liu et al. [19] used only the image-text matching module, and Ide et al. [9] also used only the text-text matching module, while our study uses both image-text and text-text matching modules. Additionally, our proposed framework includes a module to filter images representing graphs or tables that are not appropriate as representative images. On the whole, in order to serve the specific purpose of automatic report generation, our study proposes an overall framework of image selection system that uses both text-text and image-text matching modules in addition to image filtering. Images to be inserted in the report in our proposed system have the purpose of representing given sentence or paragraph, and it is desirable to avoid images such as graphs and tables in which numbers or characters are the main content. The image selection framework proposed in this study is designed by considering all the aforementioned criteria in this section, confirming its effectiveness through experiments on text queries.

3 Proposed methods

In this section, we introduce an automatic image selection framework with three core modules: candidate image filtering, image scoring, and ranking module. We first examine the entire systematic structure of our proposed model and describe the workflow for each module. Detailed descriptions of each module and applied techniques are demonstrated.

3.1 Overall workflow

Figure 2 depicts the overall processing structure of the automatic image selection framework. First, it receives the summarized text phrase as a query from the automatic report generation
system. When the text query is provided, primary candidate images and correspondingly annotated captions are obtained from the database. At this point, publicly available datasets or a manually collected dataset used in the automatic report generation system can be employed. Note that the image caption should be provided in pairs with the image. In the case of using public search engines such as Google, a technique for directly extracting the image caption can be applied. The obtained primary candidate images are passed through the image filtering module to filter out inappropriate images for a report through deep-learning-based image classifier, which is described in Section 3.2. Then, filtered candidate images are input to the image scoring module that calculates the text-text and image-text matching score. In the text-text matching flow, the semantic similarities between the given text query and the caption of the candidate images are calculated. In the image-text matching flow, feature values of candidate images are used for similarity calculation. To calculate the semantic similarity between a text query and image feature values of candidate images, a technique of mapping text and image to the same vector space is required. We take an approach of the joint embedding method [26]. Detailed descriptions of the text-text and image-text matching flow of the image scoring module are provided in Section 3.3 and 3.4. Finally, ranking is performed based on the final matching score obtained by combining two semantic similarities. The candidate image with the highest similarity to the query is selected as the recommended image. The proposed image selection system can operate as an image selection module in an automatic report generation system [23], and it runs on an independent server.

### 3.2 Candidate image filtering module

Figure 3 represents the process of the image classification module. Candidate images searched for a query in specific field may include some images that are not suitable for insertion in a part of the report. The criteria of inappropriate image for reporting may vary. For example, images that are too small do not provide sufficient resolution, and images entirely covered with watermarks are not suitable. In particular, the retrieved images may include a number of images representing graphs or tables including texts and numbers, whose meanings change easily, even with small changes in values. Although images based on numerical data such as graphs or tables can be inserted in the report to improve the report’s completeness, if...
the numbers are slightly wrong or the legend of the graph or table contains unintended information, it is easy to become an image that does not fit the query even if the image has a caption similar to the query. This leads to worse results which lowers the stability of the quality of the automatic report generation system. In addition, images representing graphs or tables are mainly created to help understand numerical data and are suitable for the purpose of providing additional information rather than generally representing a certain query. Therefore, in order to secure the stability of the report quality while meeting the purpose of our study, we took an approach of removing the graphs and tables. In this study, these images are referred to as “graph images” and are primarily eliminated through a deep-learning-based classifier.

Although VGGNet [28] is a widely used deep neural network, the models trained for classifying graph images as in this study are not publicly available. Therefore, the VGG model pre-trained with ImageNet [5] dataset is retrained with manually collected graph image dataset. The total number of training images is 1894: 951 graph images and 943 non-graph images each, and the last ten layers of the VGG-16 network are trained. Candidate images are classified into graph images including graphs, charts, and tables, and non-graph images including only pictures through newly trained VGGNet. Finally, the remaining non-graph images after filtering out the graph images are transferred to the text-text matching module. The image classification module shows a classification performance of 97.78% on a validation set consisting of 50 graph images and 40 non-graph images. The image filtering module can be selectively applied according to the field and property of the reports, and has a high potential for utilization. In addition to classification of graph and non-graph images, it can be applied as a module for other appropriate image selection criteria such as filtering near-duplicate images.

### 3.3 Text-text matching

The workflow of the text-text matching module is shown in Fig. 4. The captions of the non-graph images obtained through the image filtering module are input to the text-text matching module. Prior to the semantic similarity calculation, the embedding vectors of the text query and the captions of the images should be obtained. Conventional embedding methods such as word2vec [21], fasttext [12], Glove [24], and BERT [6] shows remarkable performance in
various NLP tasks. However, we utilize universal sentence encoder (USE) [2] as an embedding model, and it is most suitable choice for our automatic image selection system. This is because USE enables embedding regardless of the input unit of the query, considering that a user may use words, sentences, or even paragraphs as an input query to search for images. In addition, USE has the advantage of relatively easy fine-tuning of a training model when necessary, because even a model trained with a relatively small data set by utilizing transfer learning has satisfactory performance in various NLP tasks. Furthermore, multilingual USE [35] models trained on various document data for 16 languages including Korean, English, Chinese, and Japanese are publicly provided, it has the advantage of increasing the degree of linguistic freedom of input queries. This is particularly useful for embedding when a query consists of more than one language.

For a given text query $Q$, and captions $t_i (i = 1, \ldots, N)$ of candidate images $I = \{I_1, \ldots, I_N\}$, a query vector for text-text matching $q$ and caption vectors $v_i (i = 1, \ldots, N)$ are obtained through USE embedding. The text-text similarities $S_{txt}(q, v_i)$ are calculated through the cosine similarity between the query vector and the ith caption vector. Because image retrieval based on text-text similarity depends only on the caption of the image, an incorrect image may be selected if the caption is incorrectly annotated. To cope with this problem, we propose to exploit image-text similarity scores and combine them with text-text similarity scores, which is described in the subsequent section.

### 3.4 Image-text matching

The overall process of the image-text embedding, and semantic similarity calculation process is shown in Fig. 5. The image-text matching network uses a joint embedding network [26] to map images and texts to the same vector space. In the training phase, for training data consisting of pairs of images and texts, text features and image features are obtained through universal sentence encoder and pre-trained convolutional neural network, respectively. Then, the two features are input to the joint embedding network to form an intermediate vector space so that the image-text pairs are closely mapped to each other. Specifically, the 2048-
dimensional image feature vector is obtained through a pre-trained Inception-V3 [30] network, and its dimension is scaled down to 1024 dimensions through a trainable fully connected layer. For text, firstly a 512-dimensional text feature vector is first obtained through a pre-trained universal sentence encoder and mapped to 1024 dimensions through a trainable fully connected layer.

In the training phase for the joint embedding network, the entire loss function is constructed for increasing the score with the image feature vector representing the text well and decreasing it otherwise. In the test phase, the trained joint embedding network model enables the calculation of semantic similarity between the candidate images and the text query by mapping the text and the image feature to the same vector space. To elaborate, the similarity \( S_{\text{img}}(\phi_{\text{txt}}(q), \phi_{\text{img}}(u_i)) \) between the image feature vectors \( \phi_{\text{img}}(u_i) \) \( (i = 1, \ldots, N) \) of the candidate images and the query vector \( \phi_{\text{txt}}(q) \) are calculated, where \( \phi(\cdot) \) represents the embedding function for mapping to the joint embedding space. Finally, the previously calculated \( S_{\text{txt}}(q, v_i) \) and \( S_{\text{img}}(\phi_{\text{txt}}(q), \phi_{\text{img}}(u_i)) \) are combined in the ranking module, and a candidate image that has the highest combined similarity is selected as the final image, which can be written as

\[
S_{\text{cmb}}(q, I_i) = S_{\text{txt}}(q, v_i) + S_{\text{img}}(\phi_{\text{txt}}(q), \phi_{\text{img}}(u_i)),
\]

\[
\text{Selected image} = \arg\max_{I_i \in I} (S_{\text{cmb}}(q, I_i)).
\]

### 4 Experiments

In this section, we describe the experimental settings including the dataset and the queries, and then present the experimental results on two datasets. To show that our proposed method is better than using an image search engine on the Web as in a typical reporting situation, we conducted experiments using candidate images obtained from Google image search results for five queries. Similarly, experiments were performed on
manually collected Korean news article data for ten queries to demonstrate the practical applicability of our proposed image selection system. Through each experiment, we confirm that our proposed image selection system shows better performance compared to Google image search and SCOUTER [14] which is an image retrieval system for Korean news article data. Figure 6 is an example of the ranking results in which candidate images obtained from the Google image search are ranked through the proposed image selection system. In addition, through the experimental results using the text-text and image-text matching modules separately and using the combined module, we confirmed that the two matching modules can play complementary roles. Finally, through a comparative experiment with the state-of-the-art study that is most similar to our work, the applicability and utility of our proposed module was confirmed.

4.1 Characteristics of the queries

In the proposed image selection system, queries can be provided in various forms ranging from keywords to phrases, sentences, and paragraphs. The queries used in the experiment are provided in Korean, and the lists of queries introduced in Section 4.4 and 4.5 are described by translating them. Queries are mainly focused on the economy, as the candidate images for economic queries contain a fair number of graph images, which are suitable for confirming the effect of the image filtering module. In contrast to other image retrieval studies, relatively long sentences are used as queries, because the summarized sentence from the text summary module of the automatic report generation system is given as an input to our proposed image selection module [23]. For a query to be economical, it needs to be detailed to some extent rather than simple to select an appropriate image to be inserted into the report. However, in section 4.7, which shows the comparative experimental results with state-of-the-art study and our work, relatively short word-level queries are chosen due to the limitations of the database.

Fig. 6 Sample of filtered images and ranking results by text-text, image-text and combined matching score
4.2 Datasets

In the experiment using Google image search results, the candidate image and the caption pairs were manually crawled through data parsing technology. For each of the five queries (Q1 to Q5), 50 candidate images and caption pairs were collected. The Korean news article dataset used in this study was obtained from a data search system called scalable document repository manager (SCOUTER) [14], a system that supports scalable and efficient search system for collected large-scale news article data. SCOUTER collects raw documents on various fields such as politics, society, and the economy, and preprocesses them using a json scheme through morphological analysis. Subsequently, SCOUTER stores over two million document numbers, image URLs, image captions, and original documents of articles. The text query from the text summary module of the automatic report generation system is divided into keyword units through morphological analysis, and data are retrieved by scoring through TF-IDF with the original document of the article. To train the image-text matching module, we used approximately 10,000 image and caption pairs obtained from SCOUTER. In the test phase, each of the ten queries (Q6 to Q15) contained ten candidate image and caption pairs. The data used in the comparative experiment with the state-of-the-art study were obtained through a demo system provided by Liu et al. [19]. The details are covered in Section 4.7.

4.3 Evaluation criteria

For the quantitative evaluation of our proposed method, the normalized discounted cumulative gain (NDCG), which is widely used in the field of information retrieval [17, 33, 36] is used to measure image ranking accuracy. NDCG is an index of evaluation that can evaluate whether top K candidate image ranking through the model is sorted well, in the order of high relevance score. An NDCG value closer to 1 indicates that the images are ideally ranked. NDCG is calculated as follows:

\[
DCG_K = \sum_{i=1}^{K} \frac{2^{rel(i)} - 1}{\log_2(i + 1)},
\]

\[
IDCG_K = \sum_{i=1}^{K} \frac{2^{rel^*(i)} - 1}{\log_2(i + 1)},
\]

\[
NDCG_K = \frac{DCG_K}{IDCG_K},
\]

where K indicates the number of top-ranked images, \(rel(i)\) represents the relevance score corresponding to the \(i\)th image of the results ranked through the model, and \(rel^*(i)\) represents the relevance score of the \(i\)th image, based on the ideal ranking order obtained through user evaluation. Three graduate students majoring in computer science participated in the user evaluation, and they evaluated each candidate images in three levels: three points for suitable, two points for ambiguous, and one point for not suitable base on the image quality and whether the image represents the query well. The relevance score regarded as the ground truth is obtained by averaging the
evaluation points of three users. In Figs. 7, 8, 10 and 11, the obtained relevance score for each image is shown under the corresponding image.

4.4 Experiments on Google image search

This section discusses the experimental results for five queries using Google image search. In Fig. 7, the first row of each query represents the top 10 images in search order among 50 candidate images from Google, and the second row shows the top 10 ranked images in order of the score from the proposed image selection system. The left-most image marked with a red
box is the selected image with the highest combined similarity, and the combined similarity of the candidate image reduces as it moves to the right. The number indicated below each image is the relevance score through user evaluation.

We chose five queries (Q1 to Q5) among world events for the experiments using Google image search results. The list of queries is as follows.

Q1. US-China trade negotiations amid Trump’s threat of additional tariffs.
Q2. China falls into Trump’s overturning tactics, even negotiating is a problem.

Fig. 8 Experimental results on Korean news article data for ten queries Q6 ~ Q15
Q3. Hana bank’s housing finance corporation eases requirements for single-parent family loans and cuts interest rate.

Q4. Prime minister May, chances of empty-handed talks increase at EU summit, EU said the British parliament had to approve the agreement in response to UK’s request to postpone Brexit deadline.

Q5. EU lowered GDP growth rate in Eurozone 1.3%

In Fig. 7, when observing the candidate images shown in the first row for each of Q1 to Q5, the top ten candidate images according to the Google search order include several graph images. Contrastingly, there are no graph images in the top ten candidate images displayed in the second row of each query, which implies that the candidate image filtering module has successfully removed them. Observing the selected images marked with red boxes for Q1 to Q5, the selected images may appear ambiguous unlike in other studies that have experimented with queries in keyword units that specify objects such as “bus” or “white tiger” [17, 33]. This is because sentence queries that are somewhat difficult to describe the situation with pictures were used for the experiment. Therefore, in the user evaluation stage, rather than the degree of matching between the words included in the query and the objects in the image, the criterion for a well selected image would be whether the situation is well expressed. For instance, the image selected for Q4 represents the scene wherein Jean-Claude Junker, President of the European Commission (left) and Donald Tusk, President of the European council (right), were talking at a meeting held in the Europa building, and this is the image used in the actual Internet news article for the query.

Table 1 shows the average NDCG for Q1 to Q5. NDCG@1 to NDCG@10 show that the images ranked through the proposed image selection system are well sorted by priority according to the user evaluation. On the other hand, the original candidates include a number of graph images with low relevance score evaluated by users. From the results, we can see the effect of the proposed overall framework on improving user’s preference.

4.5 Experiments on Korean news article data

The experimental results for ten queries using Korean news article data are shown in Fig. 8. The images displayed in the row of each query represent ten candidate images obtained in order of high matching score by the internal matching algorithm of SCOUTER. The left-most image shows the image with the highest SCOUTER matching score, and the matching score decreases toward the right. In the same way as the Google image search experiment, the selected image is marked with a red box, and the number indicated below each image is the relevance score through user evaluation. However, unlike Fig. 7 in section 4.4 which is the experimental result for Google image search data, the ranking result through the proposed image selection system was not included due to space problems, which can be confirmed in the appendix.

| NDCG          | @1     | @2     | @3     | @4     | @5     | @6     | @7     | @8     | @9     | @10    |
|---------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Proposed Method | 0.9520 | 0.9192 | 0.8993 | 0.8845 | 0.8995 | 0.9118 | 0.9185 | 0.9440 | 0.9616 | 0.9621 |
| Google        | 0.7326 | 0.7001 | 0.7193 | 0.7151 | 0.7283 | 0.7700 | 0.7975 | 0.8193 | 0.8599 | 0.8765 |
We chose ten queries in consideration of five sub-fields related to economy, two each for finance (Q6, Q7), price (Q8, Q9), export (Q10, Q11), COVID-19 (Q12, Q13), and aviation (Q14, Q15). The list of queries is as follows.

Q6. Financial authorities tighten regulations on credit loans.
Q7. The Bank of Korea lowered base rate to 0%.
Q8. National gasoline prices fell below 1300 won.
Q9. Service prices rose only 0% last month.
Q10. Exports continue to decline due to the decrease in semiconductor exports.
Q11. Exports are showing a recovery trend.
Q12. The number of COVID-19 confirmed cases shows an increase in Korea.
Q13. Domestic economic activity shrinks due to COVID-19.
Q14. Japanese government resumes flights to Korea.
Q15. The government reduced rental fees for airport stores.

Although there are variations depending on the query, it can be observed that a fair number of graph images are also included in the candidate images obtained from SCOUTER. In addition, there are cases wherein images are duplicated. This is because the data storing method of SCOUTER is crawling news articles on the Internet and storing them based on the title and ID number of the news article. In practice, news articles on the Internet often use the same image even if the article has a different title.

The experimental results on Korean news article data also show the same context as the experiments on Google image search. All graph images from SCOUTER are removed through the candidate image filtering module, and the graph images have the lowest score in the user evaluation stage. Observing the image finally selected for Q13: “Domestic economic activity shrinks due to COVID-19,” it represents situation in which there are no people on the street because of the sharp reduction in domestic consumption owing to COVID-19. In evaluating the image selected for another query, Q7: “The Bank of Korea lowered base rate to 0%,” indicates that the governor of the Bank of Korea is pounding the gavel as he presides over a meeting, whose agenda is lowering the base rate at a plenary session of the Monetary Policy Board. As such, the overall selected images well represent the situation of each query compared to other candidate images.

Table 2 shows the average NDCG for Q6 ~ Q15. The values from NDCG@1 to NDCG@10 shows that the ranking result through the proposed image selection system is better than that of SCOUTER’s internal image retrieval algorithm. The difference between the values of NDCG@1 for SCOUTER and for the proposed method can be analyzed by comparing the image selected for each query with the image retrieved from SCOUTER with the first priority. For example, in case of Q14 in Fig. 8, the selected image marked with red box shows the appearance of Incheon airport, where flights to Japan were suspended owing to COVID-19, which is more appropriate than the image simply representing airplane floating in

| NDCG       | @1     | @2     | @3     | @4     | @5     | @6     | @7     | @8     | @9     | @10    |
|------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Proposed Method | 0.9757 | 0.9147 | 0.8915 | 0.8915 | 0.8875 | 0.9316 | 0.9477 | 0.9581 | 0.9644 | 0.9661 |
| Google     | 0.6932 | 0.6918 | 0.6847 | 0.6784 | 0.6997 | 0.7181 | 0.7490 | 0.7646 | 0.8167 | 0.8642 |
the sky. This result shows the effect of text-text matching module utilizing image captions that sometimes contain richer information than we can get from image analysis. In another example, the first displayed image of Q12 shows a graph in which the number of COVID-19 confirmed cases in Korea changed from a declining to an increasing trend. However, this does not indicate the cumulative number of confirmed cases, and at this point, the cumulative number of confirmed cases in Korea would have already exceeded 15,000. Therefore, it is inappropriate to represent the Q12, and the selected image representing the COVID-19 testing center better matches with the given query, which shows the effectiveness of the graph-filtering module.

4.6 Experiments on the effectiveness of the combined matching module

One may raise a question that “why combine text-text and image-text matching module?”. As mentioned in Section 3.3, captions of candidate images used in the text-text matching module often include sentences that do not indicate a query at all. In this case, although the image itself is very suitable for representing a given query, it is placed at a lower rank in the ranking process. In addition, due to the characteristic of our study, named entities representing specific person or company are quite often included in queries which is different from other image retrieval studies, and image-text matching module alone cannot retrieve images considering these proper nouns. Therefore, we tried to build an automatic image selection system in which two matching modules, text-text and image-text, complement each other without depending only on one matching module.

Figure 9 shows the experimental results according to matching module. When the image-text matching module was used alone, the average NDCG value was measured very low. However, when the image-text matching module used in combination with the text-text matching module, it shows higher NDCG values than the case where the text-text matching module was used alone. As a result, it was found that complementary synergies can be obtained by combining the matching scores calculated through the image-text and text-text matching modules, and they play a positive role in the decision-making to finally select an image that well represents the query.

4.7 Comparative experiments with state of the art work

Since our study has a specific purpose of report generation which is different from conventional image retrieval studies in related fields, comparative experiments are not easy. A study
most similar to ours was published by Liu et al. [19], which is a study to retrieval images suitable for article content to assist photo editors when writing news articles. A demo of their research is made available as a Web page https://modemos.epfl.ch/article, and we used it for comparison.

A total of 10 queries were used in the comparative experiments. Unlike the relatively long sentence queries (Q1 to Q15) used in our work, they were composed of short word-level queries. The provided demo Web page by Liu et al. [19] states that a limited database is used, so there is some variation in the quality and the number of retrieved images depending on the queries entered by the user. We selected queries that can obtain a sufficient number of candidate images through experiments. The 10 queries that has been selected are given below.

Q16. President Trump Speech.
Q17. UK Brexit.
Q18. More Lake Zurich visitors in summer.
Q19. Elon Musk Tesla.
Q20. Iphone Apple.
Q21. Beer.
Q22. Swiss football.
Q23. Airliner.
Q24. Drone.
Q25. Bus.

In the study of Liu et al. [19], it was noted that their proposed system may not perform well in retrieval for queries including named entities, but performs well for general queries such as “car accident” or “traffic jam”. As a result of checking the retrieved images using a general query through several trials, almost all of them were confirmed to be suitable for the query. However, when performance is evaluated with NDCG as in our study, if all images are given maximum scores in the user evaluation stage, it is impossible to compare the adequacy between candidate images, so such queries were not used in comparative experiments. Therefore, the queries (Q16 to Q25) consists of a query representing a general object, a very famous person or company such as “Donald Trump” or “Apple”, and one sentence query used in the study of Liu et al. [19].

In the Web page provided by Liu et al. [19], retrieved images are shown in ranking order, but unlike the purpose of our study of recommending the most appropriate single image in the end, the purpose of their study is to help the user choose among retrieved images. In Liu et al. [19], it is mentioned that the ranking order was not significantly meaningful. Although the purpose of our and their study are different, comparative experiments were conducted as a study most similar to ours in that it recommends images to be inserted to users when writing a report (or article). Top 20 candidate images are taken for each query, and only images with captions were taken in the order they are displayed in the demo Web page. As for the experimental results, only the top ten images were taken in the form of Figs. 7 and 8. In Figs. 10 and 11, the order of images retrieved through demo Web page is expressed as ‘Newsroom search order’ in the first row for each query and is briefly expressed as ‘Newsroom’ in the description of the experimental results.

As shown in Fig. 10, images retrieved through the Newsroom for “President Trump Speech” can be seen that images of two people other than President Trump were
retrieved in the first and second rankings. However, when ranking is performed through our proposed system, it can be seen that the two images do not belong to the top ten images. Likewise, in the case of Q20, 3rd and 4th images that are not related to iPhone or Apple are retrieved in high ranking in image search through Newsroom, whereas only images related to Apple or iPhone are sorted in high ranking in ranking through the proposed system. In the experiment on Q25 in Fig. 11, some images representing subways, not buses, were placed in a high rank, whereas when ranking was performed through the proposed system, it can be seen...
that images representing buses were ranked high. Among 10 queries, the ranking order through the Newsroom had better results for 4 queries: “Drone”, “More Lake Zurich visitors in summer”, “Swiss football”, and “Elon Musk Tesla”, and for the remaining queries, our proposed system showed better ranking results. Figure 12 shows the average NDCG for 10 queries in the proposed system and Newsroom experiment. Although there is a difference according to the query, it was found that our proposed system has a relatively better average NDCG value even for a short word-level query.
5 Conclusions and future works

In this study, an image selection system that is specially designed for automatic report generation is proposed. This system consists of detailed modules to complement the drawbacks of conventional image retrieval studies. The candidate image filtering module plays a role in reducing the selection range by primarily filtering out images that are not appropriate as representative images among candidates obtained from the database. In this study, we trained the module to filter out graph images where numbers or letters are the main elements of the image. In the image scoring module, text-text and image-text similarities were calculated using both textual and visual features of the candidate image to improve the dependency problem of text-based and content-based image retrieval. Finally, the ranking for final image selection was performed by combining the two similarities through the ranking module. Through experiments on Google image search and the Korean news article dataset, it was found that the proposed image selection system recommends images that suit the user’s intention compared to the image search on the Web and confirmed the practical applicability of the proposed image selection system.

For future work, a function for handling duplicate images is expected to be added to the candidate image filtering module, and the module will be refined by establishing a firmer criterion for representative images through experiments on various datasets. We plan to develop a more sophisticated image selection system by improving the matching function of the image scoring module and the combining function of the ranking module, and a user-centric system that allows the user to participate in the image selection process to reduce the risk of unintentional automatic image selection. In addition, by expanding the purpose of our study, it is also worth investigating the problem of inserting images providing additional information such as graphs or tables into the report.
Appendices

Detailed experimental results on SCOUTER data

Appendix Figs. 13 and 14 are figures including the ranking results through the proposed image selection system shown in Fig. 8, which is the experimental result for SCOUTER data in section 4.5 of the main text. Again, due to space problems, Appendix Figs. 13 and 14 are shown as experimental results for Q6 ~ Q10 and Q11 ~ Q15, respectively.

Fig. 13  Experimental results on Korean news article data for queries Q6 ~ Q10
Acknowledgements  This work was supported by the Human Resources Program in Energy Technology of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea (No. 20204010600060). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-02068, Artificial Intelligence Innovation Hub).

Declarations

Conflicts of interest / competing interest  The authors claims no conflict or competing interest.
References

1. Alkhawlani M, Elmogy M, El Bakry H (2015) Text-based, content-based, and semantic-based image retrievals: a survey. Int J Comput Inf Technol 4(01):58–66
2. Cer D, Yang Y, Kong SY et al (2018) Universal sentence encoder. arXiv preprint arXiv:1803.11175
3. Chaudhary C, Goyal P, Goyal N, Chen YPP (2020) Image retrieval for complex queries using knowledge embedding. ACM Trans Multimedia Comput Commun Appl (TOMM) 16(1):1–23. https://doi.org/10.1145/3375786
4. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40(2):1–60. https://doi.org/10.1145/1348246.1348248
5. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition 248–255. https://doi.org/10.1109/cvpr.2009.5206848
6. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
7. Gudivada VN, Raghavan VV (1995) Content-based image retrieval systems. Computer 28(9):18–22
8. He J, Li M, Zhang HJ, Tong H, Zhang C (2006) Generalized manifold-ranking-based image retrieval. IEEE Trans Image Process 15(10):3170–3177. https://doi.org/10.1109/tip.2006.877491
9. Ide I, Kawanishi Y, Kunishiro K, Nack F, Deguchi D, Murase H (2017) Automatic selection of web contents towards automatic authoring of a video biography. In: 2017 IEEE international symposium on multimedia (ISM) 304-307
10. Ji R, Yao H, Wang J, Sun X, Liu X (2008) Real-time image annotation by manifold-based biased fisher discriminant analysis. In: Visual Communications and Image Processing 2008 Vol. 6822, p. 682222. https://doi.org/10.1109/ipvc.2008.471742
11. Jing B, Xie P, Xing E (2018) On the automatic generation of medical imaging reports. In: Proceeding of the 56th annual meeting of the Association for Computational Linguistics 2577-2586. https://doi.org/10.18653/v1/p18-1240
12. Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, Mikolov T (2016) Fasttext.Zip: compressing text classification models. arXiv preprint arXiv:1612.03651
13. Latif A, Rasheed A, Sajid U, Ahmed J, Ali N, Ratyal NI, Zafar B, Dar SH, Sajid M, Khalil T (2019) Content-based image retrieval and feature extraction: a comprehensive review. Math Probl Eng 2019
14. Lee JY, Suh YK (2018) SCOUTER: a scalable document repository manager for efficient retrieval over large-scale document sets. Korean Database Conference 23–26
15. Li W, Duan L, Xu D, Tsang IWH (2011) Text-based image retrieval using progressive multi-instance learning. In: Proceedings of the 1st ACM international conference on multimedia information retrieval 141-148. https://doi.org/10.1145/1460096.1460121
16. Liu P, Guo JM, Wu CY, Cai D (2017) Fusion of deep learning and compressed domain features for content-based image retrieval. IEEE Trans Image Process 26(12):5706–5717
17. Liu F, Lebret R, Orel D, Sordet P, Aberer K (2020) Upgrading the newsroom: an automated image selection system for news articles. ACM Trans Multimedia Comput Commun Appl (TOMM) 16(3):1–28
18. Luo B, Wang X, Tang X (2003) World-wide-web-based image search engine using text and image content features. Internet Imaging IV 5018:123–130. https://doi.org/10.1107/1476329
19. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
20. Neto JL, Freitas AA, Kaestner CA (2002) Automatic text summarization using a machine learning approach. Brazilian symposium on artificial intelligence 205-215. https://doi.org/10.1007/3-540-36127-8_20
21. Noh Y, Shin Y, Park J, Kim AY, Choi SJ, Song HJ, Park SB, Park S. (2020) WIRE: an automated report generation system using topical and temporal summarization. In: Proceedings of the 43rd international ACM SIGIR conference on Research and Development in information retrieval, virtual event 2169-2172. https://doi.org/10.1145/3397271.3401409
22. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 1532-1543. https://doi.org/10.3115/v1/d14-1162
25. Qian X, Lu D, Wang Y, Zhu L, Tang YY, Wang M (2017) Image re-ranking based on topic diversity. IEEE Trans Image Process 26(8):3734–3747. https://doi.org/10.1109/tip.2017.2699623
26. Reed S, Akata Z, Lee H, Schiele B (2016) Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition 49-58. https://doi.org/10.1109/cvpr.2016.13
27. Rui Y, Huang TS, Chang SF (1999) Image retrieval: past, present, and future. J Vis Commun Image Represent 10(1):1–23
28. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
29. Su J, Hong TP, Chang YT, Tung HY (2016) Un-supervised, semi-supervised and supervised image retrieval based on conceptual features. In: 2016 IEEE second international conference on multimedia big data (BigMM). 129-133. https://doi.org/10.1109/bigmm.2016.26
30. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2818-2826. https://doi.org/10.1109/cvpr.2016.308
31. Vu XS, Vu T, Nguyen H, Ha QT (2015) Improving text-based image search with textual and visual features combination. In: Knowledge and system engineering 233-245. https://doi.org/10.1007/978-3-319-11680-8_19
32. Wan J, Wang D, Hoi SCH, Wu P, Zhu J, Zhang Y, Li J (2014) Deep learning for content-based image retrieval: a comprehensive study. In: Proceedings of the 22nd ACM international conference on multimedia. 157-166. https://doi.org/10.1145/2647868.2654948
33. Yang X, Mei T, Zhang Y, Liu J, Satoh SI (2016) Web image search re-ranking with click-based similarity and typicality. IEEE Trans Image Process 25(10):4617–4630. https://doi.org/10.1109/tip.2016.2593653
34. Yang S, Li L, Wang S, Zhang W, Huang Q, Tian Q (2019) Skeletonet: a hybrid network with a skeleton-embedding process for multi-view image representation learning. IEEE Trans Multimedia 21(11):2916–2929. https://doi.org/10.1109/tmm.2019.2912735
35. Yang Y, Cer D, Ahmad A, Guo M et al (2020) Multilingual universal sentence encoder for semantic retrieval. In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics: system demonstrations 87-94. https://doi.org/10.18653/v1/2020.acl-demos.12
36. Zhao F, Huang Y, Wang L, Tan T (2015) Deep semantic ranking based hashing for multi-label image retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition 1556-1564. https://doi.org/10.1109/cvpr.2015.7298763
37. Zhao W, Yan L, Zhang Y (2018) Geometric-constrained multi-view image matching method based on semi-global optimization. Geo-Spat Inf Sci 21(2):115–126. https://doi.org/10.1007/10095020.2018.1441754
38. Zhong SH, Liu Y, Li B, Long J (2015) Query-oriented unsupervised multi-document summarization via deep learning model. Expert Syst Appl 42(21):8146–8155. https://doi.org/10.1016/j.eswa.2015.05.034

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.